DIGITAL PHENOTYPING/DIGITAL BIOMARKERS TO MONITOR PSYCHIATRIC DISORDERS

EDITED BY: Jennifer H. Barnett, Qiang Luo, Martin J. Sliwinski and Raz Gross

PUBLISHED IN: Frontiers in Psychiatry

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-852-3
DOI 10.3389/978-2-88976-852-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

DIGITAL PHENOTYPING/DIGITAL BIOMARKERS TO MONITOR PSYCHIATRIC DISORDERS

Topic Editors:
Jennifer H. Barnett, Cambridge Cognition (United Kingdom), United Kingdom
Qiang Luo, Fudan University, China
Martin J. Sliwinski, Independent researcher
Raz Gross, Sheba Medical Center, Israel

Citation: Barnett, J. H., Luo, Q., Sliwinski, M. J., Gross, R., eds. (2022). Digital Phenotyping/Digital Biomarkers to Monitor Psychiatric Disorders. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-852-3

Table of Contents

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 12 May 2020
doi: 10.3389/fpsyt.2020.00422

Altered Temporal Variability of Local and Large-Scale Resting-State Brain Functional Connectivity Patterns in Schizophrenia and Bipolar Disorder

Yicheng Long[1,2], Zhening Liu[1,2], Calais Kin Yuen Chan[3], Guowei Wu[1,2], Zhimin Xue[1,2], Yunzhi Pan[2], Xudong Chen[1,2], Xiaojun Huang[1,2], Dan Li[4] and Weidan Pu[5]**

[1] Department of Psychiatry, The Second Xiangya Hospital, Central South University, Changsha, China, [2] Mental Health Institute of Central South University, Changsha, China, [3] Department of Psychology, The University of Hong Kong, Hong Kong, Hong Kong, [4] Department of Geriatrics, The Second Xiangya Hospital, Central South University, Changsha, China, [5] Medical Psychological Center, The Second Xiangya Hospital, Central South University, Changsha, China

OPEN ACCESS

Edited by:
Qiang Luo,
Fudan University, China

Reviewed by:
Mingrui Xia,
Beijing Normal University, China
Benjamin Becker,
University of Electronic Science and Technology of China, China

***Correspondence:**
Dan Li
lidanxy@csu.edu.cn
Weidan Pu
weidanpu@csu.edu.cn

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: 08 February 2020
Accepted: 24 April 2020
Published: 12 May 2020

Citation:
Long Y, Liu Z, Chan CKY, Wu G, Xue Z, Pan Y, Chen X, Huang X, Li D and Pu W (2020) Altered Temporal Variability of Local and Large-Scale Resting-State Brain Functional Connectivity Patterns in Schizophrenia and Bipolar Disorder. Front. Psychiatry 11:422. doi: 10.3389/fpsyt.2020.00422

Schizophrenia and bipolar disorder share some common clinical features and are both characterized by aberrant resting-state functional connectivity (FC). However, little is known about the common and specific aberrant features of the dynamic FC patterns in these two disorders. In this study, we explored the differences in dynamic FC among schizophrenia patients (n = 66), type I bipolar disorder patients (n = 53), and healthy controls (n = 66), by comparing temporal variabilities of FC patterns involved in specific brain regions and large-scale brain networks. Compared with healthy controls, both patient groups showed significantly increased regional FC variabilities in subcortical areas including the thalamus and basal ganglia, as well as increased inter-network FC variability between the thalamus and sensorimotor areas. Specifically, more widespread changes were found in the schizophrenia group, involving increased FC variabilities in sensorimotor, visual, attention, limbic and subcortical areas at both regional and network levels, as well as decreased regional FC variabilities in the default-mode areas. The observed alterations shared by schizophrenia and bipolar disorder may help to explain their overlapped clinical features; meanwhile, the schizophrenia-specific abnormalities in a wider range may support that schizophrenia is associated with more severe functional brain deficits than bipolar disorder. Together, these findings highlight the potentials of using dynamic FC as an objective biomarker for the monitoring and diagnosis of either schizophrenia or bipolar disorder.

Keywords: dynamic functional connectivity, schizophrenia, bipolar disorder, thalamus, sensorimotor, basal ganglia

INTRODUCTION

Schizophrenia and bipolar disorder are two of the most disabling psychiatric disorders worldwide, which are often misdiagnosed in clinical practice because of their overlap in clinical features. These common features entail both cognitive deficits and psychotic symptoms including hallucinations, delusions, and disorganized thinking (1–3). Over the years, neuroimaging studies using resting-state functional magnetic resonance imaging (rs-fMRI) have provided evidence for both shared and distinct disturbances in brain functions, as characterized by aberrant resting-state functional connectivity (FC), in the schizophrenia and bipolar disorder (4–7). For instance, when compared with healthy subjects, over-connectivity between the thalamus and sensorimotor cortices was commonly found in both schizophrenia and bipolar disorder patients (4, 8). On the other hand, other unique abnormalities such as hypo-connectivity within frontal–parietal areas were shown only in schizophrenia but not bipolar disorder patients (7). Appreciably, these findings have significantly advanced our understanding of the complex relationship between these severe disorders.

Most previous rs-fMRI studies were performed under the assumption that patterns of brain FC are stationary during the entire scanning period. Yet, it has been newly proven that the brain FC fluctuates over time even during the resting-sate, implying that conventional static FC methodology may be unable to fully depict the functional architecture of brain (9, 10). Therefore, the "dynamic FC" has become a hot-spot in rs-fMRI studies to capture the temporal fluctuations of brain FC patterns during the scan (11). Notably, the dynamic features of FC have been associated with a wide range of cognitive and affective processes such as learning (12), executive cognition (13), psychological resilience (14), and emotion (15), as well as multiple common psychiatric and neurological disorders such as autism (16), Alzheimer's disease (17), and major depressive disorder (18, 19). These findings thus highlight the importance of studying dynamic FC for further improving our understanding of both brain functions and dysfunctions.

Despite the accumulating knowledge on dynamic FC, it remains little known about if there are common and/or specific changes in dynamic features of FC in schizophrenia and bipolar disorder. To our knowledge, there have been only a limited number of efforts to date to differentiate schizophrenia and bipolar disorder by features of dynamic FC (20–22). Furthermore, all these studies mainly focus on the dynamic "connectivity state" changes based on the whole-brain FC profiles; therefore, although features of such global connectivity states have been reported to provide a high predictive accuracy in classifying schizophrenia and bipolar disorder (20–22), how these two disorders differ from each other in terms of dynamic connectivity profiles within particular brain regions or systems remains poorly understood, and needs to be further investigated.

The above concerns can be addressed by a novel approach, as proposed in some latest works (23, 24), to investigate dynamic FC by defining and comparing the temporal variability of FC patterns involved in specific brain regions or large-scale brain networks. This approach allows localization of those brain regions or networks showing significant group differences in FC variability, thus being helpful to identify aberrant dynamic FC patterns from the perspectives of both local and large-scale brain functional dynamics (24). In fact, using such an approach, the patients with schizophrenia have been recently found to show increased FC variabilities in sensory and perceptual systems (e.g. the sensorimotor network and thalamus) and decreased FC variabilities in high-order networks (e.g. the default-mode network) than healthy subjects at both regional and network levels (23). But to our knowledge, it remains unclear and needs to be tested whether these dynamic changes would be specific to schizophrenia, or shared with bipolar disorder.

Therefore, in this study, we aimed to explore the common and specific dynamic features of both local and large-scale resting-state FC, in terms of temporal variability, the schizophrenia and bipolar disorder. To reach this goal, groups of schizophrenia patients, bipolar disorder patients and healthy controls were recruited and scanned using rs-fMRI; applying a recently proposed novel methodological approach (23, 24), temporal variabilities of FC patterns were then compared among the groups at all the regional, intra-network, and inter-network levels. It was anticipated that our results would provide important complementary information to prior studies that mainly focused on the global dynamic FC states (20–22), and further improve our understanding about the relationship between schizophrenia and bipolar disorder from a dynamic brain functional perspective.

MATERIALS AND METHODS

Subjects and Measurements

According to the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) criteria, 78 patients with schizophrenia and 60 patients with type I bipolar disorder were recruited from the Second Xiangya Hospital of Central South University, Changsha, China; 69 age-, sex-, and education-matched healthy controls without any family history of psychiatric disorders were also recruited from the Changsha city. All participants were right-handed, Han Chinese adults with at least 9 years of education. All participants had no history of any substance abuse, any other neurological disorder, any contraindication to fMRI scanning or any history of electroconvulsive therapy. Because of excessive head motion (see *Data Acquisition and Preprocessing*), 12 schizophrenia patients, 7 bipolar disorder patients, and 3 healthy controls were excluded, and the final analyzed sample consisted of 66 schizophrenia patients, 53 bipolar disorder patients, and 66 healthy controls.

For the schizophrenia patients, severity of the current clinical symptoms was assessed using the Scale for Assessment of Positive Symptoms (SAPS) and the Scale for Assessment of Negative Symptoms (SANS) (25). For the patients with bipolar disorder, severity of the current mood and mania symptoms was assessed using the 17-item Hamilton Rating Scale for Depression (HAMD) (26) and the Young Mania Rating Scale (YMRS) (27),

respectively. Dosages of antipsychotics in all patients were converted to chlorpromazine equivalence (28). In addition, all participants completed the Information (WAIS-I) and Digit Symbol (WAIS-DS) subtests of the Wechsler Adult Intelligence Scale (29), which measure two important domains of cognitive functions, verbal comprehension and processing speed, respectively (30, 31).

The study was approved by the Ethics Committee of the Second Xiangya Hospital of Central South University, and written informed consent was obtained from all participants.

Data Acquisition and Preprocessing

The details about brain imaging data acquisition and preprocessing can be found in one of our recently published studies (14). Briefly, rs-fMRI and T1-weighted structural images were scanned for each participant on a 3.0 T Philips MRI scanner (repetition time = 2,000 ms, echo time = 30 ms, slice number = 36, field of view = 240 × 240 mm^2, acquisition matrix = 144 × 144, flip angle = 90°, and number of time points = 250 for rs-fMRI images; repetition time = 7.5 ms, echo time = 3.7 ms, slice number = 180, field of view = 240 × 240 mm^2, acquisition matrix = 256 × 200, and flip angle = 8° for T1-weighted images). Data preprocessing was performed using the standard pipeline of the DPARSF software (32, 33), including discarding the first 10 volumes, slice timing, head motion realignment, brain segmentation, spatial normalization, temporal filtering (0.01–0.10 Hz), and regressing out nuisance factors including the Friston-24 head motion parameters (34) as well as white matter and cerebrospinal fluid signals. Global signal regression (GSR) was not performed as it is still a controversial preprocessing option in rs-fMRI studies (35). Subjects with excessive head motion were excluded from the analysis, as determined by a mean framewise-displacement (FD) (36) > 0.2 mm.

Temporal Variability of FC

After preprocessing, the mean time series were extracted from each of the 116 regions of interest (ROIs) in the Automated Anatomical Labeling (AAL) atlas (37), which was validated (38, 39) and widely used in functional neuroimaging studies (40, 41). The names of all the 116 ROIs were listed in **Supplementary Table S1**.

As shown in **Figure 1**, to characterize the temporal variability of FC, all the time series were segmented into n nonoverlapping time windows with a length of l. Within each time window, a 116 × 116 pairwise Pearson correlation matrix was calculated to

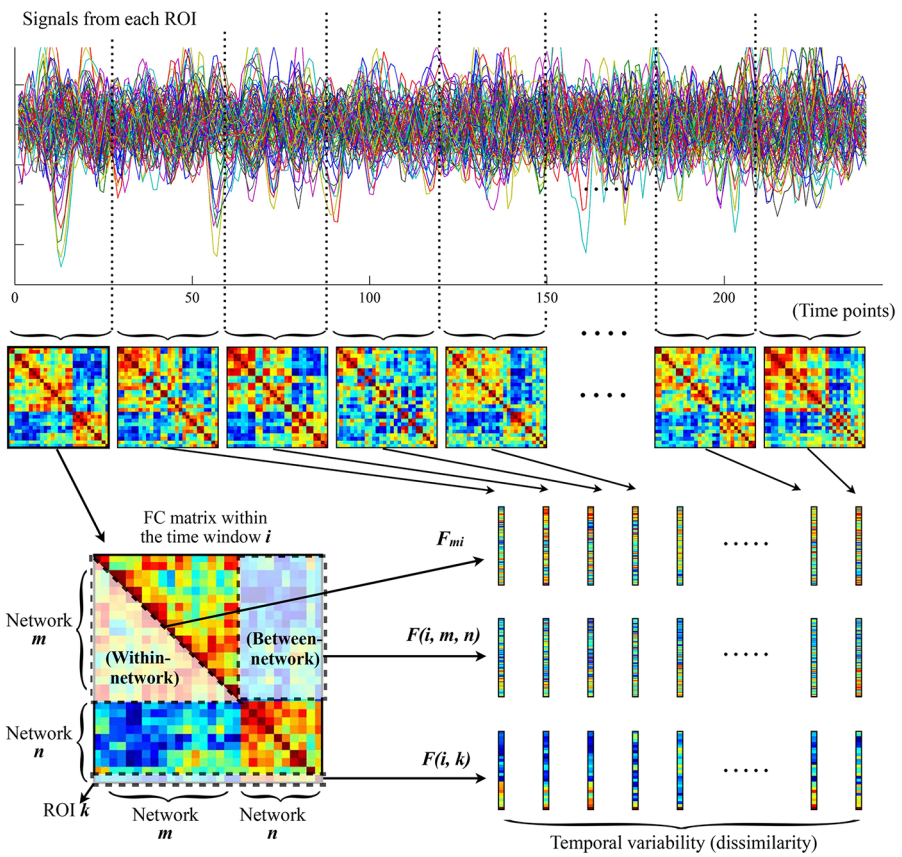


FIGURE 1 | The procedures for computing temporal variabilities of FC patterns. Refer to the section *Temporal Variability of FC* for details. FC, functional connectivity; ROI, region of interest.

represent the FC between each pair of ROIs within that window. The temporal variability of regional FC architecture in each ROI could then be estimated by computing the mean values of its dissimilarities among different windows. Briefly, temporal variability of the regional FC architecture in ROI k is defined by Equation (1):

$$V_k = 1 - \overline{\text{corrcoef}(F(i,k), F(j,k))},\ i, j = 1, 2, 3, \ldots, num;\ i \neq j, \quad (1)$$

where num is the number of time windows, and $F(i, k)$ is the vector characterizing the FC architecture between ROI k and the whole brain within the ith time window (**Figure 1**) (19, 42, 43).

The temporal variability of FC was further estimated at the network level following recently published procedures (23, 24, 44). First, all brain ROIs were assigned into 11 prior networks as defined in previous studies (45, 46), including the sensorimotor network, visual network, auditory network, default-mode network, frontoparietal network, cingulo-opercular network, salience network, attention network, subcortical network, thalamus, and cerebellum (see **Supplementary Table S1** for details about the network assignments). Note that the thalamus and cerebellum were treated as two independent networks here, given that they were poorly defined into different networks as well as their special roles in the pathophysiologic mechanisms of psychotic disorders (23, 47, 48). The temporal variabilities of intra-network and inter-network FC architectures were then calculated among the above 11 networks. Similar with the regional FC variability for each ROI, the intra-network FC variability for a network m is defined by Equation (2):

$$V_m = 1 - \overline{\text{corrcoef}\left(F_{mi}, F_{mj}\right)},\ i, j = 1, 2, 3, \ldots, num;\ i \neq j, \quad (2)$$

where num is the number of time windows, and F_{mi} is the vector characterizing the FC architecture between all ROIs belonging to the network m within the ith time window (**Figure 1**); the inter-network variability of FC between two networks m and n is defined by Equation (3):

$$V_{m-n} = 1 - \overline{\text{corrcoef}(F(i,m,n), \text{F}(j,m\,.\,n))},\ i, j = 1, 2, 3, \ldots, num;\ i \neq j, \quad (3)$$

where num is the number of time windows, and $F(i, m, n)$ is the vector characterizing the FC architecture between the networks m and n within the ith time window (**Figure 1**) (23, 24, 44).

To reduce the influences from window length and segmentation scheme, all the above temporal variabilities were calculated with a set of different window lengths (l = 21, 22, …, 30 volumes, equal to 42, 44, …, 60 seconds); moreover, a total of l – 1 times segmentations were performed for a given window length l, and each time the segmentation was started from the sth time point (s = 1, 2, …, l – 1) (42). The final values of temporal variabilities were obtained by averaging all of these values. Note that such a selection of window lengths has been used in previous studies, and was suggested to be optimal for producing robust results (49, 50). As the result, in each subject, we finally obtained the temporal variabilities of regional FC for each of the 116 ROIs, intra-network FC for each of the 11 networks, and inter-network FCs for each possible pair of networks. All these values of temporal variabilities range from 0 to 2, and a higher value suggests a higher variability.

Statistics

The demographic and clinical characteristics as well as mean FD were compared between groups using the two-sample t-test, Chi-square test or analysis of variance. Differences were considered significant at $p < 0.05$.

The temporal variabilities of FC patterns were compared between groups at all the regional, intra-network, and inter-network levels. The group differences were determined the by following statistic steps (49): 1) the analysis of covariance (ANCOVA) covarying for age, sex, education, and head motion (mean FD) was firstly applied to detect the significant main effect, with no multiple comparison corrections performed here for numbers of ROIs/networks considering the relatively small sample size; 2) *post-hoc* pairwise comparisons were adopted between all possible pairs of groups when the main effect was significant ($p < 0.05$); 3) the Bonferroni correction was applied to control the false-positive rate for multiple tests within the ANCOVA, and the groups differences were considered significant at corrected-$p < 0.05$.

For all the detected significant between-group differences, we further explored their possible relationships with the clinical and cognitive variables using Spearman's rank correlation coefficient. Here, they were correlated with the illness duration, chlorpromazine equivalence, SAPS scores, SANS scores, HAMD scores, YMRS scores, WAIS-I scores, and WAIS-DS scores in both the entire sample and each group separately. The correlations were considered significant at $p < 0.05$.

Supplementary Analyses

A number of supplementary analyses were performed to validate our findings. First, we repeated the whole analyses without the GSR, to see if the results would change without such a controversial preprocessing step. Second, to evaluate whether the results were affected by unmatched clinical characteristics between groups, the whole analyses were repeated within a subset of 48 schizophrenia patients, 30 bipolar disorder patients, and 56 healthy controls, where the illness duration and antipsychotic dosage were matched between the schizophrenia and bipolar disorder groups (see **Supplementary Table S4** for demographic and clinical characteristics of each group in the subset).

RESULTS

Demographic, Clinical, and Head Motion Characteristics

As shown in **Table 1**, there were no significant differences among the three groups in age, sex, and education (all $p > 0.05$). Shorter

illness durations but higher antipsychotic doses (both $p < 0.001$) were observed in the schizophrenia patients compared with the bipolar disorder patients. Both the schizophrenia and bipolar disorder groups showed significantly lower WAIS-I and WAIS-DS scores (all $p < 0.05$, LSD *post-hoc* comparisons) compared with healthy controls, while there was no significant difference between the schizophrenia and bipolar disorder patients in WAIS-I and WAIS-DS scores. There was no significant difference among the three groups in head motion as measured by mean FD ($F = 2.066$, $p = 0.130$).

Differences in Temporal Variability of Regional FC

As shown in **Supplementary Table S2** and **Figure 2**, for temporal variability of the regional FC, both the schizophrenia and bipolar disorder patients showed significantly higher variabilities in a number of subcortical ROIs, including the left thalamus and regions of the basal ganglia (putamen/pallidum) compared with healthy controls; the schizophrenia patients additionally showed significantly higher variabilities for a number of ROIs located in the sensorimotor (precentral gyrus and postcentral gyrus), attention (inferior parietal lobule), and limbic (hippocampus and amygdala) areas than healthy controls, as well as a significantly lower variability in the superior frontal gyrus (medial orbital) than healthy controls and a significantly lower variability in the posterior cingulate gyrus than bipolar disorder patients (all corrected-$p < 0.05$).

Differences in Temporal Variability of Intra- and Inter-Network FC

As shown in **Supplementary Table S3** and **Figure 3**, for temporal variabilities of the intra-network FC within particular networks and inter-network FC between particular pairs of networks, both the schizophrenia and bipolar disorder patients showed a significantly higher variability for inter-network FC between the sensorimotor network and thalamus compared with healthy controls; the schizophrenia patients additionally showed significantly higher variabilities of both intra-network and inter-network FC than healthy controls for several networks and pairs of networks, which mainly involved the sensorimotor, visual, and subcortical (including the thalamus) networks (all corrected $p < 0.05$).

Correlations

As shown in **Figure 4**, in the entire sample, a significant negative correlation was found between temporal variability of inter-network FC between subcortical and auditory networks and the WAIS-DS scores (Spearman's rho = −0.173, $p = 0.019$, uncorrected). Moreover, as shown in **Figure 5**, in the group of schizophrenia patients, significant correlations were found between temporal variability of regional FC for left hippocampus and the SANS scores (Spearman's rho = 0.330, $p = 0.007$, uncorrected for multiple tests), as well as between temporal variability of the inter-network FC between subcortical and auditory networks and the WAIS-I scores (Spearman's rho = −0.286, $p = 0.020$, uncorrected for multiple tests). No significant correlations were found in the groups of healthy controls and bipolar disorder patients ($p > 0.05$).

Supplementary Analyses

As shown in **Supplementary Tables S5–S7**, similar results were obtained when repeating the whole analyses with GSR, and repeating the whole analyses within a subset where the illness duration and antipsychotic dosage were matched between two patient groups: both the schizophrenia and bipolar disorder patients still showed significantly increased regional FC variabilities in subcortical areas, as well as increased inter-network FC variability between the sensorimotor cortices and thalamus; moreover, some specific significant changes were found to present only in the schizophrenia group (although

TABLE 1 | Demographic, clinical, and head motion characteristics of the three groups.

	Schizophrenia ($n = 66$)	Bipolar disorder ($n = 53$)	Healthy controls ($n = 66$)	Group comparisons
	(Mean ± SD)	(Mean ± SD)	(Mean ± SD)	
Age (years)	24.318 ± 6.127	25.340 ± 4.095	23.379 ± 4.416	$F = 2.249$, $p = 0.108$
Sex (male/female)	38/28	26/27	28/38	$\chi^2 = 3.044$, $p = 0.218$
Education (years)	13.152 ± 2.061	13.576 ± 2.578	14.076 ± 2.200	$F = 2.745$, $p = 0.067$
Illness duration (months)	22.214 ± 24.972[a]	56.987 ± 53.907	/	$t = -4.281$, $p < 0.001$
Antipsychotics (taking/not taking)	63/3	38/15	/	$\chi^2 = 12.922$, $p < 0.001$
Chlorpromazine equivalents (mg/day)	228.762 ± 155.296	108.047 ± 119.101	/	$t = 4.663$, $p < 0.001$
SAPS scores	18.231 ± 13.828	/	/	/
SANS scores	31.636 ± 27.810	/	/	/
17-item HAMD scores	/	12.660 ± 9.265	/	/
YMRS scores	/	5.113 ± 8.257	/	/
WAIS-I scores	18.174 ± 4.143	19.123 ± 4.410	20.985 ± 4.774	$F = 6.773$, $p < 0.001$[b]
WAIS-DS scores	65.182 ± 15.493	70.321 ± 14.997	88.924 ± 13.014	$F = 48.263$, $p < 0.001$[b]
Mean FD	0.095 ± 0.038	0.082 ± 0.035	0.086 ± 0.032	$F = 2.066$, $p = 0.130$

[a]*The information on illness duration was available for 56 schizophrenia patients.* [b]*The LSD post-hoc comparisons set at $p < 0.05$ showed that schizophrenia < healthy controls, and bipolar disorder < healthy controls, while there was no significant difference between the schizophrenia and bipolar disorder groups. SD, standard deviation; SAPS, Scale for Assessment of Positive Symptoms; SANS, Scale for Assessment of Negative Symptoms; HAMD, Hamilton Rating Scale for Depression; YMRS, Young Mania Rating Scale; WAIS-I, the Information subtest of the Wechsler Adult Intelligence Scale; WAIS-DS, the Digit Symbol subtest of the Wechsler Adult Intelligence Scale; FD, framewise-displacement.*

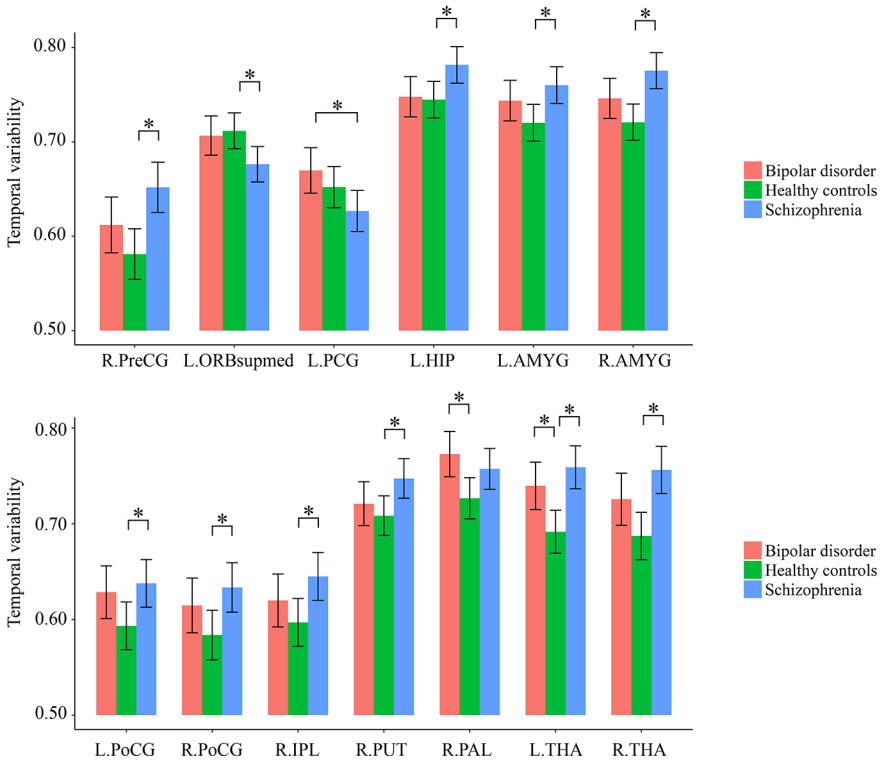


FIGURE 2 | Group differences in the temporal variabilities of regional FC patterns. The error bars present the 95% confidence intervals, and the marks "*" indicate significant between-group differences with corrected $p < 0.05$. AMYG, amygdala; FC, functional connectivity; HIP, hippocampus; IPL, inferior parietal lobule; L, left hemisphere; ORBsupmed, superior frontal gyrus (medial orbital); PAL, pallidum; PCG, posterior cingulate gyrus; PoCG, postcentral gyrus; PreCG, precentral gyrus; PUT, putamen; R, right hemisphere; THA, thalamus.

within the subset, changes in the schizophrenia group were found in a smaller range compared to those found in the entire sample).

DISCUSSION

In this study, we explored the common and specific changes in dynamic local and large-scale resting-state FC, as characterized by altered temporal variabilities, across the schizophrenia and bipolar disorder. Our results provide some innovative findings on the dynamic functional architecture of the brain for these two severe mental disorders: firstly, we found that both the schizophrenia and bipolar disorder patients showed increased regional FC variabilities in a number of subcortical areas involving the thalamus and regions of basal ganglia, as well as increased inter-network FC variability between the sensorimotor cortices and thalamus; secondly, some specific abnormalities were found to present only in the schizophrenia group, at both regional and network levels in a wider range. These findings provide valuable information for improving our insight into the neuropathology of these disorders from a dynamic brain functional perspective.

Our first important finding is that both the schizophrenia and bipolar disorder patients exhibited similar increased temporal variabilities of local FC in the thalamus (**Figure 2**), as well as of inter-network FC between the thalamus and sensorimotor cortices (**Figure 3**). It is noteworthy that shared neural disturbances in thalamo-cortical communications across schizophrenia and bipolar disorder, as characterized by similar over-connectivity between the thalamus and sensorimotor regions, have been repeatedly reported in several previous conventional static rs-fMRI studies (3, 4). Our results, therefore, may extend such findings to the context of dynamic resting-state FC for the first time to our knowledge. The thalamus is known as a "relay station" for almost all motor and sensory information flow from and to the cortex, where the information is further processed for high-order brain functions (3, 51). Specifically, aberrant communications between the thalamus and sensorimotor network were presumed to reflect a sensory gating deficit which leads to abnormal sensory information flow through the thalamus to the cortex (4, 48, 52). The observed increased temporal variability of thalamo-sensorimotor connectivity could thus point to such a sensory gating deficit, as abnormally increased temporal variability of FC was suggested to reflect excessive fluctuations in brain activities and inappropriate processing of information (23). As notably

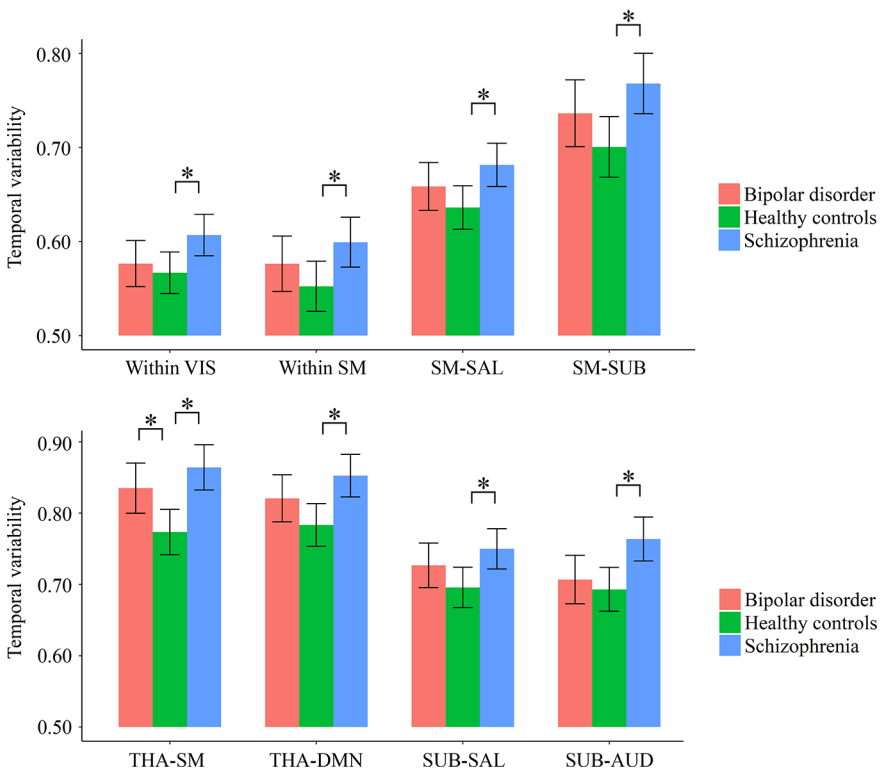


FIGURE 3 | Group differences in the temporal variabilities of intra-network and inter-network FC patterns. The error bars present the 95% confidence intervals, and the marks "*" indicate significant between-group differences with corrected $p < 0.05$. AUD, auditory network; DMN, default mode network; FC, functional connectivity; SAL, salience network; SM, sensorimotor network; SUB, subcortical network; THA, thalamus; VIS, visual network.

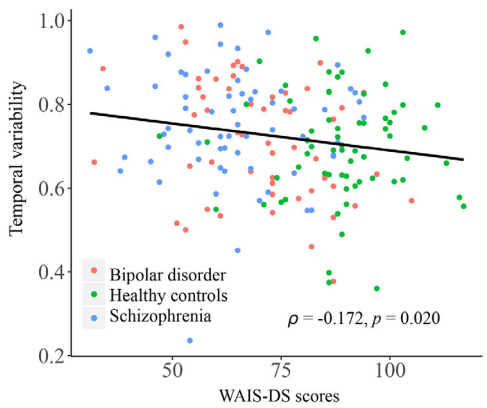


FIGURE 4 | Correlation between the temporal variability of inter-network FC between subcortical and auditory networks and the WAIS-DS scores in the entire sample. The Spearman's correlation coefficients (ρ) and p values are presented on figures. FC, functional connectivity; WAIS-DS, Digit Symbol Subtest of the Wechsler Adult Intelligence Scale.

reported in both the schizophrenia and bipolar disorder patients (53–55), the sensory gating deficit has been suggested to partly underlie the cognitive and perceptual symptoms in the disorders (3, 56). Therefore, our dynamic FC findings may further support the hypothesis that thalamo-sensorimotor connectivity disturbances and sensory gating deficits are common neurobiological features shared by schizophrenia and bipolar disorder (4, 53).

In the present study, we also found that both the schizophrenia and bipolar disorder patients showed increased local FC variability in regions of the basal ganglia (putamen and pallidum) (**Figure 2**). The basal ganglia is a group of subcortical nuclei (putamen, pallidum, caudate nucleus, substantia nigra, and subthalamic nucleus) that involves a variety of brain functions such as motor control, learning, and execution (57). The functional and structural abnormalities of basal ganglia have been widely reported to be associated with psychotic symptoms such as delusions in schizophrenia patients (58–60), and also present in psychotic bipolar disorder patients (61). Therefore, our findings of such shared alterations in the basal ganglia may be reflective of common functional deficits in both the schizophrenia and bipolar disorder. These findings, together with the observed shared

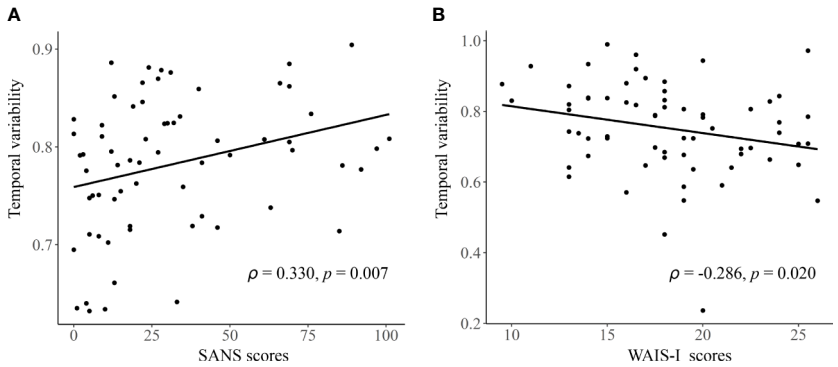


FIGURE 5 | The detected significant correlations in schizophrenia patients. **(A)** Correlation between the temporal variability of regional FC for left hippocampus and the SANS scores. **(B)** Correlation between the temporal variability of inter-network FC between subcortical and auditory networks and the WAIS-I scores. The Spearman's correlation coefficients (ρ) and p values are presented on figures. FC, functional connectivity; SANS, Scale for Assessment of Negative Symptoms; WAIS-I, Information Subtest of the Wechsler Adult Intelligence Scale.

alterations in the thalamo-sensorimotor circuit, may partly help to explain the overlap clinical features in these two disorders.

Besides the above shared alterations in both patient groups, some specific alterations in a much wider range were found to present in only the schizophrenia patients. These include widespread increased FC variabilities at both regional and network levels, involving the sensorimotor, visual, attention, limbic, and subcortical areas, as well as decreased regional FC variability in a number of areas comprising the default-mode network such as posterior cingulate gyrus and superior frontal gyrus (medial orbital part) (**Figures 2** and **3**). Generally, these results are highly consistent with the findings from another recent study (23), which reported that schizophrenia patients had significantly increased FC variabilities in sensory and perceptual systems (including the sensorimotor network, visual network, attention network, and thalamus) and decreased FC variabilities in high-order networks (including the default-mode and frontal–parietal networks) than healthy subjects at both regional and network levels. Moreover, these alterations were found to be related to patients' clinical symptoms and cognitive deficits (or relations found in the entire sample) both in the present study (**Figures 4** and **5**) and prior research (23). Therefore, our results further support the recent opinion that such widespread aberrant dynamic brain network reconfigurations may constitute a potential reliable biomarker for schizophrenia, suggestive of impaired abilities in processing inputs in sensory/perceptual systems and integrating information in high-order networks, which may underlie the perceptual and cognitive deficits in schizophrenia (23, 62). As for the bipolar disorder patients in the present study, FC variabilities in these regions and networks did not differ significantly from either of the other groups, which fell in the intermediate range between those of healthy controls and schizophrenia patients (**Figures 2** and **3**). Thus, we propose that our findings may offer support for the hypothesis of a psychosis continuum between schizophrenia and bipolar disorder, with more severe brain deficits and disabling symptoms in schizophrenia compared to bipolar disorder (63, 64); moreover, changes in dynamic FC may serve as objective biomarkers for such differences in neuropathology between these two disorders. However, future investigation with a larger sample size and a higher statistical power is required to confirm if these changes would be significant in patients with bipolar disorder, as compared to healthy controls and schizophrenia patients.

There are several issues for the present study and future research directions which should be noted. First, as mentioned before, our sample size is relatively small and the results should be further verified in future work with a larger sample to increase the reliability and statistical power (65). Second, the illness duration and doses of antipsychotics were not matched between the schizophrenia and bipolar disorder groups; moreover, the records of illness episode and age of onset were unavailable for patients. For better excluding the confounding effects of clinical distinctions on our findings, we have repeated the analyses in a subset of our sample in which the illness duration and antipsychotic dosage were matched between two patient groups and found similar results, suggesting that the observed group differences are unlikely to be mainly driven by medications or long-term hospitalizations. Despite this, however, further studies using first-episode, drug-naïve samples are warranted to exclude possible effects of all these clinical factors. Third, a number of previous studies have pointed out that the psychotic bipolar disorder may be a special phenotype from non-psychotic bipolar disorder (66, 67). In the current sample, the records of psychotic symptom histories are unavailable for most bipolar disorder patients. Future studies are necessary to replicate our results and to compare between psychotic and non-psychotic bipolar disorder patients. Fourth, although it has been proven to be reliable for analyzing the intra- and inter-network dynamic FC (39), the AAL atlas only provides a relatively coarse parcellation. However, some key brain structures such as the thalamus could be subdivided into more precise subregions with different FC patterns (47, 68). Thus, future studies to investigate the temporal variability of dynamic FC with finer parcellation schemes would further improve our understanding of its important role in differentiating schizophrenia and bipolar disorder.

In conclusion, we explored the common and specific changes in dynamic features of FC, as characterized by temporal variabilities of FC patterns involved in specific brain regions or large-scale brain networks, in schizophrenia and bipolar disorder patients. We found that both the schizophrenia and bipolar disorder patients showed significantly increased regional FC variabilities in subcortical areas including the thalamus and basal ganglia, as well as increased inter-network FC variability between the sensorimotor cortices and thalamus. More widespread significant alterations were found to present in only the schizophrenia group, including increased FC variabilities in the sensorimotor, visual, attention, limbic, and subcortical areas at both regional and network levels, as well as decreased regional FC variability in the default-mode areas. The observed alterations shared by schizophrenia and bipolar disorder may help to explain their overlap clinical features; meanwhile, the schizophrenia-specific abnormalities in a wider range could potentially support the hypothesis of a psychosis continuum between schizophrenia and bipolar disorder, that schizophrenia is associated with more severe functional brain deficits compared to bipolar disorder. Together, these findings highlight the potentials of using dynamic FC as an objective biomarker for the monitoring and diagnosis of either schizophrenia or bipolar disorder.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Second Xiangya Hospital, Central South University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Authors YL and WP designed the study and carried out the analysis. YL, ZL, GW, ZX, YP, XC, XH, and WP contributed to the data collection. YL wrote the first draft of manuscript. ZL, CC, DL, and WP contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the China Precision Medicine Initiative (grant number 2016YFC0906300) and the National Natural Science Foundation of China (grant numbers 81561168021, 81671335, 81701325).

ACKNOWLEDGMENTS

We thank Zhong He (Department of Radiology of Second Xiangya Hospital, Central South University) for his assistance in rs-fMRI data acquisition. This manuscript has been released as a pre-print at bioRxiv (69).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.00422/full#supplementary-material

REFERENCES

1. Meyer F, Meyer TD. The misdiagnosis of bipolar disorder as a psychotic disorder: Some of its causes and their influence on therapy. *J Affect Disord* (2009) 112:174–83. doi: 10.1016/j.jad.2008.04.022
2. Dezhina Z, Ranlund S, Kyriakopoulos M, Williams SCR, Dima D. A systematic review of associations between functional MRI activity and polygenic risk for schizophrenia and bipolar disorder. *Brain Imaging Behav* (2019) 13:862–77. doi: 10.1007/s11682-018-9879-z
3. Skåtun KC, Kaufmann T, Brandt CL, Doan NT, Alnæs D, Tønnesen S, et al. Thalamo-cortical functional connectivity in schizophrenia and bipolar disorder. *Brain Imaging Behav* (2018) 12:640–52. doi: 10.1007/s11682-017-9714-y
4. Anticevic A, Cole MW, Repovs G, Murray JD, Brumbaugh MS, Winkler AM, et al. Characterizing thalamo-cortical disturbances in Schizophrenia and bipolar illness. *Cereb Cortex* (2014) 24:3116–30. doi: 10.1093/cercor/bht165
5. Meda SA, Gill A, Stevens MC, Lorenzoni RP, Glahn DC, Calhoun VD, et al. Differences in resting-state functional magnetic resonance imaging functional network connectivity between schizophrenia and psychotic bipolar probands and their unaffected first-degree relatives. *Biol Psychiatry* (2012) 71:881–9. doi: 10.1016/j.biopsych.2012.01.025
6. Birur B, Kraguljac NV, Shelton RC, Lahti AC. Brain structure, function, and neurochemistry in schizophrenia and bipolar disorder- A systematic review of the magnetic resonance neuroimaging literature. *NPJ Schizophr* (2017) 3:1–15. doi: 10.1038/s41537-017-0013-9
7. Mamah D, Barch DM, Repovš G. Resting state functional connectivity of five neural networks in bipolar disorder and schizophrenia. *J Affect Disord* (2013) 150:601–9. doi: 10.1016/j.jad.2013.01.051
8. Cheng W, Palaniyappan L, Li M, Kendrick KM, Zhang J, Luo Q, et al. Voxel-based, brain-wide association study of aberrant functional connectivity in schizophrenia implicates thalamocortical circuitry. *NPJ Schizophr* (2015) 1:15016. doi: 10.1038/npjschz.2015.16
9. Chang C, Glover GH. Time-frequency dynamics of resting-state brain connectivity measured with fMRI. *Neuroimage* (2010) 50:81–98. doi: 10.1016/j.neuroimage.2009.12.011
10. Hutchison RM, Womelsdorf T, Gati JS, Everling S, Menon RS. Resting-state networks show dynamic functional connectivity in awake humans and anesthetized macaques. *Hum Brain Mapp* (2013) 34:2154–77. doi: 10.1002/hbm.22058
11. Preti MG, Bolton TA, Van De Ville D. The dynamic functional connectome: State-of-the-art and perspectives. *Neuroimage* (2017) 160:41–54. doi: 10.1016/j.neuroimage.2016.12.061
12. Bassett DS, Wymbs NF, Porter MA, Mucha PJ, Carlson JM, Grafton ST. Dynamic reconfiguration of human brain networks during learning. *Proc Natl Acad Sci U S A* (2011) 108:7641–6. doi: 10.1073/pnas.1018985108

13. Braun U, Schäfer A, Walter H, Erk S, Romanczuk-Seiferth N, Haddad L, et al. Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proc Natl Acad Sci U S A* (2015) 112:11678–83. doi: 10.1073/pnas.1422487112
14. Long Y, Chen C, Deng M, Huang X, Tan W, Zhang L, et al. Psychological resilience negatively correlates with resting-state brain network flexibility in young healthy adults: a dynamic functional magnetic resonance imaging study. *Ann Transl Med* (2019) 7:809–9. doi: 10.21037/atm.2019.12.45
15. Betzel RF, Satterthwaite TD, Gold JI, Bassett DS. Positive affect, surprise, and fatigue are correlates of network flexibility. *Sci Rep* (2017) 7:1–10. doi: 10.1038/s41598-017-00425-z
16. Harlalka V, Bapi RS, Vinod PK, Roy D. Atypical flexibility in dynamic functional connectivity quantifies the severity in autism spectrum disorder. *Front Hum Neurosci* (2019) 13:6. doi: 10.3389/fnhum.2019.00006
17. Schumacher J, Peraza LR, Firbank M, Thomas AJ, Kaiser M, Gallagher P, et al. Dynamic functional connectivity changes in dementia with Lewy bodies and Alzheimer's disease. *NeuroImage Clin* (2019) 22:101812. doi: 10.1016/j.nicl.2019.101812
18. Wise T, Marwood L, Perkins AM, Herane-Vives A, Joules R, Lythgoe DJ, et al. Instability of default mode network connectivity in major depression: A two-sample confirmation study. *Transl Psychiatry* (2017) 7:e1105. doi: 10.1038/tp.2017.40
19. Long Y, Cao H, Yan C, Chen X, Li L, Castellanos FX, et al. Altered resting-state dynamic functional brain networks in major depressive disorder: Findings from the REST-meta-MDD consortium. *NeuroImage Clin* (2020). doi: 10.1016/j.nicl.2020.102163
20. Rashid B, Arbabshirani MR, Damaraju E, Cetin MS, Miller R, Pearlson GD, et al. Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. *Neuroimage* (2016) 134:645–57. doi: 10.1016/j.neuroimage.2016.04.051
21. Rashid B, Damaraju E, Pearlson GD, Calhoun VD. Dynamic connectivity states estimated from resting fMRI Identify differences among Schizophrenia, bipolar disorder, and healthy control subjects. *Front Hum Neurosci* (2014) 8:897. doi: 10.3389/fnhum.2014.00897
22. Du Y, Pearlson GD, Lin D, Sui J, Chen J, Salman M, et al. Identifying dynamic functional connectivity biomarkers using GIG-ICA: Application to schizophrenia, schizoaffective disorder, and psychotic bipolar disorder. *Hum Brain Mapp* (2017) 38:2683–708. doi: 10.1002/hbm.23553
23. Dong D, Duan M, Wang Y, Zhang X, Jia X, Li Y, et al. Reconfiguration of Dynamic Functional Connectivity in Sensory and Perceptual System in Schizophrenia. *Cereb Cortex* (2019) 29:3577–89. doi: 10.1093/cercor/bhy232
24. Zhu H, Huang J, Deng L, He N, Cheng L, Shu P, et al. Abnormal dynamic functional connectivity associated with subcortical networks in Parkinson's disease: A temporal variability perspective. *Front Neurosci* (2019) 13:80. doi: 10.3389/fnins.2019.00080
25. Andreasen NC. Methods for assessing positive and negative symptoms. *Mod Probl Pharmacopsychiatry* (1990) 24:73–88. doi: 10.1159/000418013
26. Williams JBW. A Structured Interview Guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry* (1988) 45:742–7. doi: 10.1001/archpsyc.1988.01800320058007
27. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: Reliability, validity and sensitivity. *Br J Psychiatry* (1978) 133:429–35. doi: 10.1192/bjp.133.5.429
28. Andreasen NC, Pressler M, Nopoulos P, Miller D, Ho BC. Antipsychotic Dose Equivalents and Dose-Years: A Standardized Method for Comparing Exposure to Different Drugs. *Biol Psychiatry* (2010) 67:255–62. doi: 10.1016/j.biopsych.2009.08.040
29. Yao-xian G. Revision of Wechsler's Adult Intelligence Scale in China. *Acta Psychol Sin* (1983) 15:121–9. Available at: http://118.145.16.229:81/Jweb_xlxb.
30. Long Y, Ouyang X, Liu Z, Chen X, Hu X, Lee E, et al. Associations among suicidal ideation, white matter integrity and cognitive deficit in first-episode schizophrenia. *Front Psychiatry* (2018) 9:391. doi: 10.3389/fpsyt.2018.00391
31. Deng M, Pan Y, Zhou L, Chen X, Liu C, Huang X, et al. Resilience and Cognitive Function in Patients With Schizophrenia and Bipolar Disorder, and Healthy Controls. *Front Psychiatry* (2018) 9:279. doi: 10.3389/fpsyt.2018.00279
32. Chao-Gan Y, Yu-Feng Z. DPARSF: A MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. *Front Syst Neurosci* (2010) 4:13. doi: 10.3389/fnsys.2010.00013
33. Yan CG, Wang XD, Zuo XN, Zang YF. DPABI: Data Processing & Analysis for (Resting-State) Brain Imaging. *Neuroinformatics* (2016) 14:339–51. doi: 10.1007/s12021-016-9299-4
34. Friston KJ, Williams S, Howard R, Frackowiak RSJ, Turner R. Movement-related effects in fMRI time-series. *Magn Reson Med* (1996) 35:346–55. doi: 10.1002/mrm.1910350312
35. Murphy K, Fox MD. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage* (2017) 154:169–73. doi: 10.1016/j.neuroimage.2016.11.052
36. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* (2002) 17:825–41. doi: 10.1016/S1053-8119(02)91132-8
37. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* (2002) 15:273–89. doi: 10.1006/nimg.2001.0978
38. Termenon M, Jaillard A, Delon-Martin C, Achard S. Reliability of graph analysis of resting state fMRI using test-retest dataset from the Human Connectome Project. *Neuroimage* (2016) 142:172–87. doi: 10.1016/j.neuroimage.2016.05.062
39. Zhang C, Baum SA, Adduru VR, Biswal BB, Michael AM. Test-retest reliability of dynamic functional connectivity in resting state fMRI. *Neuroimage* (2018) 183:907–18. doi: 10.1016/j.neuroimage.2018.08.021
40. Cao H, Bertolino A, Walter H, Schneider M, Schafer A, Taurisano P, et al. Altered functional subnetwork during emotional face processing a potential intermediate phenotype for schizophrenia. *JAMA Psychiatry* (2016) 73:598–605. doi: 10.1001/jamapsychiatry.2016.0161
41. Piel JH, Lett TA, Wackerhagen C, Plichta MM, Mohnke S, Grimm O, et al. The effect of 5-HTTLPR and a serotonergic multi-marker score on amygdala, prefrontal and anterior cingulate cortex reactivity and habituation in a large, healthy fMRI cohort. *Eur Neuropsychopharmacol* (2018) 28:415–27. doi: 10.1016/j.euroneuro.2017.12.014
42. Zhang J, Cheng W, Liu Z, Zhang K, Lei X, Yao Y, et al. Neural, electrophysiological and anatomical basis of brain-network variability and its characteristic changes in mental disorders. *Brain* (2016) 139:2307–21. doi: 10.1093/brain/aww143
43. Mueller S, Wang D, Fox MD, Yeo BTT, Sepulcre J, Sabuncu MR, et al. Individual Variability in Functional Connectivity Architecture of the Human Brain. *Neuron* (2013) 77:586–95. doi: 10.1016/j.neuron.2012.12.028
44. Sun J, Liu Z, Rolls ET, Chen Q, Yao Y, Yang W, et al. Verbal creativity correlates with the temporal variability of brain networks during the resting state. *Cereb Cortex* (2019) 29:1047–58. doi: 10.1093/cercor/bhy010
45. Cao H, Chung Y, McEwen SC, Bearden CE, Addington J, Goodyear B, et al. Progressive reconfiguration of resting-state brain networks as psychosis develops: Preliminary results from the North American Prodrome Longitudinal Study (NAPLS) consortium. *Schizophr Res* (2019). doi: 10.1016/j.schres.2019.01.017
46. Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, et al. Functional Network Organization of the Human Brain. *Neuron* (2011) 72:665–78. doi: 10.1016/j.neuron.2011.09.006
47. Ji JL, Spronk M, Kulkarni K, Repovš G, Anticevic A, Cole MW. Mapping the human brain's cortical-subcortical functional network organization. *Neuroimage* (2019) 185:35–57. doi: 10.1016/j.neuroimage.2018.10.006
48. Cao H, Chén OY, Chung Y, Forsyth JK, McEwen SC, Gee DG, et al. Cerebello-thalamo-cortical hyperconnectivity as a state-independent functional neural signature for psychosis prediction and characterization. *Nat Commun* (2018) 9:1–9. doi: 10.1038/s41467-018-06350-7
49. Hou Z, Kong Y, He X, Yin Y, Zhang Y, Yuan Y. Increased temporal variability of striatum region facilitating the early antidepressant response in patients with major depressive disorder. *Prog Neuropsychopharmacol Biol Psychiatry* (2018) 85:39–45. doi: 10.1016/j.pnpbp.2018.03.026
50. Shirer WR, Ryali S, Rykhlevskaia E, Menon V, Greicius MD. Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb Cortex* (2012) 22:158–65. doi: 10.1093/cercor/bhr099

51. Sherman SM, Guillery RW. The role of the thalamus in the flow of information to the cortex. *Philos Trans R Soc B: Biol Sci* (2002) 357:1695–708. doi: 10.1098/rstb.2002.1161
52. Çetin MS, Christensen F, Abbott CC, Stephen JM, Mayer AR, Cañive JM, et al. Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia. *Neuroimage* (2014) 97:117–26. doi: 10.1016/j.neuroimage.2014.04.009
53. Sánchez-Morla EM, García-Jiménez MA, Barabash A, Martínez-Vizcaíno V, Mena J, Cabranes-Díaz JA, et al. P50 sensory gating deficit is a common marker of vulnerability to bipolar disorder and schizophrenia. *Acta Psychiatr Scand* (2008) 117:313–8. doi: 10.1111/j.1600-0447.2007.01141.x
54. Cheng CH, Chan PYS, Liu CY, Hsu SC. Auditory sensory gating in patients with bipolar disorders: A meta-analysis. *J Affect Disord* (2016) 203:199–203. doi: 10.1016/j.jad.2016.06.010
55. Hazlett EA, Rothstein EG, Ferreira R, Silverman JM, Siever LJ, Olincy A. Sensory gating disturbances in the spectrum: Similarities and differences in schizotypal personality disorder and schizophrenia. *Schizophr Res* (2015) 161:283–90. doi: 10.1016/j.schres.2014.11.020
56. Cromwell HC, Mears RP, Wan L, Boutros NN. Sensory gating: A translational effort from basic to clinical science. *Clin EEG Neurosci* (2008) 39:69–72. doi: 10.1177/155005940803900209
57. Haber SN. The primate basal ganglia: Parallel and integrative networks. *J Chem Neuroanat* (2003) 26:317–30. doi: 10.1016/j.jchemneu.2003.10.003
58. Huang X, Pu W, Li X, Greenshaw AJ, Dursun SM, Xue Z, et al. Decreased left putamen and thalamus volume correlates with delusions in first-episode schizophrenia patients. *Front Psychiatry* (2017) 8:245. doi: 10.3389/fpsyt.2017.00245
59. Raij TT, Mäntylä T, Kieseppä T, Suvisaari J. Aberrant functioning of the putamen links delusions, antipsychotic drug dose, and compromised connectivity in first episode psychosis-Preliminary fMRI findings. *Psychiatry Res - Neuroimaging* (2015) 233:201–11. doi: 10.1016/j.pscychresns.2015.06.008
60. Tao H, Wong GHY, Zhang H, Zhou Y, Xue Z, Shan B, et al. Grey matter morphological anomalies in the caudate head in first-episode psychosis patients with delusions of reference. *Psychiatry Res - Neuroimaging* (2015) 233:57–63. doi: 10.1016/j.pscychresns.2015.04.011
61. Karcher NR, Rogers BP, Woodward ND. Functional Connectivity of the Striatum in Schizophrenia and Psychotic Bipolar Disorder. *Biol Psychiatry Cognit Neurosci Neuroimaging* (2019) 4:956–65. doi: 10.1016/j.bpsc.2019.05.017
62. Zhang Y, Guo G, Tian Y. Increased temporal dynamics of intrinsic brain activity in sensory and perceptual network of schizophrenia. *Front Psychiatry* (2019) 10:484. doi: 10.3389/fpsyt.2019.00484
63. Pearlson GD. Etiologic, Phenomenologic, and Endophenotypic Overlap of Schizophrenia and Bipolar Disorder. *Annu Rev Clin Psychol* (2015) 11:251–81. doi: 10.1146/annurev-clinpsy-032814-112915
64. Sorella S, Lapomarda G, Messina I, Frederickson JJ, Siugzdaite R, Job R, et al. Testing the expanded continuum hypothesis of schizophrenia and bipolar disorder. Neural and psychological evidence for shared and distinct mechanisms. *NeuroImage Clin* (2019) 23:101854. doi: 10.1016/j.nicl.2019.101854
65. Cao H, McEwen SC, Forsyth JK, Gee DG, Bearden CE, Addington J, et al. Toward leveraging human connectomic data in large consortia: Generalizability of fmri-based brain graphs across sites, sessions, and paradigms. *Cereb Cortex* (2019) 29:1263–79. doi: 10.1093/cercor/bhy032
66. Glahn DC, Bearden CE, Barguil M, Barrett J, Reichenberg A, Bowden CL, et al. The Neurocognitive Signature of Psychotic Bipolar Disorder. *Biol Psychiatry* (2007) 62:910–6. doi: 10.1016/j.biopsych.2007.02.001
67. Mazzarini L, Colom F, Pacchiarotti I, Nivoli AMA, Murru A, Bonnin CM, et al. Psychotic versus non-psychotic bipolar II disorder. *J Affect Disord* (2010) 126:55–60. doi: 10.1016/j.jad.2010.03.028
68. Hua J, Blair NIS, Paez A, Choe A, Barber AD, Brandt A, et al. Altered functional connectivity between sub-regions in the thalamus and cortex in schizophrenia patients measured by resting state BOLD fMRI at 7T. *Schizophr Res* (2019) 206:370–7. doi: 10.1016/j.schres.2018.10.016
69. Long Y, Liu Z, Chan CK, Wu G, Xue Z, Pan Y, et al. Altered Temporal Variability of Local and Large-scale Resting-state Brain Functional Connectivity Patterns in Schizophrenia and Bipolar Disorder. *bioRxiv* (2020). doi: 10.1101/2020.02.04.934638

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 14 August 2020
doi: 10.3389/fpsyt.2020.00830

Depressive Emotion Detection and Behavior Analysis of Men Who Have Sex With Men *via* Social Media

Yong Li[1†], *Mengsi Cai*[2,3†], *Shuo Qin*[4] *and Xin Lu*[2,5*]

[1] College of Economy and Management, Changsha University, Changsha, China, [2] College of Systems Engineering, National University of Defense Technology, Changsha, China, [3] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands, [4] National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu, China, [5] School of Business, Central South University, Changsha, China

OPEN ACCESS

Edited by:
Qiang Luo,
Fudan University, China

Reviewed by:
Linyuan Lu,
University of Electronic Science and Technology of China, China
Hongfei Lin,
Dalian University of Technology, China

***Correspondence:**
Xin Lu
xin.lu@flowminder.org

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: 05 April 2020
Accepted: 30 July 2020
Published: 14 August 2020

Citation:
Li Y, Cai M, Qin S and Lu X (2020) Depressive Emotion Detection and Behavior Analysis of Men Who Have Sex With Men via Social Media. Front. Psychiatry 11:830. doi: 10.3389/fpsyt.2020.00830

Background: A large amount of evidence has indicated an association between depression and HIV risk among men who have sex with men (MSM), but traditional questionnaire-based methods are limited in timely monitoring depressive emotions with large sample sizes. With the development of social media and machine learning techniques, MSM depression can be well monitored in an online and easy-to-use manner. Thereby, we adopt a machine learning algorithm for MSM depressive emotion detection and behavior analysis with online social networking data.

Methods: A large-scale MSM data set including 664,335 users and over 12 million posts was collected from the most popular MSM-oriented geosocial networking mobile application named Blued. Also, a non-MSM Benchmark data set from Twitter was used. After data preprocessing and feature extraction of these two data sets, a machine learning algorithm named XGBoost was adopted for detecting depressive emotions.

Results: The algorithm shows good performance in the Blued and Twitter data sets. And three extracted features significantly affecting the depressive emotion detection were found, including depressive words, LDA topic words, and post-time distribution. On the one hand, the MSM with depressive emotions published posts with more depressive words, negative words and positive words than the MSM without depressive emotions. On the other hand, in comparison with the non-MSM with depressive emotions, the MSM with depressive emotions showed more significant depressive symptoms, such as insomnia, depressive mood, and suicidal thoughts.

Conclusions: The online MSM depressive emotion detection using machine learning can provide a proper and easy-to-use way in real-world applications, which help identify high-risk individuals at the early stage of depression for further diagnosis.

Keywords: depressive emotion detection, men who have sex with men, behavior analysis, Blued, Twitter

INTRODUCTION

Depression is a prevalent but potentially treatable health problem and is a leading cause of disability worldwide, with more than 264 million people affected (1). Depressed people have various symptoms of depression manifested by distinguishing behaviors, such as persistent sadness, loss of interest, changes in appetite, low concentration, sleep problems, feelings of guilt or hopelessness, decreased energy, suicidal thoughts, etc. This emotional disturbance not only affects daily functions but also increases the global burden.

Men who have sex with men (MSM) are disproportionately affected by and infected with HIV/AIDS (2). Suffering from disproportionate rates of stigma and discrimination (3, 4), MSM experience remarkably poorer mental health (5). And evidence has demonstrated that HIV risk (6–9), amphetamine-type-stimulants (ATS) use and homosexuality-related stigma (10), and sexual risk behaviors (11, 12) are associated with depression among MSM population. Therefore, it is of crucial importance to develop reliable and efficient methods for detection and early warning of the depressive emotions or mental health among MSM population, and enable those suffering from depression to be more proactive about their mental health.

Current scales to assess depression have included: *Beck's Depression Inventory (BDI)* (13) with 21 questions about users' mental and physiological states; *Beck Depression Inventory Fast-Screen (BDI-FS)* (14), *CES-D Scale* (15) with 20 questions about mental conditions such as users' feelings of guilt and sleep conditions; *the Patient Health Questionnaire-9 (PHQ-9)* (16) with nine questions about depressive symptoms, such as little interest and suicidal thoughts. In addition, the *Diagnostic and Statistical Manual of Mental Disorders (DSM)* (17) provides standard criterion for diagnosing depression and describes nine kinds of depressive indicators, such as depressed mood and diminished interest, etc. Although these methods have achieved effectiveness for depression diagnosis, the aforementioned criteria may not comprehensively cover newly emerging depressive behaviors and symptoms as the symptoms of depressive disorders evolve over time. n addition, people are somehow ashamed or unaware of depression such that more than 70% of people in the early stage of depression do not consult psychological doctors, and their conditions were deteriorated (18).

On the other hand, social media platforms have become an indispensable part of everyday life for MSM and others. Studies have revealed that it is possible to monitor mental health and predict depression *via* social media (19, 20). Automated analysis of social media potentially provides methods for early detection of depression (21). As a preliminary research, Park et al. (22) explored the use of language in describing depressive moods using real-time moods captured from Twitter users; they then explored the depressive behaviors of 14 active Twitter users through face-to-face interviews (23). Predictive models, which use features or variables that have been extracted from social media data, have achieved many successes in identifying depression *via* social media. For example, commonly extracted and used features include user profile, linguistic signals, sentiment, social network, social interaction, domain-specific features, time of posts, and so on (24–27). Features are then treated as independent variables in an algorithm, such as support vector machine (SVM) (28, 29), naive Bayes methods (30), random forest (31), multi-model depressive dictionary learning model (MDDL) (18), etc. Deep learning models, which have recently achieved outstanding results in many classification tasks due to their ability to learn complex non-linear functions, have also been applied in detecting depression by feeding n-gram features or word embedding of social media data as model input (32, 33). However, deep learning models generally lack interpretability and transparency in the feature extraction and decision processes, limiting their value in clinical diagnosis (34, 35). In addition to using constituent learning algorithms alone, ensemble methods, which can obtain better and consistent predictive performance, have attracted much attention in the community. For example, the Inverse Boosting Pruning Trees (IBPT), based on the AdaBoost algorithm, demonstrated a strong depression classification ability among Twitter users (25).

With massive user-generated content, social media provide a means for capturing depressive behavioral attributes relevant to an individual's thinking, mood, communication, activities and socialization (36). The emotion and language used in social media postings may indicate feelings of worthlessness, guilt, helplessness, and self-hatred that characterize depression or high risk of depression (29). The analysis of social networking data, thus, potentially provides methods for early detection of depression (21). Therefore, in this study, we aim to predict multidimensional depressive emotions among MSM through social media with machine learning methods, which can achieve early detection of depressive emotions, and then complement and extend traditional approaches to diagnose depression. To predict depressive emotions for depression detection, we collect massive social networking data from the most popular MSM-oriented geosocial networking mobile application named Blued, and adopt a decision tree-based ensemble machine learning algorithm named XGBoost, which has been widely applied in various classification tasks with outstanding performance and can provide explainable results. In addition, we comprehensively evaluate the classification capability of XGBoost algorithm on a publicly accessible Twitter data set (18), and compare the differences in online behaviors between MSM with depressive emotions (depressed MSM) and MSM without depressive emotions (non-depressed MSM) and the differences between depressed MSM and non-MSM with depressive emotions (depressed non-MSM).

MATERIALS AND METHODS

Data Sources

Regarding MSM and non-MSM populations, we constructed two main databases: Blued database and Twitter database, representing MSM users and non-MSM users, respectively. Blued is the largest gay social networking application in the world with over 40 million registered users. Twitter is a popular social networking platform with over 300 million active users per

month and 500 million tweets per day. Blued is mainly used for seeking partners and dating, and Twitter is mainly used for social purposes and information consumption.

To predict depressive emotions (depressed or non-depressed), we need to get a batch of labeled depression and non-depression data sets to train the classification model. Also, to expand the size of depressed users, we construct an unlabeled data set named depression-candidate data set for depressive emotion detection and for behavior analysis. For each user, we obtain the profile information and an anchor post to infer the mental state. An anchor post is a post that can help identify whether or not a user has severe depressive emotions which indicate depression. Considering that each user might have published a number of posts related to depression identification, we select only one post as the final anchor post to construct our data sets. For depressed users, we select their first anchor post to ensure timely depressive emotion detection; and for depression candidates, we select their last anchor post to cover more posts for depressive emotion detection. According to clinical experience, an observation period is required for depression diagnosis; therefore, users' posts published within one or two months before the anchor post were also obtained. Finally, our data sets contain the users' profile information, an anchor post for each user, and all the other posts published within one or two months before the anchor post.

Blued Database

Targeting the MSM community, we designed elaborate crawlers *via* Scrapy, a fast web-crawling framework, to collect public social networking records from Blued. We collected about 13 million posts on Blued, concerning 664,335 users (a sub sample from Blued). These posts were published from 1st January 2012 to 31st March 2019 and saved in a local MongoDB database. To make depressive emotion detection, the Blued database was divided into three parts:

(1) Depression Data Set B_D1: Users were labeled as depressed if their anchor posts indicated that they were diagnosed with depression, and they had suicidal thoughts, e.g. "I have depression; I want to suicide!" Posts published within two months before the anchor post of the users were included in this data set.

(2) Non-Depression Data Set B_D2: Users were labeled as non-depressed if they had never posted any posts containing the character string "depress." As the Blued data set collected in this paper contains 664,335 active users and about 13 million posts, we select the posts published from 1st February 2019 to 31st March 2019.

(3) Depression-candidate Data Set B_D3: To expand the sample size of depressed MSM users, we constructed an unlabeled large data set. Users were identified as depression candidates if their anchor posts loosely contained the character string "depress." Posts that published within two months before the anchor post of the users were included in this data set.

Twitter Database

To get non-MSM samples for comparison, we obtained a publicly available large-scale benchmark Twitter database (18, 37), which is also used in this paper to validate the performance of the XGBoost algorithm. The Twitter database contains three data sets:

(1) Depression Data Set T_D1: Users were labeled as depression if their anchor tweets satisfied a strict pattern "(I'm/I was/I am/I've been) diagnosed depression." Tweets published within one month before the anchor tweet were included in this data set.

(2) Non-Depression Data Set T_D2: Users were labeled as non-depressed if they had never posted any tweet containing the character string "depress." Tweets published on December 2016 of these users were selected in this data set.

(3) Depression-candidate Data Set T_D3: Based on the tweets in December 2016, users were selected as depression candidates in this unlabeled data set if their anchor tweets loosely contained the character string "depress."

Data Description

To further clean the MSM data set, we removed noisy samples from the Blued database for which the number of postings was less than five. Inspired by the work of Shen et al. (18), we also removed accounts that had 15,000 or more followers, which may be organizations or bots. Then we manually checked whether the user was MSM according to his posting content, and removed the samples of non-MSM users. Finally, we obtained 346 depressed MSM users, 8,552 non-depressed MSM users, and 2,627 depression candidates. On the other hand, to construct a non-MSM data set based on the Twitter database, we removed 815 accounts, whose tweets satisfied the strict pattern "(I'm/I was/I am/I've been) gay." Then, we obtained 2,364 depressed non-MSM users, 5,041 non-depressed non-MSM users, and 44,536 depression candidate non-MSM users.

As presented in **Table 1**, the Blued database contains 11,525 users and 237,927 posts, and the Twitter database contains 51,101 users and 41,461,047 tweets. The time range of posts on Blued database is two months, while the time range of tweets on Twitter database is one month. It shows that Twitter users were more active online than Blued users, such that a Twitter user published 798.2 tweets per month but a Blued user only published 10.3 posts per month.

Data Preprocessing

Before feature extraction, data preprocessing is necessary since there are flexible and variant words in the raw data of social media. To deal with the difficulties in word matching and

TABLE 1 | Description on the Blued and Twitter data sets.

Database	Data set	Number of users	Number of posts	Time range of posts
Blued	B_D1	346	19,457	Jan 2012 to Mar 2019
	B_D2	8,552	155,138	Feb 2019 to Mar 2019
	B_D3	2,627	63,332	Jan 2012 to Mar 2019
	Total	11,525	237,927	Two-month time window
Twitter	T_D1	2,353	480,631	Jan 2009 to Dec 2016
	T_D2	4,990	3,956,077	Dec 2016
	T_D3	43,668	37,024,339	Dec 2016
	Total	51,011	41,461,047	One-month time window

semantic analysis, we carried out the following data preprocessing procedures on the two data sets:

1. Emoji processing. Emojis are incompatible with many text processing algorithms, so we removed the useless emojis from the emoji library (see **Additional file 1**), and then counted emoticons (emojis with sentiment characteristics) separately.
2. Topics processing. Many posts contain particular topic tags, which are noisy for word matching and text processing. Topic tags are published between two "#" in Blued, e.g. "#love#." and for Twitter, topic tags are published with one "#" like "#love." Therefore, we removed all topic tags contained in posts, and then counted the number of topic tags separately.
3. Mention processing. There are many mention marks in the texts of posts, e.g. "@someone"; thus; we removed all such character strings, and counted the times of mentions separately.
4. Stemming. We used the stemming algorithm of the Scikit-learn module in Python to unify word representations. For example, "am/is/are/was/were" should be represented as "be" uniformly.
5. Word segmentation. Unlike English texts, Chinese texts should be segmented into words before text analysis. The word segmentation processing was done using the Jieba module in Python.
6. Filtering stop-words. We removed stop-words from the posts, for example, "a" and "an." The stop-words in Chinese were filtered using the Jieba module in Python after word segmentation. English stop-words were removed by the Scikit-learn module in Python.

Feature Extraction

Using references from the psychological and behavioral sciences (18, 25), we finally defined and extracted 19 features based on the Blued and Twitter data sets (see **Table 2**), and classified them into four feature groups to describe multidimensional characteristics of users. As presented in **Table 2**, user-level features were extracted from the Blued users and Twitter users, and post-level features were extracted from the posts on Blued (two-month period) and the tweets on Twitter (one-month period).

User Profile Features

In this paper, three different user profile features are defined: the numbers of *followings* and *followers* describe the online egocentric social networks of users during their accounts'

TABLE 2 | Features used in the XGBoost model.

Level	Group	Feature Name	Definition
Use-level	User Profile Features	followings	The total number of users who followed me.
		followers	The total number of users who I followed.
		listNum	The total number of interested groups that I participated in.
Post-level	Social Interaction Features	favorNum	The average number of times each post was favored by others, $favorNum = totalFavorNum/postNum$, where *totalFavorNum* is the total number of times my posts was favored.
		mentionNum	The average number of times I was mentioned in other's posts (e.g. @me), $mentionNum = totalMentionNum/postNum$, where *totalMentionNum* is the total number of times I was mentioned by others.
		repostNum	The average number of times each post was reposted/retweeted by others, $repostNum = totalRepostNum/postNum$, where *totalRepostNum* is the total number of times my posts were reposted by others.
		topicNum	The average number of topic tags contained in each post (e.g. #fun# on Blued or #fun on Twitter), $topicNum = totalTopicNum/postNum$, where *totalTopicNum* is the total number of topics included in my posts.
		postNum	The number of my posts in the data sets.
		timeDist	The average numbers of my posts during 24 h, $timeDist = [postNum_0, postNum_1,\ldots, postNum_{23}]$, where , $postNum_1 = totalPostNum_1/postNum$ is the proportion of posts published at i^{th} o'clock.
	Emotion Features	posWordNum	The average number of positive words in each post, $posWordNum = totalPosWordNum/postNum$, where *totalPosWordNum* is the total number of positive words included in my posts.
		negWordNum	The average number of negative words in each post, $negWordNum = totalNegWordNum/postNum$, where *totalNegWordNum* is the total number of negative words included in my posts.
		emoNum	The total number of emoticons in my posts.
		posEmoNum	The average number of positive emoticons contained in my posts, $posEmoNum = totalPosEmoNum/postNum$, where *totalPosEmoNum* is the total number of positive emoticons included in my posts.
		negEmoNum	The average number of negative emoticons contained in my posts, $negEmoNum = totalNegEmoNum/postNum$ where *totalNegEmoNum* is the total number of negative emoticons included in my posts.
	Linguistic Features	LDATopicWords	The top 15 LDA topic words extracted from my posts, $LDATopicWords = [word_1, word_2,\ldots, word_{15}]$.
		antidepressNum	The average number of antidepressant drug names in each post, $antidepressNum = totalAntidepressNum/postNum$, where *totalAntidepressNum* is the total number of antidepressant drug names mentioned in my posts.
		depressWordNum	The average number of times the character string "depress" appeared in each post (named depressive word), $depressWordNum = totalDepressWordNum/postNum$, where *totalDepressWordNum* is the total number of depressive words in my posts.
		picNum	The average number of pictures in each post, $picNum = totalPicNum/postNum$, where *totalPicNum* is the total number of pictures in my posts.
		videoNum	The average number of videos in each post, $videoNum = totalVideoNum/postNum$, where *totalVideoNum* is the total number of videos in my posts.

lifetime; and *listNum* is the number of interested groups/lists that the user participated in.

Social Interaction Features

Based on the number of posts (*postNum*), we obtained the number of times for which each post was favored (*favorNum*) and reposted/retweeted (*repostNum*) by other users, the number of topics mentioned in each post (*topicNum*), and the number of times a user was mentioned in others' posts (*mentionNum*), to describe the interaction behaviors among users. In addition, the time distribution of postings (*timeDist*) was collected to characterize the active time of users on social networks.

Emotion Features

There are many differences in emotional status between depressed users and non-depressed users; thus, emotional features are beneficial for depressive emotion detection. We counted the numbers of positive words (*posWordNum*) and negative words (*negWordNum*) in each post. These positive and negative words in English and in Chinese are extracted by the Scikit-learn module in Python and dictionary-based methods, respectively. The number of emoticons (*emoNum*) from the posting content was counted as well as the numbers of positive emoticons (*posEmoNum*) and negative emoticons (*negEmoNum*). Six volunteers voted on the sentiment (positive, negative, and neutral) of the emoticons, as shown in **Additional file 1**.

Linguistic Features

With respect to the content characteristics of users' postings, firstly, the topics expressed by depressed users and non-depressed users are likely to differ significantly. Therefore, we applied the unsupervised latent Dirichlet allocation (LDA) model to extract 15 salient LDA topic words (*LDATopicWords*) from users' posts. Second, there are some domain-specific features from depressed users, such as the number of antidepressant drug names (*antidepressNum*) and the number of depressive words which containing the character string "depress" (*depressWordNum*). The names of the antidepressants in Chinese and in English are presented in **Additional file 2**. Finally, the media types of posts were also collected, such as the numbers of pictures (*picNum*) and videos (*videoNum*).

Classification Method

Let $D = \{(x_i, y_i)\}(|D| = n, x_i \in R^n, y_i \in R^n)$ represents a data set with n examples and m features. The classification method aims to find a relationship between an input $X = \{x_1, x_2, \ldots, x_N\}$ and an output Y. In our case, we determined whether a user was depressed based on four feature groups, and we aimed to determine the best combinations of features that would show the most predictive power in this binary classification. We used the XGBoost (eXtreme Gradient Boosting) (38) algorithm for this purpose, which is a well-designed gradient-boosted decision tree (GBDT) (39) algorithm that has demonstrated its state-of-the-art advantages in scientific research for machine learning and data mining. XGBoost belongs to a group of widely used tree learning algorithms, which do not require linear features or linear interactions between features. A decision tree allows for making predictions on an output variable based on a series of rules arranged in a tree-like structure. In this paper, all users (samples) were described by the set of 19 features that were classified into four groups. The XGBoost algorithm was implemented on the training data sets (D1 and D2) using the Scikit-learn Python (40) libraries for machine learning processes. To prevent overfitting and to make a good assessment of model validity, we employed a stratified five-fold cross validation to conduct the experiments with 12 randomized experimental runs in order to reduce variance.

Experiment Setting

The optimal values of parameters for the XGBoost algorithm were carefully tuned by a grid search with small but adaptive step size to enumerate the classification accuracy rates in different XGBoost parameter settings. We choose six estimator values of 50, 100, 200, 300, 400, and 500, five minimum child weight values of 1, 2, 3, 4, and 5, and five regularization alpha values of 1e-5, 1e-3, 1e-2, 0.1, and 1. The search ranges for learning rate, maximum tree depth, subsample and colsample_bytree were [0.05,0.2], (3, 11), [0.5,1] and [0.5,1], respectively. We carefully tuned the parameters of the XGBoost algorithm to obtain the best performance, the final XGBoost model parameter settings are described in **Table 3**.

Metrics

We evaluate the detection performance of XGBoost using four widely used metrics, i.e. accuracy, macro-averaged recall (Recall), macro-averaged precision (Precision), and macro-averaged F1 score (F1 score). For the classification task in this paper, the terms *true depressed*, *true non-depressed*, *false depressed* and *false non-depressed* compare the results of the classifier.

These four metrics can be counted by the following formulas:

$$Accuracy = \frac{\Sigma\, true\ depressed + \Sigma\, true\ non\_depressed}{\Sigma\, total\ population}$$

$$Recall = \frac{\Sigma\, true\ depressed}{\Sigma\, true\ depressed + \Sigma\, false\ non\_depressed}$$

$$Precision = \frac{\Sigma\, true\ depressed}{\Sigma\, true\ depressed + \Sigma\, false\ depressed}$$

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

TABLE 3 | Parameters of XGBoost model for the Blued and Twitter data sets.

	Parameter	Blued	Twitter
XGBoost	Estimators	50	200
	Maximum tree depth	4	4
	Learning rate	0.06	0.06
	Minimum child weight	2	1
	Subsample	0.45	0.8
	Colsample_bytree	0.65	0.8
	Regularization alpha	1	0.001

RESULTS

We first validated the effectiveness of the XGBoost model on the labeled Blued data sets (B_D1 and B_D2) and Twitter data sets (T_D1 and T_D2) and compared the importance of each extracted feature for the model. Next, we applied the trained XGBoost model to the unlabeled depression-candidate data sets (B_D3 and T_D3), in order to obtain a large number of depressed users. Finally, based on the massive social networking data of MSM and non-MSM users, we analyzed the differences in online behaviors between depressed MSM and non-depressed MSM, as well as the differences between depressed MSM and depressed non-MSM.

Model Performance

The classification accuracy rates on the Blued and Twitter data sets are displayed in **Table 4**. It can be seen that the XGBoost achieved outstanding performance on both the Blued data sets and the Twitter data sets. Given that certain online behavior characteristics and patterns of users can be found on online social networks, similar characteristics might be found on different kinds of social networks. The high performance on both the Blued data set and the Twitter data set also indicates that depressive emotion detection methods *via* online behavioral feature learning techniques possess strong applicability and generality for different social networking platforms. Thus, other sources of online social networks, such as Facebook and Instagram, would also provide multidimensional online information for depressive emotion detection.

In order to study the effectiveness of different feature combinations, we then constructed an experiment to feed our model with one feature group removed each time based on the Blued data sets. Specifically, we first used all feature groups, denoted as XGB. We then removed the four feature groups separately and denoted them as XGB-U (for removing user profile features), XGB-S (for removing social interaction features), XGB-E (for removing emotion features) and XGB-L (for removing linguistic features), respectively. From the results shown in **Figure 1**, we can see that the XGB-L performed worse than the others, indicating that linguistic features are more significant than the other feature groups. Furthermore, the social interaction features also contribute much to the performance, which shows that depressed users usually have different social networking behaviors.

Feature Importance

Feature importance gives a score for each feature of our data, and the higher the score, the more important or relevant the feature is towards the output (depressive emotion detection result). In this paper, the feature importance of each feature in the XGBoost algorithm is a fraction of the number of times the feature was used to split the data across all decision trees among the number of times all features are used to split the data across all decision trees. **Figure 2** shows the importance of each feature in the XGBoost model for classification on the Blued data sets (B_D1 and B_D2) and Twitter data sets (T_D1 and T_D2). The three most important features in both Blued and Twitter data sets are *depressWordNum*, *LDATopicWords* and *timeDist*, which held about 84% of feature importance in total. It suggests that these three features are more significant than other features in reflecting the different online behaviors of depressed users and, thus, contribute most to depressive emotion detection.

There are a number of differences in feature importance between Blued and Twitter data sets. The most crucial feature in Blued was *depressWordNum*, which comprised nearly half of feature importance, however, in Twitter, *depressWordNum* only held less than 20% of feature importance. There is an obvious difference in the usage of depressive word between depressed and non-depressed Blued users, such that each depressed Blued user in B_D1 has published at least one post that contained the character string "depress" and 99.2% of non-depressed Blued users in B_D2 have never published depressive words. Different from the depressed Blued users, 7.18% of depressed Twitter users in T_D1 have never posted depressive words and 81.4% of non-depressed Twitter users in T_D2 have never published depressive words. Affected by the outstanding classification power of *depressWordNum*, the *LDATopicWords* became the second crucial feature in Blued but held about a half of feature importance in Twitter, though the difference between depressed and non-depressed users are similar in Blued and Twitter data sets (see **Figure 4**). The feature importance of *timeDist* is larger in Twitter than in Blued, which can be explained by the more obvious differences observed between depressed and non-depressed users in Twitter, as shown in **Figures 3A, D**.

An interesting finding is that *videoNum* was the fourth important feature in Twitter data sets, but was an unimportant

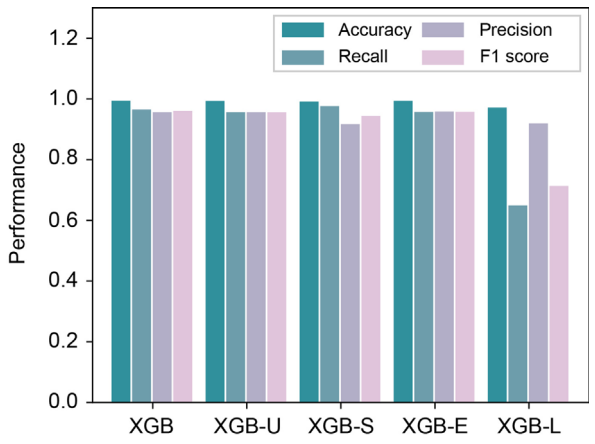


FIGURE 1 | Performance comparison of different feature combinations in the Blued data sets.

TABLE 4 | Classification performance of the XGBoost algorithm on the Blued and Twitter data sets.

Data set	Accuracy	Recall	Precision	F1 score
Blued	0.9940	0.9648	0.9563	0.9602
Twitter	0.9671	0.9591	0.9649	0.9619

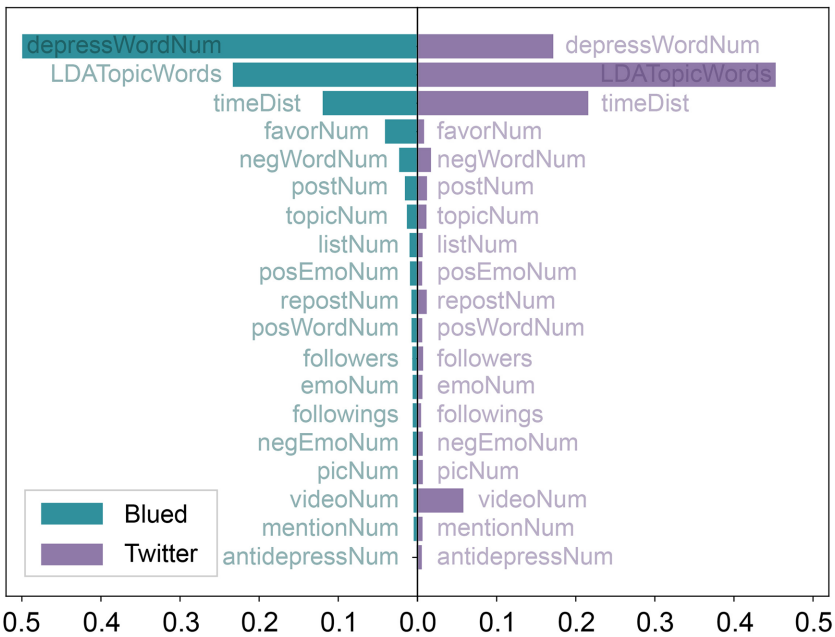


FIGURE 2 | Feature importance of XGBoost algorithm in the Blued and Twitter data sets.

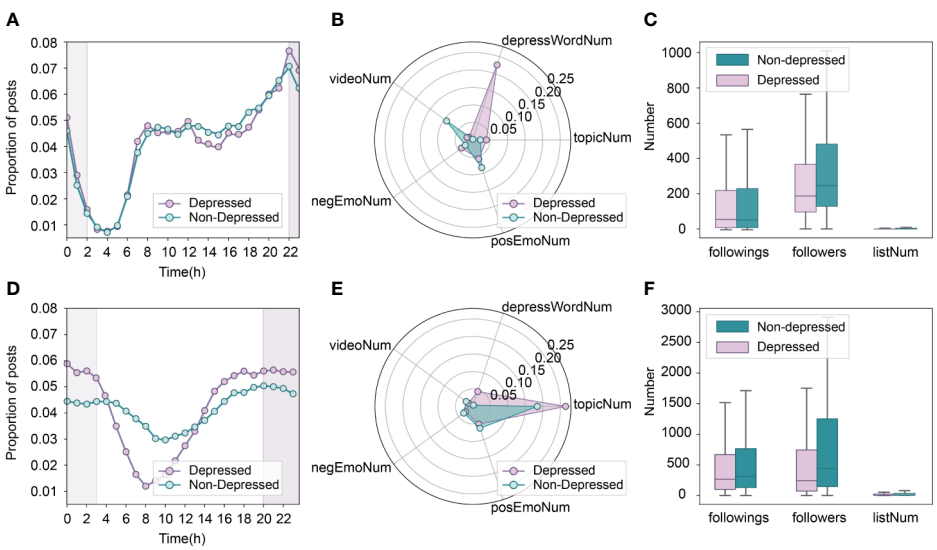


FIGURE 3 | Online behavior characteristics of depressed and non-depressed users for MSM population on Blued (top row) and non-MSM population on Twitter (bottom row). **(A**, **D)**: active time comparison; **(B**, **E)**: posting custom of each post with regard to five representative behaviors; **(C**, **F)**: distributions of user profile features.

feature in Blued data sets. However, we find that the difference in *videoNum* between depressed and non-depressed users is more obvious in Blued data sets (mean diff = 0.07) than in Twitter data sets (mean diff = 0.02). One reason for the inconsistency might be the relatively lower feature importance of *videoNum* (5.7%) compared with the three most crucial features (83.7%). That is, *videoNum* is the fourth important feature in Twitter, however, compared with the difference between depressed and non-depressed users in *LDATopicwords*, *depressWordNum* and *timeDist*, the difference in *videoNum* become less obvious. In general, the results of feature importance in Blued and Twitter data sets indicate that on the one hand, the dominating

FIGURE 4 | Salient LDA topic words for **(A)** depressed MSM users, **(B)** depressed non-MSM users, **(C)** non-depressed MSM users, and **(D)** non-depressed non-MSM users. The larger the word, the more frequent it appears in the posts.

explainable features share many similarities, i.e., both data sets have *depressWordNum*, *LDATopicWords* and *timeDist* as the most important features; on the other hand, there are considerable differences between MSM and non-MSM users regarding the order of the most crucial features and the *videoNum* feature.

Behavior Analysis

All users on both Blued (MSM users) and Twitter (non-MSM users) can be classified into four user groups: depressed MSM, non-depressed MSM, depressed non-MSM, and non-depressed non-MSM. After applying the XGBoost algorithm to the depression-candidate data sets B_D3 and T_D3, we detected 1,445 depressed MSM users and 24,748 depressed non-MSM users. Then, the numbers of depressed MSM users and non-depressed MSM users are 1,791 and 9,734, respectively. And the numbers of depressed non-MSM users and non-depressed non-MSM users are 27,101 and 23,910, respectively. We analyzed the difference of online behaviors between different categories of users, the mean results of the extracted features are presented in **Table 5**, **Figures 3** and **4**.

Comparison Between Blued Users and Twitter Users

There are many differences in online behaviors between Blued users and Twitter users, since the purposes of users in Blued and Twitter may show some differences to some extents. As shown in **Table 5**, with an enormous number of active users, Twitter users had more followings and followers, and attended more interested groups. Regarding the posting behaviors, Twitter users posted more tweets than Blued users, and each tweet contained a larger number of topics, pictures, videos, antidepressant drug names, and depressive words. In comparison with Twitter users, Blued users posted more emoticons, positive emoticons, positive words, and negative words. In addition, while tweets on Twitter are more likely to be retweeted than the posts on Blued caused by the purpose of information sharing, posts on Blued are more likely to be favored by others, driven by the purpose of seeking partners. Concerning LDA topics, we compared the top 30 salient LDA topic words between Blued and Twitter data sets, and found that 30% of the words were similar. These similar words were commonly used in daily life, such as "feel," "go," "life," "like," "love," "really," "repost," "time," and "year."

The difference between Blued users and Twitter users, to a certain extent, reveals that MSM and non-MSM users show different online social behaviors. For example, the active online interaction among Twitter users might indicate that non-MSM users are more active online than MSM users and have more diversified forms to express their feeling and emotion, such as pictures and videos. In addition, MSM users prefer directly expressing their emotions and feelings, thus their posts contained more emoticons and emotion words.

Behavior Analysis of Depressed and Non-Depressed MSM

First, we examined the users' posting proportions at different time periods, as shown in **Figure 3A**. It was found that MSM users tend to publish more posts between 6 pm and 1 am, indicating that the MSM population prefers frequently using social networks at night. Another possible explanation is that they are likely to suffer from insomnia. It also appears that the degree of going to bed late or suffering from insomnia is higher

TABLE 5 | Mean of features on Blued data sets and Twitter data sets.

Feature	MSM users (Blued)			Non-MSM users (Twitter)		
	Depressed	Non-depressed	Total	Depressed	Non-depressed	Total
followings	254.491	306.236	298.2 (0~29,798)	676.427	748.8731	710.384 (0~16,890)
followers	514.511	550.949	545.3 (0~82,409)	888.671	1237.16	1052.01 (0~14,994)
listNum	1.9	2.758	2.6 (0~20)	35.313	52.8384	43.53 (0~9,313)
postNum	28.675	19.167	20.6 (1~1,826)	392.887	1214.16	777.834 (1~3,260)
repostNum	0.0227	0.0798	0.07 (0~57.19)	1201.23	1558.16	1368.53 (0~422,814)
mentionNum	0.0004	0.0028	0.002 (0~2.19)	0.605	0.6466	0.624 (0~8.636)
topicNum	0.0394	0.022	0.025 (0~6.8)	0.265	0.184	0.227 (0~10.535)
picNum	0.966	1.371	1.31 (0~9)	0.17	0.21	0.189 (0~3.06)
videoNum	0.0129	0.0921	0.08 (0~1)	0.0155	0.024	0.019 (0~1)
antidepressNum	0.0003	3.453	0.00007 (0~0.25)	0.013	0.0122	0.0126 (0~1.01)
depressWordNum	0.225	0.0019	0.037 (0~4)	0.0447	0.0039	0.025 (0~2.988)
emoNum	0.098	0.112	0.109 (0~2.5)	0.0916	0.112	0.101 (0~3.286)
posEmoNum	0.057	0.084	0.079 (0~2)	0.0533	0.0655	0.059 (0~1.971)
negEmoNum	0.039	0.026	0.028 (0~1.5)	0.026	0.0325	0.029 (0~2.5)
posWordNum	2.368	1.41	1.559 (0~33)	0.434	0.339	0.39 (0~6)
negWordNum	2.079	0.992	1.161 (0~24)	0.339	0.266	0.305 (0~6)
favorNum	7.618	23.622	21.135 (0~1353.73)	1.0065	0.976	0.992 (0~597.875)

among depressed MSM than non-depressed MSM, since depressed MSM published more posts between 10 pm and 2 am.

Next, the user posting patterns and user profile features were analyzed. As shown in **Figure 3B**, depressed MSM users posted on average 0.225 depressive words and 0.039 negative emoticons per post, which surpassed the non-depressed MSM users by 0.223 and 0.013, respectively. This suggests that depressed MSM users may express their depressive emotions and complain more about their bad moods. In addition, the depressed MSM users published more posts that containing more topics but fewer videos than non-depressed MSM users. With regard to *negWordNum* and *posWordNum*, as presented in **Table 5**, we found that depressed MSM users posted more negative words (1.087) and positive words (0.958) than non-depressed MSM users. From **Figure 3C**, we can see that the differences between the average numbers for the followings and interest groups are small between depressed MSM and non-depressed MSM. In addition, the non-depressed MSM has more followers than the depressed MSM, indicating that the non-depressed MSM users are more likely to be favored or followed by others.

Finally, we constructed the word cloud of the top 150 salient terms of LDA topics, as shown in **Figures 4A, C**. The content of the posts was so casual that both the depressed MSM users and non-depressed MSM users published posts containing many commonly used words, such as "love," "like," "live." In addition, their posts included words such as "eat," "go," "life," and "friend," indicating that both classes share their daily lives on Blued. However, depressed MSM users are more likely to be under pressure, as their posts included more occurrences of "work" and "don't," while the posts of non-depressed MSM users contained more words like "movie" and "sleep." In addition, in contrast to the posts of non-depressed MSM users, the posts of depressed MSM users involved words such as "depressed," "depression," and "die," which implies that depressed MSM users tend to post their mental status on Blued and they have significant mental health problems.

For the remaining features, in comparison with the depressed MSM users, we discovered that the non-depressed MSM users' posts are more likely to be favored and reposted by others. Furthermore, the non-depressed MSM users publish more posts

containing more pictures, and they also prefer mentioning other users in their posts.

Behavior Analysis of Depressed MSM and Non-MSM Users

As shown in **Figures 3** and **4**, the non-MSM users show a clear posting pattern of being active at nighttime and silence during the daytime. Compared to non-MSM users, MSM users maintain a high level of activity most of the time except in the small hours. This shows that MSM users almost cannot survive without online social networks, and they have a stronger virtual-social-dependence than non-MSM users.

Similar to depressed MSM users, the depressed non-MSM users posted more depressive words than non-depressed non-MSM users, and the content of their posts contained more topic tags. An obvious difference is that the posting content of depressed MSM users contained 0.18 more depressive words than that of depressed non-MSM users, indicating that depressed MSM users complain more and they are more likely to express their negative emotions on social networks. In addition, depressed non-MSM users published posts with 0.226 more topic tags than depressed MSM users, which is related to the differences in the target users and design principles between Twitter and Blued. The former has a-broad-audience and a stronger information dissemination capacity. Similarly, non-MSM users have more followers and followings than MSM users, since Twitter has more than 500 million users while Blued only has about 40 million users.

An interesting finding is that depressed MSM users posted 2.079 negative words and 2.368 positive words, surpassing that of depressed non-MSM users by 1.74 and 1.934, respectively (see **Table 5**). The same results can be seen between non-depressed MSM users and non-depressed non-MSM users. This again shows that the MSM population prefers to express their emotions on social networks compared to the non-MSM population. Another interesting finding is that non-depressed non-MSM users posted more negative emoticons than depressed non-MSM users, which differs from that of the MSM users. In comparison with depressed MSM users, depressed non-MSM users published posts with more positive LDA topic words, such as "good," "new," "thank," and less negative LDA topic words, such as "depression," "depressed," and "dont," as shown in **Figures 4A, B**.

DISCUSSION

Overall, we aimed to investigate the feasibility of automated detection of depressive emotions among the MSM population using social networking data on online mass media. With the well-labeled data sets and well-defined depression-oriented features, the XGBoost algorithm achieved good performance on detecting depressed users for both MSM users on Blued and non-MSM users on Twitter. We further analyzed the contribution of the feature groups and the importance of each feature in the XGBoost algorithm. In addition, based on the depressive emotion detection results, we analyzed the differences between Blued users and Twitter users, the differences between depressed and non-depressed MSM users, and the differences between depressed MSM and non-MSM users. Compared to non-MSM users, MSM users complain more and are more likely to express their negative emotions on social networks.

In our study, we adopted a machine learning model named XGBoost to implement depressive emotion detection. In comparison with deep learning models which provide powerful predictive capability but generally lack interpretability, the XGBoost model is a kind of decision tree-based ensemble method which can provide interpretability to some extent in feature extraction and decision processes as well as achieve outstanding performance in data mining fields and classification tasks. Since our work focused on the depressive emotion detection and behavior analysis of MSM population, the XGBoost algorithm is capable of achieving these goals. And based on the results achieved by the XGBoost model, we can further analyze the importance of each feature in the depressive emotion detection task.

The linguistic feature group is more significant than the other feature groups, when using XGBoost algorithm to detect depressive emotions. The linguistic features include LDA topic words, depressive words, the number of pictures, and the number of videos in the posts, which reflect the content characteristics of the users' online posts. Similar with the linguistic features in reality, the online linguistic feature is also a potential reference for identifying the mental health of users (41). Additionally, in the depressive emotion detection task, the most important features for determining the detection accuracy are depressive words, LDA topic words, and posting time distribution. These features are useful for describing depressive symptoms, such as depressed mood, users' feelings, social behaviors, and sleep disturbance, which are also used as diagnosing indicators in many scales to assess depression (17). Therefore, the XGBoost model can extract multi-dimensional depressive symptoms of users through massive online social networking data, and then help detect depression in the early stage.

In addition, while most studies conducted by questionnaire and interview were limited by the sample size of participants and multidimensional characteristics of social behaviors, our work suggests that the automated depressive emotion detection using online mass media data with a machine learning manner, is a potential and easy-to-use way to achieve both large-scale samples and multidimensional depressive symptoms detection. Evidence has shown that many depressive symptoms that are used for clinical depression diagnoses are reflected in online social behaviors, such as insomnia, suicidal thoughts, and depressive mood (21, 29, 41). Thus, further diagnosis or medication of depression is needed for depressed users, as they are found to have significant mental health problems.

Regarding MSM population who are at high risk of HIV and experience remarkably poorer mental health, few studies have focused on the depressive emotion detection of MSM, and the

differences between depressed and non-depressed MSM, as well as between MSM population and non-MSM population. In our work, we provide an innovative approach for the monitoring of depressive emotions or mental health among MSM population *via* social media. It is of significant importance in this at-risk population as it has been reported that depression is associated with increased risk for HIV in MSM population. The digital monitoring of depressive symptoms of MSM population in this paper can complement and extend traditional approaches to diagnose depression, and enable those suffering depressive emotions to be more proactive about their mental health. Additionally, compared to survey methods to recruit MSM populations with depression, we expect our findings to provide more perspectives and insights for MSM-related researches in online data collection and big data analysis.

However, in our study, depressed MSM users were identified if they published posts saying they are depressed and have suicidal thoughts; this kind of depressive emotion identification method is poor and inaccurate compared to the clinical depression diagnoses and the depression diagnosis measure in the benchmark Twitter data set. Furthermore, the non-MSM data set was constructed from the Twitter data set by removing the accounts of those who published posts describing themselves as gay; this kind of identification measure is also weak to a certain extent. Fortunately, based on the well-labeled Twitter data set, the rationality of the algorithm was proved with a good detection performance.

CONCLUSION

In summary, this is a new attempt to detect depressive emotions among MSM population using massive online social networking data with a machine learning algorithm. An effective and easy-to-use method is provided here for monitoring depressive emotions, which can help identify at-risk individuals in the early stage of depression for further clinical diagnosis. In addition, this is a novel analysis of the differences between MSM population and non-MSM population with or without depressive emotions. Automated depressive emotion screening *via* social media is a feasible and efficient measure for both the general population and hard-to-access populations. In the future, we expect to improve the representativeness of MSM population samples from online social media data and research the association between depression and stigma, and the sexual risk behaviors in MSM with or without HIV *via* online recruitment methods.

DATA AVAILABILITY STATEMENT

The Twitter data used in this article are available at http://depressiondetection.droppages.com and the Blued data can be accessed under reasonable requests from the authors.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements.

AUTHOR CONTRIBUTIONS

XL and MC designed the research. MC and YL performed the experiments and wrote the paper. SQ designed all the data visualizations. YL, XL, and SQ helped to revise the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

XL acknowledges the Natural Science Foundation of China (91846301, 71771213, and 71790615) and the Hunan Science and Technology Plan Project (2017RS3040, 2018JJ1034). YL is supported by the Natural Science Foundation of Hunan Province (2019JJ40328, 2019GK2131). MC is supported by the Natural Science Foundation of China (71690233, 71774168) and China Scholarship Council (CSC201903170182). The funding bodies had no role in the study design, data collection, analysis, interpretation of the data, and in writing the manuscript. The opinions expressed here represent those of the authors and do not necessarily reflect the views of the funders.

ACKNOWLEDGMENTS

We are grateful to use two benchmark well-labeled depression and non-depression data sets on Twitter, which are released online (http://depressiondetection.droppages.com). The authors would also like to thank Hui Guan, Yuxin Chen, Wo liu, Ling Tan, Yuxuan Tong and Shangyao Guo from the College of Economy and Management at Changsha University for participating as volunteers to vote the sentiment of emoticons used in this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.00830/full#supplementary-material

ADDITIONAL FILE 1 | emoji_and_emoticon_library.xlsx.

ADDITIONAL FILE 2 | antidepressant_table.xlsx.

REFERENCES

1. World Health Organization. *Depression*. (2020). https://www.who.int/en/news-room/fact-sheets/detail/depression
2. Singh S, Mitsch A, Wu B. HIV care outcomes among men who have sex with men with diagnosed HIV infection—United States, 2015. *MMWR Morb Mortal Wkly Rep* (2017) 66(37):969. doi: 10.15585/mmwr.mm6637a2
3. Hatzenbuehler ML, Nolen-Hoeksema S, Erickson SJ. Minority stress predictors of HIV risk behavior, substance use, and depressive symptoms: results from a prospective study of bereaved gay men. *Health Psychol* (2008) 27(4):455. doi: 10.1037/0278-6133.27.4.455
4. Wohl AR, Galvan FH, Myers HF, Garland W, George S, Witt M, et al. Do social support, stress, disclosure and stigma influence retention in HIV care for Latino and African American men who have sex with men and women? *AIDS Behav* (2011) 15(6):1098–110. doi: 10.1007/s10461-010-9833-6
5. King M, Semlyen J, Tai SS, Killaspy H, Osborn D, Popelyuk D, et al. A systematic review of mental disorder, suicide, and deliberate self harm in lesbian, gay and bisexual people. *BMC Psychiatry* (2008) 8(1):70. doi: 10.1186/1471-244X-8-70
6. Jeffries IV WL. Beyond the bisexual bridge: sexual health among US men who have sex with men and women. *Am J Prev Med* (2014) 47(3):320–9. doi: 10.1016/j.amepre.2014.05.002
7. Reisner SL, Mimiaga MJ, Skeer M, Bright D, Cranston K, Isenberg D, et al. Clinically significant depressive symptoms as a risk factor for HIV infection among black MSM in Massachusetts. *AIDS Behav* (2009) 13(4):798–810. doi: 10.1007/s10461-009-9571-9
8. Salomon EA, Mimiaga MJ, Husnik MJ, Welles SL, Manseau MW, Montenegro AB, et al. Depressive symptoms, utilization of mental health care, substance use and sexual risk among young men who have sex with men in EXPLORE: implications for age-specific interventions. *AIDS Behav* (2009) 13(4):811. doi: 10.1007/s10461-008-9439-4
9. Olatunji BO, Mimiaga MJ O, Cleirigh C, Safren SA. Review of treatment studies of depression in HIV. *Top HIV Med* (2006) 14(3):112–24.
10. Vu NT, Holt M, Phan HT, La LT, Tran GM, Doan TT, et al. Amphetamine-type-stimulants (ATS) use and homosexuality-related enacted stigma are associated with depression among men who have sex with men (MSM) in two major cities in Vietnam in 2014. *Subst Use Misuse* (2017) 52(11):1411–9. doi: 10.1080/10826084.2017.1284233
11. Brickman C, Propert KJ, Voytek C, Metzger D, Gross R. Association between depression and condom use differs by sexual behavior group in patients with HIV. *AIDS Behav* (2017) 21(6):1676–83. doi: 10.1007/s10461-016-1610-8
12. Alvy LM, McKirnan DJ, Mansergh G, Koblin B, Colfax GN, Flores SA, et al. Project MIX Study Group. Depression is associated with sexual risk among men who have sex with men, but is mediated by cognitive escape and self-efficacy. *AIDS Behav* (2011) 15(6):1171–9. doi: 10.1007/s10461-010-9678-z
13. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry* (1961) 4(6):561–71. doi: 10.1001/archpsyc.z1961.01710120031004
14. Beck AT, Guth D, Steer RA, Ball R. Screening for major depression disorders in medical inpatients with the Beck Depression Inventory for Primary Care. *Behav Res Ther* (1997) 35(8):785–91. doi: 10.1016/S0005-7967(97)00025-9
15. Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. *Appl Psychol Meas* (1977) 1(3):385–401. doi: 10.1177/014662167700100306
16. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* (2001) 16(9):606–13. doi: 10.1046/j.1525-1497.2001.016009606.x
17. Whooley O. Diagnostic and statistical manual of mental disorders (DSM). *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society* (2014) 381–4. doi: 10.1002/9781118410868.wbehibs011
18. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, et al. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne: IJCAI (2017). p. 3838–44. doi: 10.24963/ijcai.2017/536
19. Lin H, Jia J, Nie L, Shen G, Chua TS. What does social media say about your stress? In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. Palo Alto, California: IJCAI (2016).
20. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in twitter. In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland: ACL (2014). p. 51–60. doi: 10.3115/v1/W14-3207
21. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Curr Opin Behav Sci* (2017) 18:43–9. doi: 10.1016/j.cobeha.2017.07.005
22. Park M, Cha C, Cha M. Depressive moods of users portrayed in Twitter. In: *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*. Beijing: ACM (2012). p. 1–8.
23. Park M, McDonald DW, Cha M. Perception differences between the depressed and non-depressed users in twitter. In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. Cambridge, Massachusetts: AAAI (2013). p. 476–85.
24. Hussain A, Heidemann J, Heidemann J, Papadopoulos C. A framework for classifying denial of service attacks. In: *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communications*. Karlsruhe: ACM (2003). p. 99–110. doi: 10.1145/863955.863968
25. Tong L, Zhang Q, Sadka A, Li L, Zhou H. Inverse boosting pruning trees for depression detection on Twitter. *arXiv* (2019). arXiv:1906.00398.
26. Xu R, Zhang Q. Understanding online health groups for depression: social network and linguistic perspectives. *J Med Internet Res* (2016) 18(3):e63. doi: 10.2196/jmir.5042
27. Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen VA, Boyd-Graber J. Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado: ACL (2015). p. 99–107.
28. Shuai HH, Shen CY, Yang DN, Lan YF, Lee WC, Philip SY, et al. A comprehensive study on social network mental disorders detection via online social media mining. *IEEE Trans Knowl Data Eng* (2017) 30(7):1212–25. doi: 10.1109/TKDE.2017.2786695
29. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* Cambridge, Massachusetts: AAAI (2013). p. 128–37.
30. Nadeem M. Identifying depression on Twitter. *arXiv Preprint* (2016). arXiv:1607.07384.
31. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. *Sci Rep* (2017) 7(1):1–11. doi: 10.1038/s41598-017-12961-9
32. Orabi AH, Buddhitha P, Orabi MH, Inkpen D. Deep learning for depression detection of twitter users In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. New Orleans: ACL (2018). pp. 88–97.
33. Benton A, Mitchell M, Hovy D. Multi-task learning for mental health using social media text. *arXiv Preprint* (2017). arXiv:1712.03538.
34. Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, et al. Interpretability of deep learning models: a survey of results. In: *Proceedings of IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*. San Francisco Bay Area, California: IEEE (2017). p. 1–6. doi: 10.1109/UIC-ATC.2017.8397411
35. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In: *Proceeding of IEEE 5th International Conference on data science and advanced analytics (DSAA)*. Turin: IEEE (2018). p. 80–9. doi: 10.1109/DSAA.2018.00018
36. Liu C, Lu X. Analyzing hidden populations online: topic, emotion, and social network of HIV-related users in the largest Chinese online community. *BMC Med Inform Decis Mak* (2018) 18(1):2. doi: 10.1186/s12911-017-0579-1
37. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, et al. *Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution*. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne: IJCAI (2017) p. 3838–44. doi: 10.24963/ijcai.2017/536
38. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining (2016). p. 785–94. doi: 10.1145/2939672.2939785
39. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* (2001) 29(5):1189–232. doi: 10.1214/aos/1013203451
40. Raschka S, Mirjalili V. *Python machine learning*. Birmingham: Packt Publishing (2017).
41. ODea B, Boonstra TW, Larsen ME, Nguyen T, Venkatesh S, Christensen H. The relationship between linguistic expression and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study of blog content. *arXiv Preprint* (2018). arXiv:1811.02750.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 16 September 2020
doi: 10.3389/fpsyt.2020.574375

Deep Learning-Based Human Activity Recognition for Continuous Activity and Gesture Monitoring for Schizophrenia Patients With Negative Symptoms

Daniel Umbricht[1]*, *Wei-Yi Cheng*[2], *Florian Lipsmeier*[1], *Atieh Bamdadian*[1] *and Michael Lindemann*[1]

[1] Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland, [2] Roche Innovation Center New York, Roche TCRC Inc, NY, United States

OPEN ACCESS

Edited by:
Jennifer H. Barnett,
Cambridge Cognition,
United Kingdom

Reviewed by:
Sebastian Walther,
University of Bern, Switzerland
Mark Opler,
MedAvante-ProPhase Inc,
United States

***Correspondence:**
Daniel Umbricht
daniel.umbricht@roche.com

Specialty section:
This article was submitted to Psychological Therapies, a section of the journal Frontiers in Psychiatry

Received: 26 June 2020
Accepted: 26 August 2020
Published: 16 September 2020

Citation:
Umbricht D, Cheng W-Y, Lipsmeier F, Bamdadian A and Lindemann M (2020) Deep Learning-Based Human Activity Recognition for Continuous Activity and Gesture Monitoring for Schizophrenia Patients With Negative Symptoms. Front. Psychiatry 11:574375. doi: 10.3389/fpsyt.2020.574375

Background: We aimed to develop a Human Activity Recognition (HAR) model using a wrist-worn device to assess patient activity in relation to negative symptoms of schizophrenia.

Methods: Data were analyzed in a randomized, three-way cross-over, proof-of-mechanism study (ClinicalTrials.gov: NCT02824055) comparing two doses of RG7203 with placebo, given as adjunct to stable antipsychotic treatment in patients with chronic schizophrenia and moderate levels of negative symptoms. Baseline negative symptoms were assessed using the Positive and Negative Syndrome Scale (PANSS) and Brief Negative Symptom Scale (BNSS). Patients were given a GeneActiv™ wrist-worn actigraphy device to wear over a 15-week period. For this analysis, actigraphy data and behavioral and clinical assessments obtained during placebo treatment were used. Motivated behavior was evaluated with a computerized effort-choice task. A trained HAR model was used to classify activity and an activity–time ratio was derived. Gesture events and features were inferred from the HAR-detected activities and the acceleration signal.

Results: Thirty-three patients were enrolled: mean (±SD) age 36.6 ± 7 years; mean (±SD) baseline PANSS negative symptom factor score 23.0 ± 3.5; and mean (±SD) baseline BNSS total score 36.0 ± 11.5. Activity data were collected for 31 patients with a median monitoring time of 1,859 h per patient, equating to ~11 weeks or 74% monitoring ratio. The trained HAR model demonstrated >95% accuracy in separating ambulatory and stationary activities. A positive correlation was seen between the activity–time ratio and the percent of high-effort choices (Spearman r = 0.58; $P = 0.002$) in the effort-choice task. Median daily gesture counts correlated negatively with the BNSS total score (Spearman r = –0.44; $P = 0.03$), specifically with the diminished expression sub-score (Spearman r = –0.42; $P = 0.03$). Gesture features also correlated negatively with the BNSS total score and diminished expression sub-scores. Activity measures showed similar correlations with PANSS negative symptom factor but did not reach significance.

Conclusion: Our findings support the use of wrist-worn devices to derive activity and gesture-based digital outcome measures for patients with schizophrenia with negative symptoms in a clinical trial setting.

Keywords: body-worn sensor, digital endpoints, digital health technology, digital outcome measures, gesture detection, human activity recognition, negative symptoms, schizophrenia

INTRODUCTION

Negative symptoms are a key psychopathologic dimension and an important driver of functional disability in schizophrenia with up to 60–70% of patients exhibiting at least one such symptom (1, 2). Despite the high unmet medical need, there is currently no approved treatment for negative symptoms of schizophrenia in the United States.

Factor analyses of negative symptoms have demonstrated at least two dimensions: one consisting of apathy, amotivation, avolition, asociality, and anhedonia (referred to as 'avolition'); and expressive deficits (including affective flattening and poverty of speech and diminished use of gestures). The first dimension has been shown to be a key driver of functional impairment (3).

Currently, negative symptoms of schizophrenia are primarily assessed with clinician administered rating scales (4). Known problems with rating scales include challenges in establishing interrater reliability in large multinational studies, reliance on patients' reports for symptoms that are not directly observable in the interview, and expectation bias. These factors reduce the likelihood of signal detection in clinical trials of novel therapies for schizophrenia, increase the risk and cost of drug development, and diminish the chance of finding a treatment for this debilitating disease (5).

The development of alternative methods to objectively assess negative symptoms in patients with schizophrenia, in particular avolition as a key dimension, is critical to further drug development in this disease. Continuous assessment of a patient's activity by actigraphy may offer such an assessment that provides not only objective, but also longitudinal, data usually not available directly to the clinician. Previously, correlations between actigraphic measures and clinical symptomatology have been reported in patients with schizophrenia. They have shown stability both within and between psychotic episodes (6) and have also been linked to neuroimaging markers (7, 8). In general, reduced activity has been found to correlate with higher negative symptoms, in particular apathy, but not expressive deficits. Although a few studies that obtained symptom assessment and actigraphy measures did not report any correlation, indicating that such correlations were not always found (9–12). However, to our knowledge, all studies used activity counts as the primary variable assessing activity levels. No attempts have been made to differentiate the activity signal into different kinds of activities, in particular into ambulatory and stationary activity (like gesturing while standing or sitting).

Recent developments in digital health technology, including wearable sensors, provide new opportunities to continuously and passively monitor patients over a longer duration of time, collecting rich data sets. Previous studies have found that the use of wearable devices in schizophrenia is feasible and acceptable for patients and may be used to assess heart rate, electrodermal activity, and movement in everyday life (13). Human Activity Recognition (HAR) can identify actions carried out by an individual using acceleration and gyroscope data obtained from various sources including body-worn sensors (14–17).

Here, we present data exploring the use of a wrist-worn device to assess patient activity over a 15-week period to determine how activity measures extracted from the device correlate with negative symptoms of schizophrenia. The key goal was to test new analytical approaches that allow a more fine grained classification of activities in the context of a multi-site clinical trial. In addition, we explored if and how results of this analysis relate to clinical symptoms and, importantly, to performance in the effort-choice task, a behavioral assay probing the reward system and motivated behavior. A relatively large number of studies using effort-choice paradigms have shown that negative symptoms, particularly avolition, are associated with reduced motivation or willingness to expend high efforts for highly rewarded outcomes in such tasks (18–20).

MATERIALS AND METHODS

Study Design

The data were collected in a randomized, three-way cross-over, proof-of-mechanism study conducted between June 28, 2016 and April 28, 2017 (ClinicalTrials.gov: NCT02824055; protocol BP29904). The study compared two doses of the phosphodiesterase-10 inhibitor RG7203 (5 mg and 15 mg) with placebo, given as adjunctive to stable antipsychotic treatment in patients with chronic schizophrenia and moderate levels of negative symptoms. Outpatients were recruited through referral, direct contacts, and advertisements.

Patients were randomized to one of six treatment sequences using a central randomization system. Patients received once-daily placebo, 5 mg RG7203, or 15 mg RG7203 (matching oral capsules). To reach the 15 mg dose, treatment was up-titrated during Week 1. Each treatment period lasted for 3 weeks, followed by a 2-week washout period. For activity monitoring, study participants were provided with a GeneActiv™ (Activinsights Ltd, Cambridge, UK) wrist-worn actigraphy device to record data. Patients were asked to wear the device for 24 h each day throughout the entire 15-week trial period. The study was conducted in accordance with the principles of the Declaration of Helsinki and Good Clinical Practice guidelines. All patients provided written informed consent for study participation.

Primary results from the study will be published separately (Umbricht et al. Manuscript submitted for publication). The current analysis did not determine drug efficacy, but leveraged the placebo periods of the study.

Participants

Eligible patients were aged 18–50 years with a Diagnostic and Statistical Manual of Mental Disorders-5 diagnosis of schizophrenia and a Positive and Negative Syndrome Scale (PANSS) negative symptom factor score (NSFS) ≥18 (21) at screening. Patients were to be symptomatically stable and receiving antipsychotic treatment not exceeding a dose equivalent to 6 mg risperidone. Additional requirements for symptom severity at screening included: a Clinical Global Impression-Severity (CGI-S) score ≥3 (at least mildly ill); a score ≤4 (moderate or less) for PANSS items of hostility (P7) and uncooperativeness (G8); a PANSS depression score (G6) ≤4 (moderate or less); and a score ≤8 on the Calgary Depression Scale for Schizophrenia.

Exclusion criteria included: patients with a score >2 (mild) for any of the four CGI-S items of the Extrapyramidal Symptom (EPS) Rating Scale; electroconvulsive treatment within 6 months of screening, and olanzapine or clozapine within 3 months of screening; use of more than one antidepressant, or a change in dose of antidepressant within 4 weeks of screening; strong/moderate inhibitor or inducer of cytochrome P350 (CYP) 3A or CYPC8 within 14 days of screening; presence of a substance use disorder; positive urine screen for amphetamines, methamphetamines, opiates, buprenorphine, methadone, cannabinoids, cocaine, or barbiturates; a movement disorder that might affect ratings on the EPS scale; or prior or current medical conditions that could impair cognition or psychiatric function.

Clinical Assessments

Patients were assessed during inperson study visits. Baseline negative symptoms were assessed using the PANSS and Brief Negative Symptom Scale (BNSS) (22, 23). For correlational analysis, the PANSS NSFS, PANSS positive symptom factor score (PSFS) (21), the BNSS total score, and BNSS apathy index and expressive deficits factors (24) were used.

Motivated behavior was assessed with a computerized effort-choice task where patients were given a choice of an easy task with a lower reward or a more difficult task with a higher reward (25). Patients had the option to press a blue balloon 20 times until it popped for which they would receive one point, or to press a green balloon 100, 120, or 150 times until it popped to receive three, five, or seven points. The patients were informed about the maximum possible reward and the probability of receiving the reward (50 or 100%). Each set of cumulated 20 points convert to a $1 bonus. The percentage of high-effort choices across all effort levels and reward levels of five and seven points at 100% probability of receiving the reward was measured to determine the participant's motivated behavior.

The GeneActiv™ actigraphy device recorded the acceleration of wrist movement at 20 Hz to assess patient activity throughout the trial period. For this analysis, actigraphy data and behavioral and clinical assessments obtained during the placebo treatment period were used. The monitoring ratio was calculated by dividing the total number of hours with sensor data collected by the total number of hours in the study.

Sensor-Based Features

A 9-layer convolutional recurrent neural network (26) was trained using two public annotated data sets [Reiss et al. (27) and Stisen et al. (28)] containing wrist-worn acceleration data for nine subjects each to infer patients' activities. Performance evaluation of the trained HAR model was conducted using held-out testing data and internally collected sensor data from patients with multiple sclerosis who performed balance tests (stationary) and 2-min walking tests (ambulatory) (29). Two subjects from each of the Reiss and Stisen data sets were left out during training, while sensor data from 14 subjects were used as input to train an activity recognition model. One held-out subject from each data set formed the validation set, which was used to tune the hyper-parameters and determine convergence of the model during training, and one held-out subject from each data set formed the testing set, which was used only for final performance evaluation.

The trained neural network was used to infer a classification of patient activity collected on the actigraphy devices. Patient activity was categorized as either ambulatory (*i.e.* walking, climbing stairs, cycling, jogging) or stationary (*i.e.* sitting, standing, lying down, doing hand work). An activity ratio was derived for each patient based on the activities determined using the HAR and was defined as the total active time involving gait (*i.e.* walking, climbing stairs, running, cycling) divided by the total monitoring time.

Gesture events were inferred from the HAR model-predicted activities, combined with the standard deviation (SD) of the magnitude of acceleration signal from the wrist, using a 0.01 g threshold within a 1-s moving window, inspired by a previously published method by Rai et al. (30). In Rai et al. the authors observed that when the SD of the magnitude of the accelerometer signal was below 0.01 g, the user was idle with 99% probability. Therefore, a gesture event was defined as the time when the patient was not moving (*i.e.* sitting, standing, lying down, doing hand work) according to the HAR model, while the SD of the acceleration signal was >0.01 g. As gesture events were identified by a 1-s moving window across the accelerometer signal, the start and end time of a gesture event was defined by the continuous moving windows that fit the SD criteria, with a maximal gap between eligible windows smaller than 1 s. From the defined gesture events, we calculated gesture count and gesture power. Gesture count was calculated as the total number of gesture events per day. Gesture power was calculated during the detected gesture events by integrating the squared magnitude of the acceleration signal: $\Sigma_t \quad (m_t - \bar{m})^2$, where m_t is the magnitude of accelerometer signal at time t, and $\bar{m}$ is the mean of magnitude across time in a gesture event.

Statistical Analysis

Correlations between clinical scores and sensor-based features were evaluated using Spearman's correlation coefficient.

RESULTS

Study Participants

In total, 33 patients with negative symptoms of schizophrenia were enrolled at three study centers in the United States. Study participants had a mean (±SD) age of 36.6 ± 7 years, and the majority were male (30/33) and Black (21/33) (**Table 1**). The mean (±SD) baseline PANSS NSFS was 23.0 ± 3.5, the mean (± SD) baseline BNSS total score 36.0 ± 11.5, and the mean CGI-S score 3.7 ± 0.5.

Activity Data

Overall, 31 patients agreed to wear a GeneActiv™ wrist-worn actigraphy device to record actigraphy data. Median collected monitoring data per patient was 1,859 h equating to around 11 weeks or 74% monitoring ratio. There was no significant correlation between baseline PANSS NSFS, PANSS PSFS, or BNSS total score and monitoring ratio.

TABLE 1 | Demographics and clinical characteristics at screening.

	Cohort (N = 33)
Mean age ± SD (years)	36.6 ± 7.0
Male gender, n (%)	30 (91)
Race, n (%)	
Black	21 (64)
White	9 (27)
Asian	3 (9)
Mean BNSS total score ± SD	36.0 ± 11.5
Mean PANSS NSFS ± SD	23.0 ± 3.5
Mean PANSS PSFS ± SD	19.2 ± 4.8

BNSS, Brief Negative Symptom Scale; NSFS, Negative Symptom Factor Score; PANSS, Positive and Negative Syndrome Scale; PSFS, Positive Symptom Factor Score; SD, standard deviation.

Validation of Human Activity Recognition Model

Based on the held-out validation data, the trained HAR model demonstrated >95% accuracy in separating ambulatory (*i.e.* walking, climbing stairs, cycling, jogging) and stationary activities (*i.e.* sitting, standing, lying down, doing hand work). The model showed 94.9 and 95.5% accuracy in identifying stationary and ambulatory activities, respectively.

Correlation of Actigraphy-Derived Features With Clinical Scores

The activity–time ratio correlated positively with the percent of high-effort choices at the end of the placebo period (Spearman's r = 0.58; P = 0.002; **Figure 1**).

Median daily gesture counts were negatively correlated with the BNSS total score (Spearman's r = −0.44; P = 0.03; **Figure 2A**) at the end of the placebo period, specifically with the diminished expression sub-score (Spearman's r = −0.42; P = 0.03; **Figure 2B**). No correlation was observed with the BNSS apathy sub-score or the PANSS NSFS. Gesture power and activity–time ratio correlated negatively with the diminished expression sub-score, but not with other measures of negative symptoms (**Table 2**). Notably, all significant correlations were in the *a priori* expected direction supporting the convergent validity of the proposed novel digital measures with the established clinical scales.

Performance in the effort-choice task did not correlate with any clinical measure of negative symptoms.

DISCUSSION

The results of this analysis demonstrate that the use of a wrist-worn actigraphy device is feasible to support continuous monitoring of clinically relevant behavior in a multi-site

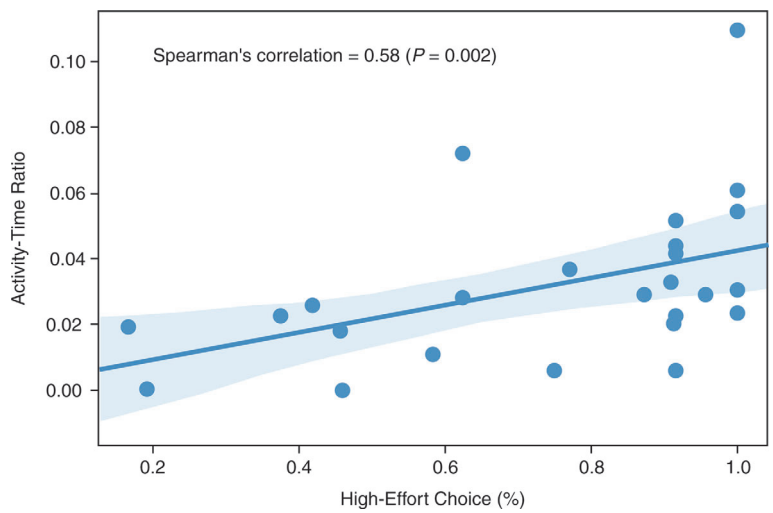


FIGURE 1 | Spearman's correlation between activity–time ratio and high-effort choice (Spearman's r = 0.58; P = 0.002).

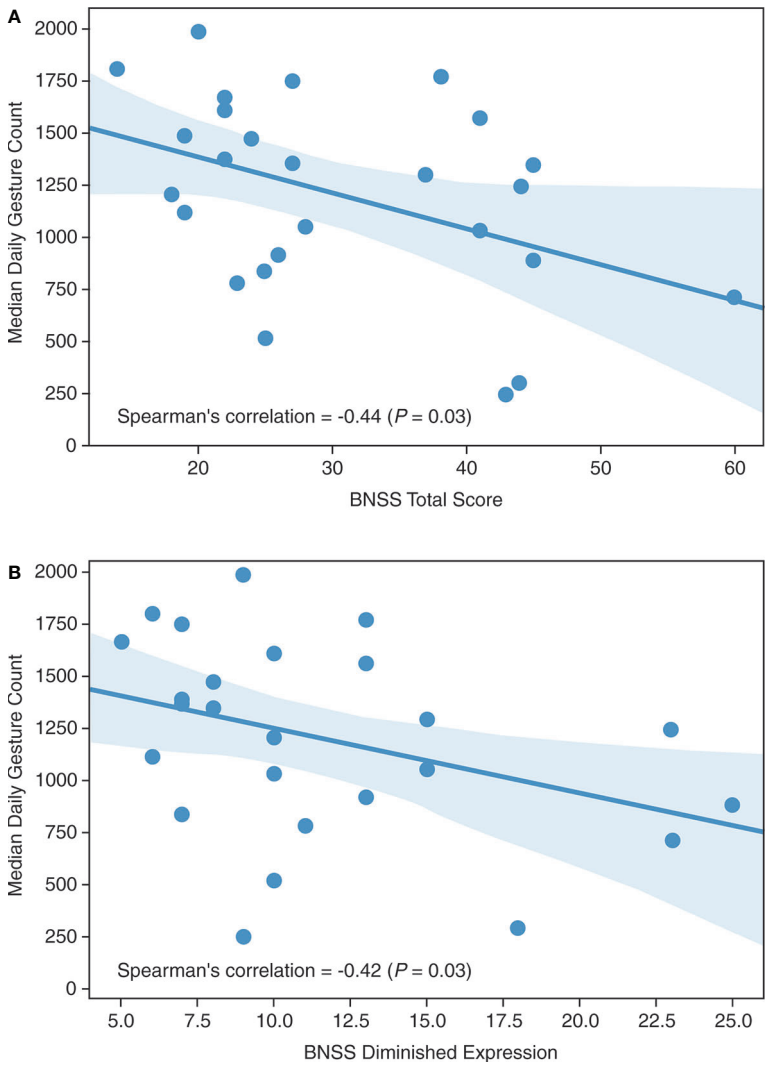


FIGURE 2 | Median daily gesture count *versus* **(A)** BNSS total score and **(B)** BNSS diminished expression. BNSS, Brief Negative Symptom Scale.

TABLE 2 | Spearman's correlation between high-effort choice, activity, and gesture features with clinical scores.

	High-effort choice in effort-choice task (N = 27)	Activity ratio (N = 26)	Gesture power (N = 26)	Gesture count (N = 26)
BNSS Apathy Index	0.12	−0.025	−0.178	−0.277
BNSS Diminished Expression	0.06	−0.210	−0.423*	−0.424*
BNSS Total Score	0.12	−0.080	−0.312	−0.438*
PANSS NSFS	0.02	−0.256	0.080	−0.251
PANSS PSFS	–	−0.149	0.164	−0.026

**P < 0.05. Higher BNSS scores indicate greater disease severity; cells showing significant correlations are shaded in gray.*
BNSS, Brief Negative Symptom Scale; NSFS, Negative Symptom Factor Score; PANSS, Positive and Negative Syndrome Scale; PSFS, Positive Symptom Factor Score.

clinical trial setting over an extended period of time. The patient monitoring ratio was acceptable throughout the trial, allowing for a high volume of data to be collected by the end of the placebo period. This suggests that patients were comfortable wearing the device and that the device is suitable for continuous assessment over a number of weeks.

Previous studies have investigated the feasibility and use of wearable devices in schizophrenia (10, 31). Cella et al. examined the use of a novel mobile health method using wearable technology to determine illness severity in patients with schizophrenia. The device was acceptable to patients and provided accurate and reliable measures of everyday activity and behavior, including assessment

of physiological measures, functioning symptoms, and levels of medication (13). Meyer et al. utilized a combination of wrist-worn devices and smartphones to continuously monitor sleep and rest-activity profiles in people with schizophrenia over a 2-month period. All study participants exceeded the 70% threshold for feasibility of the wearable device with a mean wear time of 21.8 h per day or 91% of the total study duration (31).

Passive monitoring has also been used to quantify behavioral changes in several neurological conditions, including Parkinson's disease (32, 33), multiple sclerosis (29), and Huntington's disease (34, 35). Lipsmeier et al. assessed the feasibility, reliability, and validity of smartphone-based digital biomarkers of Parkinson's disease over a 6-month period in a phase I clinical trial (32). Adherence was acceptable and sensor-based features showed moderate to excellent test–retest reliability (32). The use of smartphone- and smartwatch-based remote patient monitoring was also employed by Midaglia et al. in a 24-week pilot study in patients with multiple sclerosis (29). Adherence to passive monitoring was 70.8% with patient satisfaction rated as good to excellent, which remained stable throughout the study (29). An ongoing Digital-HD study is investigating the tolerability and feasibility of smartphone-based technology to passively monitor motor and non-motor manifestations of Huntington's disease (35). These studies indicate the potential for the measurement of disease-relevant features from daily life in a clinical trial setting.

While most, if not all, previous actigraphy studies in schizophrenia have used simple activity counts as a measure of patient activity, our approach used machine learning to extend beyond a simple count, to identification specifically of gesture events during non-locomotive activities and hence, eliminates the situation where hand movements are caused by walking or running. Our HAR model was validated using previously published data, with a high level of accuracy, demonstrating that this model is reliable and robust for the detection of ambulatory *versus* stationary activity. There was a significant positive correlation ($P = 0.002$) between the activity–time ratio and the percent of high-effort choices, indicating an association of avolition and lower activity in daily life. As expected, gesture features derived from the HAR model were associated with expressive deficits, supporting the validity of activity and gesture-based digital outcome measures for negative symptoms in patients with schizophrenia.

Our study also highlights well-known problems with clinician administered rating scales. We did not find correlations between percentage of high effort-choice in the effort-choice task and the various clinical assessments of negative symptoms, in particular apathy, as previously reported by others (12). This may be due to differences in the patient sample. However, the correlations between the effort-choice performance and the activity index would indicate that both measures indeed capture an aspect of reduced motivation. The lack of an association between the clinical assessments of apathy and the effort-choice performance thus suggests that the clinical assessments capture apathy unreliably, which may be the key reason for the lack of association between activity measures and apathy. This is not surprising, as the key features of apathy cannot be observed in the interview but have to be elicited by the clinician and rely primarily on the patient's memory and report, which has been shown to negatively affect the accuracy of symptom reporting (36). Not surprisingly, we found the highest correlation between the expressive deficit scores and gesture count and power. Expressive deficits are directly observable in the interview and hence can be assessed more reliably.

Our study has several limitations. Firstly, we could not validate our findings in a larger cohort of patients. This will be a critical step in establishing our analytical approach as a tool to assess negative symptoms. Secondly, our study did not allow the establishment of test–retest reliability, *i.e.*, to establish the stability of these measures in stable patients. Both issues are key for implementation of these measures in future clinical trials. Also, our approach focused on stationary periods and did not include gesture events during non-stationary activity which of course occur as well. Identification and inclusion of these additional gesture events should be attempted in future studies and may increase the correlation with negative symptoms. Also, it is possible that among events counted as gestures some may have been included that comprised movements that did not represent gestures such as playing instruments, doing crafts, cooking, other household chores. Excluding such activities in future studies may increase the sensitivity of our approach. In addition, although our approach is a step forward in differentiating gestures from non-gesture activity, it does not allow the differentiation of communicative and socially relevant gestures and gestures that do not have such characteristics. It would require the establishment of an 'alphabet' of such gestures in healthy volunteers in terms of actigraphy features that could then be used to detect the presence or absence of communicative gestures in patients—a relevant aspect for the assessment of negative symptoms. Also, previous studies have investigated the use, perception and imitation of gestures by patients in much more detail by direct observation or video-based studies (37, 38) and found abnormalities in all three aspects. Obviously a passive monitoring system like actigraphy is not able to provide these kinds of data. Finally, differences in publicly available data sets of healthy volunteers that were used for validation, including the method of detection, of activities measured, was also a limitation. We did not have a comparison sample of healthy volunteers in whom data were obtained with the same device. It is also conceivable that the very nature of gestures differs between healthy volunteers and patients; however, we are not aware of evidence supporting this assumption. Furthermore, the HAR model has not yet been tested in a drug-based clinical trial or in patient subgroups with baseline characteristics other than negative symptoms.

CONCLUSIONS

Overall, our findings support the use of wrist-worn devices to derive activity and gesture-based digital outcome measures for patients with schizophrenia with negative symptoms in a clinical trial setting. We present initial evidence of convergent validity of sensor-based features with established clinical outcome measures. This could, in the future, enable the objective measurement of behavioral changes in schizophrenia and pave

the way towards novel ways to evaluate treatments for the negative symptoms of schizophrenia, thereby supporting essential drug development for these patients.

PRIOR PRESENTATION OF THE DATA

Umbricht D, Cheng W-Y, Lipsmeier F, Bamdadian A, Tamburri P, Lindemann M. Deep learning-based human activity recognition for continuous activity and gesture monitoring for schizophrenia patients with negative symptoms. ACNP 57th Annual Meeting 2018. Abstract T210.

DATA AVAILABILITY STATEMENT

Qualified researchers may request access to individual patient level data through the clinical study data request platform (https://vivli.org/). Further details on Roche's criteria for eligible studies are available here (https://vivli.org/members/ourmembers/). For further details on Roche's Global Policy on the Sharing of Clinical Information and how to request access to related clinical study documents, see here (https://www.roche.com/research_and_development/who_we_are_how_we_work/clinical_trials/our_commitment_to_data_sharing.htm).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Copernicus Group IRB, PO Box 110605, Research Triangle Park, NC 27709; Washington University in St Louis Human Protection Office, 660 South Euclid Ave. Campus Box 8089 St Louis, MO 63110; Alpha IRB, 1001 Avenida Pico, Suite C#497 San Clemente, CA 92673; Integ Review IRB, 3815 S. Capital of Texas Hwy, Suite 320, Austin TX 78704. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DU, FL, and ML contributed to the conception and design of the study. W-YC performed the statistical analysis. All authors interpreted the data and commented on the draft manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by F. Hoffmann-La Roche Ltd.

ACKNOWLEDGMENTS

We thank the participating patients, their families, research coordinators, and nurses. Third-party medical writing assistance, under the direction of the authors, was provided by Victoria Eyre-Brook, PhD, at Gardiner-Caldwell Communications and was funded by F. Hoffmann-La Roche Ltd.

REFERENCES

1. Harvey PD, Strassnig MT, Silberstein J. Prediction of disability in schizophrenia: symptoms, cognition and self-assessment. *J Exp Psychopathol* (2019) 10:1–20. doi: 10.1177/2043808719865693
2. Bobes J, Arango C, Garcia-Garcia M, Rejas J, Group CSC. Prevalence of negative symptoms in outpatients with schizophrenia spectrum disorders treated with antipsychotics in routine clinical practice: findings from the CLAMORS study. *J Clin Psychiatry* (2010) 71:280–6. doi: 10.4088/JCP.08m04250yel
3. Foussias G, Remington G. Negative symptoms in schizophrenia: avolition and Occam's razor. *Schizophr Bull* (2010) 36:359–69. doi: 10.1093/schbul/sbn094
4. Kumari S, Malik M, Florival C, Manalai P, Sonje S. An assessment of five (PANSS, SAPS, SANS, NSA-16, CGI-SCH) commonly used symptoms rating scales in schizophrenia and comparison to newer scales (CAINS, BNSS). *J Addict Res Ther* (2017) 8:324. doi: 10.4172/2155-6105.1000324
5. Hyman SE. Psychiatric drug development: diagnosing a crisis. *Cerebrum* (2013) 5.
6. Walther S, Stegmayer K, Horn H, Rampa L, Razavi N, Müller TJ, et al. The longitudinal course of gross motor activity in schizophrenia - within and between episodes. *Front Psychiatry* (2015) 6:10:10. doi: 10.3389/fpsyt.2015.00010
7. Bracht T, Horn H, Strik W, Federspiel A, Razavi N, Stegmayer K, et al. White matter pathway organization of the reward system is related to positive and negative symptoms in schizophrenia. *Schizophr Res* (2014) 153:136–42. doi: 10.1016/j.schres.2014.01.015
8. Walther S, Stegmayer K, Federspiel A, Bohlhalter S, Wiest R, Viher PV. Aberrant hyperconnectivity in the motor system at rest is linked to motor abnormalities in schizophrenia spectrum disorders. *Schizophr Bull* (2017) 43:982–92. doi: 10.1093/schbul/sbx091
9. Walther S, Koschorke P, Horn H, Strik W. Objectively measured motor activity in schizophrenia challenges the validity of expert ratings. *Psychiatry Res* (2009) 169:187–90. doi: 10.1016/j.psychres.2008.06.020
10. Docx L, Morrens M, Bervoets C, Hulstijn W, Fransen E, De Hert M, et al. Parsing the components of the psychomotor syndrome in schizophrenia. *Acta Psychiatr Scand* (2012) 126:256–65. doi: 10.1111/j.1600-0447.2012.01846.x
11. Wee ZY, Yong SWL, Chew QH, Guan C, Lee TS, Sim K. Actigraphy studies and clinical and biobehavioural correlates in schizophrenia: a systematic review. *J Neural Transm (Vienna)* (2019) 126:531–58. doi: 10.1007/s00702-019-01993-2
12. Kluge A, Kirschner M, Hager OM, Bischof M, Habermeyer B, Seifritz E, et al. Combining actigraphy, ecological momentary assessment and neuroimaging to study apathy in patients with schizophrenia. *Schizophr Res* (2018) 195:176–82. doi: 10.1016/j.schres.2017.09.034
13. Cella M, Okruszek L, Lawrence M, Zarlenga V, He Z, Wykes T. Using wearable technology to detect the autonomic signature of illness severity in schizophrenia. *Schizophr Res* (2018) 195:537–42. doi: 10.1016/j.schres.2017.09.028
14. Moschetti A, Fiorini L, Esposito D, Dario P, Cavallo F. Recognition of daily gestures with wearable inertial rings and bracelets. *Sensors (Basel)* (2016) 16:1341. doi: 10.3390/s16081341
15. Shoaib M, Bosch S, Incel OD, Scholten H, Havinga PJ. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors (Basel)* (2016) 16:426. doi: 10.3390/s16040426
16. Cheng W, Scotland A, Lipsmeier F, Kilchenmann T, Jin L, Schjodt-Eriksen J, et al. Human Activity Recognition from Sensor-Based Large-Scale Continuous Monitoring of Parkinson's Disease Patients. In: *2017 IEEE/ACM International*

Conference on Connected Health. Philadelphia, PA: Applications, Systems and Engineering Technologies (CHASE (2017). p. 249–50. doi: 10.1109/CHASE.2017.87
17. Totty MS, Wade E. Muscle activation and inertial motion data for noninvasive classification of activities of daily living. *IEEE Transl BioMed Eng* (2018) 65:1069–76. doi: 10.1109/TBME.2017.2738440
18. Barch DM, Treadway MT, Schoen N. Effort, anhedonia, and function in schizophrenia: reduced effort allocation predicts amotivation and functional impairment. *J Abnorm Psychol* (2014) 123:387–97. doi: 10.1037/a0036299
19. Treadway MT, Peterman JS, Zald DH, Park S. Impaired effort allocation in patients with schizophrenia. *Schizophr Res* (2015) 161:382–5. doi: 10.1016/j.schres.2014.11.024
20. McCarthy JM, Treadway MT, Bennett ME, Blanchard JJ. Inefficient effort allocation and negative symptoms in individuals with schizophrenia. *Schizophr Res* (2016) 170:27884. doi: 10.1016/j.schres.2015.12.017
21. Marder SR, Davis JM, Chouinard G. The effects of risperidone on the five dimensions of schizophrenia derived by factor analysis: combined results of the North American trials. *J Clin Psychiatry* (1997) 58:538–46. doi: 10.4088/jcp.v58n1205
22. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull* (1987) 13:261–76. doi: 10.1093/schbul/13.2.261
23. Kirkpatrick B, Strauss GP, Nguyen L, Fischer BA, Daniel DG, Cienfuegos A, et al. The brief negative symptom scale: psychometric properties. *Schizophr Bull* (2011) 37:300–5. doi: 10.1093/schbul/sbq059
24. Strauss GP, Hong LE, Gold JM, Buchanan RW, McMahon RP, Keller WR, et al. Factor structure of the Brief Negative Symptom Scale. *Schizophr Res* (2012) 142:96–8. doi: 10.1016/j.schres.2012.09.007
25. Gold JM, Strauss GP, Waltz JA, Robinson BM, Brown JK, Frank MJ. Negative symptoms of schizophrenia are associated with abnormal effort-cost computations. *Biol Psychiatry* (2013) 74:130–6. doi: 10.1016/j.biopsych.2012.12.022
26. Ordonez JF, Roggen D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* (2016) 16:115. doi: 10.3390/s16010115
27. Reiss A, Stricker D. *Introducing a new benchmarked dataset for activity monitoring. Proceedings of the 16th Annual International Symposium on Wearable Computers (ISWC)*. ResearchGate (2012). pp. 108–9.
28. Stisen A, Blunck H, Bhattacharya S, Siiger Prentow T, Baun Kjaergaard M, Dey A, et al. *Smart devices are different: assessing and mitigating mobile sensing heterogeneities for activity recognition. Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys 2015)*. Seoul, South Korea: ACM (2015). p. 127–40. doi: 10.1145/2809695.2809718
29. Midaglia L, Mulero P, Montalban X, Graves J, Hauser SL, Julian L, et al. Adherence and satisfaction of smartphone- and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: nonrandomized interventional feasibility study. *J Med Internet Res* (2019) 21:e14863. doi: 10.2196/14863
30. Rai A, Chintalapudi KK, Padmanabhan VN, Sen R. *Zee: zero-effect crowdsourcing for indoor localization. Proceedings of the 18th Annual International Conference on Mobile Computing and Networking, MobiCom" 12*. New York, NY, USA: ACM (2012) p. 293–304. doi: 10.1145/2348543.2348580
31. Meyer N, Kerz M, Folarin A, Joyce DW, Jackson R, Karr C, et al. Capturing rest-activity profiles in schizophrenia using wearable and mobile technologies: development, implementation, feasibility, and acceptability of a remote monitoring platform. *JMIR Mhealth Uhealth* (2018) 6:e188. doi: 10.2196/mhealth.8292
32. Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Mov Disord* (2018) 33:1287–97. doi: 10.1002/mds.27376
33. Chen O, Lipsmeier F, Phan H, Prince J, Taylor K, Gossens C, et al. Building a machine-learning framework to remotely assess Parkinson's disease using smartphones. *IEEE Trans BioMed Eng* (2020). doi: 10.1109/TBME.2020.2988942
34. Lipsmeier F, Simillion C, Bamdadian Atieh A, Smith A, Schobel S, Gossens C, et al. Reliability, feasibility and validity of a novel digital monitoring platform assessing cognitive and motor symptoms in people with Stage I and II Huntington's disease (HD). *Mov Disord* (2019) 34(Suppl 2). Abstract 26 and poster presented at the American Academy of Neurology 2019 meeting.
35. Tortelli R, Simillion C, Lipsmeier F, Kilchenmann T, Rodrigues FB, Byrne LM, et al. The Digital-HD study: smartphone-based remote testing to assess cognitive and motor symptoms in Huntington's disease. *Neurology* (2020) 94(Suppl 15).
36. Moran EK, Culbreth AJ, Barch DM. Ecological momentary assessment of negative symptoms in schizophrenia: Relationships to effort-based decision making and reinforcement learning. *J Abnorm Psychol* (2017) 126:96–105. doi: 10.1037/abn0000240
37. From the American Association of Neurological Surgeons (AANS), American Society of Neuroradiology (ASNR), Cardiovascular and Interventional Radiology Society of Europe (CIRSE), Canadian Interventional Radiology Association (CIRA), Congress of Neurological Surgeons (CNS) and European Society of Minimally Invasive Neurological Therapy (ESMINT), et al. Multisociety consensus quality improvement revised consensus statement for endovascular therapy of acute ischemic stroke. *Int J Stroke* (2018) 13:612–32. doi: 10.1177/1747493018778713
38. Walther S, Stegmayer K, Sulzbacher J, Vanbellingen T, Müri R, Strik W, et al. Nonverbal social communication and gesture control in schizophrenia. *Schizophr Bull* (2015) 41:338–45. doi: 10.1093/schbul/sbu222

Conflict of Interest: All authors are employed by company F. Hoffmann - La Roche Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

METHODS
published: 17 September 2020
doi: 10.3389/fpsyt.2020.572938

Assessing Saccadic Eye Movements With Head-Mounted Display Virtual Reality Technology

Yu Imaoka[1]*, *Andri Flury*[1] and *Eling D. de Bruin*[1,2]

[1] Motor Control & Learning Laboratory, Institute of Human Movement Sciences and Sport, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland, [2] Division of Physiotherapy, Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Stockholm, Sweden

As our society is ageing globally, neurodegenerative disorders are becoming a relevant issue. Assessment of saccadic eye movement could provide objective values to help to understand the symptoms of disorders. HTC Corporation launched a new virtual reality (VR) headset, VIVE Pro Eye, implementing an infrared-based eye tracking technique together with VR technology. The purpose of this study is to evaluate whether the device can be used as an assessment tool of saccadic eye movement and to investigate the technical features of eye tracking. We developed a measurement system of saccadic eye movement with a simple VR environment on Unity VR design platform, following an internationally proposed standard saccade measurement protocol. We then measured the saccadic eye movement of seven healthy young adults to analyze the oculo-metrics of latency, peak velocity, and error rate of pro- and anti-saccade tasks: 120 trials in each task. We calculated these parameters based on the saccade detection algorithm that we have developed following previous studies. Consequently, our results revealed latency of 220.40 ± 43.16 ms, peak velocity of 357.90 ± 111.99°/s, and error rate of 0.24 ± 0.41% for the pro-saccade task, and latency of 343.35 ± 76.42 ms, peak velocity of 318.79 ± 116.69°/s, and error rate of 0.66 ± 0.76% for the anti-saccade task. In addition, we observed pupil diameter of 4.30 ± 1.15 mm (left eye) and 4.29 ± 1.08 mm (right eye) for the pro-saccade task, and of 4.21 ± 1.04 mm (left eye) and 4.22 ± 0.97 mm (right eye) for the anti-saccade task. Comparing between the descriptive statistics of previous studies and our results suggests that VIVE Pro Eye can function as an assessment tool of saccadic eye movement since our results are in the range of or close to the results of previous studies. Nonetheless, we found technical limitations especially about time-related measurement parameters. Further improvements in software and hardware of the device and measurement protocol, and more measurements with diverse age-groups and people with different health conditions are warranted to enhance the whole assessment system of saccadic eye movement.

Keywords: saccadic eye movement, virtual reality, head mounted display (HMD), dementia, neurological disorder, ageing, pupillary response, saccade

OPEN ACCESS

Edited by:
Jennifer H. Barnett, Cambridge Cognition, United Kingdom

Reviewed by:
Jan Drewes, University of Trento, Italy
Iva Georgieva, Institute for Advanced Study Varna, Bulgaria

***Correspondence:**
Yu Imaoka
yu.imaoka@hest.ethz.ch

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: *15 June 2020*
Accepted: *18 August 2020*
Published: *17 September 2020*

Citation:
Imaoka Y, Flury A and de Bruin ED (2020) Assessing Saccadic Eye Movements With Head-Mounted Display Virtual Reality Technology. Front. Psychiatry 11:572938. doi: 10.3389/fpsyt.2020.572938

INTRODUCTION

Our society is ageing worldwide. According to the United Nations (UN), the number of people aged 65 years and over is projected to increase from 0.7 billion (9% of the global population) in 2019 to 1.5 billion (16%) in 2050 (1). In addition, the elderly people aged 60 years and over are expected to outnumber the children under 5 years in 2020 for the first time in our history (2). Because of this development, the World Health Organization (WHO) launched The Decade of Healthy Ageing (2020–2030), an opportunity to bring together relevant stakeholders to improve the lives of older people in our entire society (2). Particularly, WHO emphasizes the importance of developing guidance and measurement tools for primary care providers so that they can assess health status of the elderly more comprehensively to slow and/or reverse the declines in their physical and mental capacities (3, 4).

Among various relevant diseases and disorders confronting us in the growing ageing society, neurodegenerative diseases are becoming more prevalent (5). It is reported that neurological disorders were the leading cause of Disability-Adjusted Life Year (DALY) in 2015, accounting for 10.2% of global DALYs, and were the second leading cause of deaths, comprising 16.8% of global deaths (6). Detecting early disease-related symptoms would facilitate implementation of preventive measures. In this context, it has been hypothesized that assessment of eye movements can be invaluable for healthcare providers because eye tracking provides indirect access to the neural and cognitive processing in a simple manner (7–9) and associates with neurodegeneration (10). Ocular movements can be subcategorized into two clauses: 1) fixation, blink, vergence, smooth pursuit, vestibuloocular reflexes, optokinetic nystagmus, and pupillary responses, and 2) saccade (7, 9). While it is important to evaluate eye movements with combining these subcategorized events (9), saccadic eye movement is often assessed in various neurodegenerative disorders; e.g., in dementia (11).

A saccade refers to a rapid and conjugate eye movement that voluntarily shifts the eyes from one target to another (12). We usually perform the saccades, initiating the movement within 250 ms, shifting our eyes at the speed up to 700°/s, and completing the shift in 30 ∼ 100 ms (12, 13). Saccades are generally subdivided into two main classes: 1) saccades that are made in response to an external guide of visual stimuli and 2) saccades that are performed without visual target (14). These two types of saccades work differently in terms of the brain processing. One of the frequently used methods for the former type of saccade is the visually guided saccade (VGS). The subjects first look at a central fixed target and then have to move their eyes to another target that appears at some point outside the center of their visual field after the central target disappears. An example paradigm of the latter type of saccade is called the memory-guided saccade (MGS). The saccade procedure is similar to the VGS; however, the subjects need to remember for a short time the location of another target toward which they have to make a saccade. In performing a saccade, several cortical areas play an important role: frontal eye field (FEF), supplementary eye field (SEF), parietal eye field (PEF) or posterior parietal cortex (PPC), and superior colliculus (SC) (14). Specifically, FEF and PPC are important to initiate a saccade. The parietal cortex generally contributes to triggering automatic or reflexive saccades, while the frontal lobe, including the FEF, is affected by more cognitively demanding eye movement tasks. After SC receives the signals from FEF, SEF, and PEF (or PPC), it sends out next signals to the saccade generators in the brainstem. Finally, the brainstem saccade generator sends its outputs to the motoneurons of the oculomotor nuclei. To investigate saccades in detail, oculo-metrics such as latency (response time), velocity, amplitude, frequency response, duration, and error rate are often analyzed (9).

Saccadic eye movement disorders can be categorized, depending on the characteristics of measured oculometrics, to 1) hypokinetic movement disorders often seen in Parkinson's Disease (PD), Multiple System Atrophy (MSA), Progressive Supranuclear Palsy (PSP), and Corticobasal Degeneration (CBD) and 2) hyperkinetic movement disorders observed in, for example, Huntington's Disease (HD) and Spinocerebellar Ataxia (SCA) (15). For example, people suffering from PD with mild cognitive impairment (PD-MCI) showed a longer latency in comparison with a control group (8, 14). The saccade latency was associated with the brain regions that were affected in people with PD (9). Those with Alzheimer's Disease (AD) and amnestic MCI (aMCI) also exhibited significantly longer response time in the anti-saccade task compared to non-aMCI population and healthy controls. More errors were also observed in people with AD and aMCI (11). Obliquely oriented microsaccades were often seen in people with AD and aMCI (8). Most people with MSA presented abnormally large square wave jerks (SWJ) (7, 15). The population with PSP revealed more frequent and larger SWJ, slower saccades, prolonged latencies, and impaired pursuit ocular movement in comparison with the healthy population (7, 13–15). Those with HD also showed dysfunction in fixing their eye stably, impaired initiation and inhibition of saccadic eye movements, and longer response time and decreased velocity in saccades (7, 8, 13, 15). People with SCA manifested unusually large SWJ, slowed saccades, and slightly increased latency of saccades (7, 14). Hence, as described above, saccadic eye movement could be a practical useful biomarker to understand the symptoms of diverse neurological disorders.

There are several methods of eye movement measurement (7, 9, 14). A simple method is to use Frenzel goggles to disable the user from visually fixating on an object while the examiner investigates the eyes of the user. However, since the method does not provide quantitative outcomes, it is not possible to acquire oculo-metrics such as latency of saccadic eye movement. Instead, computer-based recording techniques can quantify ocular movements. Electrooculography (EOG) is one of the traditional techniques and has been employed since the 1970s. Electrodes are placed on the skin around the eyes to record the changes in eye position from the differences in electrical potential between the two electrodes. While EOG provides a good temporal resolution, saccade measurement is sometimes affected by artifacts such as electromyography signals. Another

method is to put on an eye directly a modified contact lens with a search coil embedded. When the eye with the search coil moves, a current is induced on the coil and the electrical change can be used to measure the eye position. It is the most accurate method, however, is invasive and painful for the users. Recently, video-based eye tracking is becoming more prevalent. Generally, infrared light is emitted from the light source on the eye tracker to the cornea of eyes. Then, infrared cameras record the positions of reflected light from the cornea and the center of pupil. The relative difference between these two positions is used to calculate the pupil position and gaze direction. The users usually conduct the calibration process to look at several points in the visual field and then the system compares the predetermined points and the measured gaze data to adjust the system configuration depending on the user. The method is non-invasive, offers a high spatial resolution of 0.25° to 0.5°, and provides data of pupillary responses in addition to gaze data. On the other hand, since sampling frequency is a key factor for accurate saccadic eye movement assessment, it will cost more if the eye tracking device with higher sampling frequency is selected. Nonetheless, video-based eye tracking technique is more promising to characterize ocular motions in the absence of absolute biomarkers for ocular assessment.

As the technology of video-based eye tracking has advanced, developers have integrated eye tracking technique into a virtual reality (VR) technology with head-mounted display (HMD) in recent years. The combined system enables us to measure eye movements while showing VR animations. VR is also an effective tool for both diagnosis and intervention in the research field of neurodegenerative disorders (16–18). Previous studies reported that interventions with using VR technology improved motor and cognitive functions of people with disorders such as stroke, MCI, AD, and PD (19, 20). Other studies also found that VR-based measures were more useful as an assessment tool in detecting cognitive impairments and evaluating self-awareness (21–23).

To summarize, the previous studies have found that assessment of saccadic eye movement with video-based eye tracking bears the potential to evaluate various neurodegenerative disorders. In addition, VR technology would have the potential to enhance the diagnosis and interventions of neurodegenerative disorders. We assume that combination of video-based eye tracking technique and HMD-based VR technology could improve the assessment of saccadic eye movement by taking advantages of immersive environments created by VR technology. Therefore, this study aims at investigating a combined device, HTC VIVE Pro Eye, for the purpose of using it for saccadic eye movement assessment and report the measured data of saccadic eye movement and technological findings of the device.

MATERIALS AND EQUIPMENT

The main component of this study is the VR-based HMD (HTC VIVE Pro Eye, HTC Corporation). **Table 1** shows the technical specifications of hardware and software components. **Table 1** also lists the main measurement parameters and explains how to interpret the output value of validity of measured eye data. In addition, **Table 2** shows the main technical specifications of the computer used in this study. VR environments are designed on the computer and the VIVE Pro Eye is controlled from the same computer.

We develop VR environments and implement an eye tracking software algorithm on Unity, following a software development kit (SDK) called SRanipal provided by HTC Corporation. SRanipal includes functions to measure ocular movements and time-related data. The detailed guideline for the program development is found in the SDK; when the SRanipal SDK is installed into a computer, a folder automatically named SRanipal_SDK_1.1.0.1 is created. The guideline is found in the folder: SRanipal_SDK_1.1.0.1\02_Unity\Document\Eye \Document_Unity.html.

Figure 1 shows the coordinate system of VIVE Pro Eye. Recorded data of gaze origin and gaze direction are three-dimensional and based on the right-handed coordinate system. Gaze direction data are normalized to between −1 and 1. Pupil position data are also normalized to between 0 and 1. The origin (0, 0) of pupil position data is at the top left of the sensor area from the user perspective, (0.5, 0.5) is at the center of view field, and (1, 1) is at the bottom right of the sensor area.

RESEARCH METHODS

Study Design

The aim of our study is to evaluate whether the VR headset, HTC VIVE Pro Eye, could be used as an assessment tool for saccadic eye movement. As a first step, we developed a simple VR environment and measurement protocol for saccadic eye movement assessment on Unity design platform, following a previously proposed protocol of saccade evaluation (24), to simulate the environment similar to the design that had been often used on a monitor-based assessment system. Subsequently, we measured saccadic eye movement of healthy young adults. We processed the measured data to calculate oculo-metrics such as peak velocity and latency, analyzed them, and compared the results with those from previous studies. Finally, we summarized the technical findings, limitations, and improvements that were observed in this study. The project was organized at ETH Zurich, Switzerland from August 2019 to February 2020. The ethics was approved by ETH Zurich Ethics Commission (registration number 2019-N-181). We recruited healthy young (between 18 and 35 years old) adults in Switzerland.

Experimental System

Figure 2 illustrates the experimental system developed for the assessment of saccadic eye movement using HTC VIVE Pro Eye. The VR environment is designed on the computer of NUC8i7HVK and is output to the VIVE Pro Eye headset *via* the link box. The eye tracker embedded in the VR headset records the ocular movement and the measured data are stored in the computer storage. The base stations are necessary to detect the VR headset.

TABLE 1 | Technical specifications of investigated HTC VIVE Pro Eye.

Item		Specification	
VR headset	Screen	Dual OLED 3.5" diagonal	
	Resolution	1440 x 1600 pixels per eye (2880 x 1600 pixels combined)	
	Refresh rate	90 Hz	
	Field of view	110°	
	Audio	High resolution	
	Input	Dual integrated microphones	
	Interface	USB-C 3.0, DP 1.2, Bluetooth	
	Sensors	SteamVR tracking	
		Accelerometer	
		Gyroscope	
		Proximity	
		Interpupillary distance (IPD) sensor	
		Near-infrared (NIR 850nm) LED (9 for each eye)	
		Infrared camera (1 for each eye)	
Eye tracker	Sampling frequency (binocular)	120 Hz	
	Accuracy (within FOV 20)	0.5°~ 1.1°	
	Calibration	5 points	
	Trackable field of view	110°	
	Major measurement parameters	Frame sequence	
		Timestamp (ms)	
		Gaze origin (mm)	
		Gaze direction (normalized to between −1 and 1)	
		Pupil position (normalized to between 0 and 1)	
		Pupil diameter (mm)	
		Eye openness (normalized to between 0 and 1)	
		Validity of eye data (The details are below.)	
		Enumerator	Binary digit if valid	Decimal digit if valid
		Gaze origin	00001	1
		Gaze direction	00010	2
		Pupil diameter	00100	4
		Eye openness	01000	8
		Pupil position	10000	16
Software	HTC VIVE Pro Eye setup	version 1.0.8.161	
	HTC software development kit for eye tracking	SRanipal version 1.1.0.1	
	SR Runtime	version 1.1.2.0	
	Steam VR	version 1.11.11	
	Unity, VR design platform	version 2019.2.5f1	

TABLE 2 | Technical specifications of computer.

Item	Manufacturer	Model and specification
Computer	Intel	NUC8i7HVK
CPU	Intel	core i7-8809G
Memory	Kingston	ValueRAM SO-DDR4-RAM 2400 MHz 16 GB, SO-DIMM 260 Pins
Storage	Samsung	SSD MZ-V6E500BW, M.2 500GB
Graphic	Intel/AMD	Intel HD Graphics 630/Radeon RX Vega M GH graphics
OS	Microsoft	Windows 10 Education, version 1903

Research Protocol

Protocol of Saccadic Eye Movement Assessment

We designed our assessment system of saccadic eye movement, following a previously proposed standardized saccade protocol (24). Oculo-motor scientists and clinicians with diverse experience in saccadic eye movement analysis developed the standardized protocol, with the aim of enhancing the clinical and scientific outcomes. **Figure 3** illustrates the detailed measurement protocol visually, and the following are the points that are recommended in the proposed standardized protocol (24).

Saccade task:

We prepared pro- and anti-saccade tasks in a horizontal direction. Anti-saccade task is a reliable and sensitive measure to evaluate the processes involved in resolving the conflict between volitional and reflexive behavioral responses and can provide important insights on the conditions of various neurodegenerative disorders (25). In general, the subjects look at a target at the center first before performing the saccade tasks. After the target at the center disappears, another target appears on either the right or the left side. The

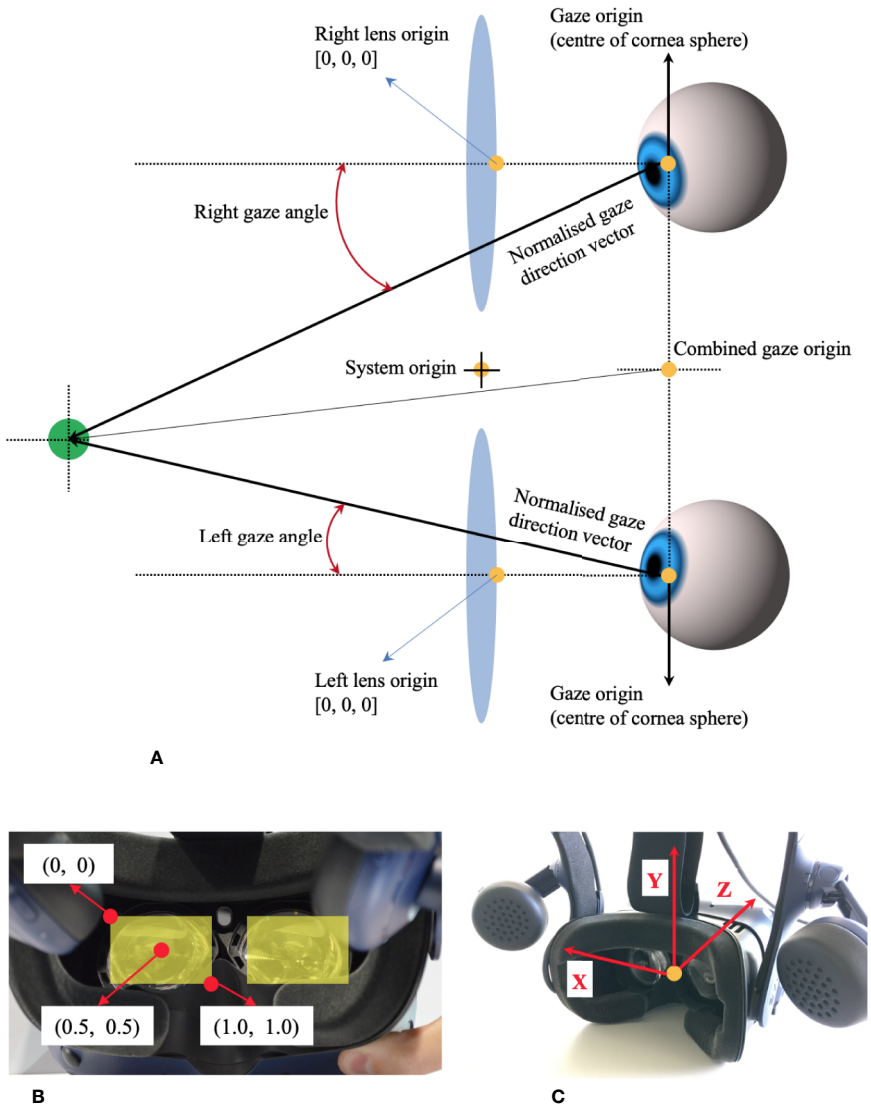


FIGURE 1 | Coordinate system of HTC VIVE Pro Eye, based on the manual of SRanipal SDK. **(A)** Coordinate system of eye tracking on VIVE Pro Eye. **(B)** Coordinate system of pupil position data from user's view. **(C)** Coordinate system of gaze direction vector from user's view.

subjects need to move their eyes toward the new target on the side for the case of pro-saccade and toward the opposite direction of the target for the case of anti-saccade.

Protocol of a saccade trial on time domain:

One saccade trial consisted of two phases: 1) a white circle appeared at the center of visual field for between 1 and 3.5 s with the mean of 1.5 s averaged over each set of saccade trials and 2) a red circular target appeared on either the right or the left of the white circle for 1 s.

Time interval between the two phases:

Some measurement protocols set a time interval between after the white target disappears and before the red target appears. However, following the recommendation, we removed the gap phase between these two phases, setting the interval to 0.

Direction and amplitude of stimuli (red target):

As the literature recommends the saccade task in only a horizontal direction and an amplitude of 8°–10° for the red target, we created the VR environment, where the red target came out on the right or left with the amplitude of 8° from the center.

Contrast of targets:

The contrast of the targets was clear enough with over 50%.

Size of targets:

While the proposed protocol recommended the diameter of 0.5° for the system with a screen display, it also stated that the size

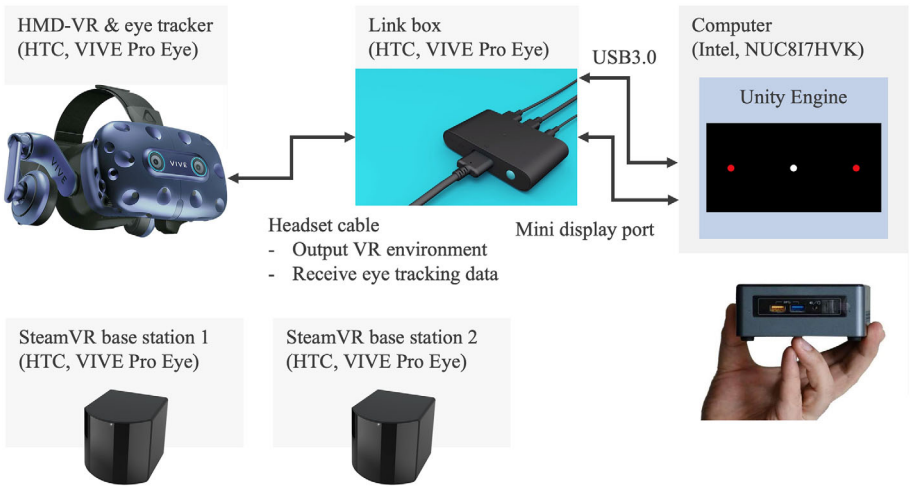


FIGURE 2 | Experiment system of saccadic eye movement assessment with HTC VIVE Pro Eye.

and shape of the targets were not important. Therefore, we designed the targets with the diameter of 1°. In addition, as shown in **Figure 3**, we set the distance of 7 m between the user's view and each target. With the predetermined parameters of the distance and the amplitude and size of the targets, we calculated the diameter of targets in meter.

Saccade task flow:

The measurement flow was composed of five main phases. The subjects started with 60 trials of pro-saccade task after practicing the pro-saccade task for 10 times for customization. After a break for 1 min, they practiced anti-saccade task for four times and then moved to 40 trials of anti-saccade task. They performed the second and third anti-saccade tasks consecutively for 40 trials in each with having 1-min short break between the phases and finally ended with the second pro-saccade task for 60 trials. Thus, they conducted 120 trials in each of the pro- and anti-saccade tasks. In each saccade task, the subjects gazed at the red target on the right or left at the equal number of times (i.e., 30 or 20 times on the right and 30 or 20 times on the left for pro-saccade task or anti-saccade task). The fore-period, a duration to display the white target at the center of visual field, varied at random in each saccade trial and the direction of displayed red targets also changed randomly. However, all the subjects experienced the same randomized fore-periods and directions of red targets. The whole measurement took less than 20 min.

Programming Algorithm of Eye Tracking

The program was developed using C# programming language on Unity. The programming code is openly published on GitHub: https://github.com/MotorControlLearning, with the detailed explanation of the algorithm for the saccade measurement.

Data Processing

List of Recorded Parameters

We measured the following parameters especially to calculate the important oculo-metrics of latency, peak velocity, and error rate of saccadic eye movement (26). We recorded time information with timestamp in SRanipal SDK and DateTime.Now.Ticks on Unity system. The SDK also provided frame_sequence to record the frame sequence. The following data of ocular movement in each of the left and right eyes were read from the VerboseData in the struct data of ViveSR.anipal.Eye.EyeData_v2 in SRanipal SDK: validity of eye data, eye openness level, pupil diameter, pupil position, gaze origin, and gaze direction. All the data are stored in text files. However, we did not use the data of timestamp because the current version of SRanipal SDK did not provide correct time stamp. The details are described in the following discussion part.

Detection Algorithm of Saccades

Figure 4 explains the flow of data processing for eye movement data and detection algorithm of saccadic eye movement visually.

First, we performed a data cleaning. After we loaded the raw data of ocular motions from the text files on MATLAB R2019b (MathWorks, MA, U.S.), we checked the value of validity of eye data to eliminate the invalid data with the value less than 31 and kept the valid data with the value equal to 31 (see **Table 1**).

Second, we processed the cleaned data to calculate the gaze direction in degrees with the formula (1). We quantified how far eyes moved from the point of $x = 0$ (see **Figure 1**). Since the recorded data of gaze direction were normalized to between −1 and 1, we calculated the angle in radian by applying the arc tangent and then converted the data from radian to degrees. Here, we have to note that the sign of the calculated data in degrees is negative when eyes move toward the right and is positive when eyes move toward the left (see the coordinate

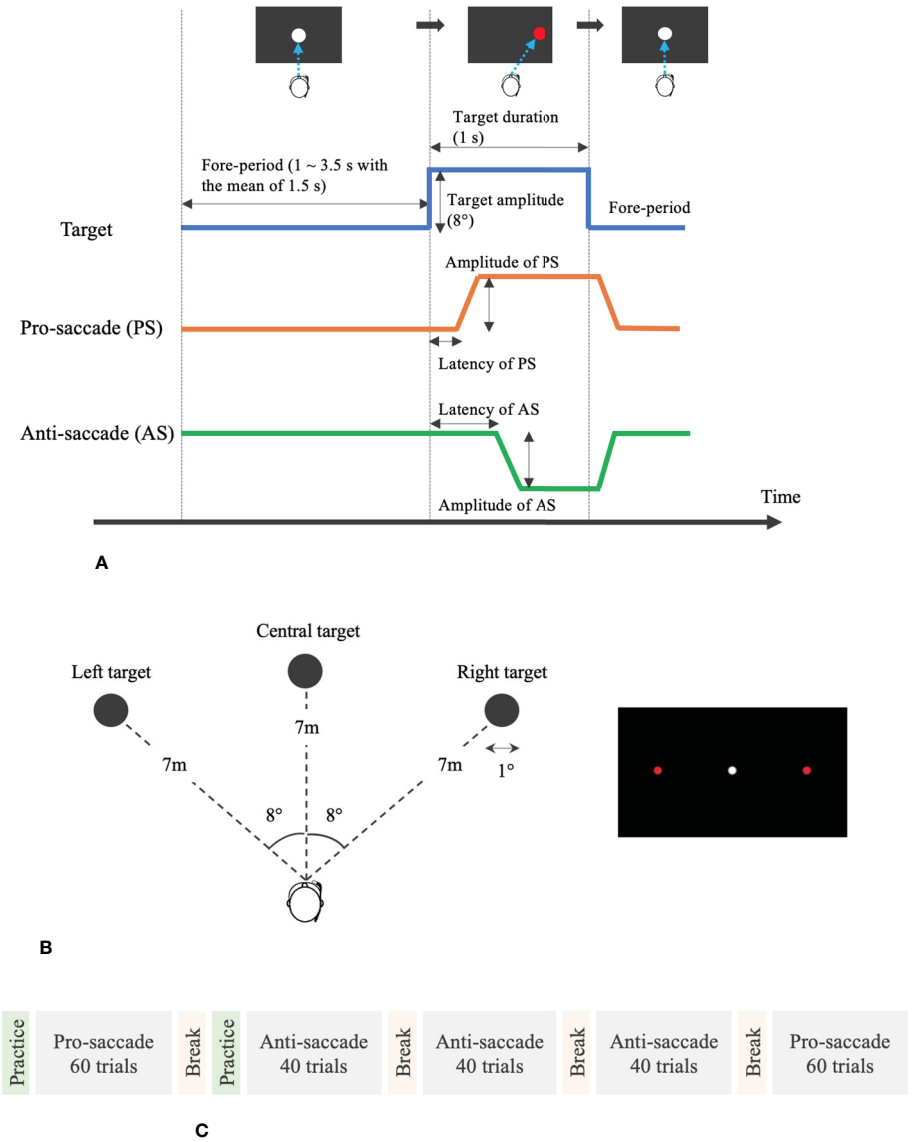


FIGURE 3 | Measurement protocol and VR design for the assessment of saccadic eye movement. **(A)** Protocol of pro- and anti-saccade tasks per trial. **(B)** Designed VR environment for saccade assessment. **(C)** Measurement flow of saccade assessment.

system in **Figure 1**). In addition, we identified the time when the red circular target appeared on either the right or the left in each trial, by reading the Unix time that we recorded in every saccade trial when the red target was displayed.

$$E_x = \frac{tan^{-1}\left(\frac{G\vec{D}_x}{G\vec{D}_z}\right)}{\pi} 180^{\circ} \quad (1)$$

E_x : Gaze direction in degrees on X axis

$G\vec{D}_x$: Normalized gaze direction on X axis

$G\vec{D}_z$: Normalized gaze direction on Z axis

Third, we implemented a program to detect the saccades in each trial. Since spike noise possibly confounded the saccade detection, we removed the spike noise with a median filter with an order of 10 (27). We then extracted the filtered gaze direction data of first saccade trial in a specific time range within the period of 1 s for which the red target was being displayed. Since we knew when the red target appeared based on the Unix time as explained above, we extracted the gaze direction data in a time range from t_i + 100 to t_i + 500 ms for the pro-saccade task and from t_i + 125 to t_i + 750 ms for the anti-saccade task, where t_i was the time when the i_{th} (i = 1, 2,…, 240) red target appeared. We set the time range, assuming that saccadic eye movement would occur in the period by referring to the results of saccade

latency in the previous studies (28, 29). Subsequently, we drew velocity from the gaze direction data to understand the time when the velocity changed sharply by inspecting the peaks of velocity as visually explained in **Figure 4**. We used a velocity-based algorithm to detect saccadic eye movements (30). Investigating the waveform of the filtered gaze direction data, we considered five potential cases to detect saccades as illustrated in **Figure 4**. In case #1, gaze at the white central target (fore-period), initiation of saccade, and gaze at the red target are seen. Case #2 observes the return period to move eyes from the red target to the white target in addition to the movement illustrated in case #1. Case #3 is often found in the anti-saccade task because of reflexive eye movement induced by the red target. Thus, another peak velocity could be observed before the correct saccadic eye movement occurs. Similar to case #2, case #4 includes the return phase in addition to the movement in case #3. The final case of #5 is seen if the subjects do not perform the saccadic eye movement task properly. If the gaze direction changed within 1° or 0.001 for normalized pupil position in the inspection period, we applied the case #5. After normalizing the velocity data to between −1 and 1, we found the peaks of velocity that were over the threshold of 0.5 or below −0.5, according to the five cases. We assumed that first velocity peaks in cases #3 and #4 were due to reflexive responses if the

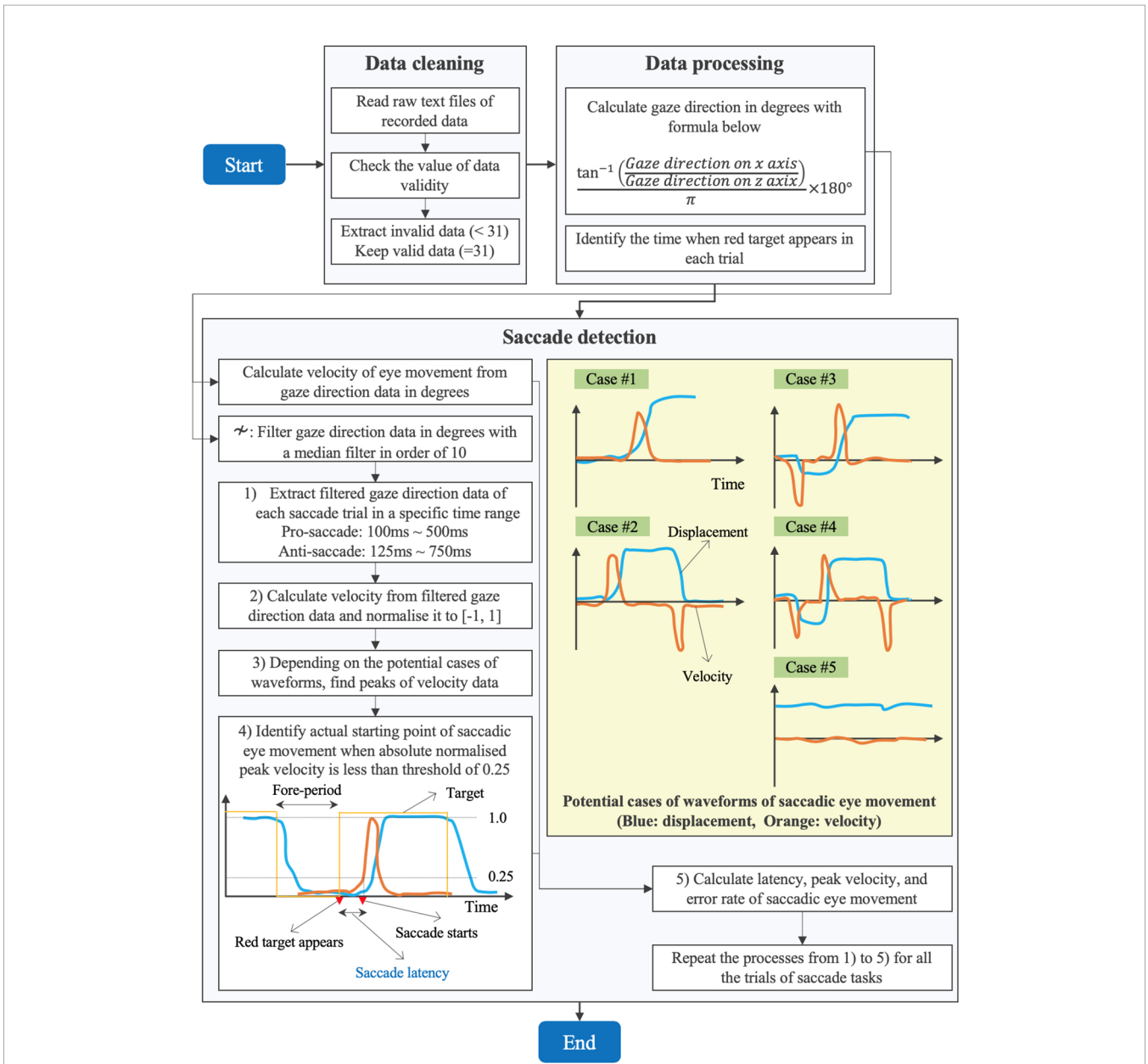


FIGURE 4 | Data processing of eye movement data and detection algorithm of saccades.

signs of peaks did not correspond to those of the predefined protocol. After fixing the peak of saccadic eye movement, we took the absolute value of velocity and set the threshold of 0.25 to define when the saccadic eye movement started in certain response time (i.e., saccade latency) after the red target appeared. Finally, we calculated the latency, peak velocity, and error rate of saccadic eye movement of the trial. We estimated the errors of saccadic eye movement by comparing the sign of peak velocity and the predetermined protocol explaining where the red target appeared in each saccade trial. For example, when the target appears on the right in the pro-saccade task, the peak velocity should be negative, whereas the velocity is supposed to be positive when the target comes out on the left in the visual field. We needed to invert the sign of peak velocity or the target direction of predefined protocol for the analysis of anti-saccade task. We repeated the same procedures explained above for all the saccade trials.

Data Analysis

Following the algorithm described in the previous sections, we calculated latency, peak velocity, and error rate of saccadic eye movement using gaze direction data. We also evaluated latency and error rate of saccade tasks using pupil position data based on the same saccade detection algorithm explained in **Figure 4**. However, peak velocity was not derived from the pupil position data since the conversion from normalized values to degrees was not available. Specifically, we visualized waveforms of gaze direction and pupil position data on X axis and data distribution of latency, peak velocity, and error rate in box plots. We also compared the calculated data between the data types (i.e., gaze direction data in degrees and normalized pupil position data), between pro- and anti-saccade tasks, and between left and right eyes to evaluate whether the data types, the saccade types, and the individual ocular measurement on each eye affected the results of computed oculo-metrics respectively. The statistical analysis was performed on a data analysis tool of R version 3.6.3. In addition, we investigated the data related to time by visualizing timestamp data recorded with SRanipal SDK and Unix time data recorded on Unity system, in particular to explore the sampling interval of the eye tracking device. Finally, we showed the measured data of pupil diameter as supplementary information.

RESULTS

Participants

Seven healthy young adults joined the experiment: four men and three women, 29 ± 4 years old (range 25–36 years).

Displacement of Gaze Direction in Degrees and Pupil Position

Figure 5 shows the displacement of gaze direction on X axis from the origin (see **Figure 1** for the coordinate system of VIVE Pro Eye). Separating the processed data depending on eyes (i.e., left and right), saccade tasks (i.e., pro- and anti-), and data types (i.e., gaze direction in degrees and normalized pupil position), we extracted each saccade trial of each participant and overlap each displacement waveform of each trial in a single figure.

Oculo-Metrics of Latency, Peak Velocity, and Error Rate of Saccadic Eye Movement, and Pupillary Response

Figure 6 illustrates the data distribution of the parameters: latency, peak velocity, and error rate of each pro- and anti-saccade task of each eye from all the participants. The first row of the figure is the result from gaze direction data in degrees and the second row is the result from normalized pupil position data. **Figure 7** shows the changes of pupil diameter of left and right eyes during each pro- and anti-saccade task in each participant. Each sub-figure illustrates the changes of pupillary response over the 120 saccade trials. **Table 3** shows the mean and standard deviation (SD) of each ocular parameter averaged over all the participants.

Sampling Interval of Eye Tracking

Figure 8 shows the sampling interval of eye tracking recorded with Unix time on Unity and with time stamp from SRanipal SDK of each participant. We particularly visualized the sampling interval of the first 3600 samples that we recorded; we assumed 30 s of recording time at the sampling frequency of 120 Hz. X axis is the order of samples and Y axis is the sampling interval in milliseconds between the samples. Each sub-figure shows two types of data: raw data and filtered data. We filtered the data of Unix time with a median filter with an order of 10 and the data of SRanipal time stamp with a moving average filter with a window size of 5. We then estimated the point where the sampling interval became smaller than 8 ms, as highlighted with red dash lines, by using the filtered data.

DISCUSSION

Displacement Data of Gaze Direction and Pupil Position

The first four sub-figures in **Figure 5** visualize the displacement of gaze direction in degrees both for each left and right eye in each pro- and anti-saccade task of all the participants. A clear difference between the eyes is not observed in both saccade tasks. On the other hand, we see differences between the saccade tasks visually. While the subjects moved their eyes almost precisely toward the red target at ±8° from the origin that was positioned as designed in **Figure 3** in the pro-saccade task: 7.3 ± 1.3° for left eye, 7.4 ± 1.2° for right eye, we observe a larger variability ranging from around ±3° to ±20° in the anti-saccade task: 8.3 ± 4.8° for left eye, 8.2 ± 4.9° for right eye. In addition, we visually see longer latency and more reflexive responses between 200 and 300 ms in the anti-saccade task compared to pro-saccade task; latency is 220.40 ± 43.16 ms in the pro-saccade task and 343.35 ± 76.42 ms in the anti-saccade task, and reflexive responses are observed for 1.4% of all the trials from all the participants in the pro-saccade task and for 13.2% in the anti-saccade task when we define the

TABLE 3 | Oculo-metrics of saccadic eye movement and pupillary response.

Parameters		Mean ± SD	
		Pro-saccade	Anti-saccade
Gaze direction data: Left eye	Latency	220.36 ± 43.31 [ms]	343.29 ± 76.22 [ms]
	Peak velocity	353.64 ± 108.02 [°/s]	316.15 ± 115.49 [°/s]
	Error rate	0.24 ± 0.41 [%]	0.60 ± 0.79 [%]
Gaze direction data: Right eye	Latency	220.45 ± 44.01 [ms]	343.41 ± 76.67 [ms]
	Peak velocity	362.15 ± 115.73 [°/s]	321.43 ± 117.89 [°/s]
	Error rate	0.24 ± 0.41 [%]	0.71 ± 0.75 [%]
Gaze direction data: Both eyes combined	Latency	220.40 ± 43.16 [ms]	343.35 ± 76.42 [ms]
	Peak velocity	357.90 ± 111.99 [°/s]	318.79 ± 116.69 [°/s]
	Error rate	0.24 ± 0.41 [%]	0.66 ± 0.76 [%]
Pupil position data: Left eye	Latency	220.08 ± 43.82 [ms]	342.55 ± 76.67 [ms]
	Error rate	0.24 ± 0.41 [%]	0.48 ± 0.66 [%]
Pupil position data: Right eye	Latency	220.05 ± 43.75 [ms]	342.60 ± 76.73 [ms]
	Error rate	0.24 ± 0.41 [%]	0.48 ± 0.66 [%]
Pupil position data: Both eyes combined	Latency	220.07 ± 43.77 [ms]	342.58 ± 76.67 [ms]
	Error rate	0.24 ± 0.41 [%]	0.48 ± 0.66 [%]
Pupillary data: Left eye	Pupil diameter	4.30 ± 1.15 [mm]	4.21 ± 1.04 [mm]
Pupillary data: Right eye	Pupil diameter	4.29 ± 1.08 [mm]	4.22 ± 0.97 [mm]

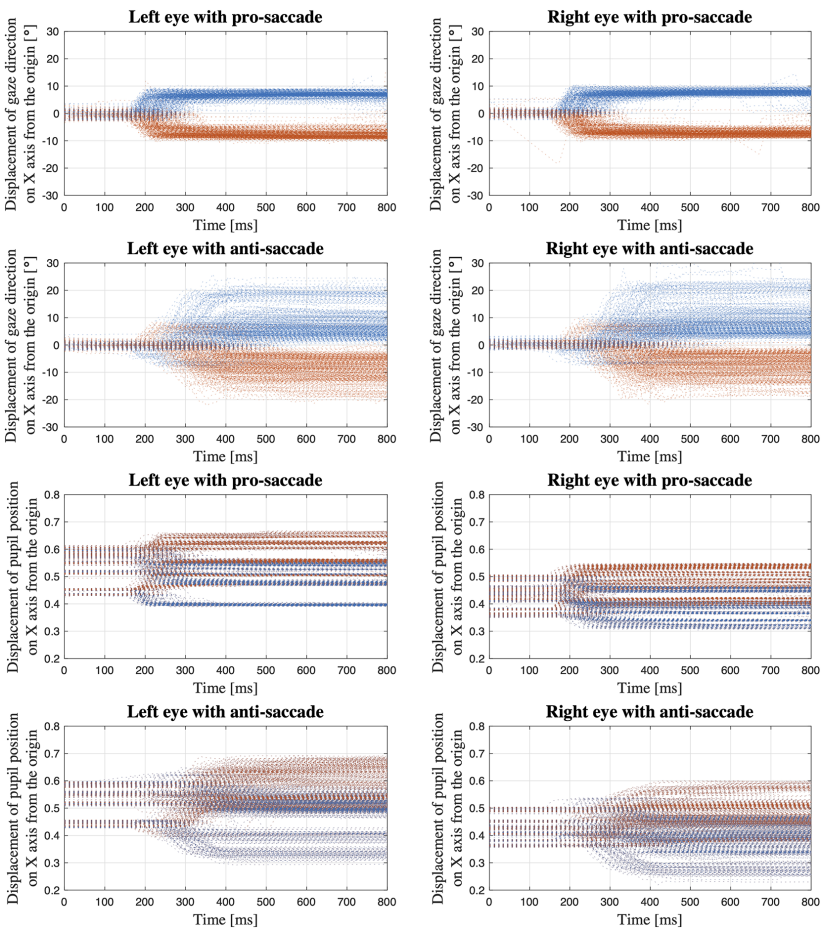


FIGURE 5 | Displacement of gaze direction in degrees and normalized pupil position on X axis from the origin; The data of all the saccade trials of all the participants are overlapped within the same time period from the time when the red target appears to the time 800 ms after the target appears (Blue dots: gaze toward the left; Orange dots: gaze toward the right).

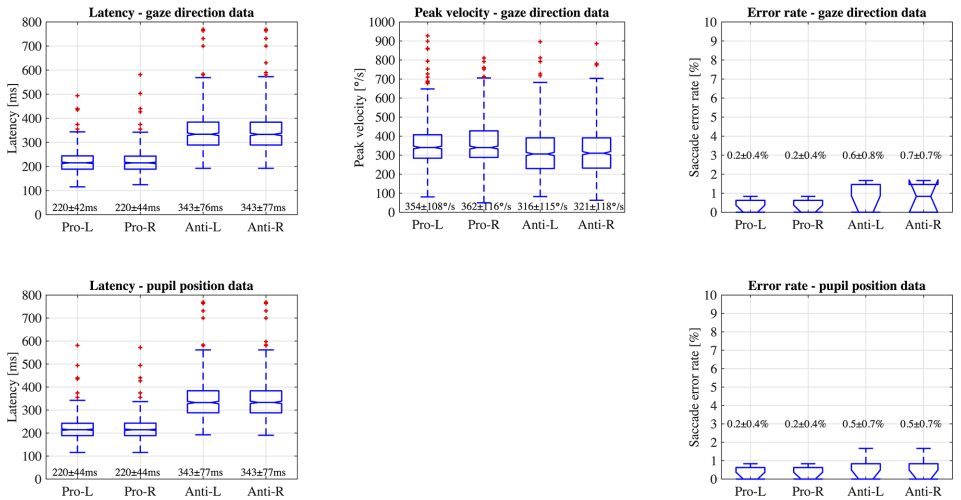


FIGURE 6 | Data distribution of parameters; latency, peak velocity, and error rate of saccadic eye movement of each eye in each pro- and anti-saccade task of all the participants (Pro, Pro-saccade; Anti, Anti-saccade; L, Left eye; R, Right eye).

reflexive response as a movement of eyes that shift from the central white target toward the erroneous direction over 3°. The reflexive response is usually observed in anti-saccade task because we automatically respond to the red stimulus target when it appears, despite being aware that we need to gaze at the opposite direction to the target (31).

The last four sub-figures show the displacement of normalized pupil position. When the waveforms are compared between the eyes, the values are generally larger in the left eye: the data range from 0.4 to 0.65 in the pro-saccade task and from 0.3 to 0.68 in the anti-saccade task for the left eye, whereas we observe the data between 0.33 and 0.58 in the pro-saccade task and between 0.22 and 0.6 in the anti-saccade task for the right eye. Moreover, when we look at the waveforms in the first 100 ms, we see the pupil position varies depending on the participants in comparison with the gaze direction data. These may be caused by a computational process embedded in the SRanipal SDK to convert the raw gaze data to the normalized pupil position data. Similar to the gaze direction waveforms, the pupil position waveforms also illustrate larger variability and longer latency in the anti-saccade task, although the reflexive responses are less distinguishable.

Latency, Peak Velocity, and Error Rate of Saccadic Eye Movement

Comparison Between Data Types, Saccade Types, and Eyes

We statistically compared the calculated data of latency and error rate of saccadic eye movement between two data types: gaze direction data type in degrees and normalized pupil position data type. Wilcoxon signed-rank test showed no significant difference at the significance level of 5% between the data types: $P = 0.75$ for latency, $P = 0.98$ for SD of latency, and $P > 0.999$ for error rate in the pro-saccade task, and $P = 0.73$ for latency, $P = 0.77$ for SD of latency, and $P = 0.56$ for error rate in the antisaccade task. In addition, we found significant differences between pro- and anti-saccade tasks in latency ($P < 0.001$), SD of latency ($P < 0.001$), and peak velocity ($P < 0.001$), but not in SD of peak velocity ($P = 0.77$) and error rate ($P = 0.07$). We also compared the data between left and right eyes, finding no significant differences: $P = 0.97$ for latency, $P = 0.99$ for SD of latency, $P = 0.16$ for peak velocity, $P = 0.51$ for SD of peak velocity, and $P = 0.82$ in error rate. Thus, if latency and error rate are main oculo-metrics in saccadic eye movement assessment, both of gaze direction data in degrees and normalized pupil position data can be used. Only gaze direction data can compute peak velocity of saccades.

Comparison With the Results of Previous Studies

Previous studies investigated saccadic eye movement of healthy populations at different ages using a 2D monitor-based assessment system. We compared our results with those studies.

Pro-saccadic eye movement was assessed in 100 healthy subjects at the age between 6 and 76 years, using an infrared video-based eye tracking technique with a sampling frequency of 220 Hz (28). The stimulus target was shown at ±5°, ± 15°, and ±30° horizontally. The criteria of saccade detection were 1) the velocity was over ±100°/s, 2) the evaluation duration was 500 ms after the stimulus target appeared, and 3) the amplitude was more than 0.5 of the relevant target displacement. The result showed that the latency of pro-saccade task ranged approximately from 110 and 260 ms for the population at the age between 20 and 39 years. Peak velocity increased as the position of target was placed farther from the center of visual field. The peak velocity was almost between 80 and 290°/s when the target appeared at ±5°, whereas the velocity ranged from 250 to 500°/s when the target was positioned at ±15°. Another study also researched age-related changes of eye movement in 250 healthy people at the age between 18 and 70 years, using an infrared-based video-oculography (VOG) (32). The study defined the criteria of saccade detection: 1) the amplitude of

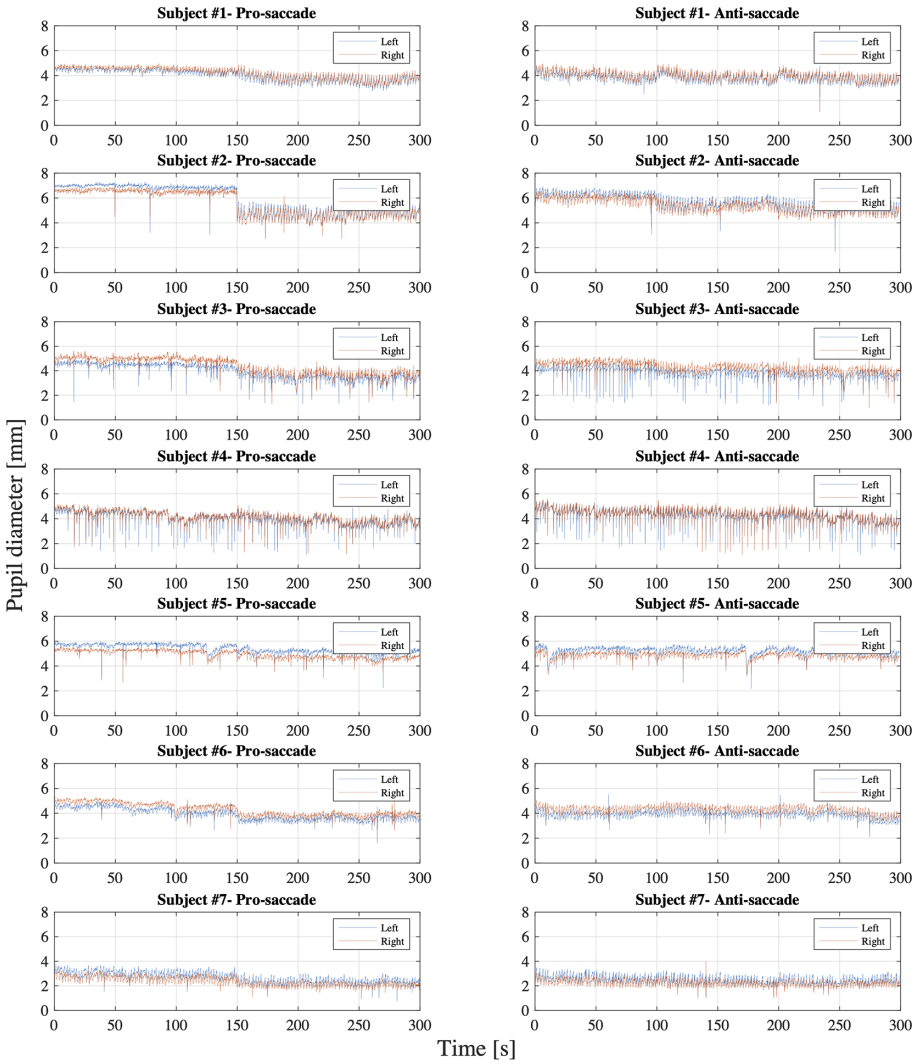


FIGURE 7 | Pupil diameter of each left and right eye in each pro- and anti-saccade task of all the participants.

eye movement was over 10°, 2) the evaluation duration was within 500 ms after the appearance of stimulus target, and 3) the saccade amplitude was over 10% of the target amplitude. The study found that the latency of pro-saccade task was 237.24 ± 18.23 ms in the leftward saccade and 265.34 ± 35.84 ms in the rightward saccade for the subjects at the age between 18 and 30 years (Group A), and was 241.44 ± 24.80 ms in the leftward saccade and 252.26 ± 37.65 ms in the rightward saccade for the population group at the age between 31 and 40 years (Group B). The research also reported the peak velocity: 245.30 ± 78.20°/s in the leftward saccade and 237.52 ± 75.64°/s in the rightward saccade for the Group A and 228.28 ± 66.29°/s in the leftward saccade and 226.50 ± 67.06°/s in the rightward saccade for the Group B.

On the other hand, another study investigated both pro- and anti-saccadic eye movement tasks in 1,058 healthy young adults at the age between 16 and 40 years, with an infrared-based oculography recording the ocular motions at 1 kHz (33). The saccade task started with presenting a target at the center of visual field for a random period between 500 and 1,500 ms and then a stimulus target at one of the ten horizontal positions: ± 3°, ± 6°, ± 9°, ± 12°, and ±15° for 600 ms in the pro-saccade task and 1,000 ms in the anti-saccade task. The pro-saccade task consisted of 200 trials in total and the anti-saccade task consisted of 50 trials. After the recorded data were filtered with a 300 Hz low-pass filter, the saccade in each trial was detected based on both eye acceleration and eye velocity criteria. Specifically, the presence of saccade was detected if the eye acceleration data exceeded a threshold: six times the median value of the SD of the acceleration data in the first 80 ms of all the trials in each person, or if the absolute value of eye velocity exceeded 50°/s. Then, the study defined borders of the saccades as the areas where the eye velocity was less than three times the median value of the SD of the eye velocity data measured in the first 80 ms of all the trials in

each participant. The research observed the latency of 177.2 ± 18.52 ms (range: 142 ∼ 322 ms) for the pro-saccade case and of 305.5 ± 43.06 ms (range: 113 ∼ 539 ms) for the anti-saccade case. The error rate of the anti-saccade task was 37.7 ± 21.5%. Saccadic eye movement was also inspected with an EOG device recording ocular movement at 500 Hz sampling frequency in 168 healthy subjects at the age between 5 and 79 years (34). The study arranged two different types of saccade task conditions: 1) overlap condition where a target appearing at the center of visual field remained illuminated when a stimulus target appeared and 2) gap condition where after the central target disappeared, no target was shown for 200 ms and then the stimulus target came out. In each condition, the stimulus target appeared at ±20° from the center and remained illuminated for 1 s. 120 trials of pro-saccade test and 240 trials of anti-saccade test were tested. The saccade was detected when the eye velocity exceeded 30°/s in the evaluation duration between 90 and 1,000 ms. The results revealed that the mean latency of saccadic eye movement over the participants was 224.71 ms for the pro-saccade and 307.14 ms for the anti-saccade in the gap condition, while the mean latency was 280.01 ms for the pro-saccade and 357.77 ms for the anti-saccade in the overlap condition. However, the latency was clearly shorter for the young population in the study. The error rate of anti-saccade task ranged from 0 to 48% for the participants at the age between 20 and 40 years. Moreover, a regression analysis was performed in 327 healthy subjects aged between 9 and 88 years, using an infrared reflection devise (35). The study measured both pro- and anti-saccadic eye movement for 200 trials in each. Overlap condition was utilized in the pro-saccade test where 1.2 s after a central fixed point was shown in the middle of a monitor, a stimulus target appeared at ±4° from the center and then remained brightened for 1 s. Gap condition was used in the anti-saccade test where the central target disappeared 0.2 s before the onset of the stimulus target and then the stimulus target remained visible for 1 s. Saccade onset was defined by an eye velocity threshold of 20°/s and latency of saccadic eye movement was computed in the time range of 136 ∼ 700 ms. In addition, pro-saccadic eye movement during the anti-saccade task was considered as an error if the latency of eye movement was longer than 80 ms. The study results indicated that the latency ranged nearly from 125 to 290 ms (SD: 18 ∼ 102 ms) for the pro-saccade task and from 150 to 380 ms for the anti-saccade task in the participants aged between 25 and 37 years.

In summary, while the prior studies discussed above used different measurement protocols in terms of position of stimulus target, configuration on time domain (i.e., gap or overlap condition, duration of illumination of targets on monitor), saccade detection algorithm, sampling frequency of eye tracking device, monitor specification, target population, and sample size, the descriptive statistics show that our results of latency, peak velocity, and error rate are within the range of or close to the results of these previous studies. Therefore, this seems to indicate that the HTC VIVE Pro Eye could be useful as an assessment tool of saccadic eye movement for the specific oculo-metrics.

Pupillary Response

On the whole, **Figure 7** shows that the pupil diameter did not fluctuate widely during each pro- and anti-saccade task for both eyes, while we observe some differences between the participants. Specifically, subject #7 showed smaller pupil diameter than the others. This may be because subject #7 wore glasses during the measurement. Iris color might not cause the differences in pupil diameter between the participants as reported in (36). A previous research measured changes of pupil diameter under the different luminance conditions in 155 healthy people with the mean age of 29.7 ± 17.8 years and the age range of 6 ∼ 64 years (37). The measurement was conducted under four different illumination levels: scotopic (0.1 cd/m^2), mesopic (1 cd/m^2), low photopic (10 cd/m^2), and high photopic (100 cd/m^2) visions. The study observed the pupil diameter of 5.4 ± 0.7 mm and 3.9 ± 0.4 mm for the subjects aged between 21 and 30 years and of 4.4 ± 0.5 mm and 3.5 ± 0.5 mm for the subjects aged between 31 and 40 years in mesopic and low photopic visions respectively. Another study also inspected the changes of pupillary response at five luminance levels: 0, 0.5, 4, 32, and 250 cd/m^2 in 245 healthy population (mean age: 51.9 ± 18.3 years; age range: 6 ∼ 87 years) (38). The research found a mean pupil diameter of 5.39 ± 1.04 mm at 0 cd/m^2, 5.20 ± 1.00 mm at 0.5 cd/m^2, 4.70 ± 0.97 mm at 4 cd/m^2, 3.74 ± 0.78 mm at 32 cd/m^2, and 2.84 ± 0.50 mm at 250 cd/m^2 and decreasing pupil size with increasing age. Moreover, a different research evaluated reliability of pupil diameter measurement in 416 healthy participants across diverse demographics with the mean age of 42 ± 8.7 years and the age range from 18 to 73 years (39). The research used an infrared eye tracking technique with a 2D monitor with 85 cd/m^2 luminance in a room at 344 cd/m^2 luminance level, revealing that the mean pupil diameter was 3.53 ± 0.26 mm for right eye and 3.54 ± 0.28 mm for left eye and that the test-to-test reliability was strong. When we compare our experiment results (see **Table 3**) with those in the previous studies mentioned above, we conclude that the eye tracking device in HTC VIVE Pro Eye also measures pupillary response properly. While it is difficult to compare the results precisely, our data are in the range of or close to the measured mean pupil diameter of the prior studies. Future studies should, however, determine the test-retest reliability of our approach in predefined target populations to substantiate this assumption.

Sampling of Eye Tracking

The eye tracker on VIVE Pro Eye samples eye movement at the maximum frequency of 120 Hz. When ocular motions are sampled at 120 Hz, the sampling interval is 8.33 ms. We report the following findings related to the sampling of the eye tracker.

First, we observe that the sampling interval is longer than 8.33 ms for most of the first 2,000 ∼ 3,000 samples for all the participants in both time measurement parameters: Unix time and SRanipal time stamp as shown in **Figure 8**. Specifically, we find that the sampling interval decreased to the expected value of around 8 ∼ 9 ms, 19 ∼ 28 s after the sampling started as highlighted with the red dash lines in **Figure 8**. The time when the sampling interval changes varies depending on the

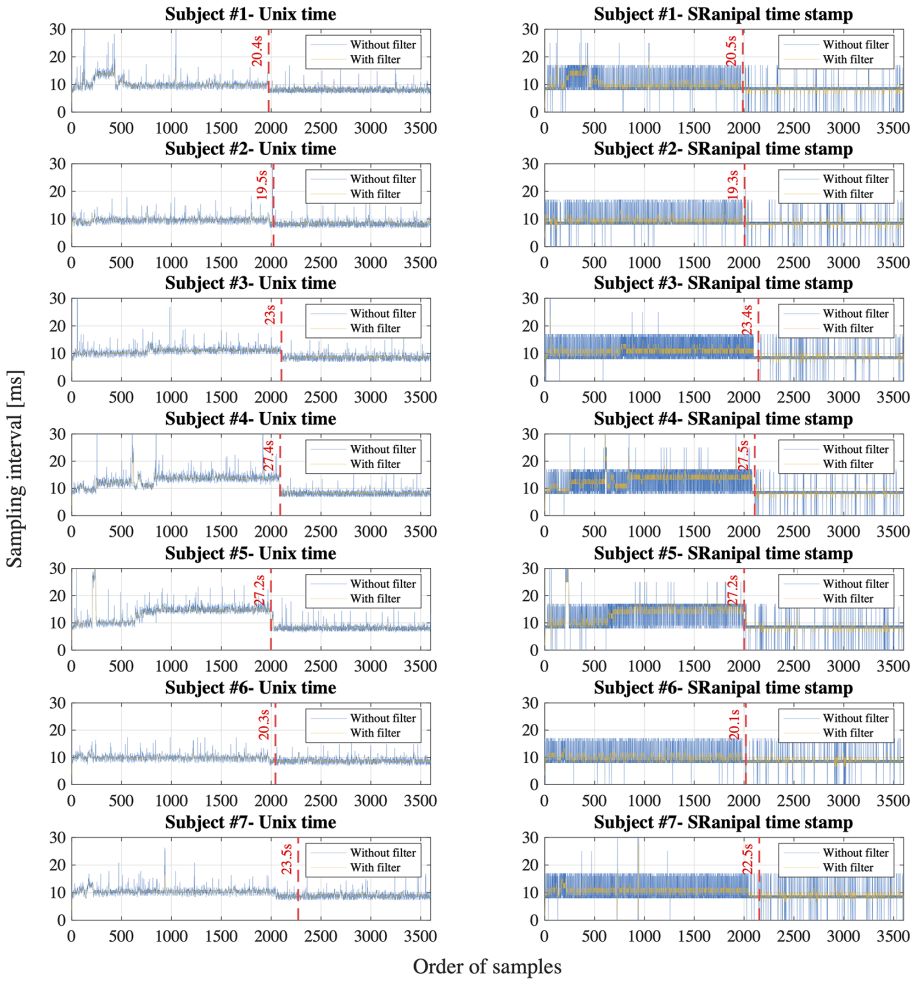


FIGURE 8 | Sampling interval calculated with Unix time on Unity and time stamp from SRanipal SDK of all the participants; Blue line shows the unfiltered data and yellow line shows the filtered data.

participants, though the timing is similar or quite close between the two time measurement parameters. We suppose that the longer sampling interval may occur because of processing speed of C# programming on Unity. As explained on our GitHub website: https://github.com/MotorControlLearning we recorded eye movement with using the callback function. More specifically, we implemented a while loop in the callback function to keep recording ocular motions while the saccadic eye movement task was performed. Assuming that the processing speed of the while loop might cause the longer sampling interval (40), we further investigated the computation of while loop. We created a simple while loop on C# programming and measured the processing speed of the simple while loop of 12,000 times in four different configurations: 1) Unity on Windows operating system (OS) that was used in the measurement of saccadic eye movement in this study, 2) Visual Studio on Windows OS (without Unity), 3) Unity on Mac OS (Apple MacBook Pro, Retina, 15-inch, Late 2013, 16-GB memory, 2.3-GHz Quad-Core Intel Core i7-4850HQ), and 4) Visual Studio on Mac OS. As a result, we found that the processing time of the simple while loop was longer than 8.33 ms for the configuration of Unity on Windows OS at some samples in the first 29.4 s after the program was executed (see **Figure 9**). In addition, we observe that the processing speed decreased to less than 8 ms, 29.4 s after the while loop program started. On the other hand, while the longer processing time than 8.33 ms was also observed for the configuration of Visual Studio on Windows OS, we saw the situation less frequently than the configuration of Unity on Windows OS as the total processing time explained: 54.9 s for Unity on Windows OS and 47.4 s for Visual Studio on Windows OS. Interestingly, we discovered that the processing time was much shorter on Mac OS than Windows OS as shown in **Figure 9**. The processing time of each while loop was less than 5 ms for the configurations with Mac OS and the total processing speed was 2.8 s for Unity on Mac OS and 1.3 s for Visual Studio on Mac OS despite that the technical specification of the Mac laptop was lower than the computer with Windows OS. Since we needed to use the computer with

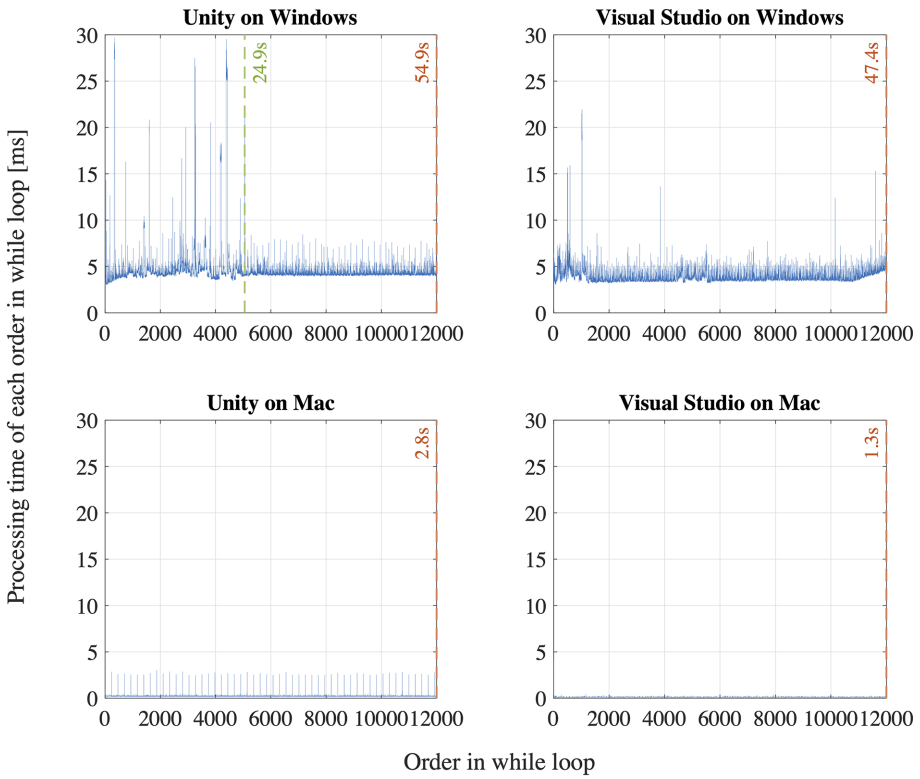


FIGURE 9 | Evaluation of a while loop for 12,000 times on Unity and Visual Studio on Windows and Mac operating systems.

Windows OS to provide enough power to support HTC VIVE Pro Eye and to sample eye movements at the maximum sampling frequency of 120 Hz, we intentionally had a non-assessment period over 30 s after the recording of eye movement started to wait for the sampling interval to get close to the expected value of 8.33 ms. Therefore, we recommend setting the non-assessment period for over 30 s in the initial measurement phase if Windows OS is used and the maximum sampling of 120 Hz is necessary. However, it is also important to note that the duration of 30 s may not be enough occasionally since we have observed that the longer processing time continued even after the 30 s with around 3 ∼ 5% possibility. If a laptop with Mac OS meets the requirements of HTC VIVE Pro Eye, the issue would not be encountered, though we recommend evaluating whether the problem occurs on any configurations.

Second, we find that the time stamp recorded in SRanipal SDK sometimes becomes zero while the time recorded with Unix time does not reach zero as visually confirmed in **Figure 8**. We consider two reasons for the finding. First reason is a software bug of SRanipal SDK. When we inspected the recorded data of time stamp, frame, and pupil diameter specific to two consecutive samples, we found that the pupil diameter of left eye changed from 4.857666 to 4.847290 mm and the pupil diameter of right eye also changed from 4.549805 to 4.549103 mm, whereas the values of time stamp and frame did not change. If the time stamp value had been the same for these two samples, the same pupillary response should have been recorded. We reported the finding on HTC developer community forum and HTC confirmed the problem as of the 3rd of December in 2019. Due to the issue, we did not use the time data of time stamp recorded by SRanipal SDK in the data analysis of saccadic eye movement and used the data of Unix time instead. HTC plans to fix the issue in the next version of SRanipal SDK. Second reason is processing speed of C# programming on Unity. As illustrated in **Figure 9**, the processing time measured with the configuration of Unity on Windows OS fluctuates between 3.9 and 8.5 ms. Since the sampling interval of eye tracker is 8.33 ms, it is possible for the computer to record the same sample twice in the while loop (40). For instance, if the computer records the data of eye movements in the processing time of 4 ms, the computer possibly activates the clock signal twice within the period of the sampling interval of eye tracking, 8.33 ms, as illustrated in Case B of **Figure 10**. This means that the computer records the same sample (i_{th} sample in Case B of **Figure 10**) twice continuously. The situation happens more frequently especially when the timing of clock activation and the sampling of eye tracker are close to each other. To summarize, if the next version of SRanipal SDK solves the issue of time stamp and we eliminate the duplicated data due to recording the same sample twice, we could utilize the time stamp data recorded by SRanipal SDK for the data analysis of saccadic eye movement with more accuracy. A consequence of our observation is also related to the replicability of existing research. It seems important for researchers to pay attention to

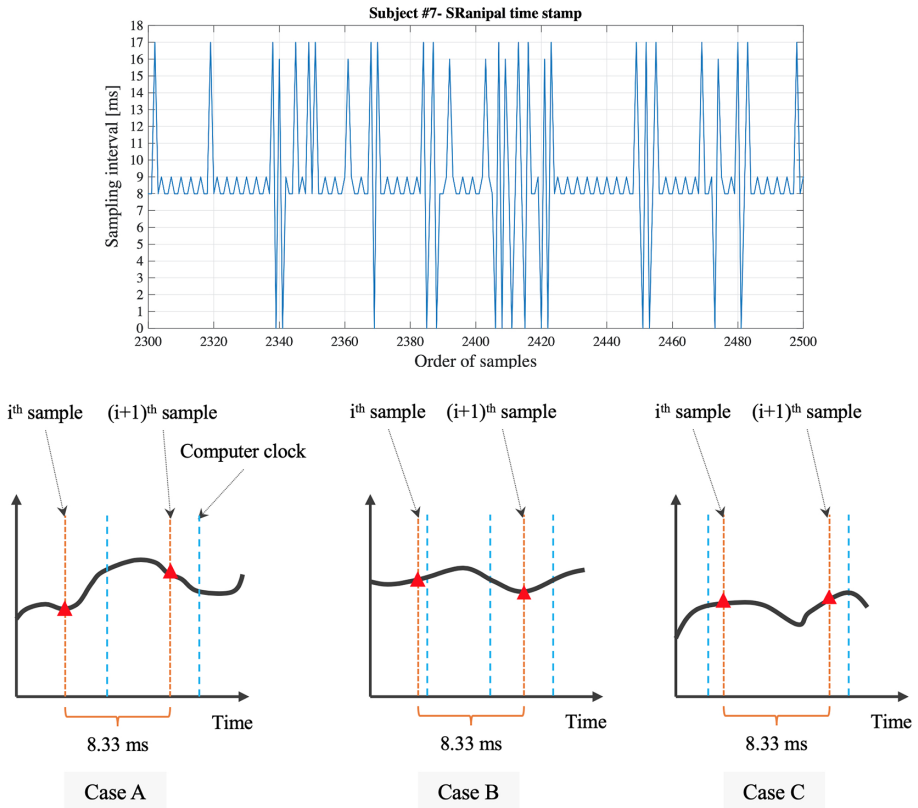


FIGURE 10 | Detailed sampling interval based on the time stamp recorded with SRanipal SDK and visual explanation of relation between sampling of eye tracker and clock signal of computer.

the hardware configuration of their experimental set-up when one of the aims of their research is to replicate findings that were previously reported for saccadic eye movement behaviors.

Third, while the time stamp recorded by SRanipal SDK needs to be improved as mentioned above, we notice that the time stamp values fluctuate; the calculated sampling interval from the time stamp data recorded with SRanipal SDK becomes 8, 9, 16, or 17 ms as a portion of the data of subject #7 shows in **Figure 10**. We subcategories these four values into two sets: 1) 8 and 9 ms, and 2) 16 and 17 ms. We suppose that the two sets are observed because of the relation of timing between the sampling of the eye tracker and the clock signal of the computer (40). In particular, it is possible to see the first set of 8 and 9 ms if Case A occurs as explained in **Figure 10**, while we possibly encounter Case C where we cannot record the i_{th} sample because of longer processing time of the computer than the sampling interval of 8.33 ms, causing the calculated sampling interval to become 16 or 17 ms. In addition, we see the difference of 1 ms in each set. The difference might be caused by the computational calculation of time stamp data of Tobii eye tracker as discussed in the white paper published from Tobii Technology (41). Although the paper refers to previous Tobii products using an eye tracker with sampling frequency of 60 Hz, it states that there is the uncertainty in the time stamp data with a nominal value of 1 ms. Therefore, if the Tobii eye tracker integrated in HTC VIVE Pro Eye also has the similar uncertainty, we could also observe the 1 ms difference in our measured data.

In this section, we have reported our findings regarding the sampling technique of eye tracking used in HTC VIVE Pro Eye. From what we have investigated, we recommend evaluating the capability of sampling of VIVE Pro Eye in the development phase especially if temporal parameters of eye movement are investigated and the maximum sampling frequency at 120 Hz is required. Specifically, if the similar experimental system to ours using Unity on Windows OS is used, we would suggest setting a non-assessment period more than 30 s after the eye tracker starts to record. The waiting could lead the system to record the data at the interval of 8.33 ms more frequently. Nonetheless, the non-assessment period may not be needed depending on the OS and hardware configuration of the computer. Finally, since the next version of SRanipal SDK is going to fix the problem of time stamp data as explained above, we propose trying the next version when it is released. If the new version provides the time stamp data correctly, we could improve the precision of the calculated oculo-metrics of saccadic eye movement by using the time stamp instead of Unix time because the time stamp data are supposed to provide more accurate time data with considering the image delivery time from

the tracking sensor to the eye tracker firmware as discussed in Tobbi Technology (41).

Limitations and Improvements

Measurement Protocol of Saccadic Eye Movement Assessment

The main purpose of this study was to evaluate whether HTC VIVE Pro Eye could be used as an assessment tool of saccadic eye movement. While we have observed that the device can function as an assessment tool by comparing our results with the descriptive statistics of previous studies, one of the aspects in our study that could be perceived as a limitation is that our sample size is small and the age-group is specific to a young and healthy population. The assessment ultimately targets individuals prone to develop cognitive impairments that expectedly are older than the assessed convenience sample. In these elderly people, issues of attitudes and acceptance with VR headsets might result in differing findings. In that sense, our results are first indicative findings that should be replicated with older individuals. Recent research revealed, however, that the approach of wearing VR headsets seems feasible also in older adults with cognitive and/or physical impairments (42). Further measurement with an increasing sample size and diverse demographic groups, and the direct comparison with the data obtained with a highly reliable and high-spec VOG device would provide more detailed, broad, and reliable results. In addition, as research has been undertaken to discuss the saccade detection algorithm, our algorithm could be also enhanced for more accurate saccade detection. Finally, sampling frequency of eye tracking device can be also an important factor as discussed in (43, 44). While higher sampling frequency costs more in terms of price and power consumption of eye tracking device, lower sampling frequency may lead us to misestimate the saccade detection. It would be important to find the appropriate sampling frequency, considering the proper balance of cost and reliability of a device.

Measurement Parameters

As discussed in the previous sections, the time stamp data recorded with the present version of SRanipal SDK (version 1.1.0.1) needs to be improved. In addition to the time stamp, the following parameters have not been supported yet, as far as we have confirmed with HTC. The future version of SRanipal SDK would implement further update to validate these parameters.

- int timestamp
- bool convergence_distance_validity (combined eye data)
- float convergence_distance_mm (combined eye data)
- float eye_squeeze (eye expression data)
- float eye_frown (eye expression data)

Finally, we were not able to calculate the peak velocity of saccadic eye movement when we used the normalized pupil position data. Nonetheless, if a conversion mapping between normalized screen coordinate and degrees is available, we could also calculate the peak velocity from the pupil position data.

Evaluation of Ocular Dominance

Ocular dominance is one of the important factors that could affect the results of saccadic eye movement. A previous study observed that eye dominance influenced saccade amplitude as the participants with strong ocular dominance reached more accurate saccades toward the target in the hemifield opposite to the side of dominant eye than in the same side (45). While our investigation has shown that both gaze direction and pupil position data can be used to evaluate latency and error rate of saccadic eye movement, ocular dominance as well as peak velocity of saccades could be evaluated with only gaze direction data. As discussed in the previous section, we observed visually the difference in the range of eye movement between left and right eyes in the normalized pupil position data, whereas the clear visual difference was not seen in the gaze direction data (see **Figure 5**). The difference seen in the pupil position data could not be due to the factor of ocular dominance because we should have seen the similar difference in the gaze direction data if eye dominance caused the difference. Therefore, we suppose that gaze direction data would be a proper data set when ocular dominance is inspected.

Potential of HMD-VR Technology

While we designed a simple VR environment to simulate a situation of conventional 2D graphic display, the VR headset has a unique and novel potential in providing immersive VR environments. A previous review suggested the benefits of VR-based measures in its sensitivity of detecting cognitive impairments (21). Another study also emphasized the importance of taking advantage of immersive VR environments for the assessment of AD (46). Assessment of saccadic eye movement with various levels of immersion in VR environments could lead us to new research findings. Moreover, while we target at using the VR headset for the saccadic eye movement of neurodegenerative disorders, the device could become an alternative assessment tool in different fields. For example, if the VR headset is combined with a postural assessment device, we could stimulate postural sway with immersive VR environments and achieve a cost-efficient and portable assessment system to measure in a small space, instead of being restricted to dedicated laboratories with non-transportable systems for the balance assessment (47). Future studies integrating the VR device bear potential for assessments of diverse research fields.

CONCLUSION

In our growing ageing society globally, neurodegenerative disorders are becoming a more relevant problem. Assessment of saccadic eye movement in those suffering from the disorders could be a promising way to diagnose them in a simple, time-efficient, and low-cost manner. Along with the advanced technologies of video-based eye tracking and VR, HTC launched a VR headset, VIVE Pro Eye, consisting of an

infrared-based eye tracker and HMD. The purpose of this study was therefore to evaluate whether the VIVE Pro Eye could be used as an assessment tool of saccadic eye movement. We measured saccadic eye movement of seven healthy young adults using the product, following an internationally proposed standard protocol of saccade measurement. Our investigation results suggest that VIVE Pro Eye can function as an assessment device of saccadic eye movement and record the data necessary to calculate the oculo-metrics of latency, peak velocity, and error rate of saccades, and pupillary response. We also found technical limitations related to the eye tracking, especially about time-related measurement parameters. Further improvements in the provided SDK and the measurement protocol including saccade detection algorithm, and more measurements with diverse age-groups and those with different health conditions are necessary to enhance the whole assessment system of saccadic eye movement.

DATA AVAILABILITY STATEMENT

All data can be shared upon reasonable request to the first author. Programming background information is available on GitHub: https://github.com/MotorControlLearning.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by ETH Zurich Ethics Commission. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Each of the authors has contributed to developing the research concept and experimental designs. YI and AF developed the assessment system and performed the measurement. All authors conducted the first analysis and interpretation of the data. In addition, all authors contributed in drafting and revising the article to bring it to its current state. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We would like to acknowledge the support from the people who kindly joined the experiment and thank AF for his great contribution especially to the development work and EB for his supervision of the whole project.

REFERENCES

1. United Nations. *World population ageing 2019*. Tech. rep., New York, the USA: United Nations (2019).
2. World Health Organization. *Decade of healthy ageing 2020-2030*. Tech. rep., Geneva, Switzerland: World Health Organisation (2020).
3. World Health Organization. *10 priorities for a decade of action on healthy ageing*. Tech. rep., Geneva, Switzerland: World Health Organisation (2017a).
4. World Health Organization. *Global strategy and action plan on ageing and health*. Tech. rep., Geneva, Switzerland: World Health Organisation (2017b).
5. Heemels M-T. Neurodegenerative diseases. *Nature* (2016) 539(7628):179–9. doi: 10.1038/539179a
6. Feigin VL, Abajobir AA, Abate KH, Abd-Allah F, Abdulle AM, Abera SF, et al. Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the global burden of disease study 2015. *Lancet Neurol* (2017) 16:877–97. doi: 10.1016/s1474-4422(17)30299-5
7. Gorges M, Pinkhardt EH, Kassubek J. Alterations of eye movement control in neurodegenerative movement disorders. *J Ophthalmol* (2014) 2014:11. doi: 10.1155/2014/658243
8. MacAskill MR, Anderson TJ. Eye movements in neurodegenerative diseases. *Curr Opin Neurol* (2016) 29:61–8. doi: 10.1097/wco.0000000000000274
9. Marandi RZ, Gazerani P. Aging and eye tracking: in the quest for objective biomarkers. *Future Neurol* (2019) 14:22. doi: 10.2217/fnl-2019-0012
10. Molitor RJ, Ko PC, Ally BA. Eye movements in alzheimer's disease. Journal of. *Alzheimers Dis* (2015) 44:1–12. doi: 10.3233/jad-141173
11. Wilcockson TDW, Mardanbegi D, Xia BQ, Taylor S, Sawyer P, Gellersen HW, et al. Abnormalities of saccadic eye movements in dementia due to alzheimer's disease and mild cognitive impairment. *Aging-Us* (2019) 11:5389–98. doi: 10.18632/aging.102118
12. Goffart L. *Saccadic Eye Movements*. Academic Press: Oxford (2009). p. 437–44. doi: 10.1016/B978-008045046-9.01101-3
13. Ramat S, Leigh RJ, Zee DS, Optican LM. What clinical disorders tell 751 us about the neural control of saccadic eye movements. *Brain* (2007) 130:10–35. doi: 10.1093/brain/awl309
14. Terao Y, Fukuda H, Hikosaka O. What do eye movements tell us about patients with neurological disorders? - an introduction to saccade recording in the clinical setting. *Proc Jpn Acad Ser B-Physical Biol Sci* (2017) 93:772–801. doi: 10.2183/pjab.93.049
15. Termsarasab P, Thammongkolchai T, Rucker JC, Frucht SJ. The diagnostic value of saccades in movement disorder patients: a practical guide and review. *J Clin Mov Disord* (2015) 2:14. doi: 10.1186/s40734-015-0025-4
16. Liu Y, Tan WJ, Chen C, Liu CY, Yang JZ, Zhang YC. A review of the application of virtual reality technology in the diagnosis and treatment of cognitive impairment. *Front Aging Neurosci* (2019) 11:280. doi: 10.3389/fnagi.2019.00280
17. Tarnanas I, Tsolakis A, Tsolaki M. Assessing Virtual Reality Environments as Cognitive Stimulation Method for Patients with MCI. In: Brooks A, Brahnam S, Jain L, editors. *Technologies of Inclusive Well-Being. Studies in Computational Intelligence, vol 536*. Berlin, Heidelberg: Springer (2014). doi: 10.1007/978-3-642-45432-5_4
18. Benyoucef Y, Lesport P, Chassagneux A. The emergent role of virtual reality in the treatment of neuropsychiatric disease. *Front Neurosci* (2017) 11:491. doi: 10.3389/fnins.2017.00491
19. Moreno A, Wall KJ, Thangavelu K, Craven L, Ward E, Dissanayaka NN. A systematic review of the use of virtual reality and its effects on cognition in individuals with neurocognitive disorders. *Alzheimer's Dementia (New York N Y)* (2019) 5:834–50. doi: 10.1016/j.trci.2019.09.016
20. Dockx K, Bekkers EMJ, Van den Bergh V, Ginis P, Rochester L, Hausdorff JM, et al. Virtual reality for rehabilitation in parkinson's disease. *Cochrane Database Syst Rev* (2016) 12(12):CD010760. doi: 10.1002/14651858.CD010760.pub2
21. Negut A, Matu SA, Sava FA, David D. Virtual reality measures in neuropsychological assessment: a meta-analytic review. *Clin Neuropsychol* (2016) 30:165–84. doi: 10.1080/13854046.2016.1144793
22. Muratore M, Tuena C, Pedroli E, Cipresso P, Riva G. Virtual reality as a possible tool for the assessment of self-awareness. *Front Behav Neurosci* (2019) 13:62. doi: 10.3389/fnbeh.2019.00062
23. Nolin P, Banville F, Cloutier J, Allain P. *Virtual Reality as a New Approach to Assess Cognitive Decline in the Elderly, vol. 2 of 2013*. Rome, Italy: Mediterranean Center of Social and Educational Research (2013).
24. Antoniades C, Ettinger U, Gaymard B, Gilchrist I, Kristjansson A, Kennard C, et al. An internationally standardised antisaccade protocol. *Vision Res* (2013) 84:1–5. doi: 10.1016/j.visres.2013.02.007

25. Hutton SB, Ettinger U. The antisaccade task as a research tool in psychopathology: A critical review. *Psychophysiology* (2006) 43:302–13. doi: 10.1111/j.1469-8986.2006.00403.x
26. Bijvank JAN, Petzold A, Balk LJ, Tan HS, Uitdehaag BMJ, Theodorou M, et al. A standardized protocol for quantification of saccadic eye movements: Demons. *PLoS One* (2018) 13:19. doi: 10.1371/journal.pone.0200695
27. Larsson L, Nystrom M, Stridh M. Detection of saccades and postsaccadic oscillations in the presence of smooth pursuit. *IEEE Trans Biomed Eng* (2013) 60:2484–93. doi: 10.1109/tbme.2013.2258918
28. Hopf S, Liesenfeld M, Schmidtmann I, Ashayer S, Pitz S. Age dependent normative data of vertical and horizontal reflexive saccades. *PLoS One* (2018) 13:13. doi: 10.1371/journal.pone.0204008
29. Magnusdottir BB, Faiola E, Harms C, Sigurdsson E, Ettinger U, Haraldsson HM. Cognitive measures and performance on the antisaccade eye movement task. *J Cognit* (2019) 2:3. doi: 10.5334/joc.52
30. Andersson R, Larsson L, Holmqvist K, Stridh M, Nystrom M. One algorithm to rule them all? an evaluation and discussion of ten eye movement event-detection algorithms. *Behav Res Methods* (2017) 49:616–37. doi: 10.3758/s13428-016-0738-9
31. Coe BC, Munoz DP. Mechanisms of saccade suppression revealed in the anti-saccade task. *Philos Trans R Soc B-Biol Sci* (2017) 372:10. doi: 10.1098/rstb.2016.0192
32. Seferlis F, Chimona TS, Papadakis CE, Bizakis J, Triaridis S, Skoulakis C. Age related changes in ocular motor testing in healthy subjects. *J Vestib Res-Equilib Orientation* (2015) 25:57–66. doi: 10.3233/ves-150548
33. Bargary G, Bosten JM, Goodbourn PT, Lawrance-Owen AJ, Hogg RE, Mollon JD. Individual differences in human eye movements: An oculomotor signature? *Vision Res* (2017) 141:157–69. doi: 10.1016/j.visres.2017.03.001
34. Munoz DP, Broughton JR, Goldring JE, Armstrong IT. Age-related performance of human subjects on saccadic eye movement tasks. *Exp Brain Res* (1998) 121:391–400. doi: 10.1007/s002210050473
35. Klein C, Foerster F, Hartnegg K, Fischer B. Lifespan development of pro- and anti saccades: Multiple regression models for point estimates. *Dev Brain Res* (2005) 160:113–23. doi: 10.1016/j.devbrainres.2005.06.011
36. Bradley JC, Bentley KC, Mughal AII, Bodhireddy H, Young RSL, Brown SM. The effect of gender and iris color on the dark-adapted pupil diameter. *J Ocular Pharmacol Ther* (2010) 26:335–40. doi: 10.1089/jop.2010.0061
37. Tekin K, Sekeroglu MA, Kiziltoprak H, Doguizi S, Inanc M, Yilmazbas P. Static and dynamic pupillometry data of healthy individuals. *Clin Exp Optom* (2018) 101:659–65. doi: 10.1111/cxo.12659
38. Rickmann A, Waizel M, Kazerounian S, Szurman P, Wilhelm H, Boden KT. Digital pupillometry in normal subjects. *Neuro-Ophthalmology* (2017) 41:12–8. doi: 10.1080/01658107.2016.1226345
39. Murray NP, Hunfalvay M, Bolte T. The reliability, validity, and normative data of interpupillary distance and pupil diameter using eye-tracking technology. *Trans Vision Sci Technol* (2017) 6:12. doi: 10.1167/tvst.6.4.2
40. Andersson R, Nystrom M, Holmqvist K. Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. *J Eye Mov Res* (2010) 3:12. doi: 10.16910/jemr.3.3.6
41. Tobii Technology. *Timing Guide for Tobii Eye Trackers and Eye Tracking Software*. Tech. rep., Danderyd, Sweden: Tobii Technology (2010).
42. Appel L, Appel E, Bogler O, Wiseman M, Cohen L, Ein N, et al. Older adults with cognitive and/or physical impairments can benefit from immersive virtual reality experiences: A feasibility study. *Front Med* (2020) 6(329). doi: 10.3389/fmed.2019.00329
43. Leube A, Rifai K. Sampling rate influences saccade detection in mobile eye tracking of a reading task. *J Eye Mov Res* (2017) 10:11. doi: 10.16910/jemr.10.3.3
44. Wierts R, Janssen MJA, Kingma H. Measuring saccade peak velocity using a low frequency sampling rate of 50 hz. *IEEE Trans Biomed Eng* (2008) 55:2840–2. doi: 10.1109/tbme.2008.925290
45. Tagu J, Dore-Mazars K, Lemoine-Lardennois C, Vergilino-Perez D. How eye dominance strength modulates the influence of a distractor on saccade accuracy. *Invest Ophthalmol Visual Sci* (2016) 57:534–43. doi: 10.1167/iovs.15-18428
46. Garcia-Betances RII, Waldmeyer MTA, Fico G, Cabrera-Umpierrez MF. A succinct overview of virtual reality technology use in alzheimer's disease. *Front Aging Neurosci* (2015) 7:80. doi: 10.3389/fnagi.2015.00080
47. Harro CC, Garascia C. Reliability and validity of computerized force platform measures of balance function in healthy older adults. *J Geriatr Phys Ther* (2019) 42:E57–66. doi: 10.1519/jpt.0000000000000175

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 26 November 2020
doi: 10.3389/fpsyt.2020.538122

The Sensor-Based Physical Analogue Scale as a Novel Approach for Assessing Frequent and Fleeting Events: Proof of Concept

Stefan Stieger, Irina Schmid, Philip Altenburger and David Lewetz*

Department of Psychology and Psychodynamics, Karl Landsteiner University of Health Sciences, Krems an der Donau, Austria

New technologies (e.g., smartphones) have made it easier to conduct Experience Sampling Method (ESM) studies and thereby collect longitudinal data *in situ*. However, limiting interruption burden (i.e., the strain of being pulled out of everyday life) remains a challenge, especially when assessments are frequent and/or must be made immediately after an event, such as when capturing the severity of clinical symptoms in everyday life. Here, we describe a wrist-worn microcomputer programmed with a Physical Analogue Scale (PAS) as a novel approach to ESM in everyday life. The PAS uses the position of a participant's forearm between flat and fully upright as a response scale like a Visual Analogue Scale (VAS) uses continuous ratings on a horizontal line. We present data from two pilot studies (4-week field study and lab study) and data from a 2-week ESM study on social media ostracism (i.e., when one's social media message is ignored; $N = 53$ participants and 2,272 event- and time-based assessments) to demonstrate the feasibility of this novel approach for event- and time-based assessments, and highlight advantages of our approach. PAS angles were accurate and reliable, and VAS and PAS values were highly correlated. Furthermore, we replicated past research on cyber ostracism, by finding that being ignored resulted in significantly stronger feelings of being offended, which was more pronounced when ignored by a group compared to a single person. Furthermore, participants did not find it overly difficult to complete the assessments using the wearable and the PAS. We suggest that the PAS is a valid measurement procedure in order to assess fleeting and/or frequent micro-situations in everyday life. The source code and administration application are freely available.

Keywords: experience sampling method (ESM)/ecological momentary assessment (EMA), wearable devices, cyber ostracism, digital phenotyping, indirect assessment

OPEN ACCESS

Edited by:
Jennifer H. Barnett, Cambridge Cognition, United Kingdom

Reviewed by:
Simone Verhagen, Maastricht University, Netherlands
Anna Beukenhorst, School of Public Health and Harvard University, United States

****Correspondence:***
Stefan Stieger
stefan.stieger@kl.ac.at

Specialty section:
This article was submitted to Psychological Therapies, a section of the journal Frontiers in Psychiatry

Received: *26 February 2020*
Accepted: *29 October 2020*
Published: *26 November 2020*

Citation:
Stieger S, Schmid I, Altenburger P and Lewetz D (2020) The Sensor-Based Physical Analogue Scale as a Novel Approach for Assessing Frequent and Fleeting Events: Proof of Concept. Front. Psychiatry 11:538122. doi: 10.3389/fpsyt.2020.538122

INTRODUCTION

The Experience Sampling Method (ESM)—that is, the collection of longitudinal data from participants in their everyday lives—has not only contributed to psychologists' understanding of how people behave in the real world (1), but has also enhanced understanding in other disciplines, such as psychiatry [see special issue (2)] and economics (3). Relative to cross-sectional and laboratory studies, ESM reduces recall bias, provides temporally-dense profiles of each participant, and results may be more externally valid because they capture psychological

phenomena in participants' natural environments. Smartphones and wearable microcomputers have recently made it easier to conduct ESM studies (4, 5). However, limiting interruption burden (i.e., the strain of being pulled out of everyday life) remains a challenge, especially when participants must make frequent and/or immediate assessments, such as when capturing the severity of symptoms in everyday life [i.e., (5, 6)]. High interruption burden might result in low compliance, which results in missing data, which in turn may introduce measurement error and selection bias, i.e., data quality decreases and the advantages of ESM studies vanish. Here, we describe how a wearable microcomputer programmed with a Physical Analogue Scale (PAS) can be used to assess fleeting and/or frequent events in everyday life. We utilize data from two pilot studies and an ESM study on social media ostracism to demonstrate the advantages of our approach for reliable and accurate event- and time-based assessments.

Experience Sampling Method With Smartphones and Wearables

Smartphones are now commonly used in ESM studies (7, 8), with wearable microcomputers (typically worn on the wrist) also increasingly used [e.g., (5, 9, 10); for a review, see (11)]. Although smartphones and wearables have a number of advantages over traditional paper-and-pencil diaries [e.g., the possibility of timestamping data, (12)], limiting interruption burden remains a challenge. Participants are sometimes unwilling to go through the onerous process of removing and unlocking a smartphone, opening the application, carrying out the assessment, closing the application, turning off the smartphone, and replacing it; that is, their commitment may sometimes dramatically decrease when they have to respond very frequently (9). Furthermore, participants may sometimes find it difficult to comply with lengthy procedures (e.g., in job situations; while driving), resulting in delayed assessments.

Relative to smartphones, wearables offer shorter access time (13), are more comfortable (14), and allow researchers to use more reliable tactile vs. auditory signals (10). However, many existing wearables are still rather bulky, expensive, and need to be recharged frequently (15). Some wearables only work in combination with a smartphone (e.g., many commercially-available smartwatches) and use proprietary software. Furthermore, their small displays make it difficult to use text-based instructions or response scales.

A Novel Approach: The Sensor-Based Physical Analogue Scale

To address the disadvantages of existing approaches and take advantage of sensor-based data (16), we programmed a wrist-worn wearable with a PAS. The PAS uses the position of a participant's forearm as a continuous response scale. Specifically, participants indicate a response by positioning their forearm flat (0° = lowest scale value), in a fully upright position (90° = highest scale value), or somewhere in-between (see **Figure 1**). Participants press the wearable's button to record their response, at which point the built-in accelerator sensor determines, timestamps, and locally stores the angle. The PAS thus makes it possible to quickly and intuitively conduct assessments without questionnaires or visual response scales. Because interruption burden is low, the PAS also makes it possible to assess even very fleeting phenomena and/or conduct very frequent assessments, which may be especially useful for clinical psychologists and psychiatrists, for whom symptoms could be assessed longitudinally rather than requiring memory of symptomatology in anamnesis (6). Because the PAS also allows for continuous measurement, it can be compared to Visual Analogue Scales (VAS), which are psychometric response scales where participants indicate a position on a graphically presented continuous line between two end-points.

Validating the Physical Analogue Scale: First Pilot Studies

In the first pilot study, we were interested in how well people are able to estimate a certain angle. With VAS, people can usually easily estimate, say, the middle of the scale, e.g., the middle of a graphically presented line. But how good are people when instructed to estimate the middle of a 90° angle? In an ESM field study (4 weeks duration), we instructed participants to estimate 45° when prompted by the wearable using a time-based sampling procedure with a haptic prompt.

In the second pilot study, we were interested in whether a PAS measurement comes to the same result as a VAS measurement.

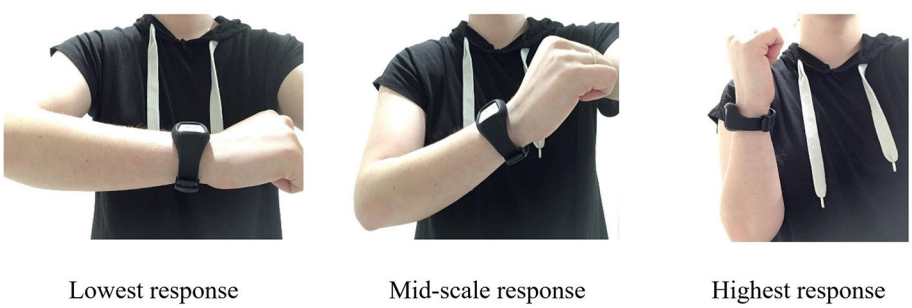


FIGURE 1 | The physical analogue scale.

In a laboratory study, we asked participants to judge their extraversion using eight items, first in an online questionnaire using a VAS and later doing the same assessment using the wearable and the PAS.

Using the Physical Analogue Scale to Assess Fleeting and Frequently-Occurring Phenomena: The Example of Social Media Ostracism

To demonstrate both the feasibility of our approach for reliable and accurate event- and time-based assessments and the advantages of the wearable/PAS approach, we conducted an ESM study on the effects of social media ostracism (17). Many social media platforms now integrate a so-called "seen-function" for outgoing messages in their software, which signals to users when a recipient has seen their message (e.g., WhatsApp uses a gray tick-mark to indicate when a message has left the sender's device, two gray tick-marks to indicate when the message has been delivered, and two blue tick-marks once the recipient has seen the message). Knowing that someone has seen but not responded to a message (i.e., social media ostracism) may cause the sender to experience negative emotions, such as feeling ostracized and/or offended (18). Given the current permanently-online, permanently-connected *zeitgeist* (19), social media ostracism represents a frequently-occurring daily life event with immediate effects. Although social media ostracism occurs in people's everyday lives (18, 20, 21), thus far almost all studies on its effects have either been conducted in the laboratory or are cross-sectional [for exceptions, see (22, 23)]. Social media ostracism is thus an ideal micro-situation for showcasing the advantages of the wearable/PAS method.

Our main aim was to demonstrate the feasibility of our approach for reliable and accurate event- and time-based assessments and the advantages of the wearable/PAS method: if the method is valid, we should be able to replicate the results of previous research. Based on existing research (20, 22, 24), we had two hypotheses:

Hypothesis 1: People feel offended after experiencing social media ostracism (compared to their own personal baseline).

Hypothesis 2: People feel more offended when a group vs. a single recipient ignores a message.

We also analyzed the extent to which feeling offended generally (i.e., participants' personal baselines) was associated with several personal characteristics. We expected that feeling offended would be positively correlated with Neuroticism, narcissism, and perceived text message dependency, and negatively correlated with self-esteem (22) and collective self-esteem related to online groups (CSE-OG). We had no specific hypotheses about the relationships between feeling offended and the personality traits of Agreeableness, Conscientiousness, Openness to Experience, and Extraversion; their inclusion was purely exploratory (for study preregistration, see https://osf.io/7j3e9/).

METHODS—FIRST PILOT STUDY

Participants and Procedure

Eight subjects (75% female; $M_{age} = 33.3$, $SD_{age} = 9.59$, range = 21–49 years) participated in this study. Six subjects used the wearable on their left arm and the other two on the right one (i.e., mostly the non-dominant hand). Data collection started on the same day for each participant and ended after 4 weeks. Participants were instructed to hold the forearm at an angle of 45° and press the button once whenever they were signaled by a haptic stimulus elicited by the wearable itself (time-based sampling procedure). During the data collection phase, participants were in diverse field settings, ranging from the office to leisure activities, such as 2-week hiking trips and journeys abroad.

Wearable

We developed software for a commercially available, openly programmable wearable from mbientlab (MMR+ wristband kit including microcomputer, eight MB memory, re-chargeable battery, case, elastic band, and coin vibrator motor: ∼100$; https://mbientlab.com/metamotionr/). Participant data were stored on the wearable and uploaded at the end of the study onto the researchers' smartphone or tablet via a Bluetooth connection (without Internet). Although an integrated infrastructure with servers, databases, and administration interfaces would allow data to be processed in near real-time (25), we elected to use a different approach that does not require additional data security measures (e.g., firewalls, encryption). The wearable had one button and several built-in sensors (e.g., light intensity, acceleration, air pressure, gyroscope). For this study, only the button and the acceleration sensor were enabled and used. The source code and administration application (Android) is freely available (see Open Practices Section). We used the following configuration, which represents a time-based ESM study with three time-points per day. Random signal time points within the following time frames: 8 a.m. to 11 a.m., 11 a.m. to 2 p.m., 2 p.m. to 5 p.m.

Wearable: Estimate 45° (Dependent Measure)

When participants pressed the wearable's button, the built-in accelerator sensor determined, timestamped, and stored its position in 3-dimensional space, and also saved the number of button presses[1]. The values for x-, y-, and z- were then transformed into an angle between 0° and 90° using the following formula:[2]

$$\text{Angle}\left[\text{degree}\right] = \arctan\left(\frac{|y|}{\sqrt{\left(x^2 + z^2\right)}}\right) * \frac{180}{\pi}$$

[1] For technical specifications, see https://mbientlab.com/metamotionr/.

[2] We assumed that participants would wear the wearable on the wrist of their non-dominant hand. However, some participants did not or changed the wearable from one hand to another during the data collection phase. We therefore had to use the absolute value of *y*, which effectively mirrors negative angles into positive angles.

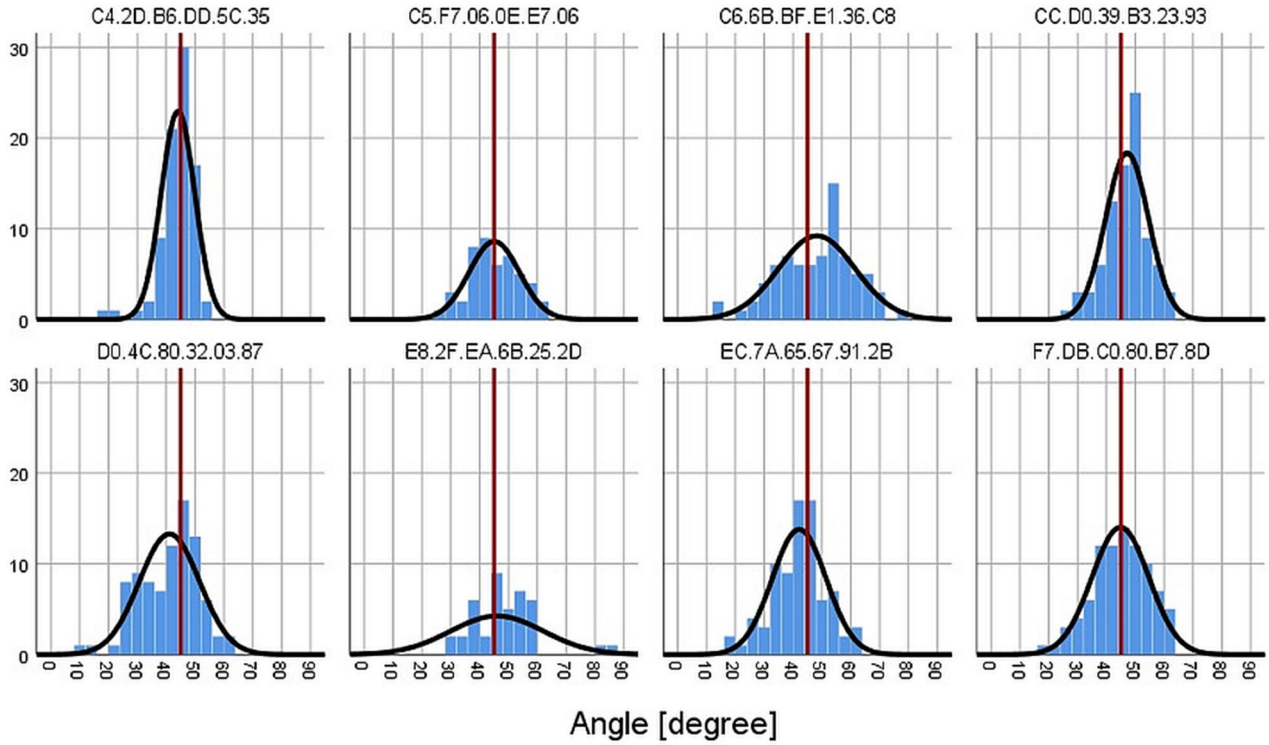


FIGURE 2 | Frequency histogram of all estimated 45° angles for each participant (represented by the wearable's MAC address). The red line marks the 45° from the instruction.

RESULTS—FIRST PILOT STUDY

The mean angle over all measurements ($n = 592$) was 44.7° on average ($SD = 10.38$, Median = 45.3; one-sample t-test: $t = -0.75$, $p = 0.454$, reference value = 45). The mean angle across participants ranged from 41.1° to 48.2° (range of $SD = 5.85$–16.18). In general, participants were quite accurate in estimating the angle of 45° during the field phase of 4 weeks (see **Figure 2**), bearing in mind that the field settings were very diverse (from office settings to leisure activities like hiking).

METHODS—SECOND PILOT STUDY

Participants and Procedure

Sixteen subjects (87.5% female; $M_{age} = 22.7$, $SD_{age} = 2.6$, range 20–28 years) participated in this study. Fourteen subjects used the wearable on their left arm and the other two on the right one (i.e., mostly the non-dominant hand). Data collection was realized as a group administration in a classroom. In addition to some test measurements that are not part of this study, participants completed the Extraversion subscale of the BFI (26) on a smartphone using a VAS and in parallel using the PAS on the wearable. For the PAS, participants were instructed to hold the forearm in the desired angle and press the button once. A short haptic feedback was elicited by the wearable when the angle was successfully saved.

Material–Big Five Inventory [BFI: German Version (26)]

To keep the validation study short, we only assessed the Extraversion subscale (8 Items) of the BFI. For the smartphone administration, we used a VAS (0: *does not apply at all*, 100: *applies very well*) and for the wearable administration the PAS (0°: *does not apply at all*, 90°: *applies very well*).

Statistical Analyses

We used SPSS (v. 26) to conduct all statistical analyses. We calculated Cronbach α, Pearson correlations, and curve estimation regression analyses to check for linearity between VAS and PAS.

RESULTS—SECOND PILOT STUDY

All extraversion items were highly intercorrelated between the PAS and VAS (rs between 0.63 and 0.94, for more details and descriptives, see **Supplementary Table 1**). Furthermore, reliabilities were all very good (Cronbach α: $\alpha_{VAS} = 0.93$; $\alpha_{PAS} = 0.83$) but descriptively slightly lower for the PAS. Extraversion mean scores of the VAS and PAS were highly correlated ($r = 0.95$, $p < 0.001$). To analyze if the relationship between PAS and VAS scores was linear, we calculated a curve estimation regression analysis. It could be that participants when using the PAS were better at differentiating at lower PAS angles but coarser at higher

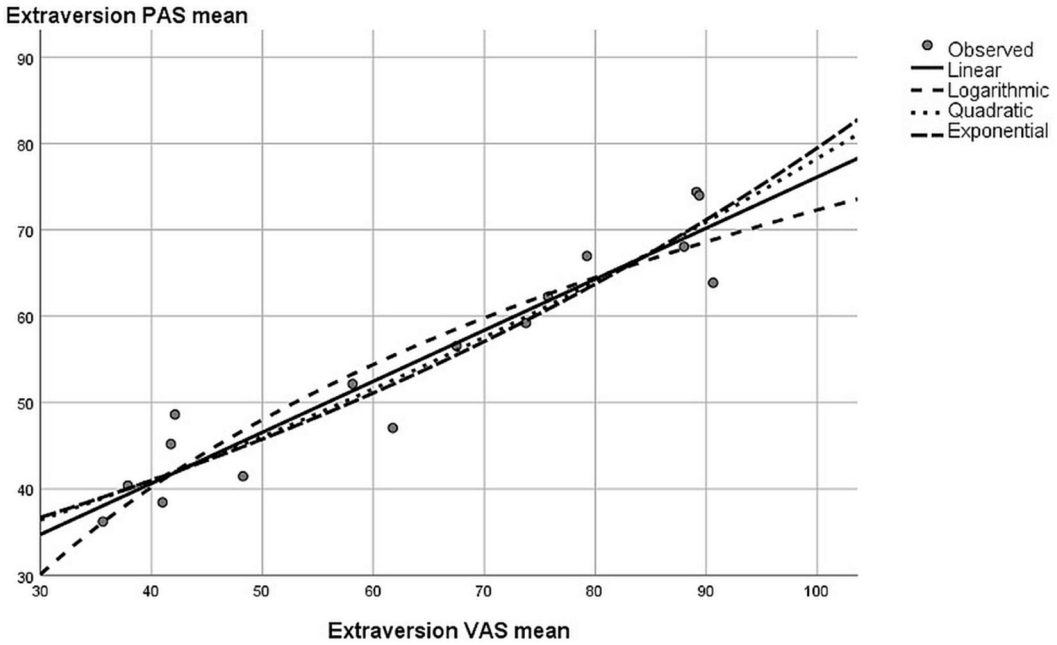


FIGURE 3 | Graphical results from the curve estimation regression—second pilot study.

angles (lower angles might be easier to establish than higher angles were the forearm is in an almost upright position). If this was the case, the relationship between PAS and VAS should not be linear. A curve estimation regression did not find substantial differences between linear ($R^2 = 0.907$), logarithmic ($R^2 = 0.886$), quadratic ($R^2 = 0.910$), and exponential ($R^2 = 0.905$) curve estimations regarding their explained variance levels (see also **Figure 3**).

METHODS—OSTRACISM STUDY

Participants and Procedure

The research project was conducted in February through April 2019 and participants were recruited on a rolling basis. Based on a power analysis for multi-level designs [(27), p. 123ff], we determined that a minimum sample size of 23 participants with 14 observations each would be sufficient for revealing medium-sized effects of social media ostracism on offendedness [based on (24), we conservatively assumed an effect size of 0.3; $\alpha = 5\%$]. To account for drop-out, technical problems, and so forth, we aimed to recruit 60 participants.

We used a project homepage and a detailed information sheet to provide interested individuals with information about the study objectives and design. Participants were met face-to-face in order to provide them with a fuller description of the assessment procedures. After providing consent, research assistants handed out a wearable that could be easily worn on the wrist. Participants were asked to familiarize themselves with the wearable by clicking once and then waiting for the haptic confirmation (i.e., vibration), then twice, and finally three times. The principal investigators directly answered any questions. Participants also received a credit card-sized laminated information sheet with visual instructions for the PAS, a definition of feeling offended as the variable to be assessed ("Feeling offended: the experience that one's honor, values, and/or feelings having been disregarded or violated, and especially the feeling that one has been insulted"; definition of *Kränkung* from the German-language Wikipedia), and the researchers' contact information. Although the wearable's battery lasts up to 4 weeks, we instructed participants to charge the battery weekly with a USB charger.

Participants then wore the wearable in their everyday lives for 14 days. We collected both event-sampling and time-based sampling data. First, participants were instructed to press the button whenever a single recipient or a group had seen but not responded to their social media message, provided that the participant expected a response. Participants then used the PAS to indicate how offended they felt (event-based sampling). In addition, in order to assess how offended participants generally felt, participants used the PAS to indicate how offended they felt twice every day (time-based sampling). Specifically, the wearable was programmed to issue a vibration signal once between 10 a.m. and 4 p.m. and again sometime between 4 p.m. and 10 p.m., at which point participants used the PAS to indicate how offended they felt at that time. These assessments were used as a baseline measure (i.e., how offended they normally felt during the day). To submit a response, participants positioned their forearm to the appropriate position (from 0° = *not at all offended* to 90° = *extremely offended*) and then pressed the button: once when their message had been ignored by a single recipient, twice when their message had been ignored by a group, and three times for the time-based baseline assessments.

After the field phase, participants returned to the lab and completed an online questionnaire on a laptop computer. Participants entered the first four digits of the wearable's unique media access control address[3] (printed on the case) so that questionnaire data could be matched with the wearable data while preserving anonymity. After completing the questionnaire, participants deposited the wearable into a box. Participants were thanked and debriefed, and those who wanted to be included in the raffle were asked to provide an email address.

We recruited a total of 59 participants from our social networks. To be eligible for participation, participants had to: (1) own a smartphone, (2) use social media (e.g., WhatsApp, Facebook, Twitter, Snapchat, etc.), and; (3) already use or be willing to activate the "seen function" for outgoing messages for the duration of the study. Six participants (10.2%) dropped out before the end of the study or refused to complete the final questionnaire. Reasons for the drop-out were assessed orally (e.g., unable to use wearables at work, infrequent use of social media, and lost wearables). The final sample size thus consisted of $N = 53$ participants.

Participants in the final sample were $M = 26.5$ years old ($SD = 9.56$, range = 18–57), predominantly women (81%), and from Austria. Most had completed secondary school (63.5%) or had a tertiary degree (26.3%) as their highest level of completed education. Participants were treated in accordance with the World Medical Association Declaration of Helsinki and with local ethical guidelines. They gave informed consent prior to participating. As an incentive, participants could voluntarily enter a raffle with the chance of winning two prizes worth 100 Euro each.

Wearable: Feeling Offended (Dependent Variable)

When participants pressed the wearable's button, the built-in accelerator sensor determined, timestamped, and stored its position in 3-dimensional space, and also saved the number of button presses. The values for x-, y-, and z- were then transformed into an angle between 0° (*not at all offended*) and 90° (*extremely offended*).

Online Questionnaire Measures

Measures are described in the order of their presentation in the questionnaire.

Demographics

Participants reported their age, sex (female/male/other), and highest level of completed education (categories).

Social Media Use

Participants responded to four items about their social media use during the field phase: "How many messages have you read in the last 14 days per day on average?," "How many messages have you sent in the last 14 days per day on average?," "For how many minutes per day on average have you passively used social media (e.g., reading tweets and Facebook posts, looking at Snapchat pictures, watching YouTube videos)?" and "For how many minutes per day on average have you actively used social media (e.g., writing posts)?" Underlining was used to stress the difference in item wording.

Interruption Burden

Participants indicated how difficult it was to conduct the assessments (1: *not at all difficult*, 9: *very difficult*) and how often they forgot to submit a rating over the 2-week period.

Self-Esteem

We used the German-version (28) of the Rosenberg Self Esteem Scale [RSES; (29)] to assess participants' global self-esteem. Participants used a 4-point scale (1: *totally disagree*, 4: *totally agree*) to respond to 10 items. Answers to the 10 items were averaged ($\alpha = 0.85$).

Big Five Personality Traits

We used the German-version (26) of the BFI as in the second pilot study (30) to assess participants' personality traits. Participants used a 7-point scale (1: *totally disagree*; 7: *totally agree*) to respond to 44 items. Responses to items related to each trait were averaged (Neuroticism: 8 items, $\alpha = 0.63$; Extraversion: 8 items, $\alpha = 0.80$; Conscientiousness: 9 items, $\alpha = 0.69$; Agreeableness: 9 items, $\alpha = 0.69$; Openness: 10 items, $\alpha = 0.77$).

Perceived Text Message Dependency

We used the Self-perception of Text-message Dependency Scale (31) to assess the extent to which participants perceived themselves as being psychologically dependent on receiving text messages. Scores on the original scale have adequate psychometric properties (32, 33). We translated the scale into German using the parallel blind technique (34). Participants used a 7-point scale (1: *totally disagree*; 7: *totally agree*) to respond to 15 items about three dimensions of text message dependency: Emotional Reaction (5 items, $\alpha = 0.83$, e.g., "I feel disappointed if I don't get a reply to my message immediately"), Excessive Use (5 items, $\alpha = 0.85$, e.g., "I consider myself a quick typist on mobile phones"), and Relationship Maintenance (5 items, $\alpha = 0.81$, e.g., "I feel disappointed if I don't receive any text messages"). In the current study, we analyzed only overall scores ($\alpha = 0.68$) to reduce the number of predictors included in the model (for intercorrelations between the subscales, see **Supplementary Table 2**).

Narcissism

We used a short version of the Narcissistic Admiration and Rivalry Questionnaire (35) to assess narcissism. Scores on the scale are reliable and valid (36). Participants used a 6-point scale (1: *not agree at all*, 6: *agree completely*) to indicate their agreement with six items. Scores on the scale can be computed to reflect overall narcissism score or two subdimensions of narcissism (35). Because the reliability for scores on one of the subdimensions was low (Admiration: $\alpha = 0.62$; Rivalry dimension: $\alpha = 0.41$), we analyzed only overall scores ($\alpha = 0.68$).

[3]A media access control address is a unique identifier assigned to a network interface controller, such as a Bluetooth network device. Because wearables had a Bluetooth connection, each had a unique address.

Collective Self-Esteem Related to Online Groups

We used a translated and modified version of the Collective Self-Esteem scale [CSE; (37)] to assess participants' CSE-OG. We first translated the CSE into German using the parallel blind technique (34). We then modified the scale so that all items referred specifically to "online social groups" as opposed to "social groups" in general (i.e., CSE-OG). Participants used a 7-point scale (1: *strongly disagree*; 7: *strongly agree*) to respond to 16 items (e.g., "I am a worthy member of the online social groups I belong to"). Scores on the CSE can be calculated in terms of an overall or in terms of four subdimensions of collective self-esteem. Because the reliability of one subdimension of the adapted scale was low (Importance to Identity: $\alpha = 0.45$; reliability of all other subdimensions ≥ 0.62), we analyzed only overall scores ($\alpha = 0.80$).

General Comments and Comments About the Wearable

At the end of the online questionnaire, participants had the option of providing open comments ("Do you have any general comments about this wearable study?;" "Do you have comments about the wearable itself, e.g., the signals, which were sent out twice a day using vibration, and so forth?"). For results see, **Supplementary Material**.

Statistical Analyses

We used *R* [package lme4 (38), sjstats (39)] to conduct all statistical analyses (40). After a first inspection of the data, we did not exclude any participants even if participation in the longitudinal part stopped before the end of the study. First, we analyzed descriptive statistics (e.g., *M*, *SD*) and intercorrelations of all study variables. Next, we used a random-intercept, random-slope multi-level regression analysis to analyze the effect of social media ostracism (by either a single-recipient or a group) on how offended participants felt. The multi-level model accounts for the nested design of our study with measurement occasions (level 1) nested within persons (level 2). We created dummy variables for sex (female = 1, male = 2), being ignored by a single recipient (*Single-chat*), and being ignored by a group (*Group-chat*). We ran a baseline model without any predictors to determine the overall intraclass correlation (ICC, i.e., the extent to which how offended participants felt varied between people as opposed to across measurement occasions). We similarly calculated ICCs as indicators of test-retest reliability (i.e., the consistency of responses across measurement occasions; see **Supplementary Material**).

We then ran a model in which age, sex, self-esteem, Extraversion, Openness, Neuroticism, Agreeableness, Consciousness, text message dependency, narcissism, and CSE-OG were all simultaneously entered on level 2 [all were grand-mean centered except sex; (41)]. The saturated model is displayed below:

Level 1 (within person): $\text{Offendedness}_{ti} = \pi_{0i} + \pi_{1i}\ \text{Single-chat}_{ti} + \pi_{2i}\ \text{Group-chat}_{ti} + e_{ti}$

Level 2 (between people): $\pi_{0i} = \beta_{00} + \beta_{01}\ \text{Sex}_i + \beta_{02}\ \text{Age}_i + \beta_{03}\ \text{Self-Esteem}_i + \beta_{04}\ \text{Extraversion}_i + \beta_{05}\ \text{Neuroticism}_i + \beta_{06}\ \text{Openness}_i + \beta_{07}\ \text{Agreeableness}_i + \beta_{08}\ \text{Consciousness}_i + \beta_{09}\ \text{Text-message Dependency}_i + \beta_{010}\ \text{Narcissism}_i + \beta_{011}\ \text{CSE-OG}_i + r_{0i}$

Level 2: $\pi_{1i} = \beta_{10} + r_{1i}$

Level 2: $\pi_{2i} = \beta_{20} + r_{2i}$

We used Ω^2–a generalized R^2 for linear mixed effect models (42)—as a measure of explained variance, with $\Omega^2 \geq 0.01$, 0.09, and 0.25, respectively, indicating small, medium, and large shares of explained variance, respectively.

RESULTS—OSTRACISM STUDY

A total of 2,588 responses were recorded (i.e., the angle from the accelerometer in combination with a single-, double-, or triple-button press). We deleted test responses (see Participants and Procedure section) as well as 55 (2.1%) responses that were followed by more than three button presses[4]. This resulted in a final sample of 2,272 assessments, of which 1,031 were time-based (triple-button press) and 1,241 followed an event of social media ostracism (991 times by a single recipient, i.e., single-button press; 250 times by a group, i.e., double-button press). The compliance rate (i.e., whether participants responded to the time-based assessment signals)[5] dropped slightly over time ($M = 59.4\%$; range: 39.8–75.4%; see **Figure 4**), whereas the final days showed the largest drop, probably due to participants erroneously assuming that the study had ended (i.e., non-response; from 53.4% on day 13 to 39.8% on day 14—the last day of study). If we correct the compliance rate by non-response, the drop in compliance rate is less steep. Drop-out attrition (i.e., leaving the study before the end) was 10% ($n = 6$) and no non-response attrition occurred (i.e., taking part in the study but not pressing the button). **Supplementary Table 2** displays the variable intercorrelations and **Supplementary Figure 1** displays the distribution of all responses (angles) separated by category (time-based, message ignored by a group, message ignored by a single-recipient). Both the event- and time-based PAS responses were highly consistent across measurement occasions (ICCs > 0.91, see **Supplementary Material**).

Interruption Burden

Participants did not find it difficult to conduct the assessments ($M = 2.4$, $SD = 1.9$, Median = 2, range: 1–9; possible scale range: 1–9). Participants estimated that they forgot to submit $M = 3.8$ assessments ($SD = 3.4$; Median = 3; range: 0–15) during the 14-day data collection phase.

Social Media Use

During the 14 days of data collection, participants reported using social media actively for $M = 41$ minutes ($SD = 30.8$, Median = 30, range: 1–120) and passively for $M = 80$ minutes ($SD = 65.3$, Median = 60, range: 1–240) every day. They read $M = 62$ ($SD = 137.8$, Median = 30, range: 3–1,000) and wrote $M = 34$ ($SD = 46.5$, Median = 20, range: 2–300) social

[4]Some participants indicated difficulty feeling the resistance of the button. To reduce this error, we have adapted the software so that participants receive haptic feedback after every button press.

[5]Compliance rate = (responded to time-based signals / scheduled time-based signals) × 100.

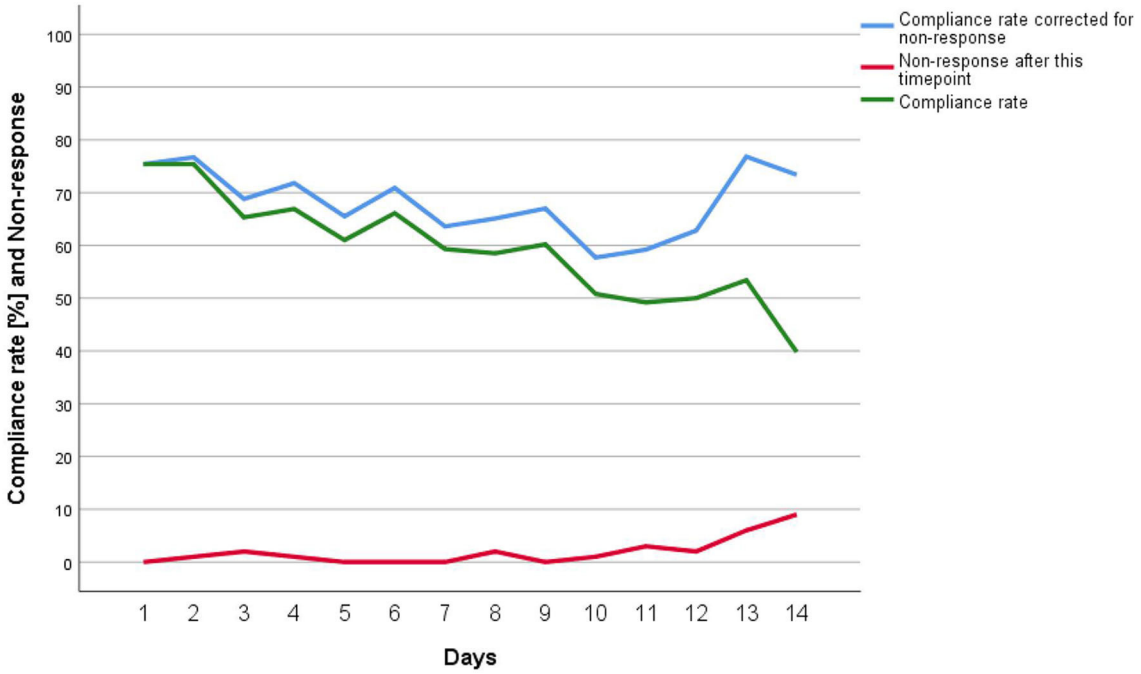


FIGURE 4 | Compliance rate of the time-based assessments and non-response before end of study.

media messages each day. Thus, participants estimated that they wrote a total of 25,144 messages within the 14-day timeframe, implying that ∼4.9% or one out of twenty of their messages was ignored.

People Feel More Offended After Experiencing Social Media Ostracism

The ICC for the null model was 8.9%, indicating that 8.9% of the observed variance in how offended participants felt was associated with differences between people, while 91.1% of the variance was associated with within-person differences across measurement occasions.

Table 1 displays the results of the multi-level analysis. The overall mean level of offendedness (intercept) was 13.5°. Participants felt significantly more offended when a single recipient ignored their message (7.0° more than their own baseline) and even more offended when they were ignored by a group of recipients (11.5° more than their own baseline). The difference between the increase in how offended participants felt after their message was ignored by a group vs. single-recipient was also significant (see **Supplementary Table 3**)[6]. This supports Hypothesis 1, as well as Hypothesis 2. In sum, the included predictors explained a substantial proportion of variance in how offended participants felt ($\Omega^2 = 0.21$).

[6]When taking the sign of the angle positive/ negative into account, the mean baseline level of offendedness is lower and the effect sizes of being ignored by a single-recipient and group are higher than reported here.

Between-Person Differences

None of the level 2 personal characteristics were significantly related to how offended participants generally felt (i.e., across measurement occasions). In follow-up analyses, we also checked whether any of the level 2 variables moderated the increase in how offended participants felt after experiencing social media ostracism (i.e., whether any of the level 2 variables predicted the *slope* of either the level 1 dummy variables "single-chat" or "group-chat"). Again, none of the interactions was significant (for details, see R-code in the online repository).

GENERAL DISCUSSION

Wearables programmed with a PAS have the potential to improve the assessment of frequently occurring and/or fleeting events in participants' everyday lives (5, 9, 10), which may be especially useful in clinical psychology and psychiatric settings where symptoms could be assessed longitudinally (6). In the first pilot study, we showed that participants could accurately estimate an angle of 45° using the PAS in a 4-week field setting. In the second pilot study (lab setting), we confirmed the validity of the PAS by comparing mean extraversion values between the PAS and VAS. Furthermore, the PAS had also comparable reliability to the VAS when assessing extraversion and both formed linear relationships. This suggests that differences in angles are probably equidistant along the measurement scale. Finally, we used data from an ESM study on social media ostracism as an example of a micro-situation that can be difficult to assess in laboratory settings or with traditional cross-sectional questionnaires, but feasible with wearables and the

TABLE 1 | Results of the multi-level analysis.

Dependent variable: offendedness predictors	Fixed						Random	
	Coeff.	β	95% CI	*B*	*SE*	*t*	Coeff.	*SD*
Intercept	β_{00}			13.5	0.79	17.01***	r_{0i}	3.29
Within-person								
Single chat	β_{10}	0.20	0.13–0.26	7.0	1.13	6.21***	r_{1i}	5.43
Group chat	β_{20}	0.21	0.14–0.27	11.5	1.85	6.18***	r_{2i}	8.88
Between-person								
Sex	β_{01}	−0.06	-0.15–0.02	−2.7	1.83	−1.48		
Age	β_{02}	0.07	-0.02–0.17	0.1	0.09	1.53		
Self-esteem	β_{03}	−0.03	-0.12–0.07	−1.1	1.95	−0.58		
Extraversion	β_{04}	−0.01	-0.09–0.09	−0.8	1.07	−0.07		
Neuroticism	β_{05}	0.01	-0.10–0.12	0.3	1.37	0.25		
Openness	β_{06}	−0.01	-0.09–.07	−0.3	1.13	−0.26		
Agreeableness	β_{07}	−0.01	-0.11–0.09	−0.3	1.38	−0.25		
Consciousness	β_{08}	> −0.01	-0.11–0.09	−0.2	1.46	−0.14		
Text message dependency	β_{09}	0.07	-0.06–0.20	1.1	1.11	1.03		
Narcissism	β_{010}	−0.03	-0.15–0.09	−0.7	1.38	−0.51		
CSE-OG	β_{012}	−0.02	-0.13–0.08	−0.6	1.29	−0.46		

All level 2 variables were grand mean centered except for sex. CI, Confidence Interval; CSE-OG, Collective self-esteem related to online groups.
****p < 0.001.*

PAS. The wearable/PAS approach worked well. We successfully replicated past research on ostracism (21–23), which found negative effects on emotional states, belongingness (24), and heightened negative affect [for a review, see (43)]. In the present study, we also found negative effects of ostracism, i.e., being ignored online led to feelings of being offended in one-to-one chat situations (Hypothesis 1) and more so when ignored by a group (group chat: Hypothesis 2). Furthermore, our findings of a negative effect of ostracism are in line with other ESM studies on ostracism (22), although still, ESM research is rare (22, 23). Participants also did not find it difficult to complete the assessments and general comments suggested that most participants felt positive or at least neutral about the usage of a wearable (see also the results to the wearable-specific open questions in the **Supplementary Material**). Nevertheless, compliance rates gradually dropped during the study, with lowest compliance on the last day (see **Figure 4**). Future research needs to analyze the reasons for this in more detail. In the present study, it may have been an effect of the tactile vibration alarm (e.g., frequency and duration of vibration) or other problems (e.g., time-based signals too early or too late for some participants; see results in **Supplementary Material**).

Test-retest consistency of subsequent button presses (i.e., two- and three- button presses) was high. Although we did not investigate whether this extrapolates to button-presses with more time in between, this means that the sensors' measurement accuracy was high and, furthermore, that participants did not substantially change the angle of their forearm when pressing the button more than once. Although we investigated validity in rather small sample sizes, the findings suggest that usage of the PAS is feasible, well-accepted by participants, and easy-to-use (5, 9, 10).

Our main focus was on demonstrating the feasibility of our approach for reliable and accurate event- and time-based assessments and advantages of the wearable/PAS approach. Nevertheless, our results also make some important contributions to clinical research on ostracism. By using an ESM design, we were able to assess how often participants experienced social media ostracism in their everyday lives. We found that approximately every 20th message was ignored, causing our participants to feel offended several times a day. Given that people use social media all around the world (∼65 billion messages are currently sent each day), the impact of social media ostracism may be a highly relevant experience for people around the globe. It therefore seems worthwhile to further analyze the short- and long-term consequences of social media ostracism.

Interestingly, offendedness differed predominantly within as opposed to between participants, and we found no evidence that personal characteristics (e.g., self-esteem, Big Five traits, text message dependency) explained differences in how offended people generally felt. This does not mean, however, that personal characteristics are completely unrelated to experiences of social media ostracism. Personal characteristics might, for example, matter more in the longer- than in the short-term [e.g., participants with high emotional stability might immediately feel offended by social media ostracism just like their peers, but might return to their baseline level faster; (22)]. Future studies on how the effects of social media ostracism unfold over time would be fruitful [for a similar approach to well-being, see (44)].

Potentials and Limitations of the Physical Analogue Scale and Other Sensor-Based Data Collection Procedures

With low interruption burden, long battery life, smartphone independence, and relatively low price (∼100$), our wearable/PAS approach overcomes several of the challenges

associated with previous data collection procedures. Although further validation studies are needed such as the accuracy of the sensors or study compliance in comparison with smartphones, we believe that the wearable/PAS approach offers not only psychologists and psychiatrists but also researchers in other disciplines (e.g., medicine, sociology) a valuable combination for studying micro-events in everyday life (e.g., clinical symptoms). It is another example of how computer science can extend the methods of other sciences, such as psychology (45) or physiology (5). At present, the wearable/PAS can only be used to assess a few items; however, applications could potentially be developed so that additional items could be presented on the touchscreen of existing smartwatches [e.g., using Android Wear; for example, see (5, 10)].

We see large potential for sensor-based scales like the PAS. We think that sensor-based scales are particularly well-suited to capture frequent and/or short-lived phenomena because of the low interruption burden. Furthermore, due to the unobtrusive assessment procedure of the PAS, we think our approach is suitable for the assessment of sensitive topics (e.g., sexuality, racism, suicidal thoughts, self-harming behavior like "cutting"). Aside from that, other sensor-based assessments could be developed in the future, such as using hand tilts as a response scale (46) or the acceleration with which one punches one's own fist into one's open hand as an intuitive measure of aggression. Of course, wearables and sensor-based data do not replace but rather complement more traditional methods. Furthermore, our approach probably will not work for every population and should be thoroughly thought out when planning a study based on the wearable/PAS approach. For example, Vega et al. (4) found, that paper/pencil diaries worked better than several digital measurement procedures in a sample of patients with Parkinson's disease.

Conclusion

Although further in-depth validation studies are needed, wearables might offer researchers the possibility of delving into participants' everyday lives more deeply than ever before (5, 6, 10) by being unobtrusive and inconspicuous. We have described how an inexpensive wearable programmed with the PAS can be used to assess frequent and/or fleeting events, supplementing past wearable developments. Our validity studies and application of the PAS suggest that the sensor-based PAS is an intuitive, easy-to-use scale for collecting data on how people feel and behave in the real world.

DATA AVAILABILITY STATEMENT

De-identified data along with the analysis scripts and all materials are posted at https://osf.io/7j3e9/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SS, IS, and PA developed the ostracism study concept. Testing and data collection were performed by IS and PA for the ostracism study. SS and DL collected the data for the pilot studies and developed the wearable concept. DL programmed the wearable and gave technical support. SS performed the data analysis and interpretation. SS drafted the manuscript and IS, PA, and DL provided critical revisions. All authors contributed to the study design and approved the final version of the manuscript for submission.

FUNDING

This work was supported by the Austrian Science Fund (FWF) under the Grant No. P31800-N38.

ACKNOWLEDGMENTS

We thank Catherine Bowen and Viren Swami for their useful comments and Selina Volsa for here support in data collection for the pilot studies.

OPEN PRACTICES

The study was preregistered; the preregistration can be assessed at https://osf.io/76kru. De-identified data along with the analysis scripts and all materials are posted at https://osf.io/7j3e9/.

Source code of the administration app can be found on GitHub at https://github.com/KL-Psychological-Methodology/ESM-Board-Admin/.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.538122/full#supplementary-material

REFERENCES

1. Conner TS, Tennen H, Fleeson W, Barrett LF. Experience sampling methods: a modern idiographic approach to personality research. *Soc Personal Psychol Compass*. (2009) 3:292–313. doi: 10.1111/j.1751-9004.2009.00170.x
2. Ebner-Priemer UW, Trull TJ. Ambulatory assessment: an innovative and promising approach for clinical psychology. *Eur Psychol*. (2009) 14:109–19. doi: 10.1027/1016-9040.14.2.109
3. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. *Annu Rev Clin Psychol*. (2008) 4:1–32. doi: 10.1146/annurev.clinpsy.3.022806.091415

4. Vega J, Couth S, Poliakoff E, Kotz S, Sullivan M, Jay C, et al. Back to analogue: self-reporting for Parkinson's disease progression. In: *Proceedings of the CHI 2018 Conference on Human Factors in Computing Systems - Association for Computing Machinery*. New York, NY: ACM (2018).
5. Beukenhorst AL, Howells K, Cook L, McBeth J, O'Neill TW, Parkes MJ, et al. Engagement and participant experiences with consumer smartwatches for health research: longitudinal, observational feasibility study. *JMIR Mhealth Uhealth*. (2020) 8:e14368 doi: 10.2196/14368
6. Insel TR. Digital phenotyping: technology for a new science of behavior. *JAMA*. (2017) 318:1215–6. doi: 10.1001/jama.2017.11295
7. Mehl MR, Conner TS. *Handbook of Research Methods for Studying Daily Life*. New York, NY: Guilford (2012).
8. Miller G. The smartphone psychology manifesto. *Perspect Psychol Sci*. (2012) 7:221–37. doi: 10.1177/1745691612441215
9. Intille S, Haynes C, Maniar D, Ponnada A, Manjourides J. μEMA: micro-interactions based ecological momentary assessments (EMA) using a smartwatch. In: *Proceedings of the ACM International Conference on Ubiquitous and Pervasive Computing (UbiComp' 16)*. New York, NY: ACM (2016).
10. Ponnada A, Haynes C, Maniar D, Manjourides J, Intille S. Microinteraction ecological momentary assessment response rate: effect of microinteractions or the smartwatch? In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. New York, NY: ACM (2017).
11. Al-Eidan RM, Al-Khalifa H, Al-Salman AM. A review of wrist-worn wearables: sensors, models, and challenges. *J Sensors*. (2018) 2018:id5853917. doi: 10.1155/2018/5853917
12. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient noncompliance with paper diaries. *Br Med J*. (2002) 324:1193–4. doi: 10.1136/bmj.324.7347.1193
13. Ashbrook DL, Clawson JR, Lyons K, Starner TE, Patel N. Quickdraw: the impact of mobility and on-body placement on device access time. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '8)*. New York, NY, ACM (2008).
14. Pizza S, Brown B, McMillan D, Lampinen A. Smartwatch *in vivo*. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY: ACM (2016). p. 5456–69.
15. Shih PC, Han K, Poole ES, Rosson MB, Carroll JM. Use and adoption challenges of wearable activity trackers. In: *iConference Proceedings 2015*. Newport Beach, CA (2015).
16. de Barbaro K. Automated sensing of daily activity: a new lens into development. *Dev Psychobiol*. (2019) 61:444–64. doi: 10.1002/dev.21831
17. Vorderer P., Schneider FM. Social media and ostracism. In: Williams KD, Nida SA, Editors. *Ostracism, Exclusion, and Rejection*. New York, NY: Psychology Press (2016). p. 240–57.
18. Hoyle R, Das S, Kapadia A, Lee AJ, Vaniea K. Was my message read? privacy and signaling on Facebook messenger. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. New York, NY: ACM (2017). p. 3838–42.
19. Klimmt C, Hefner D, Reinecke L, Rieger D, Vorderer P. The permanently online and permanently connected mind. mapping the cognitive structures behind mobile internet use. In: Vorderer P, Hefner D, Reinecke L, Klimmt, C, editors. *Permanently Online, Permanently Connected. Living and Communication in a POPC World*. New York, NY: Routledge (2018) p. 18–28.
20. Mai LM, Freudenthaler R, Scheider FM, Vorderer P. "I know you've seen it!" Individual and social factors for users' chatting behavior on facebook. *Comput Hum Behav*. (2015) 49:296–302. doi: 10.1016/j.chb.2015.01.074
21. Tobin SJ, Vanman EJ, Verreynne M, Saeri AK. Threats to belonging on facebook: lurking and ostracism. *Soc Influen*. (2015) 10:31–42. doi: 10.1080/15534510.2014.893924
22. Nezlek JB, Wesselmann ED, Wheeler L, Williams KD. Ostracism in everyday life. *Group Dyn Theory Res Pract*. (2012) 16:91–104. doi: 10.1037/a0028029
23. Nezlek JB, Wesselmann ED, Wheeler L, Williams KD. Ostracism in everyday life: the effects of ostracism on those who ostracize. *J Soc Psychol*. (2015) 155:432–51. doi: 10.1080/00224545.2015.1062351
24. Schneider FM, Zwillich B, Bindl MJ, Hopp FR, Reich S. Social media ostracism: the effects of being excluded online. *Comput Hum Behav*. (2017) 73:385–93. doi: 10.1016/j.chb.2017.03.052
25. Kheirkhahan M, Nair S, Davoudi A, Rashidi P, Wanigatunga AA, Corbett DB, et al. A smartwatch-based framework for real-time and online assessment and mobility monitoring. *J Biomed Inform*. (2019) 89:29–40. doi: 10.1016/j.jbi.2018.11.003
26. Lang FR, Lüdtke O, Asendorpf JB. Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen [validity and psychometric equivalence of the German version of the big five inventory in young, middle-aged and old adults]. *Diagnostica*. (2001) 47:111–21. doi: 10.1026//0012-1924.47.3.111
27. Twisk JWR. *Applied Multilevel Analysis*. Cambridge: Cambridge University Press (2006).
28. von Collani G, Herzberg PY. Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. *Zeitsch Different Diagno Psychol*. (2003) 24:3–7. doi: 10.1024//0170-1789.24.1.3
29. Rosenberg M. *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press (1965).
30. John OP, Srivastava S. The big five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin LA, John OP, editors. *Handbook of Personality: Theory and Research*. New York, NY: Guilford (1999). p. 102–38.
31. Igarashi T, Motoyoshi T, Takai J, Yoshida T. No mobile, no life: Self-perception and text-message dependency among Japanese high school students. *Comput Hum Behav*. (2008) 24:2311–24. doi: 10.1016/j.chb.2007.12.001
32. Liese BS, Benau EM, Atchley P, Reed D, Becirevic A, Kaplan B. The self-perception of text-message dependency scale (STDS): psychometric update based on a United States sample. *Am J Drug Alcohol Abuse*. (2018) 45:42–50. doi: 10.1080/00952990.2018.1465572
33. Lu X, Katoh T, Chen Z, Nagata T, Kitamura T. Text messaging: are dependency and excessive use discretely different for Japanese university students? *Psychiatry Res*. (2014) 216:255–62. doi: 10.1016/j.psychres.2013.12.024
34. Behling O, Law KS. *Translating Questionnaires and Other Research Instruments: Problems and Solutions*. Thousand Oaks, CA: Sage (2000).
35. Back MD, Küfner ACP, Dufner M, Gerlach TM, Rauthmann JF, Denissen JJA. Narcissistic admiration and rivalry: disentangling the bright and dark sides of narcissism. *J Person Soc Psychol*. (2013) 105:1013–37. doi: 10.1037/a0034431
36. Leckelt M, Wetzel E, Gerlach TM, Ackerman RA, Miller JD, Chopik WJ, et al. Validation of the narcissistic admiration and rivalry questionnaire short scale (NARQ-S) in convenience and representative samples. *Psychol Assess*. (2018) 30:86–96. doi: 10.1037/pas0000433
37. Luhtanen R, Crocker J. A collective self-esteem scale: self-evaluation of one's social identity. *Person Soc Psychol Bull*. (1992) 18:302–18. doi: 10.1177/0146167292183006
38. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. (2015) 67:1–48. doi: 10.18637/jss.v067.i01
39. Lüdecke D. *sjstats: Statistical Functions for Regression Models (Version 0.18.0)* (2020).
40. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing (2014).
41. Enders CK, Tofighi D. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol Methods*. (2007) 12:121–38. doi: 10.1037/1082-989X.12.2.121
42. Xu R. Measuring explained variation in linear mixed effects models. *Stat Med*. (2003) 22:3527–41. doi: 10.1002/sim.1572
43. Williams KD. Ostracism. *Annu Rev Psychol*. (2007) 58:425–52. doi: 10.1146/annurev.psych.58.110405.085641
44. Stieger S, Götz FM, Gehrig F. Soccer results affect subjective well-being, but only briefly: a smartphone study during the 2014 FIFA World Cup. *Front Psychol*. (2015) 6:497. doi: 10.3389/fpsyg.2015.00497
45. Yarkoni T. Psychoinformatics: new horizons at the interface of the psychological and computing sciences. *Curr Dir Psychol Sci*. (2012) 21:391–7. doi: 10.1177/0963721412457362
46. Xiao R, Laput G, Harrison C. Expanding the input expressivity of smartwatches with mechanical pan, twist,

tilt and click. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM (2014). p. 193–6.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 15 December 2020
doi: 10.3389/fpsyt.2020.607890

Identifying Psychological Symptoms Based on Facial Movements

Xiaoyang Wang[1,2], *Yilin Wang*[1,2], *Mingjie Zhou*[1,2], *Baobin Li*[2,3], *Xiaoqian Liu*[1,2]* and *Tingshao Zhu*[1,2]

[1] Institute of Psychology, Chinese Academy of Sciences, Beijing, China, [2] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China, [3] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China

Background: Many methods have been proposed to automatically identify the presence of mental illness, but these have mostly focused on one specific mental illness. In some non-professional scenarios, it would be more helpful to understand an individual's mental health status from all perspectives.

Methods: We recruited 100 participants. Their multi-dimensional psychological symptoms of mental health were evaluated using the Symptom Checklist 90 (SCL-90) and their facial movements under neutral stimulation were recorded using Microsoft Kinect. We extracted the time-series characteristics of the key points as the input, and the subscale scores of the SCL-90 as the output to build facial prediction models. Finally, the convergent validity, discriminant validity, criterion validity, and the split-half reliability were respectively assessed using a multitrait-multimethod matrix and correlation coefficients.

Results: The correlation coefficients between the predicted values and actual scores were 0.26 and 0.42 ($P < 0.01$), which indicated good criterion validity. All models except depression had high convergent validity but low discriminant validity. Results also indicated good levels of split-half reliability for each model [from 0.516 (hostility) to 0.817 (interpersonal sensitivity)] ($P < 0.001$).

Conclusion: The validity and reliability of facial prediction models were confirmed for the measurement of mental health based on the SCL-90. Our research demonstrated that fine-grained aspects of mental health can be identified from the face, and provided a feasible evaluation method for multi-dimensional prediction models.

Keywords: mental health, psychological symptoms, SCL-90, facial movements, machine learning, multitrait-multimethod matrix

OPEN ACCESS

Edited by:
Jennifer H. Barnett, Cambridge Cognition, United Kingdom

Reviewed by:
Aniruddha Sinha, Tata Consultancy Services, India
Jose A. Piqueras, Miguel Hernández University of Elche, Spain

***Correspondence:**
Xiaoqian Liu
liuxiaoqian@psych.ac.cn

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: *18 September 2020*
Accepted: *17 November 2020*
Published: *15 December 2020*

Citation:
Wang X, Wang Y, Zhou M, Li B, Liu X and Zhu T (2020) Identifying Psychological Symptoms Based on Facial Movements. Front. Psychiatry 11:607890. doi: 10.3389/fpsyt.2020.607890

INTRODUCTION

Mental illnesses have a significant impact on an individual's physical health (1), achievements (2, 3), and life satisfaction (4). In addition to scales, behavioral recognition methods have been developed to judge the existence (5) or degree (6, 7) of specific mental illnesses. However, identifying an individual's mental health status from a range of perspectives may be more helpful in non-professional scenarios such as self-monitoring or large-scale monitoring.

Many studies have found that the physiological and behavioral indicators of individuals with mental illnesses differ, including brain activity (8, 9), galvanic skin response (10), eye contact (11, 12), voice (13, 14), and facial movements (15). Moreover, people with different mental health disorders behave differently (16, 17). For example, patients with schizophrenia can be distinguished from those with depression by analyzing their non-verbal behavior during medical consultation (16). More granularly, neural activity in response to different emotional faces can help distinguish bipolar depression from unipolar depression. Such differences make it possible for machine learning models to diagnose the multi-dimensional psychological symptoms of mental illnesses. Meanwhile, the Symptom Checklist 90 (SCL-90) (18) provides a simple way for researchers to obtain a series of quantitative indicators to comprehensively describe an individual's mental health.

Of all the non-verbal cues related to mental health, facial expressions are relatively stable (19) and easy to obtain. Consequently, we used facial prediction models based on SCL-90 to assess the psychological symptoms of mental illnesses. Given that this is a multi-dimensional research, one model should predict the same symptomatic dimension as assessed by the corresponding subscale, meaning that the depression model and the depression subscale should measure the same thing. Existing model evaluation methods, such as accuracy or mean square error, cannot evaluate such convergent validity. Therefore, we applied the assessment method of scales to machine learning models. The development and application of scales are typically accompanied by tests of reliability and validity. Researchers use the correlation between the scores of a certain scale with those of other scales to evaluate the criterion validity, convergent validity, and discriminant validity, and use the correlation between the scores of the two half items in the scale to evaluate the reliability (20, 21). Similarly, we used the correlation between the predicted scores from models and actual scores from scales to calculate validity, and used the correlation between predicted scores from models based on the two halves of the facial data to test reliability.

In summary, we obtained facial movements and SCL-90 scores, built facial prediction models to identify psychological symptoms, and calculated reliability and validity by way of evaluation. The results showed that our method has fair reliability and validity, and revealed the possibility for machine learning models to recognize more detailed aspects of mental health status, not just at the disease level.

MATERIALS AND METHODS

Participants

We recruited participants at a large event in Wuhan in July 2019, most of whom were coach drivers. The exclusion criteria for this study included: (1) participants whose scale scores were all minimum or maximum; (2) participants whose facial data recorded by Kinect were <700 frames. After balancing gender and normalizing the SCL-90 score distribution, 100 participants were included in the final analysis, including 60 males and 40 females.

Instruments

Demographic information. Basic demographic information such as the gender, age, number of children, education level, and marital status of each participant was obtained.

Symptom check list. The SCL-90 (18) is a 90-item self-report scale with responses made on a 5-point Likert scale. It was first used in China in 1984 (22). The SCL-90 assesses mental health status over the past seven days, using 10 subscales reflecting 10 physical and psychological symptoms. Since the SCL-90 assesses a wide range of psychiatric features and can measure multiple physical and psychological symptoms, it has been widely used in the mental health assessment of various groups (23). Due to the limited data collection time available, we chose the six symptomatic dimensions of the SCL-90 which contribute most to people's mental status (24–29), and are also known to affect the non-verbal expression of individuals (30–34). Those dimensions were: interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, and psychoticism. A brief descriptive summary of each of the six symptoms is provided in **Table 1** (35). It is generally believed that when the factor scores of the SCL-90 are >=2, the individual suffers from negative mental health symptoms (factor score = subscale score/number of items). As a result, the threshold of the total score of six symptomatic dimensions was equal to 110 points in this study.

Kinect. Kinect is a low cost, convenient, and reliable depth sensor with an RGB image camera developed by Microsoft. Unlike traditional planar image characteristics, Kinect can record the movement of facial key points in 3D space (36). Therefore,

TABLE 1 | Details of the six dimensions of SCL-90 used in this study.

Dimension	Number of items	Score interval	Dimension description
Interpersonal sensitivity	9	9–45	This scale assesses feelings of insufficient personal abilities and negative expectations of interpersonal interactions.
Depression	13	13–65	This scale assesses a depressed mood, loss of motivation, and suicidal tendency.
Anxiety	10	10–50	This scale reflects symptoms and behaviors related to anxiety.
Hostility	6	6–30	This dimension includes thoughts or behaviors of an emotional state of anger.
Phobic anxiety	7	7–35	This scale refers to persistent anxiety about a specific person, place, object, or sitting posture.
Psychoticism	10	10–50	This dimension provides a graduate rank from mild interpersonal alienation to serious mental illness.

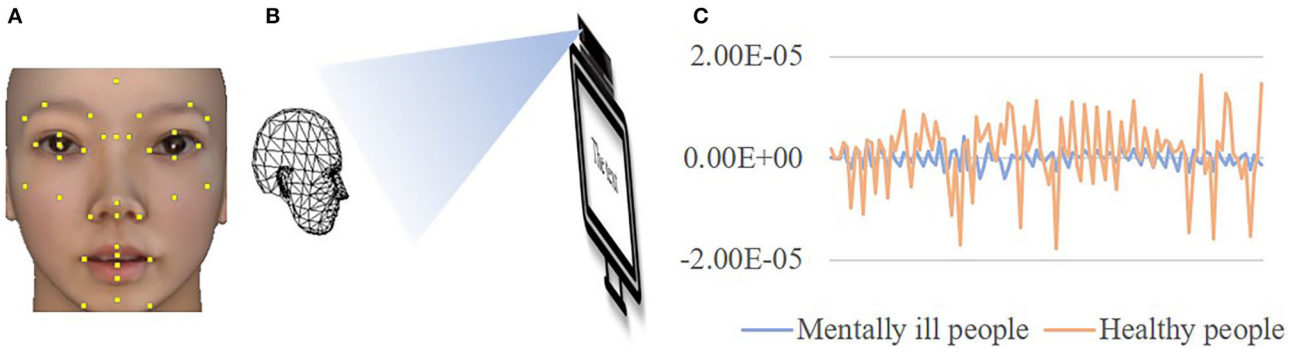


FIGURE 1 | The participant's facial movements were recorded by Kinect. **(A)** The positions of the 36 facial key points on faces. **(B)** A Kinect camera recorded the facial movements of participants when they were reading a neutral text. **(C)** Time-series characteristics were extracted to discover the difference in facial movements between individuals with mental illnesses and those who were healthy. The face image was virtualized.

comprehensive information about facial movements can be extracted. In this study, Kinect was purchased and Kinect for Windows SDK v2.0 was installed to record the 3D coordinates of key points on the face. Kinect can recognize 1,347 key points on the face, and key points near the facial features were considered to be the most closely related to mental illnesses such as depression (37). On this basis, we selected the points near the facial features and the center points of other parts as the key points for identifying mental health symptoms, which totaled 36. The positions on the face are shown in **Figure 1A**.

Procedure

Data collection. Participants were first asked to complete the demographic information questionnaire and the six subscales of the SCL-90. Then they read a neutral text introducing the Macro Polo bridge, during which Kinect was used to record their facial key point locations over approximately 30 s. The frame rate of Kinect is 30 HZ, the resolution of the captured image is 1,920 × 1,080 in color and 512 × 424 in depth (38). The distance between Kinect and the participant's seat was controlled to be 1.5 m to exclude the influence of distance on the intensity of facial movements. Meanwhile, we asked the participants to stay as still as possible in the instruction. The data collection for facial movements (as shown in **Figure 1B**), demographic information, and the SCL-90 were conducted according to the process shown in **Figure 2**.

Data preprocessing. After data collection, the scores of the subscales in the SCL-90 were calculated. For each participant's facial key point coordinate data, data preprocessing was conducted to eliminate the influence of noise. First, for each frame, we translated the origin of the key point coordinates to the position of key point 0 to balance the influence of the head movements. Then, for each frame, the average coordinates of the current frame, the previous frame, and the next frame were used as the coordinates of the current frame to balance the influence of noise. Next, we intercepted the data from the 100th frame to the 700th frame to eliminate the preparation time before and after reading (as seen in **Figure 3A**), which was approximately 20 s. Finally, we conducted a subtraction between the adjacent data in the time-series to obtain the coordinate changes. We named the 100th to 700th frames "whole" data, and the odd 300 frames and even 300 frames in the 600 frames "split-half" data.

Feature extraction. So that facial movements could be expressed as changes in the coordinates of key points, time-series characteristics were used to describe the movements of each key point in 3D space over time. The present study used 30 time-series characteristics as features to extract the motion information of facial key points across the entire time series. The names, types, and meanings of these 30 time-series characteristics are shown in **Table 2**. After feature extraction, we created a feature file, with each row for a participant and each column for a feature. Therefore, the feature file had 3D × 36 key points × 30 time-series characteristics = 3,240 columns. For example, a participant with mental illnesses had 108 (3 * 36) average values for the coordinate changes like the blue line in **Figure 1C**, while a healthy participant had 108 average values for the coordinate changes like the orange line in **Figure 1C**. As we can see in **Figure 1C**, some time-series characteristics can distinguish between individuals with mental illnesses and healthy individuals very well. Regardless of "whole" data or "split-half" data, the same features were extracted.

Feature selection. After extracting 3,240 features for each participant, supervised feature extraction was used to select features that were "important" for each model, which were also features related to the subscale scores. *F*-values were calculated between each feature value for "whole" data and each dimension score. Finally, we selected the 50 features with the largest *F*-value for each model. The points that changed the most with the scores for each subscale are shown in **Figure 4**. It can be seen that the left side of the face expresses more information about mental health status than the right in most symptomatic dimensions of the SCL-90. The rules for selecting features were saved and used in the "split-half" data. After that, all features were standardized to ensure that the contribution of features to models was not affected by range and distribution.

Model training. Based on prior knowledge provided by other studies, the range of nonverbal activities is mostly linear with

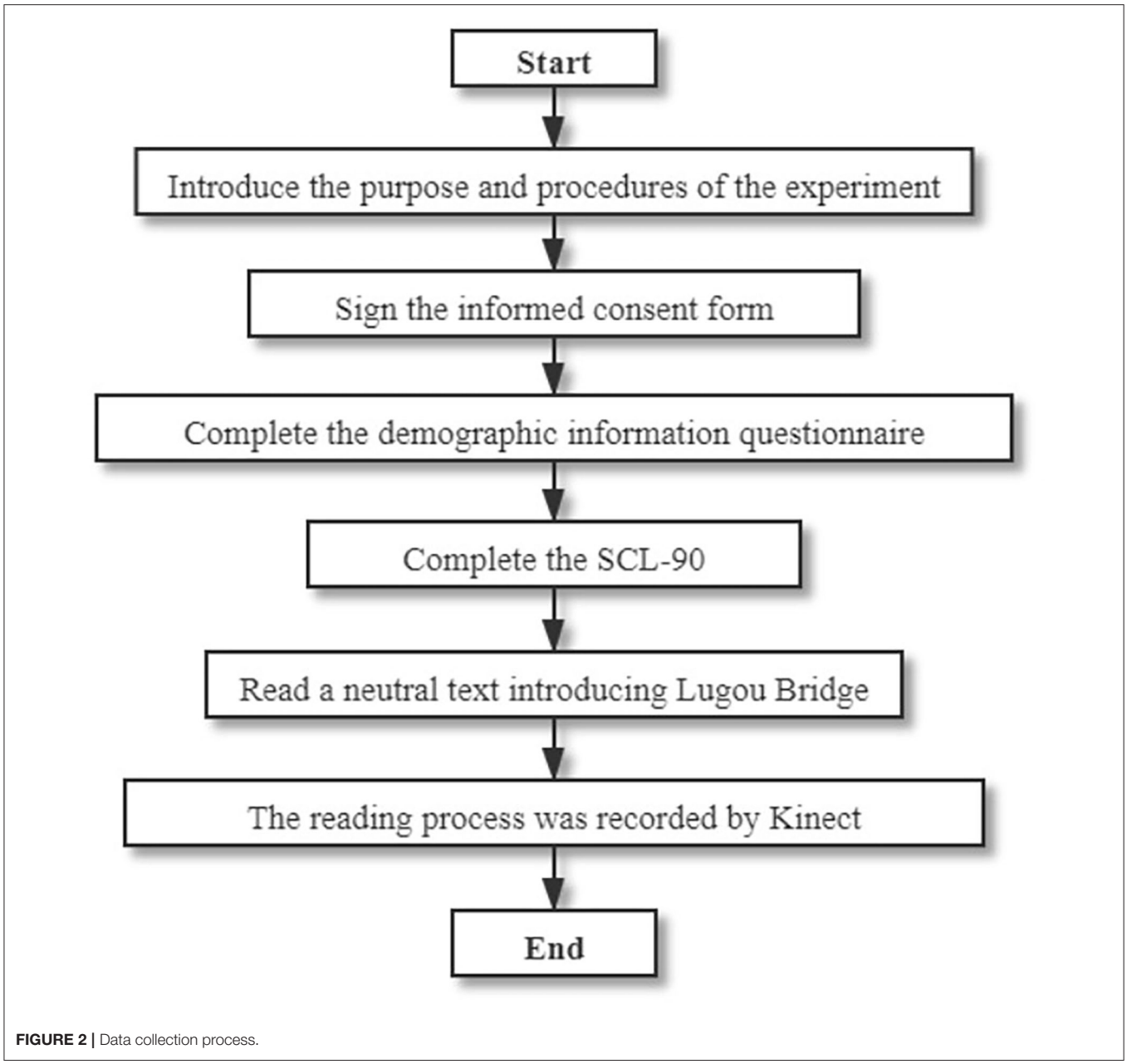


FIGURE 2 | Data collection process.

the degree of mental health (14), so the linear regression model was selected. Because too many features may lead to overfitting, we used L1 regularization to simplify the model. The least absolute shrinkage and selection operator (LASSO) (39) is an optimized technique in linear regression models which uses the L_1-norm penalty. Equation 1 is a general representation of the objective function of LASSO regression, in which y represents the outcomes and x represents the features, N and p are the numbers of samples and variables, and λ and β are the adjustment parameters and regression coefficients. Compared with traditional linear regression models, LASSO regression can enhance the generalization ability of models (40). In this study, LASSO regressions were used to fit the linear relationship between features and subscale scores, and five-fold cross-validation was used to adjust model parameters. After cross-validation, all samples were predicted once as test sets, and the results were saved as predicted values. Similarly, we first used the "whole" data to build the models for each symptomatic dimension and then applied the models to the "split-half" data. The overall process is shown in **Figure 3**. Finally, we obtained three sets of predicted values with a number of 100 based on the "whole" data and "split-half" data.

$$\sum_{i=1}^{N}\left(y_i-\sum_{j}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \quad (1)$$

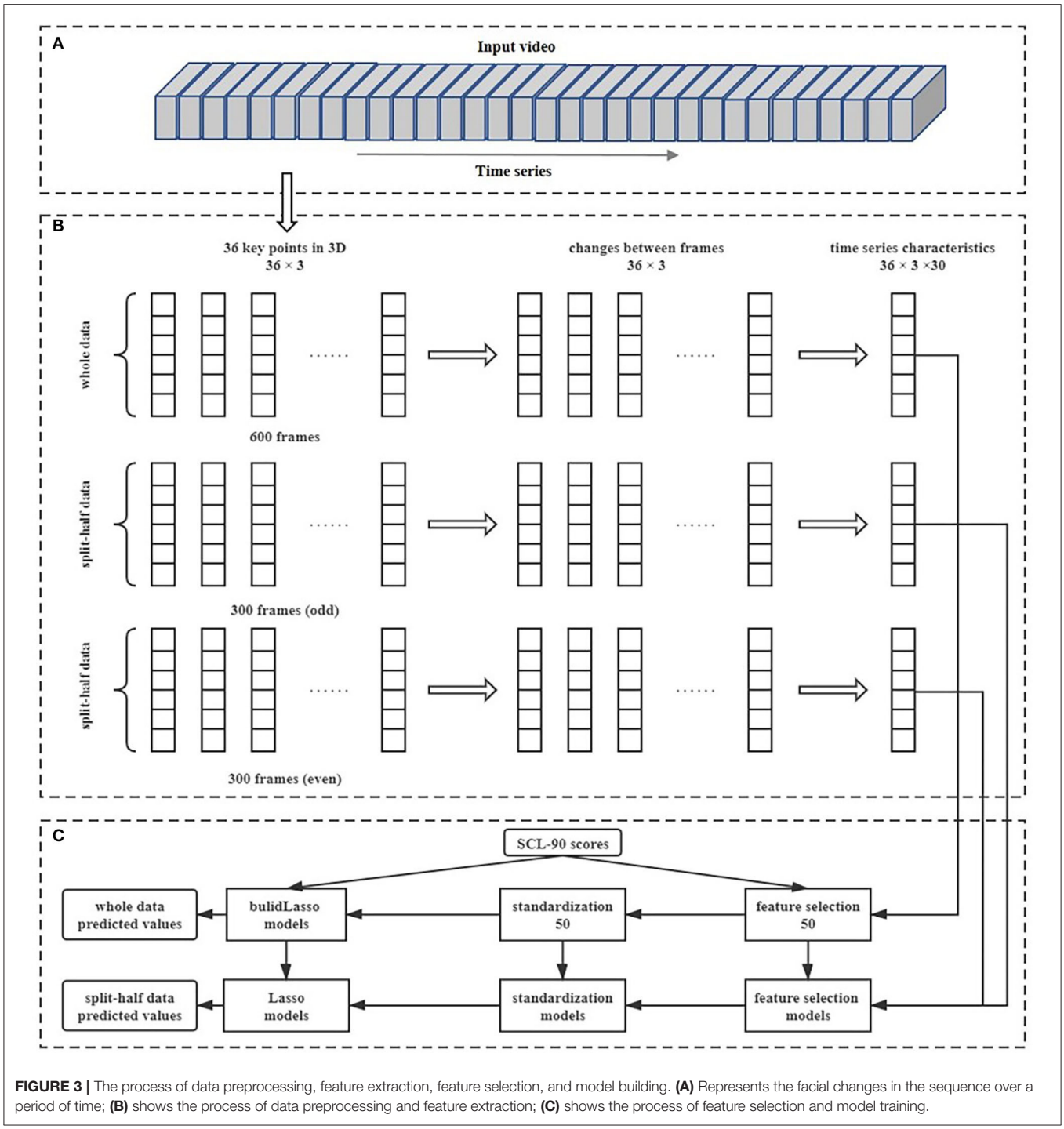


FIGURE 3 | The process of data preprocessing, feature extraction, feature selection, and model building. **(A)** Represents the facial changes in the sequence over a period of time; **(B)** shows the process of data preprocessing and feature extraction; **(C)** shows the process of feature selection and model training.

Statistical Analysis

For descriptive analyses of the quantitative variables, the mean and standard deviation values were calculated. Because of the large sample size and approximate normality distribution, a *t*-test was used to examine the differences in age and the SCL-90 scores between the mentally ill group and the healthy group. For analyses of the qualitative variables, the frequencies were used and chi-square tests were carried out to test differences in marital status, number of children, education level, and gender between the mentally ill group and healthy group. Predicted scores using "whole" data were defined as the predicted values for this method. The predicted scores of the "split-half" data were used as the "split-half" scores. The split-half reliability for each model was assessed with correlation coefficients between the "split-half" scores. Multitrait-multimethod matrix analysis and criterion validity analysis were conducted to test validity.

TABLE 2 | The names, data types, and meanings of the time-series characteristics used in this study.

Name	Type	Meaning
Maximum	Float	The highest value of x
Minimum	Float	The lowest value of x
Mean	Float	The mean value of x
Variance	Float	The variance value of x
Std	Float	The standard deviation of x
Skewness	Float	The sample skewness of x
Kurtosis	Float	The kurtosis of x
Median	Float	The median of x
Absolute energy	Float	The sum over the squared values of x
Absolute sum of changes	Float	The sum over the absolute value of consecutive changes in x
Variance larger than std	Bool	If variance is greater than std
Count above mean	Float	The number of values in x that are higher than then mean of x
Count below mean	Float	The number of values in x that are lower than then mean of x
First location of maximum	Float	The first location of the maximum value of x
Last location of maximum	Float	The relative last location of the maximum value of x
First location of minimum	Float	The first location of the minimum value of x
Last location of minimum	Float	The relative last location of the minimum value of x
Duplicated	Bool	If any value in x occurs more than once
Max duplicated	Bool	If the maximum value of x is observed more than once
Min duplicated	Bool	If the minimum value of x is observed more than once
Longest strike above mean	Float	The longest consecutive subsequence in x that is bigger than the mean of x
Longest strike below mean	Float	The longest consecutive subsequence in x that is smaller than the mean of x
Mean absolute change	Float	The mean over the absolute differences between values in x
Mean change	Float	The mean over the differences between values in x
Percentage of reoccurring datapoints	Float	The percentage of unique values that are present in x more than once
Ratio value number	Float	The percentage of unique values that are present in x only once in all values
Sum of reoccurring datapoints	Float	The sum of all data points, that are present in x more than once
Sum of reoccurring datapoints	Float	The sum of all values, that are present in x more than once
Sum values	Float	The sum of all values
Range	Float	The range value of x

x represents the time-series; std, standard deviation.

TABLE 3 | Distribution of demographic information and SCL-90 scale scores.

	Total	Healthy people (total score < 110)	Mentally ill people (total score >= 110)	*P*-value
Sample size	100	88	12	
Age	40.23 (7.582)	40.51 (7.58)	38.17 (7.66)	0.731
Sex (male)	60	49	11	0.017*
Sex (female)	40	39	1	
Marriage (yes)	87	77	10	0.708
Children (yes)	82	72	10	0.932
Higher education (yes)	57	52	5	0.253
SCL-90	88.13 (24.03)	82.28 (13.91)	131 (37.24)	0.001***
Interpersonal sensitivity	16.03 (4.72)	15.08 (3.52)	23 (6.47)	0.000***
Depression	22.02 (6.558)	20.40 (4.01)	33.92 (9.199)	0.000***
Anxiety	15.19 (4.59)	14.25 (3.00)	22.08 (7.70).	0.000***
Hostility	9.96 (3.45)	9.24 (2.51)	15.25 (4.73)	0.001***
Phobic anxiety	9.24 (2.98)	8.65 (1.90)	13.58 (5.28)	0.008**
Psychoticism	15.69 (4.68)	14.67 (3.01)	23.17 (7.47)	0.002**

*Continuous data are expressed as mean (standard deviation); discrete data are expressed as number. *P < 0.05;**P < 0.01;***P < 0.001.*

RESULTS

Demographic Information

The demographic information of individuals was collected in this study. Participants in this study were middle-aged people with an average age of 40 years, they were mostly married (87%), and had children (82%). The proportion of participants who had received higher education was 57%.

SCL-90 Score

The average value of the total scores of the SCL-90 was 88.13, and the standard deviation value was 24.03. Participants were divided into a "healthy group" ($n = 88$) and a "mentally ill group" ($n = 12$) based on the aforementioned threshold score of 110 points. Although the numbers of healthy subjects and mentally ill subjects are uneven, the data distributions of the total scores and the subscales scores are close to the normal distribution, which has less influence on the regression models. The demographic information was not distinguished between the two groups, except for gender. The scores of the six subscales were significantly different in the two groups, which was in line with expectations (see **Table 3**).

Split-Half Reliability

In this study, the original "whole" data was divided into two parts based on the parity of frames. And the Pearson correlation coefficient between the predicted values of the two split-half data was calculated as an indicator of split-half reliability. The split-half reliability of the six facial

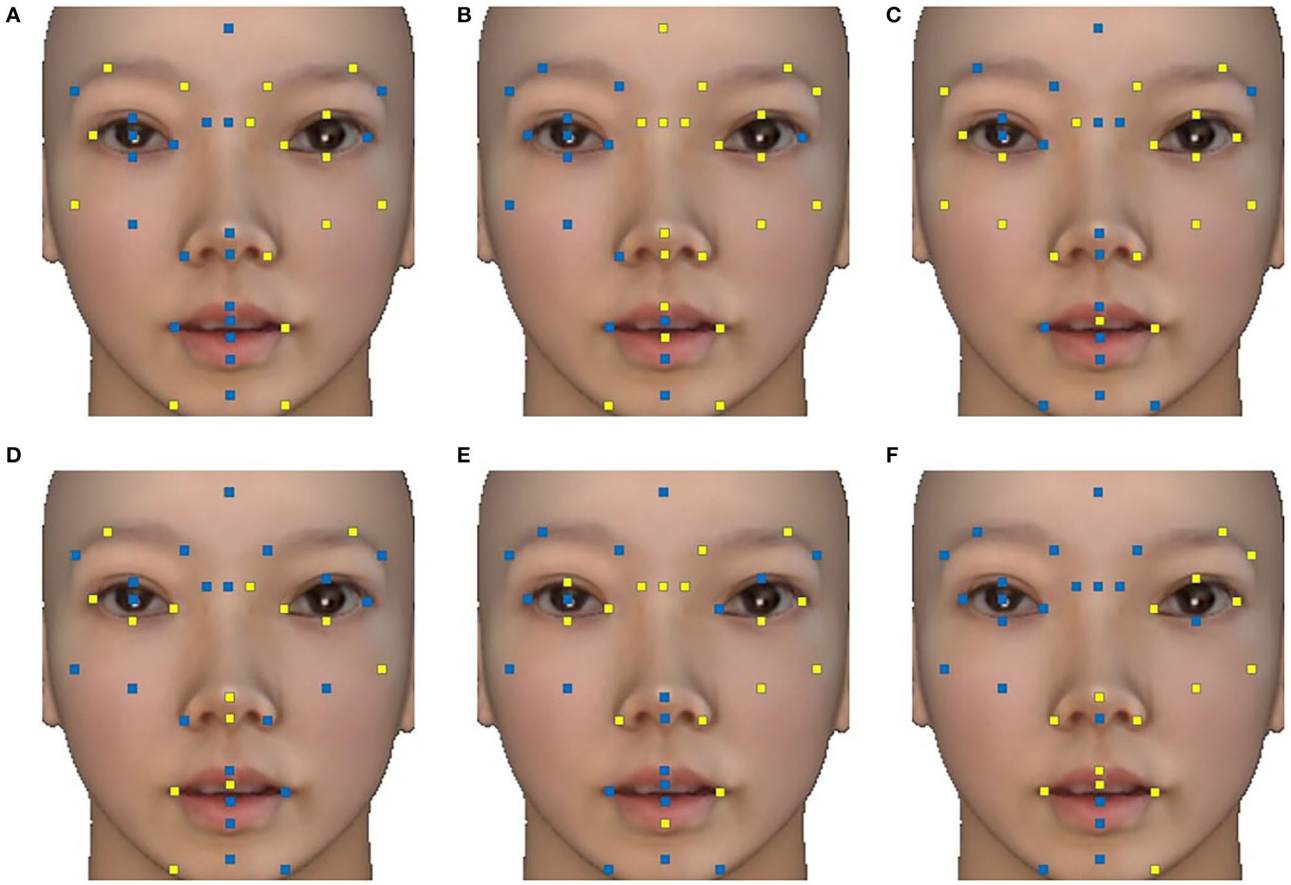


FIGURE 4 | The key points that changed the most with the scale scores in each symptomatic dimension. Key points with at least one time-series characteristic included in the 50 features are marked in blue, and others are marked in yellow. **(A)** Interpersonal sensitivity; **(B)** depression; **(C)** anxiety; **(D)** hostility; **(E)** phobic anxiety; **(F)** psychoticism. The face images were virtualized.

prediction models is shown in **Table 4**, all reaching the significance level.

Convergent Validity and Discriminant Validity

This study used a multitrait-multimethod matrix to explore the structural validity of facial prediction models. Six traits were involved in the multitrait-multimethod matrix, which were interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, and psychoticism; and two methods were involved, including the SCL-90 subscales and facial prediction models. Pearson correlation coefficients were calculated among the predicted values and the SCL-90 scores, and **Table 5** presents the zero-order correlation matrix between variables. In **Table 5**, the bold numbers on the diagonal represent the correlations between different methods measuring the same trait, the numbers in the triangles represent the correlations between different traits measured by the same method, and the numbers in the yellow area represent the correlations between different methods measuring different traits. The results indicated that the bold numbers were significantly larger than the data in the yellow area in the same column, except for the depression dimension, which meant that our models had good convergent validity. However, the bold numbers were not all greater than the corresponding values in the triangles, which meant the discriminant validity of our models was not as good.

Criterion Validity

The actual scores of each subscale were used as the effective standard, and the Pearson correlation coefficients between the predicted values of the "whole" data and the actual scores of the corresponding subscales were calculated, so as to conduct the analysis of criterion validity (as shown in **Table 4**). The results showed that the correlation coefficients had reached a significant level, which meant the models established had high criterion validity.

DISCUSSION

The present study tested the prediction of psychological symptoms based on facial movements. We collected SCL-90 scale scores as the output, and extracted the time-series characteristics of facial key points as the input, then built facial prediction models for each symptomatic dimension. Finally, we tested the stability and availability of the models by calculating the split-half reliability, criterion validity, convergent

validity, and discriminant validity. The results indicated that the facial prediction models proposed have good split-half reliability, criterion validity, and convergent validity, although the discriminant validity is lower.

Consistent with previous research on emotion-induced situations (41, 42), the high criterion validity suggests that under neutral conditions, facial movements can also be used to distinguish patients with mental illness from those who are healthy, especially the facial movements on the left side of the face. This finding is in line with previous studies that found that individuals with some mental illnesses have fewer facial movements than healthy people due to alexithymia (43, 44). An alternative explanation would be that compared with healthy people, people with poorer mental health status are more likely to produce (45) and express (46) negative emotions under neutral stimulation. Although each model had significant criterion validity, it is noteworthy that the depression model and anxiety model had lower criterion validity than the other symptomatic dimensions. Based on previous studies, we speculate that this is because comorbidity with anxiety or depression is common in people with other symptoms (47, 48). Individuals with depression and anxiety may have different subtypes, which leads to different facial movements and results in slightly lower criterion validity. Relevant studies have also pointed out that there are differences in the performance of individuals with multiple symptoms and those with only depression or anxiety (49, 50). One possible explanation for the finding that the left side of the face is more capable of expressing mental health status is that mental illness, such as depression and autism, are mainly dominated by the right hemisphere of the brain (51).

There was also fairly high convergent validity for most models except depression. Specifically, for the interpersonal sensitivity dimension, anxiety dimension, hostility dimension, phobic anxiety dimension, and psychoticism dimension, the correlations between different methods measuring the same traits were higher than all the correlations between different methods measuring different traits, which meant the two methods were measuring the same traits, consistent with our expectations. However, in the depression dimension, we did not find a higher correlation between different methods measuring the same trait, which indicates that the depression dimension may not have a specific facial expression that can be identified, and this is probably related to the complex comorbidity between depression and other negative psychological symptoms (47, 52, 53). Studies have suggested that different types of negative mental health status have different facial movements (54, 55) and the facial expressions associated with mental illness are also different from physical illness (56, 57). Our study suggests the possibility that different psychological symptoms of mental illnesses may have different facial movements that can correspond to the SCL-90 scores, which are detailed and granular. Future study is needed to explore the unique expression of each symptomatic dimension and the underlying neurological mechanisms. In addition, it is understandable that the discriminant validity is low, considering the high correlation (0.3–0.8) between the scores of the various subscales in the SCL-90 (58), and the high correlation (0.2–0.7) between the values of models which are based on scale scores.

TABLE 4 | Split-half reliability and criterion validity of each dimension.

Dimensions	R_1	R_2
Interpersonal sensitivity	0.817***	0.377***
Depression	0.755***	0.261**
Anxiety	0.551***	0.307**
Hostility	0.516***	0.423***
Phobic anxiety	0.674***	0.351***
Psychoticism	0.608***	0.376***

*R_1, the Pearson correlation coefficients between the predicted values of odd frames data and the predicted values of even frames data. R_2, the Pearson correlation coefficients between the predicted values of the "whole" data and the actual score of the dimension scale. **P < 0.01; ***P < 0.001.*

TABLE 5 | Convergent validity and discriminant validity of each dimension.

		Facial prediction models						SCL-90					
		INT_1	DEP_1	ANX_1	HOS_1	PHO_1	PSY_1	INT_2	DEP_2	ANX_2	HOS_2	PHO_2	PSY_2
Facial prediction models	INT_1												
	DEP_1	0.74											
	ANX_1	0.42	0.44										
	HOS_1	0.39	0.27	0.29									
	PHO_1	0.48	0.49	0.43	0.23								
	PSY_1	0.49	0.48	0.50	0.16	0.27							
SCL-90	INT_2	**0.38**	0.29	0.25	0.20	0.19	0.34						
	DEP_2	0.35	**0.26**	0.22	0.29	0.18	0.32	0.73					
	ANX_2	0.27	0.21	**0.31**	0.23	0.22	0.27	0.75	0.81				
	HOS_2	0.23	0.14	0.16	**0.42**	0.08	0.21	0.64	0.77	0.71			
	PHO_2	0.26	0.18	0.24	0.18	**0.35**	0.16	0.64	0.74	0.74	0.63		
	PSY_2	0.33	0.28	0.25	0.23	0.26	**0.38**	0.82	0.82	0.84	0.68	0.72	

The Pearson correlation coefficients between the predicted values of each model and the actual scores of each scale. INT, interpersonal sensitivity; DEP, depression; ANX, anxiety; HOS, hostility; PHO, phobic anxiety; PSY, psychoticism; 1, predicted values by facial prediction models; 2, actual scores by the SCL-90.

In terms of reliability, results indicate good levels of split-half reliability for all the models (from 0.52 to 0.82), which are consistent with the subscale consistency (from 0.50 to 0.90) (59–61) in previous studies examining the SCL-90. The credible split-half reliability suggests that the time-series characteristics we extracted can represent stable personal traits to some extent, rather than random factors. One previous study has explored the stability within individuals and differences between individuals in facial expressions (62). Such differences may relate to mental health status and other individual characteristics, and such stability may be the reason why the machine learning models have good reliability.

Our study indicates that the facial prediction models based on the SCL-90 have good split-half reliability, criterion validity, and convergent validity. As per the literature explored and to the knowledge of the authors, we are the first to measure the reliability and validity of machine learning models. In multi-dimensional studies, measuring the reliability and validity of machine learning models is conducive to ensuring one model can truly discover the pattern of the corresponding symptomatic dimension, which cannot be achieved by previous machine learning evaluation methods.

Our research also provides a feasible method for evaluating the performance of multi-trait machine learning models. The multi-dimensional psychological symptoms of mental health were predicted separately in this study, and most models had satisfactory convergent validity, which presents the possibility of predicting more detailed aspects of mental health through the assessment of facial movements. Furthermore, we tracked the facial movements of participants under neutral stimulation, which is close to the facial state of people during normal communication. Although the current facial prediction models cannot replace scales, existing research could be combined with monitoring technology to achieve large-scale and non-invasive mental health monitoring for appropriate occupations in practical applications.

This study also has some limitations. First, the selection of the machine learning algorithm should ensure that it can match the corresponding dataset. Selecting deep learning algorithms may slightly improve the results, but this is not the focus of this paper. Future studies based on different datasets would be needed to compare the performance of different machine learning models. In addition to regression models, classification prediction models are also of practical significance, as long as the data are balanced. Second, considering the purpose of the research, we used the SCL-90, of which the correlation among the subscales was very high. This results in low discriminant validity. Further work should take into account the comorbidity between symptoms and strive to obtain a unique facial expression for each symptom. Third, as the participants in this study were conveniently sampled at a large-scale event, although age and gender were balanced, the specific occupation of the participants may also cause some sampling bias. Moreover, due to limited time, the three symptoms of somatization, compulsion, and paranoia were not measured, and those symptoms could be explored in further studies. A further limitation may be the influence of participants' knowledge background in self-reporting methods. However, in our data acquisition and application scenarios, self-reporting was the most appropriate method. Future research can try to use the diagnosis of psychiatrists as the annotation data of prediction models. Finally, the criterion validity of the depression and anxiety models was lower compared with other models. Future research can try different data collection scenarios and feature extraction methods to better predict the psychological symptoms with many subtypes.

CONCLUSION

We proposed facial prediction models based on the SCL-90 and demonstrated that the measurement has high reliability and satisfactory validity. Furthermore, this study demonstrated that facial movements can distinguish multi-dimensional psychological symptoms, and provides a feasible method to evaluate the performance of multi-trait machine learning models.

DATA AVAILABILITY STATEMENT

The datasets generated for this article are not readily available because the raw data cannot be made public, if necessary, feature data can be provided. Requests to access the datasets should be directed to liuxiaoqian@psych.ac.cn.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the scientific research ethics committee of the Chinese Academy of Sciences Institute of Psychology (H15010). The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

TZ contributed to the conception and design of the study. XL collected the data and developed the instrument. BL provided guidance for data preprocessing and model establishment. MZ provided guidance for the reliability and validity testing plan. YW performed the statistical analysis. XW trained the facial prediction models and wrote the manuscript with input from all authors. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the Key Research Program of the Chinese Academy of Sciences (No. ZDRW-XH-2019-4).

ACKNOWLEDGMENTS

The authors wish to thank all participants for their participation in this study.

REFERENCES

1. Ohrnberger J, Fichera E, Sutton M. The relationship between physical and mental health: a mediation analysis. *Soc Sci Med.* (2017) 195:42–9. doi: 10.1016/j.socscimed.2017.11.008
2. Marques SC, Pais-Ribeiro JL, Lopez SJ. The role of positive psychology constructs in predicting mental health and academic achievement in children and adolescents: a two-year longitudinal study. *J Happiness Stud.* (2011) 12:1049–62. doi: 10.1007/s10902-010-9244-4
3. Simonton DK, Song AV. Eminence, IQ, Physical and mental health, and achievement domain: Cox's 282 geniuses revisited. *Psychol Sci.* (2009) 20:429–34. doi: 10.1111/j.1467-9280.2009.02313.x
4. Fergusson DM, McLeod GFH, Horwood LJ, Swain NR, Chapple S, Poulton R. Life satisfaction and mental health problems (18 to 35 years). *Psychol Med.* (2015) 45:2427–36. doi: 10.1017/S0033291715000422
5. Hong Q-B, Wu C-H, Su M-H, Huang K-Y. Exploring macroscopic fluctuation of facial expression for mood disorder classification. In: *2018 First Asian Conference on Affective Computing and Intelligent Interaction*. IEEE: Beijing (2018). doi: 10.1109/ACIIAsia.2018.8470337
6. de Melo WC, Granger E, Hadid A, IEEE. Depression detection based on deep distribution learning. In: *2019 IEEE International Conference on Image Processing*. IEEE: Taipei (2019). p. 4544–8. doi: 10.1109/ICIP.2019.8803467
7. Zhou X, Huang P, Liu H, Niu S. Learning content-adaptive feature pooling for facial depression recognition in videos. *Electron Lett.* (2019) 55:648–50. doi: 10.1049/el.2019.0443
8. Drevets WC, Price JL, Furey ML. Brain structural and functional abnormalities in mood disorders: implications for neurocircuitry models of depression. *Brain Struct Funct.* (2008) 213:93–118. doi: 10.1007/s00429-008-0189-x
9. Tian L, Jiang T, Liang M, Zang Y, He Y, Sui M, et al. Enhanced resting-state brain activities in ADHD patients: a fMRI study. *Brain Dev.* (2008) 30:342–8. doi: 10.1016/j.braindev.2007.10.005
10. Froese AP, Cassem NH, Hackett TP, Silverberg EL. Galvanic skin potential as a predictor of mental status, anxiety, depression and denial in acute coronary patients. *J Psychosom Res.* (1975) 19:1–9. doi: 10.1016/0022-3999(75)90044-6
11. Crawford TJ, Higham S, Renvoize T, Patel J, Dale M, Suriya A, et al. Inhibitory control of saccadic eye movements and cognitive impairment in Alzheimer's disease. *Biol Psychiatry.* (2005) 57:1052–60. doi: 10.1016/j.biopsych.2005.01.017
12. Hinchliffe MK, Lancashire M, Roberts FJ. Study of eye contact in depressed and recovered patients. *Br J Psychiatry.* (1971) 119:213–5. doi: 10.1192/bjp.119.549.213
13. Kliper R, Portuguese S, Weinshall D. Prosodic analysis of speech and the underlying mental state. In: Serino S, Matic A, Giakoumis D, Lopez G, Cipresso P, editors. *Pervasive Computing Paradigms for Mental Health*. Milan; Cham: Springer (2016). p. 52-62. doi: 10.1007/978-3-319-32270-4_6
14. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J Neurolinguistics.* (2007) 20:50–64. doi: 10.1016/j.jneuroling.2006.04.001
15. Carmines EG, Zeller RA. Reliability and validity assessment. *Beverly Hills Calif.* (1979) 33:775–80. doi: 10.4135/9781412985642
16. Dimic S, Wildgrube C, Mccabe R, Hassan I, Barnes TRE, Priebe S. Non-verbal behaviour of patients with schizophrenia in medical consultations–a comparison with depressed patients and association with symptom levels. *Psychopathology.* (2010) 43:216–22. doi: 10.1159/000313519
17. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investig Otolaryngol.* (2020) 5:96–116. doi: 10.1002/lio2.354
18. Derogatis LR, Lipman RS, Covi L. SCL-90: an outpatient psychiatric rating scale–preliminary report. *Psychopharmacol Bull.* (1973) 9:13–28.
19. Harrigan JA, Wilson K, Rosenthal R. Detecting state and trait anxiety from auditory and visual cues: a meta-analysis. *Pers Soc Psychol Bull.* (2004) 30:56–66. doi: 10.1177/0146167203258844
20. Vallejo MA, Jordan CM, Diaz MI, Comeche MI, Ortega J. Psychological assessment via the Internet: a reliability and validity study of online (vs paper-and-pencil) versions of the General Health Questionnaire-28 (GHQ-28) and the Symptoms Check-List-90-Revised (SCL-90-R). *J Med Internet Res.* (2007) 9:538. doi: 10.2196/jmir.9.1.e2
21. Zheng YP, Zhao JP, Phillips M, Liu JB, Cai MF, Sun SQ, et al. Validity and reliability of the Chinese Hamilton depression rating-scale. *Br J Psychiatry.* (1988) 152:660–4. doi: 10.1192/bjp.152.5.660
22. Zhengyu W. The symptom checklist 90 (SCL-90). *Shanghai Arch Psychiatry.* (1984) 2:68–70.
23. Wang J. Problems in application of SCL-90 in China. *Chin Mental Health J.* (2004) 18:51–2.
24. Bandelow B, Lichte T, Rudolf S, Jörg Wiltink, Beutel ME. The diagnosis of and treatment recommendations for anxiety disorders. *Deutsches Ärzteblatt Int.* (2014) 111:473–80. doi: 10.3238/arztebl.2014.0473
25. Bierman AS, Ellis BH, Drachman D. Depressed mood and mental health among elderly Medicare managed care enrollees. *Health Care Financ Rev.* (2006) 27:123–36. Available online at: https://europepmc.org/article/med/17290662
26. Erdner A, Magnusson A, Nystrom M, Lutzen K. Social and existential alienation experienced by people with long-term mental illness. *Scand J Caring Ences.* (2010) 19:373–80. doi: 10.1111/j.1471-6712.2005.00364.x
27. Farina A, Ring K. The influence of perceived mental illness on interpersonal relations. *J Abnorm Psychol.* (1965) 70:47–51. doi: 10.1037/h0021637
28. Iliffe S. The role of the GP in managing mental illness in later life. *Psychiatry-Int Biol Proc.* (2008) 4:85–8. doi: 10.1383/psyt.4.2.85.59104
29. Madathumkovilakath NB, Kizhakkeppattu S, Thekekunnath S, Kazhungil F. Coping strategies of caregivers towards aggressive behaviors of persons with severe mental illness. *Asian J Psychiatry.* (2018) 35:29–33. doi: 10.1016/j.ajp.2018.04.032
30. Duijndam S, Kupper N, Denollet J, Karreman A. Social inhibition and approach-avoidance tendencies towards facial expressions. *Acta Psychologica.* (2020) 209:103–41. doi: 10.1016/j.actpsy.2020.103141
31. Waxer P. Nonverbal cues for depression. *J Abnorm Psychol.* (2020) 209:103–41. doi: 10.1037/h0036706
32. Bruene M, Sonntag C, Abdel-Hamid M, Lehmkaemper C, Juckel G, Troisi A. Nonverbal behavior during standardized interviews in patients with schizophrenia spectrum disorders. *J Nerv Ment Dis.* (2008) 196:282–8. doi: 10.1097/NMD.0b013e31816a4922
33. Cunningham CE, McHolm AE, Boyle MH. Social phobia, anxiety, oppositional behavior, social skills, and self-concept in children with specific selective mutism, generalized selective mutism, and community controls. *Eur Child Adolesc Psychiatry.* (2006) 15:245–55. doi: 10.1007/s00787-006-0529-4
34. Newton DA, Burgoon JK. Nonverbal Conflict Behaviors: Functions, Strategies, and Tactics. New York, NY: Taylor & Francis Group (1990).
35. Derogatis LR, Cleary PA. Factorial invariance across gender for primary symptom dimensions of SCL-90. *Br J Soc Clin Psychol.* (1977) 16:347–56. doi: 10.1111/j.2044-8260.1977.tb00241.x
36. Wang Y, Liu Y, Assoc Comp M. Research on facial expression recognition based on kinect. In: *Proceedings of 2017 Vi International Conference on Network, Communication and Computing.* (2017). p. 29–33doi: 10.1145/3171592.3171639
37. Wang Q, Yang H, Yu Y. Facial expression video analysis for depression detection in Chinese patients. *J Visual Commun Image Represent.* (2018) 57:228–33. doi: 10.1016/j.jvcir.2018.11.003
38. Roland Smeenk. *Kinect V1 and Kinect V2 Fields of View Compared*. (2014). Available online at: https://smeenk.com/kinect-field-of-view-comparison (accessed November 15, 2020).
39. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc.* (1996) 58:267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x
40. Xu H, Caramanis C, Mannor S. Sparse algorithms are not stable: a no-free-lunch theorem. *IEEE Trans Pattern Anal Mach Intell.* (2012) 34:187–93. doi: 10.1109/TPAMI.2011.177
41. Barzilay R, Israel N, Krivoy A, Sagy R, Kamhi-Nesher S, Loebstein O, et al. Predicting affect classification in mental status examination using machine learning face action recognition system: a pilot study in schizophrenia patients. *Front Psychiatry.* (2019) 10:288. doi: 10.3389/fpsyt.2019.00288
42. Schwartz GE, Fair PL, Salt P, Mandel MR, Klerman GL. Facial expression and imagery in depression - electromyographic study. *Psychosom Med.* (1976) 38:337–47. doi: 10.1097/00006842-197609000-00006

43. Hesse C, Floyd K. Affectionate experience mediates the effects of alexithymia on mental health and interpersonal relationships. *J Soc Pers Relat.* (2008) 25:793–810. doi: 10.1177/0265407508096696
44. Trevisan DA, Bowering M, Birmingham E. Alexithymia, but not autism spectrum disorder, may be related to the production of emotional facial expressions. *Mol Autism.* (2016) 7:46. doi: 10.1186/s13229-016-0108-6
45. Bertocci MA, Bebko GM, Mullin BC, Langenecker SA, Ladouceur CD, Almeida JRC, et al. Abnormal anterior cingulate cortical activity during emotional n-back task performance distinguishes bipolar from unipolar depressed females. *Psychol Med.* (2012) 42:1417–28. doi: 10.1017/S003329171100242X
46. Scoralick FM, Piazzolla LP, Camargos EF, Dias Freitas MP, Guimaraes RM, Laiana C. Facial expression may indicate depression in older adults. *J Am Geriatr Soc.* (2012) 60:2371–3. doi: 10.1111/j.1532-5415.2012.04169.x
47. Bitsika V, Sharpley CF. Comorbidity of anxiety-depression among Australian university students: implications for student counsellors. *Br J Guid Counc.* (2012) 40:385–94. doi: 10.1080/03069885.2012.701271
48. Essau CA. Comorbidity of anxiety disorders in adolescents. *Depress Anxiety.* (2003) 18:1–6. doi: 10.1002/da.10107
49. Wölfling K, Beutel ME, Koch A, Dickenhorst U, Müller KW. Comorbid internet addiction in male clients of inpatient addiction rehabilitation centers: psychiatric symptoms and mental comorbidity. *J Nerv Ment Dis.* (2013) 201:934–40. doi: 10.1097/NMD.0000000000000035
50. Gilboa-Schechtman E, Foa E, Vaknin Y, Marom S, Hermesh H. Interpersonal sensitivity and response bias in social phobia and depression: labeling emotional expressions. *Cognit Ther Res.* (2008) 32:605–18. doi: 10.1007/s10608-008-9208-8
51. Wylie KP, Tregellas JR, Bear JJ, Legget KT. Autism spectrum disorder symptoms are associated with connectivity between large-scale neural networks and brain regions involved in social processing. *J Autism Dev Disord.* (2020) 50:2765–78. doi: 10.1007/s10803-020-04383-w
52. Starr LR, Hammen C, Connolly NP, Brennan PA. Does relational dysfunction mediate the association between anxiety disorders and later depression? Testing an interpersonal model of comorbidity. *Depress Anxiety.* (2014) 31:77–86. doi: 10.1002/da.22172
53. Sullivan PF, Joyce PR, Mulder RT. Borderline personality-disorder in major depression. *J Nerv Ment Dis.* (1994) 182:508–16. doi: 10.1097/00005053-199409000-00006
54. Diler RS, de Almeida JRC, Ladouceur C, Birmaher B, Axelson D, Phillips M. Neural activity to intense positive vs. negative stimuli can help differentiate bipolar disorder from unipolar major depressive disorder in depressed adolescents: a pilot fMRI study. *Psychiatry Res-Neuroimaging.* (2013) 214:277–84. doi: 10.1016/j.pscychresns.2013.06.013
55. Gavrilescu M, Vizireanu N. Predicting depression, anxiety, and stress levels from videos using the facial action coding system. *Sensors.* (2019) 19:3693. doi: 10.3390/s19173693
56. Esposito G, Venuti P, Bornstein MH. Assessment of distress in young children: a comparison of autistic disorder, developmental delay, and typical development. *Res Autism Spectr Disord.* (2011) 5:1510–6. doi: 10.1016/j.rasd.2011.02.013
57. Yirmiya N, Kasari C, Sigman M, Mundy P. Facial Expressions of affect in autistic, mentally retarded and normal children. *J Child Psychol Psychiatry.* (1989) 30:725–35. doi: 10.1111/j.1469-7610.1989.tb00785.x
58. Chen S, Li L. Re-testing reliability, validity and norm applicatility of SCL-90. *Chin J Nerv Mental Dis.* (2019) 10:493.
59. Martinez S, Stillerman L, Waldo M. Reliability and validity of the SCL-90-R with Hispanic college students. *Hisp J Behav Sci.* (2005) 27:254–64. doi: 10.1177/0739986305274911
60. Oscar Sanchez R, Daniel Ledesma R. Psychometric properties of the symptom checklist revised (SCL-90-R) in clinical population. *Revista Argentina De Clinica Psicologica.* (2009) 18:265–74.
61. Tomioka M, Shimura M, Hidaka M, Kubo C. The reliability and validity of a Japanese version of symptom checklist 90 revised. *BioPsychoSocial Med.* (2008) 2:19. doi: 10.1186/1751-0759-2-19
62. Cohn J, Schmidt K, Gross R, Ekman P. Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification. In: *Proceedings of the International Conference on Multimodal User Interfaces*. Pittsburgh (2002).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 18 December 2020
doi: 10.3389/fpsyt.2020.584711

Monitoring Changes in Depression Severity Using Wearable and Mobile Sensors

Paola Pedrelli[1†], Szymon Fedor[2†], Asma Ghandeharioun[2], Esther Howe[3], Dawn F. Ionescu[4], Darian Bhathena[2], Lauren B. Fisher[1], Cristina Cusin[1], Maren Nyer[1], Albert Yeung[1], Lisa Sangermano[1], David Mischoulon[1], Johnathan E. Alpert[5] and Rosalind W. Picard[2]*

[1] The Depression Clinical and Research Program, Massachusetts General Hospital, Boston, MA, United States, [2] The Media Lab, Massachusetts Institute of Technology, Cambridge, MA, United States, [3] Department of Psychology, University of California, Berkeley, Berkeley, CA, United States, [4] Janssen Research and Development, San Diego, CA, United States, [5] Department of Psychiatry and Behavioral Sciences, Montefiore Medical Center and Albert Einstein College of Medicine, Bronx, NY, United States

OPEN ACCESS

Edited by:
Jennifer H. Barnett,
Cambridge Cognition,
United Kingdom

Reviewed by:
Ryan S. McGinnis,
University of Vermont, United States
Hiroshi Kunugi,
National Institute of Neuroscience,
Japan

***Correspondence:**
Paola Pedrelli
ppedrelli@mgh.harvard.edu

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: 17 July 2020
Accepted: 13 November 2020
Published: 18 December 2020

Citation:
Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhathena D, Fisher LB, Cusin C, Nyer M, Yeung A, Sangermano L, Mischoulon D, Alpert JE and Picard RW (2020) Monitoring Changes in Depression Severity Using Wearable and Mobile Sensors. Front. Psychiatry 11:584711. doi: 10.3389/fpsyt.2020.584711

Background: While preliminary evidence suggests that sensors may be employed to detect presence of low mood it is still unclear whether they can be leveraged for measuring depression symptom severity. This study evaluates the feasibility and performance of assessing depressive symptom severity by using behavioral and physiological features obtained from wristband and smartphone sensors.

Method: Participants were thirty-one individuals with Major Depressive Disorder (MDD). The protocol included 8 weeks of behavioral and physiological monitoring through smartphone and wristband sensors and six in-person clinical interviews during which depression was assessed with the 17-item Hamilton Depression Rating Scale (HDRS-17).

Results: Participants wore the right and left wrist sensors 92 and 94% of the time respectively. Three machine-learning models estimating depressive symptom severity were developed–one combining features from smartphone and wearable sensors, one including only features from the smartphones, and one including features from wrist sensors–and evaluated in two different scenarios. Correlations between the models' estimate of HDRS scores and clinician-rated HDRS ranged from moderate to high (0.46 [CI: 0.42, 0.74] to 0.7 [CI: 0.66, 0.74]) and had moderate accuracy with Mean Absolute Error ranging between 3.88 ± 0.18 and 4.74 ± 1.24. The time-split scenario of the model including only features from the smartphones performed the best. The ten most predictive features in the model combining physiological and mobile features were related to mobile phone engagement, activity level, skin conductance, and heart rate variability.

Conclusion: Monitoring of MDD patients through smartphones and wrist sensors following a clinician-rated HDRS assessment is feasible and may provide an estimate of changes in depressive symptom severity. Future studies should further examine the best features to estimate depressive symptoms and strategies to further enhance accuracy.

Keywords: artificial intelligence, sensors, digital phenotyping, assessment, depression

INTRODUCTION

Depression is among the most common and disabling mental health disorders, with a worldwide prevalence of more than 300 million people (1). Despite the availability of many evidence-based treatments for Major Depressive Disorder (MDD), ∼50% of US individuals with depression are not treated (2) and only 26% of those with past year MDD receive adequate treatment (3). Barriers to treatment include stigma, limited access to specialty care, poor symptom monitoring, and limited patient insight into symptoms (4). Due to the low availability of specialized care for depression, the disorder is often identified and managed in primary care settings (5, 6). However, the time constraints of primary care physicians (PCPs) make thorough symptom monitoring difficult, which may in turn contribute to inadequate or delayed treatment (7). In the absence of resources for close symptom monitoring, many PCPs follow the American Psychiatric Association's (APA) Practice Guideline for the Treatment of Patients with MDD and wait up to 12 weeks before adjusting medications in the absence of a response (8). Delaying time to medication change may prolong suffering, reduce the likelihood of complete remission (9), and increase risk for relapse (10). Further, a high percentage of patients who experience full remission, will experience a relapse (10, 11), the treatment for which is frequently delayed due to providers' expectation that remitted patients will contact them when deteriorating. Methods enabling passive, real-time symptom monitoring may facilitate early detection of response or non-response to treatment, or of depression relapse and allow expedited delivery of relief to patients.

Over the past decade, the development of wristband and smartphone-embedded sensors has facilitated the objective measurement of numerous hallmark symptoms of depression and the passive monitoring of behavioral indices of low mood (12). Consistent with the well-established association between low levels of socialization and depressive symptoms (13), recent work shows that severe depressive symptoms are associated with shorter duration of incoming and outgoing calls, and fewer incoming and outgoing phone calls and daily text messages (13–18). Anhedonia and low energy level can manifest as reduced physical activity (19, 20), which can be measured using GPS and motion sensors embedded in smartphones or wearable sensors. It has also been documented that more severe depressive symptoms and worse mood levels are negatively correlated with a higher amount of time the phone screen is on, a higher number of entertainment applications (apps) used, and an overall time of interaction with the smartphone (18, 21). Finally, dysregulated sleep, a common problem in depression, can be measured relatively well by wearable activity sensors (22).

Preliminary studies suggest that aggregates of smartphone-based passive features are useful in predicting daily mood (23) and presence of depressive symptoms (24). While findings in the field of sensor-based assessment in psychiatry are promising, critical gaps are still present. Most studies evaluating passive monitoring have examined depressive symptoms among patients with bipolar disorders. Those studies conducted with patients with depression have mostly relied on self-report questionnaires to assign diagnoses and/or measure variation in symptoms and train the models. This has resulted in possible reliability problems, since these investigations have primarily focused on detecting presence or absence of depression rather than on assessing its severity, and they have shown overall low accuracy (14–18, 23–27). Moreover, despite evidence showing the existence of physiological indices that are markers of depressed mood (28), that can be continuously collected through wearables (29), and which can be combined with smartphone data to assess mood (27), only two studies have combined these data streams in models to monitor changing depressive symptoms (26, 27). However, both of these studies included previously described limitations such as reliance on self-report questionnaires to train their models and low accuracy. Only one relied on exclusive passive data collection (27).

The current study aimed to address these gaps by evaluating the feasibility and performance of using a machine-learning model that combines physiological features passively recorded by wearable sensors and smartphone features to assess depressive symptoms severity among patients diagnosed with Major Depressive Disorder. Models estimating depressive symptom severity only from smartphone features and only from wristband-based features, were also evaluated. Finally, we examined which features from the smartphone and wearable sensors were most informative in estimating depressive symptom severity. Based on previous reports, we hypothesized a strong correlation between estimates of depressive symptom severity from the model combining smartphone and wearable features and clinician-rated depressive symptom scores.

MATERIALS AND METHODS

Participants

Participants were recruited using standard methods (e.g., flyers). Forty-one participants with MDD were enrolled and 31 completed the study. Seven participants dropped out after the screening visit, two after visit three, and one after visit two. Participants were between the ages of 19 and 73 (M = 33.7, SD = 14), and primarily female = 23 (74%). Ethnic distribution was White = 22 (71%), Hispanic/Latino = 4 (23%), Asian = 5 (16%), Haitian/Black/African-American = 4 (12%), American Indian/Alaskan = 1 (3%), mixed-race = 2 (6%), and other = 1 (3%). At screening, participants on average had severe depressive symptoms [28-item Hamilton Depression Rating Scale (30) = 22.74; SD = 7.38].

Procedure

The study encompassed six in-person visits, daily smartphone-delivered surveys, and passive assessment over 9 weeks. The first screening visit included an informed consent procedure, a blood test to assess for potential medical contributors to depressed mood, and clinician-rated symptom assessment. During the second visit (baseline visit) the monitoring app was downloaded onto participants' phones, wristband sensors were applied, and in-person clinical assessments and self-report measures were completed. The remaining four clinical visits occurred bi-weekly over 8 weeks.

Inclusion criteria were current MDD (per the DSM-IV) (31), Hamilton Depression Rating Scale (HDRS-28) (30) score of > 18 at screening, measurable electrodermal activity, use of an Android smartphone as a primary device, ownership of a computer or tablet compatible with the wristband sensors, and daily internet access. Exclusionary criteria were drug or alcohol use disorder within the past 3 months, lifetime history of psychosis, mania, hypomania, epilepsy, or seizure disorder, current untreated hypothyroidism, unstable medical condition or cognitive impairment, acute suicide or homicide risk, current electroconvulsive therapy treatment, vagal nerve stimulation therapy, deep brain stimulation, transcranial magnetic stimulation therapy, or phototherapy, concurrent participation in other research studies involving investigational agents or blinded randomization to treatment, and inability to comprehend and communicate in English.

The protocol was approved by the Institutional Review Boards of Massachusetts General Hospital and Massachusetts Institute of Technology.

Measures

Clinician-Administered

Mini International Neuropsychiatric Interview (MINI): The MINI, a structured diagnostic interview for major psychiatric disorders, was administered during the screening visit to determine current MDD and rule out exclusionary diagnoses (32).

Hamilton Depression Rating Scale (HDRS): The HDRS-28 is a 28-item clinician-rated assessment scale to assess depressive symptoms (30). The HDRS was administered six times: during the screening visit, 1 week later during the baseline visit, and every other week from visit third to sixth. The HDRS-17 (33), one of the methods most commonly employed to measures change in depressive symptoms severity in treatment for depression clinical trials, was derived from the HDRS-28, and used as dependent variable. The HDRS-28 was administered by clinical staff at the Depression Clinical Research Program (DCRP). Staff at the DCRP has been extensively trained in the use of the HDRS by using videotapes and live interviews of patients. Recent assessment of inter-rater reliability between DCRP clinicians in diagnosing MDD and measuring severity of depression has yielded kappas > 0.75, indicating satisfactory agreement (based on internal data).

Sensors

Participants were instructed to wear two E4 Empatica (34) wristbands, one on each wrist, for 22 h a day/7 day a week (with 1 h/day for charging and 1 h/day to upload data). Participants could upload the data at any point of the day. The E4 measures electrodermal activity (EDA), peripheral skin temperature, heart rate (HR), motion from the 3-axis accelerometer and sleep characteristics using actigraphy.

Smartphone Sensor Data

Mobile-based social interactions (e.g., number of calls, texts), activity patterns (e.g., still, walking), and number of apps used were tracked through the phone app MovisensXS (35) that was downloaded onto participants' phones. For privacy reasons, no phone usage details were recorded, including content of calls/texts, app names, and internet use content. MovisensXS has been successfully used to securely and confidentially collect behavioral and self-reported mood data (36).

Data Analyses

Data analyses included evaluation of the acceptability of the E4 devices and of the performance of a model using features from smartphones and wristband sensors to estimate depressive symptoms severity assessed with the HDRS. Acceptability was evaluated by examining adherence of wearing the E4. We processed adherence for the entire study, and also after excluding the days when the data were missing because of technical problems including E4 sensors malfunctioning, problems with uploading the data to the server, or problems with the network connection. While most technical problems were promptly addressed, some resulted in the loss of data from multiple consecutive days because it took time to substitute the faulty sensors or get support from the technology providers.

The model was developed through several steps as shown in **Figure 1**. All features were preprocessed, transformed and calculated for four 6-h intervals and for daily aggregates (over the course of 24-h). We built upon and expanded our previous work (37) and encoded a comprehensive list of physiological and behavioral features including EDA, motion, sleep, phone usage, call and messaging behavior, app usage, and location change patterns (see comprehensive list in **Supplementary Table 1**). Preprocessing of EDA, motion, and sleep features was improved relative to our previous work by adding EDA features calculated during the time when there is no motion (identified by the accelerometer sensor) and by adding normalized EDA features (see additional information in Section A1 in **Supplementary Material**). Moreover, a location preprocessing step was included by down-sampling location data-points to one recording per 5 min followed by extrapolating missing location latitude and longitude values. Consistent with other investigators (18), more semantic features based on location: time spent at home, transition time, total distance traveled, and weighted stationary latitude and longitude standard deviation (A.3. in **Supplementary Material**) were added. Given that mood can be influenced by the weather (38, 39), location was used to retrieve historical weather data from the DarkSky API (40) and features related to temperature, pressure, humidity, sunrise and sunset time, cloud cover and wind were included. The final dataset included 877 features (**Supplementary Table 1**), of which 404 derived from the wearables and 473 from the mobile sensors. Similar to our previous work (37), we addressed the potential problem of overfitting by reducing the dimensionality of features using linear and non-linear transformations (see more details in A.2. in **Supplementary Material**). The resulting 25 transformed features are difficult to interpret as they are non-linearly derived from the original features. Hence, we used the Boruta algorithm described below to identify the most informative features to estimate HDRS scores.

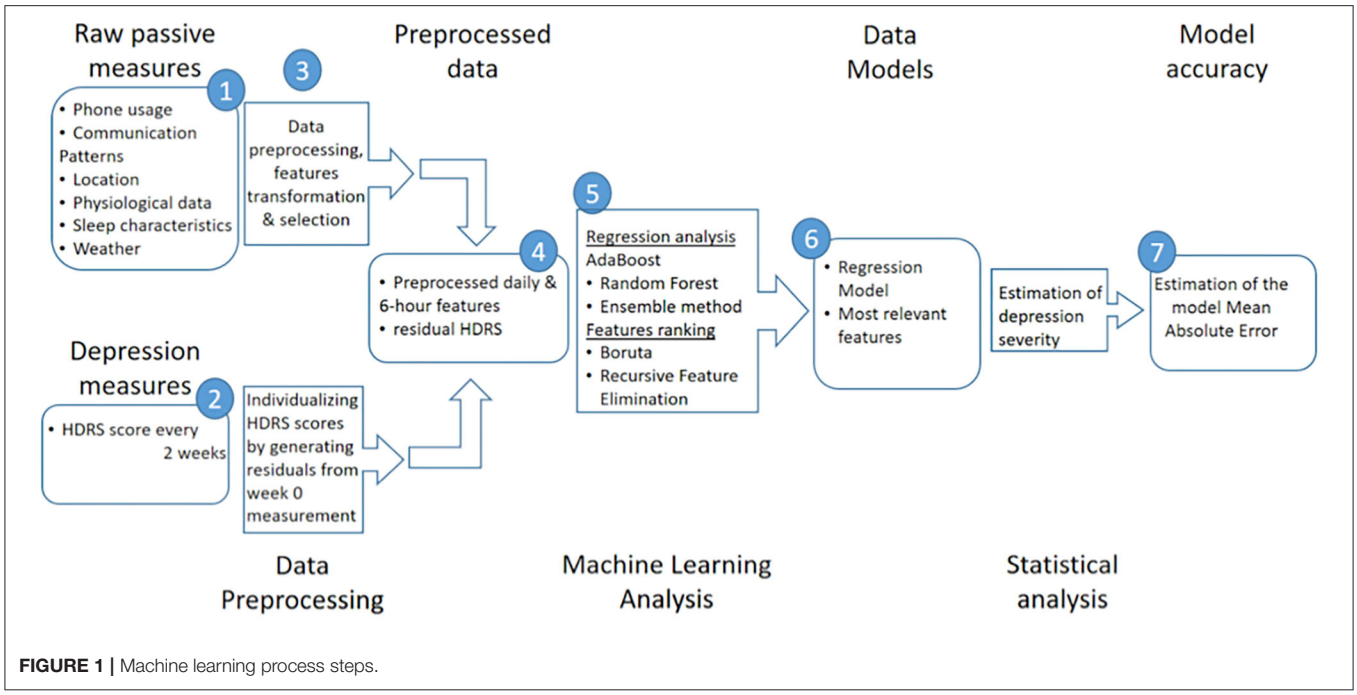


FIGURE 1 | Machine learning process steps.

Personalization

The HDRS scores included in the training and test model were the residual values obtained subtracting HDRS score of the screening visit from the HDRS scores collected during later visits (visits 2, 3, 4, 5, and 6). Residualized HDRS scores allowed to account for the heterogeneous presentation of depressive symptoms.

Multimodal Model Training and Estimation of Depressive Symptom Severity

The model to estimate residualized HDRS scores was built by using features from the wrist and smartphone sensors from the same day (midnight-to-midnight) of the HDRS administration. Machine learning techniques used to build the model were average ensemble of boosting (41) and random forest (42). To avoid overfitting, multiple dimensionality reduction and feature transformation techniques were applied to the raw features and we performed 10-fold-cross-validation (41) during training (see more details in A.2. in **Supplementary Material**); After the parameters of the model were learnt, the model was tested on a separate hold-out test set of data. The model was evaluated under two deployment scenarios resembling different clinical settings (**Figure 2**). Specifically, the data were split into training and test sets differently and, subsequently each resulting model had slightly different parameters: (1) In the *user-split scenario*, a set of 20% of participants were randomly selected as a hold-out test set and the remaining 80% of participants provided a training set. With this evaluation methodology, the performance of the model was assessed as if it were trained on specific clinic patients or a population, and then used to estimate depressive symptom variation among other patients. (2) In the *time-split scenario*, the first three HDRS scores following the screening visit from all participants were pooled along with the first HDRS score to form the training set and the remaining two HDRS scores were pooled to form the hold-out test set. With this evaluation methodology, the performance of the model was assessed as if it were trained on three visits from one patient, and then used to estimate future depressive symptom variations for the same patient.

In both scenarios HDRS scores were residualized on the HDRS from the screening visit. The performance of the model on the hold-out test set in each scenario was expressed as the mean absolute error (MAE), or the average absolute difference between the clinician-based HDRS and the model-estimated HDRS and as the root mean square error (RMSE). Per standard procedure (42), model performances in each scenario were compared to the MAE and RMSE of estimates based on: (1) group median HDRS values, (2) individual HDRS values at the screening visit, and (3) individual median value of the HDRS from the three visits following the screening visits (this was possible only in the time split scenario). Despite the simplicity of these estimates, most previous work has not outperformed individual baselines in similar settings (42). Correlations were also conducted between the clinician-based HDRS and the HDRS estimated in the testing set of the two scenarios (A.3. in **Supplementary Material**). Two additional models were developed following the procedure described above, one including only features from the wearables and one including only features from the smartphones, and their performance were compared with the model combining all the features.

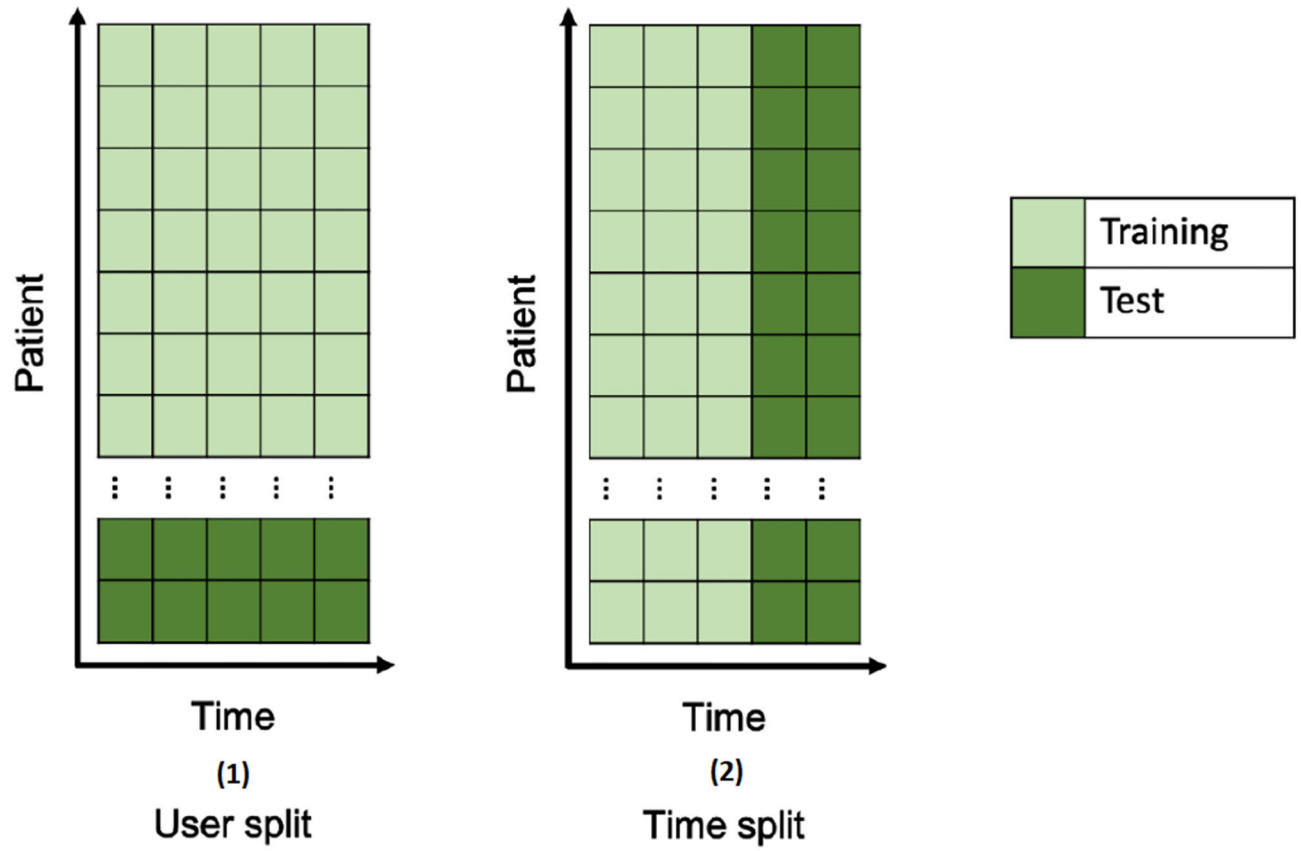


FIGURE 2 | Two evaluation scenarios for splitting the data into training and test sets.

Features Ranking

To identify the most important and non-redundant features for the estimation of the HDRS scores from passive data, the 877 features included in the regression models were examined using the Boruta algorithm (43) which uses the wrapper method around the Random Forest algorithm.

RESULTS

Acceptability

On average, participants uploaded about 17 and 15.5 h of daily data from left and right-hand sensors respectively. This corresponds to 77 and 70% adherence considering that participants were asked to use 2-h each day to charge the E4 and to upload data, which led to a maximum of 22 h of data per day. The average adherence increased to 94% and 92% from the left and right hand respectively, after we excluded the days with technical problems (**Supplementary Figures 1**, **2**).

Performance

Results of the analyses estimating HDRS score from models including passive features as well as baseline models are illustrated in **Table 1** and **Figures 3**, **4**. Overall, all of the machine learning models performed similarly with MAE ranging between 3.88 and 4.74 and correlations between the clinician-rated HDRS scores and the estimated HDRS scores ranging between 0.46 and 0.7 (**Table 1**). Of the three machine-learning models tested in the time-split scenario, the lowest mean absolute error (MAE) was obtained by the model that included only features from the mobile [F (2,12) = 19.04, $p < 0.002$]. When the three models were tested with the user-split scenario, they all performed about the same [F (2,12) = 0.55, $p < 0.59$] with the lowest MAE obtained by the model using only the features from the wearable. Thus, it is not possible to speculate as to whether one modality outperformed the others. The machine learning models provided more accurate estimates than those based on group median and individual screen models but not better than those based on individual median in the time split scenario. However, these differences were also not significant. Thus, the normalized MAE ranged between 7.5 and 9.1%, as the HDRS-17 ranges from 0 to 52. Using the Boruta algorithm (43) 39 features were defined as important for the estimation of the HDRS scores, one feature remained undecided, and the remaining features were identified as irrelevant to the outcome variable. The features that were retained were in the following categories: mobile phone engagement, activity level recorded by

TABLE 1 | Performance of all the models estimating HDRS under the User-split and Time scenarios.

	User split			Time split		
	RMSE (SD)	**MAE (SD)**	**r (95% CI)**	**RMSE (SD)**	**MAE (SD)**	**r (95% CI)**
All features Mobile + Wearable	5.43 (1.03)	4.27 (0.87)	0.57 (0.42, 0.72)	5.99 (0.14)	4.66 (0.25)	0.5 (0.45, 0.55)
Mobile	5.93 (1.45)	4.74 (1.24)	0.46 (0.18-0.74)	4.88 (0.19)	3.88 (0.18)	0.7 (0.66, 0.74)
Wearable	5.35 (1.16)	4.08 (0.9)	0.56 (0.39, 0.73)	5.76 (0.3)	4.65 (0.24)	0.54 (0.49, 0.59)
Group median baseline	6.24 (0.39)	5.06 (0.47)	NA	6.82 (0.23)	5.63 (0.24)	NA
Individual screen baseline	5.45 (1.1)	4.24 (0.99)	0.65 (0.5, 0.8)	6.64 (0.19)	5.33 (0.23)	0.42 (0.38, 0.46)
Individual median baseline	NA	NA	NA	4.13 (0.17)	3.28 (0.16)	0.81 (0.80, 0.82)

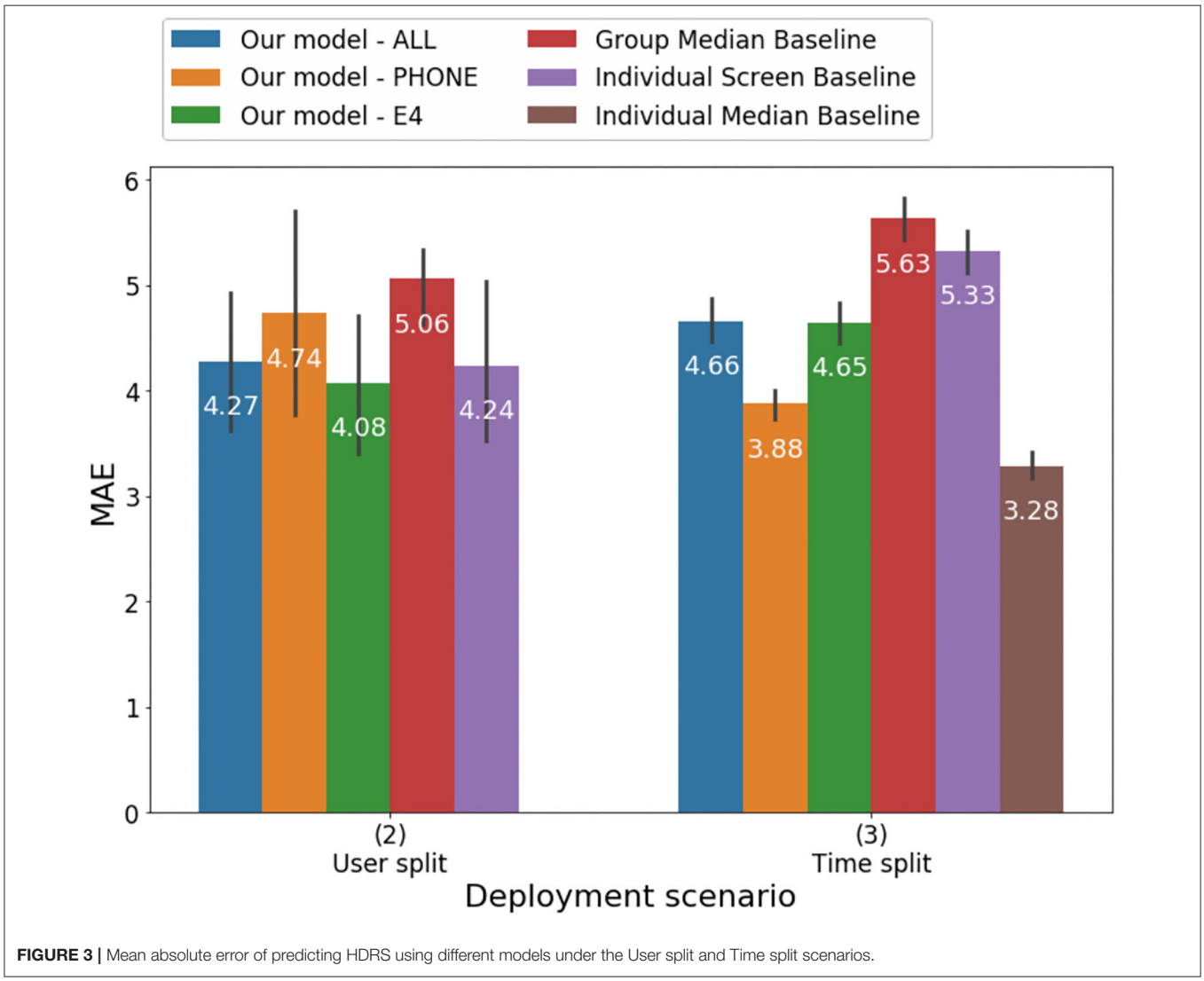


FIGURE 3 | Mean absolute error of predicting HDRS using different models under the User split and Time split scenarios.

the mobile sensors, and skin conductance and HRV features from the wearables (**Table 2**). Notably, 54% of the 39 informative features that were retained by the Boruta analyses were from the mobile phone and 46% of all the informative features were from the wearables.

DISCUSSION

The study examined the feasibility and performance of a model measuring changes in depressive symptoms severity that combined behavioral and physiological indices of depression

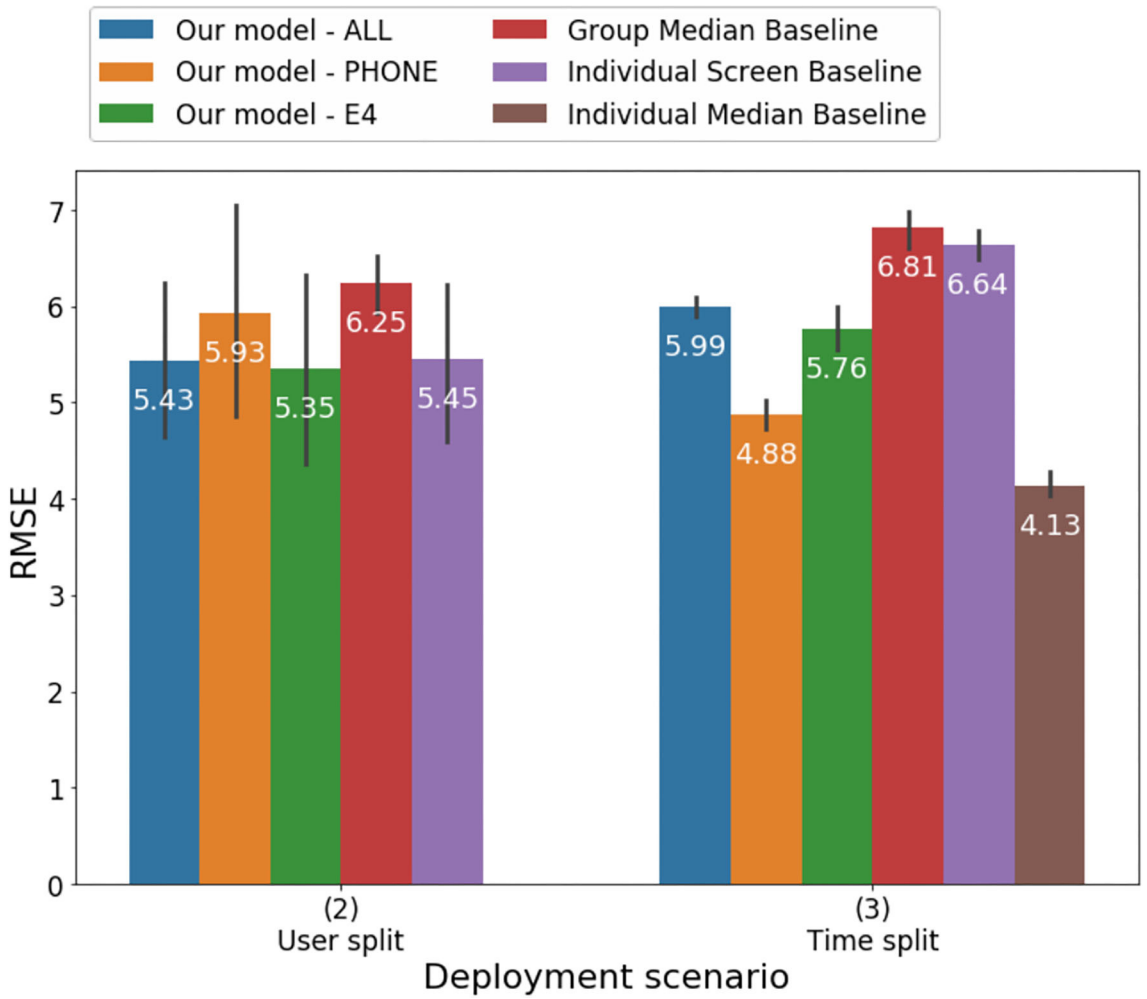


FIGURE 4 | Root mean square error of predicting HDRS using different models under the User-split and Time-split scenarios.

collected passively by smartphone and wrist sensors. Adherence was decreased by technological problems, which accounted for 17 and 22% decrease in adherence on the left-hand and right-hand wristbands respectively, a finding that also suggested that fixing the reliability of network access, connectivity, and sensor, laptop, and phone hardware would lead to more than 90% adherence.

Our study was the first to evaluate behavioral and physiological features, collected entirely passively among a sample of carefully characterized adult individuals with MDD. Previous evaluations of models to estimate depression passively have primarily relied on examining correlations between estimated and observed symptoms (18, 26, 27). However, indices of associations do not allow a granular evaluation of the accuracy of the models and of the magnitude of the difference between estimated and actual values, impacting scalability. The current study evaluated the performances of the models estimating the severity of the symptoms by using multiple indices including MAE, RMSE and correlations. Correlations between predicted and observed severity of depressive symptoms ranged from moderate to strong (r ranging between 0.46 and 0.7). The correlation between observed and estimated depression in the time-split model including features from the mobile phone ($r = 0.7$) was the strongest and was higher than the one of a previous model combining features from the fitbit and from smartphones (the best model yielded an $r^2 = 0.44$ or $r = 0.66$) (27) and the one of a model aggregating mobile-based and physiological features ($r = 0.58$) (26). Notably, despite the high magnitude of the correlations MAE ranged between 3.8 and 4.74 which may be too high of an inaccuracy for the model to be scalable.

Similarly, even though the model with mobile features in the time-split scenario performed significantly better than the others, it is unclear whether a test of significance is the most appropriate metric to compare these models and to determine whether a model is meaningfully better. In the future, criteria should be identified to evaluate when models to estimate depression severity may be deemed adoptable in clinical setting. Our models' RMSE values ranged between 4.88 and 5.99 and were higher

TABLE 2 | Illustration of the features selected by the Boruta algorithm ranked by importance.

1. Average time phone screen was on over 24 h	2. Average skin conductance response difference between right and left wrist recorded during motionless intervals over 24 h	3. Average SD of the location latitude and longitude from 12 p.m. to 6 p.m. (*location_totalStd_12 to 18*)
4. Average duration phone screen was on from 8 a.m. to 6 p.m.	5. Average location latitude over 24 h	6. Average skin conductance level mean difference between right and left wrist during motionless intervals over 24 h
7. Longitude standard deviation from 6 a.m. to 12 p.m.	8. Average location latitude from 8 a.m. to 6 p.m.	9. HRV root mean square of successive differences between normal heartbeats on the right wrist between 6 p.m. and 12 a.m.
10. Median latitude between 12 p.m. and 6 p.m.	11. Median time phone screen was on over 24 h	12. Average of skin conductance response amplitude peaks on the left wrist processed during motionless intervals over 24 h
13. HRV average of the SD of N-N intervals from 12 a.m. to 6 p.m. on the right wrist	14. Standard deviation of the Location latitude between 6 a.m. and 12 p.m.	15. Average location latitude from midnight to 6 a.m.
16. Average duration phone screen was on from 12 p.m. to 6 p.m.	17. SD of the Location latitude between 12 p.m. and 6 p.m.	18. SD of durations when phone screen was on from 12 p.m. to 6 p.m.
19. SD of the Location latitude between 8 a.m. and 6 p.m.	20. Average HRV power of the low frequency signal band over 24 h measured on right wrist	21. HRV average of the SD of N-N intervals over 24 h on the left wrist
22. HRV average of SD of N-N intervals from 6 p.m. to 12 a.m. on the right wrist	23. SD of HRV power of the low frequency signal band over 24 h measured on right wrist	24. SD of location latitude over 24 h
25. Average skin conductance level on the left wrist during motionless intervals over 24 h	26. SD of durations when phone screen was on from 8 a.m. to 6 p.m.	27. HRV Average of the SD of N-N intervals from 6 p.m. to 12 p.m. on the left wrist
28. HRV average of the SD of N-N intervals over 24 h on the right wrist	29. Skin conductance response difference right and left wrist processed during motionless intervals over 24 h	30. Number of times the phone screen was on from 12 p.m. to 6 p.m.
31. Average SD of the location latitude and longitude from 6 a.m. to 12 p.m.	32. Root mean square of successive differences between normal heartbeats over 24 h on the right wrist	33. Total time the phone screen was on from 12 p.m. to 6 p.m.
34. Average latitude 12 p.m. to 6 p.m.	35. Skin conductance response difference right and left wrist processed during motionless intervals over 24 h	36. Average power of the high frequency band of the HRV signal over 24 h on the right wrist
37. SD duration phone was on over 24 h	38. SD of the IBIs for all sinus beast from 6 p.m. to 12 p.m. on the right wrist	39. Root mean square of successive differences between normal heartbeats from 6 p.m. to 12 a.m. on the right wrist

SD, standard deviation; HRV, Heart Rate Variability.

than a model estimating self-report depressive symptoms among adolescents that included number of steps, activity level, number of SMS, and calls yielded (RMSE = 2.77) (44). The difference in results may be due to the fact that the studies employed different measures of depression. Our study, together with previous findings, suggests that including different features in the models may have significant implications on accuracy.

In order to develop a thorough method for passive assessment of depressive symptoms the study evaluated a model including behavioral as well as physiological features, which have both been described as important markers of depression. Given the high number of features that could be collected by mobile and wearable sensors the study examined which of the features included in the machine learning model were the most important, and not redundant, to estimate depressive symptoms severity. Out of the 877 features that were initially included in the model, 39 were retained that were primarily related to activity level, mobile phone engagement, skin conductance, and HRV. Our finding that activity level was among the most important features of our model estimating depressive symptoms severity is consistent with previous reports (23, 24). Saeb et al. (18) first showed that mobility pattern, recorded by phone sensors, is associated with depression, O'Brien et al. (45) documented that physical activity was low among individuals with late life depression related to healthy controls and Cao et al. (44) documented that activity level was associated with low mood among adolescents. Similarly, our finding suggesting that engagement with the phone is an important predictor of depression severity is consistent with previous studies (18).

As anticipated, physiological features were important predictors of depressive symptoms. Previously, it had been shown that wearables-based recording of skin conductance could be leveraged to detect high stress and to distinguish between high and low mental health groups of individuals (25). Our findings expand on previous reports by showing that not only commonly considered physiological features play a role in estimating depressive symptoms severity but that also features related to asymmetry of right and left skin conductance response may be important. Previously we have posited that right and left asymmetry may be a better indication of arousal than one-sided EDA measurements (46). Our finding of the role of HRV in predicting depressive symptoms is consistent and extends previous reports. Previous models estimating mood states with moderate accuracy have included measures of heart

rate (HR) (27, 44) and HRV (26). Moreover, Cao et al. (44) reported that in a model including HR features recorded by the fitbit, light exposure, and sleep one of the most important features to estimate mood variation among 18 individuals with MDD was HR. Contrary to what was anticipated, we did not find sleep being a critical feature in our model. This finding may have been due to the fact that sleep features in our model were derived from up to 48 h prior to the HDRS assessments and features capturing a longer time period may be needed to characterize depressive symptoms.

Thus, our findings suggested that behavioral as well as physiological features contributed to our model's accuracy. However, while the ubiquity of smartphones makes their use in monitoring symptoms highly scalable, passive collection of physiological indices may be less feasible due to the high cost of wearables. Given the number of behavioral and physiological features that can be collected passively, further studies are critical to examine which features, or aggregate of features, are the most critical to identify models which are the most parsimonious, feasible, and scalable.

Findings should be interpreted in the context of several limitations. Sample size was small and participants experienced low variability of depressive symptoms during the course of the study (e.g., average standard-deviation of within-user HDRS scores was 3.7 and, on average, the HDRS score from baseline to the last assessment decreased by 15%); it is unclear whether the model will have the same performance among patients with higher variability of depressive symptoms. Future studies may examine larger samples and evaluate whether other machine learning models such as Deep Neural Networks may improve performance. It is important to highlight that the Boruta method evaluates the importance of the features in the context of all the features in the model. Thus, a model including a combination of features different from ours may yield different results.

In sum, our findings highlight that machine learning may be a feasible method to estimate HDRS scores using passive monitoring based on mobile and physiological features. However, while evidence is accumulating that these models may have moderate accuracy, larger studies are needed to further evaluate them. Moreover, it is still unclear which features may be critical to develop the best models. Significant advances in the development of sensors and methodologies to analyze big data have created a new frontier of digital phenotyping, defined as the "moment-by-moment quantification of the individual-level human phenotype *in-situ* using data from smartphone and other personal devices (12)." To date, digital phenotyping has included the collection of behavioral data such as mobility patterns (via accelerometer) and socialization (via call and texts) (12). Evidence is accumulating suggesting that physiological sensing may also be included in the effort to objectively characterize changes in depression severity. Digital phenotyping of depression can be leveraged as a clinical tool and may allow a more seamless continuous treatment. In the future, sensor-based systems could privately and continuously track the symptoms of consenting patients and share this information with providers. Rather than relying on patients to reach out in times of need, providers could use these data to offer expeditious and personalized support when symptoms worsen. In addition, given considerable heterogeneity among depressed individuals to respond to different treatments, future studies should also be aimed at determining whether digital phenotyping may have a role in the prediction of response or non-response thereby allowing for more accurate initial treatment selection or more timely adjustment of treatment to optimize outcome.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because Data include sensitive information. Requests to access the datasets should be directed to ppedrelli@mgh.harvard.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by MGH and MITIRBs. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

PP, SF, DI, JA, and RP were responsible for the conceptualization of the study. PP and SF wrote the original draft. SF, AG, and DB performed the formal analysis that was supervised by SF and RP. All the authors reviewed and edited the manuscripts.

FUNDING

This study was supported by the MGH-MIT Strategic Partnership Grand Challenge Grant, by the R01MH118274 grant from the National Mental Health Institute, by European Union's FP7 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 327702 and by Abdul Latif Jameel Clinic for Machine Learning in Health.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.584711/full#supplementary-material

REFERENCES

1. World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates*. (2017). Available online at: https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf?sequence=1 (accessed October 13, 2019).
2. Wang PS, Lane M, Olfson M, Pincus HA, Wells KB, Kessler RC. Twelve-month use of mental health services in the United States: results from the

National comorbidity survey replication. *Arch Gen Psychiatry*. (2005) 62:629–40. doi: 10.1001/archpsyc.62.6.629
3. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. *Br J Psychiatry*. (2017) 210:119–24. doi: 10.1192/bjp.bp.116.188078
4. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. *Psychol Med*. (2014) 44:1303–17. doi: 10.1017/S0033291713001943
5. Simon GE, VonKorff M. Recognition, management, and outcomes of depression in primary care. *Arch Fam Med*. (1995) 4:99–105. doi: 10.1001/archfami.4.2.99
6. Unützer J, Park M. Strategies to improve the management of depression in primary care. *Prim Care*. (2012) 39:415–31. doi: 10.1016/j.pop.2012.03.010
7. Lin EH, Katon WJ, Simon GE, Von Korff M, Bush TM, Walker EA, et al. Low-intensity treatment of depression in primary care: is it problematic? *Gen Hosp Psychiatry*. (2000) 22:78–83. doi: 10.1016/S0163-8343(00)00054-2
8. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. Washington, DC: American Psychiatric Association (1994).
9. Bukh JD, Bock C, Vinberg M, Kessing LV. The effect of prolonged duration of untreated depression on antidepressant treatment outcome. *J Affect Disord*. (2013) 145:42–8. doi: 10.1016/j.jad.2012.07.008
10. Moylan S, Maes M, Wray NR, Berk M. The neuroprogressive nature of major depressive disorder: pathways to disease evolution and resistance, and therapeutic implications. *Mol Psychiatry*. (2013) 18:595–606. doi: 10.1038/mp.2012.33
11. Hardeveld F, Spijekr J, De Graaf R, Nolen WA, Beekman AT. Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatr Scand*. (2010) 122:184–91. doi: 10.1111/j.1600-0447.2009.01519.x
12. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*. (2016) 3:e16. doi: 10.2196/mental.5165
13. Courtin E, Knapp M. Social isolation, loneliness and health in old age: a scoping review. *Health Soc Care Community*. (2017) 25:799–812. doi: 10.1111/hsc.12311
14. Beiwinkel T, Kindermann S, Maier A, Kerl C, Moock J, Barbian G, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. *JMIR Ment Health*. (2016) 3:e2. doi: 10.2196/mental.4560
15. Faurholt-Jepsen M, Vinberg M, Frost M, Debel S, Margrethe Christensen E, Bardram JE, et al. Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. *Int J Methods Psychiatr Res*. (2016) 25:309–23. doi: 10.1002/mpr.1502
16. Faurholt-Jepsen M, Frost M, Vinberg M, Christensen EM, Bardram JE, Kessing LV. Smartphone data as objective measures of bipolar disorder symptoms. *Psychiatr Res Neuroimag*. (2014) 217:124–7. doi: 10.1016/j.psychres.2014.03.009
17. Grünerbl A, Muaremi A, Osmani V, Bahle G, Ohler S, Tröster G, et al. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J Biomed Heal Inform*. (2014) 19:140–8. doi: 10.1109/JBHI.2014.2343154
18. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res*. (2015) 17:e175. doi: 10.2196/jmir.4273
19. Nakonezny PA, Morris DW, Greer TL, Byerly MJ, Carmody TJ, Grannemann BD, et al. Evaluation of anhedonia with the snaith-hamilton pleasure scale (SHAPS) in adult outpatients with major depressive disorder. *J Psychiatr Res*. (2015) 65:124–30. doi: 10.1016/j.jpsychires.2015.03.010
20. Stubbs B, Vancampfort D, Firth J, Schuch FB, Hallgren M, Smith L, et al. Relationship between sedentary behavior and depression: a mediation analysis of influential factors across the lifespan among 42,469 people in low- and middle-income countries. *J Affect Disord*. (2018) 229:231–8. doi: 10.1016/j.jad.2017.12.104
21. Alvarez-Lozano J, Osmani V, Mayora O, Frost M, Bardram J, Faurholt-Jepsen M, et al. Tell me your apps and I will tell you your mood: correlation of apps usage with bipolar disorder state. In: *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments 2014*. New York, NY: ACM (2014). p. A. doi: 10.1145/2674396.2674408
22. Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep*. (2013) 36:1747–55. doi: 10.5665/sleep.3142
23. Pratap A, Atkins DC, Renn BN, Tanana MJ, Mooney SD, Anguera JA, et al. The accuracy of passive phone sensors in predicting daily mood. *Depress Anxiety*. (2019) 36:72–81. doi: 10.1002/da.22822
24. Place S, Blanch-Hartigan D, Rubin C, Gorrostieta C, Mead C, Kane J, et al. Behavioral Indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *J Med Internet Res*. (2017) 19:e75. doi: 10.2196/jmir.6678
25. Sano A, Taylor S, McHill AW, Phillips AJ, Barger LK, Klerman E, Picard R. Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study. *J Med Internet Res.* (2018) 20:e210. doi: 10.2196/jmir.9410
26. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. *Sensors*. (2020) 20:E3572. doi: 10.3390/s20123572
27. Lu J, Shang C, Yue C, Morillo R, Ware S, Kamath J, et al. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proc ACM Interact Mob Wearable Ubiquitous Technol*. (2018) 2:1–21. doi: 10.1145/3191753
28. Sarchiapone M, Gramaglia C, Iosue M, Carli V, Mandelli L, Serretti A, et al. The association between electrodermal activity (EDA), depression and suicidal behaviour: a systematic review and narrative synthesis. *BMC Psychiatry*. (2018) 18:22. doi: 10.1186/s12888-017-1551-4
29. Pantelopoulos A, Bourbakis NG. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Trans Syst Man Cybernetics C*. (2010) 40:1–12. doi: 10.1109/TSMCC.2009.2032660
30. Thase ME, Frank E, Mallinger AG, Hamer T, Kupfer DJ. Treatment of imipramine-resistant recurrent depression, III: efficacy of monoamine oxidase inhibitors. *J Clin Psychiatry*. (1992) 53:5–11.
31. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. Washington, DC: Author (2000).
32. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The mini-international neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry*. (1998) 59 (Suppl. 20):22–33.
33. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. (1960) 23:56–62. doi: 10.1136/jnnp.23.1.56
34. EmpaticaE4. *E4 Wristband Technical Specifications*. (2019). Available online at: https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications (accessed October 10, 2019).
35. MovisensXS. *eXperience Sampling for Android*. (2012). Available online at: https://xs.movisens.com (accessed April 17, 2019).
36. Fritz H, Tarraf W, Saleh DJ, Cutchin MP. Using a smartphone-based ecological momentary assessment protocol with community dwelling older African Americans. *J Gerontol B Psychol Sci Soc Sci*. (2017) 72:876–87. doi: 10.1093/geronb/gbw166
37. Ghandeharioun A, Fedor S, Sangermano L, Ionescu D, Alpert J, Dale C, et al. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In: *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII)*. San Antonio, TX (2017). doi: 10.1109/ACII.2017.8273620
38. Jiwei L, Wang X, Hovy E. What a nasty day: exploring mood-weather relationship from twitter. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. New York, NY (2014).
39. Klimstra TA, Frijns T, Keijsers L, Denissen JJ, Raaijmakers QA, Van Aken MA, et al. Come rain or come shine: individual differences in how weather affects mood. *Emotion*. (2011) 11:1495. doi: 10.1037/a0024649

40. *DarkSky*. Available online at: https://darksky.net/about (accessed October 10, 2019).
41. Géron A. *Hands-on Machine Learning With Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Boston, MA: O'Reilly Media, Inc. (2017).
42. DeMasi O, Kording K, Recht B. Meaningless comparisons lead to false optimism in medical machine learning. *PLoS ONE*. (2017) 12:e0184604. doi: 10.1371/journal.pone.0184604
43. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. (2010) 36:1–13. doi: 10.18637/jss.v036.i11
44. Cao J, Truong AL, Banu S, Shah AA, Sabharwal A, Moukaddam N. Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: development and usability study. *JMIR Ment Health*. (2020). doi: 10.2196/preprints.14045. [Epub ahead of print].
45. O'Brien JT, Gallagher P, Stow D, Hammerla N, Ploetz T, Firbank M, et al. A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression. *Psychol Med*. (2017) 47:93–102. doi: 10.1017/S0033291716002166
46. Picard RW, Fedor S, Ayzenberg Y. Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion Rev*. (2015) 8, 62–75. doi: 10.1177/1754073914565523

Conflict of Interest: RP served as cofounder and chairman of the board for Empatica, which manufactured the wearable sensors used to collect a subset of the data used in the study. She owned stock in Empatica and served as part-time consultant and chief scientist for them. PP also received royalties from MIT for patents. She was an inventor on related to wearable technology; however, none of these are directly related to this work. DM has received research support from Nordic Naturals and heckel medizintechnik GmbH. He has received honoraria for speaking from the Massachusetts General Hospital Psychiatry Academy, Harvard Blog, and PeerPoint Medical Education Institute, LLC. He also works with the MGH Clinical Trials Network and Institute (CTNI), which has received research funding from multiple pharmaceutical companies and NIMH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 22 February 2021
doi: 10.3389/fpsyt.2021.558056

Differences in Temporal Relapse Characteristics Between Affective and Non-affective Psychotic Disorders: Longitudinal Analysis

Sarah A. Immanuel[1,2]*, *Geoff Schrader*[1,3] *and Niranjan Bidargaddi*[1,2]

[1] College of Medicine and Public Health, Flinders University, Adelaide, SA, Australia, [2] Flinders Digital Health Research Centre, Flinders University, Adelaide, SA, Australia, [3] Barossa Gawler Adelaide Hills Fleurieu Local Health Network, Adelaide, SA, Australia

Objective: Multiple relapses over time are common in both affective and non-affective psychotic disorders. Characterizing the temporal nature of these relapses may be crucial to understanding the underlying neurobiology of relapse.

Materials and Methods: Anonymized records of patients with affective and non-affective psychotic disorders were collected from SA Mental Health Data Universe and retrospectively analyzed. To characterize the temporal characteristic of their relapses, a relapse trend score was computed using a symbolic series-based approach. A higher score suggests that relapse follows a trend and a lower score suggests relapses are random. Regression models were built to investigate if this score was significantly different between affective and non-affective psychotic disorders.

Results: Logistic regression models showed a significant group difference in relapse trend score between the patient groups. For example, in patients who were hospitalized six or more times, relapse score in affective disorders were 2.6 times higher than non-affective psychotic disorders [OR 2.6, 95% CI (1.8–3.7), $p < 0.001$].

Discussion: The results imply that the odds of a patient with affective disorder exhibiting a predictable trend in time to relapse were much higher than a patient with recurrent non-affective psychotic disorder. In other words, within recurrent non-affective psychosis group, time to relapse is random.

Conclusion: This study is an initial attempt to develop a longitudinal trajectory-based approach to investigate relapse trend differences in mental health patients. Further investigations using this approach may reflect differences in underlying biological processes between illnesses.

Keywords: trajectory, hospitalization, relapse, trend, psychosis, affective disorders

OPEN ACCESS

Edited by:
Qiang Luo,
Fudan University, China

Reviewed by:
Rafael Tabarés-Seisdedos,
University of Valencia, Spain
Weidan Pu,
Central South University, China

***Correspondence:**
Sarah A. Immanuel
sarah.immanuel@flinders.edu.au

Specialty section:
This article was submitted to Psychological Therapies, a section of the journal Frontiers in Psychiatry

Received: *27 May 2020*
Accepted: *28 January 2021*
Published: *22 February 2021*

Citation:
Immanuel SA, Schrader G and Bidargaddi N (2021) Differences in Temporal Relapse Characteristics Between Affective and Non-affective Psychotic Disorders: Longitudinal Analysis. Front. Psychiatry 12:558056. doi: 10.3389/fpsyt.2021.558056

INTRODUCTION

Temporal characteristics of symptom onset have played an important role in the classification of psychiatric disorders. For example, Kraepelin's original distinction between manic depression and dementia praecox was based on the idea of manic depression, now known as bipolar disorder, being a recurrent illness with periods of complete recovery alternating

with episodes of illness, while dementia praecox, now known as schizophrenia, having a chronic deteriorating course (1). With the introduction of specific treatments for bipolar disorder and schizophrenia, the outcome for both these disorders has considerably improved, although there is still debate about the extent of recovery that happens, even with treatment, particularly for schizophrenia. For example, Harrow et al. found in a 15 year follow up of patients with schizophrenia treated with contemporary interventions that while over 40% cumulatively had a period of recovery, this was followed in 60% by a period of symptom recurrence (2). Lang et al. similarly found in an extensive review of long-term outcome studies that schizophrenia had a generally poorer outcome than other diagnostic groups (3).

It has been argued that the distinction between outlook in schizophrenia and bipolar disorder may have neurological underpinnings. It has been shown that children who go on to develop schizophrenia (4) have evidence of cognitive and neurodevelopmental impairment and this is not evident in children who go on to develop bipolar disorder. These children who develop schizophrenia are also more likely to have a history of obstetric complications that could affect neurological development (5). Recently, several studies have attempted a finer grained analysis of natural history in schizophrenia in an attempt to better understand different illness trajectories and thereby better predict individual outcomes. For example, Ayesa-Arriola et al. described four different patterns of recovery in first episode psychosis based on symptoms on initial presentation (6). Velthoorst et al. compared illness trajectories and found multiple trajectories based on symptom patterns within each disorder (7). Patients with schizophrenia had more impaired trajectories, and those with mood disorders had better functioning trajectories. Such studies generally compare trajectories of distinct groups based on symptom clustering and or level of function, rather than examining any periodicity within the illness trajectory (8, 9). Periodicity, however, is important for diagnosing some affective disorders. Seasonal affective disorder and rapid cycling bipolar disorder are examples where diagnosis is made on the basis of a temporal relapse pattern characterized by predictability. Furthermore, evidence indicates people with affective disorders are more likely to have genetic polymorphisms associated with seasonal circadian disturbances (10). Additionally, within affective disorders as the number of relapses increases, there appears to be a shortening of time intervals between subsequent relapses (11). To our knowledge there have been no studies where patterns in time between relapse have been examined in non-affective psychosis.

We hypothesize, that while both affective and non-affective disorders are characterized by relapses, the temporal nature of relapse itself differs between these disorders, consistent with the differences in the underpinning biological processes and etiology. Specifically, affective disorder relapses are more likely to exhibit an inherent predictable trend pattern. Hospitalization is widely recognized as a useful proxy for reporting relapse when reporting in a naturalistic setting (12). Hence, we conducted our study on a large anonymized administrative health data set, taking rehospitalisation or presentation to the emergency department as an objective measure of relapse (13, 14). To compute the temporal nature of relapse, we applied symbolic series approach (15) on the time period between consecutive hospitalizations (gaps) for each individual. The symbolic series approach is useful for identification of predictable trend patterns within the time series while reducing inherent noise (16, 17). A predictable trend pattern is observed, when gaps are progressively increasing (relapsing less often or less frequently) or decreasing (relapsing more often or frequently) or remaining the same (regular). The output of symbolic series trend is a score between 0 to 1 with higher values indicating increasing observations of predictable patterns in the time series. We investigated for differences if any, in relapse trend score for patients with recurrent non-affective psychosis compared with patients with recurrent affective disorder.

MATERIALS AND METHODS

Dataset

In this retrospective study, anonymized longitudinal records of adult patients (18 to 65 years), with diagnosed psychiatric disorders who presented to emergency departments, or were admitted to public hospitals in South Australia between January 1, 2007, and March 10, 2017 were collected from the South Australian Mental Health Data Universe. Permission to migrate anonymized records to Flinders University for research was obtained with the approval of the South Australian Department of Health through a project agreement.

Each record of relapse contained an anonymized patient identifier, age, gender, primary ICD-10 diagnosis block documented during each hospitalization, date and time of relapse or admission and the date and time of discharge from the hospital. Other details regarding patients such as family history were not available from this data set. While most patients had a single hospitalization with no relapse within the analysis time frame, there were patients who had up to eleven relapses. While a patient may be diagnosed with an affective disorder on their first admission, the diagnosis may change to a non-affective psychotic disorder on subsequent admissions and the converse may also occur. For the purpose of this study, the primary mental health ICD-10 diagnostic block documented at each relapse to the hospital was extracted and if it had changed between relapses, the most frequently occurring diagnostic block was tagged as the primary diagnosis for each patient. Based on this diagnostic label, two subgroups of patients were extracted for analysis—patients with ICD 10 non-affective psychotic disorders, with F codes (F20–F29) and patients with ICD 10 affective disorders with F codes (F30–F39) (18). Patients with F code diagnoses F20–F29 characteristically have symptoms such as hallucinations and delusions and respond to antipsychotic medication (19), while patients with affective disorder diagnoses F 30–F39 have mood disturbances and respond to mood stabilizing medications (20). More details on the patient inclusion criteria are outlined in the **Supplementary Material**.

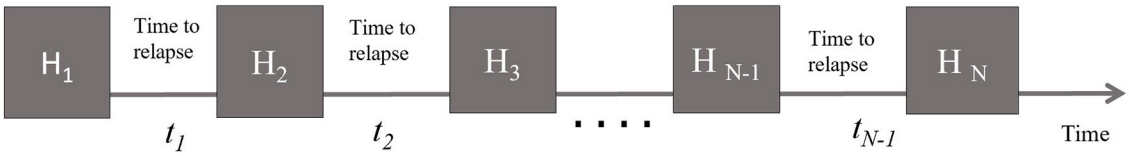


FIGURE 1 | Schematic of a MH patient trajectory indicating hospitalizations and relapses.

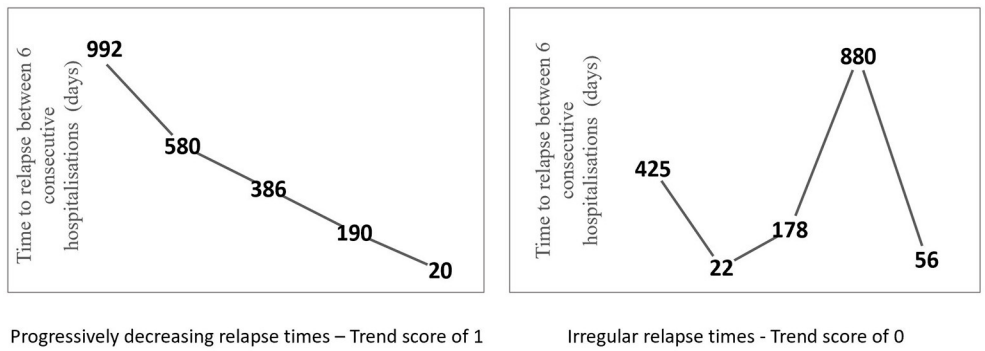


FIGURE 2 | Relapse time trajectory of a patient with five relapses where the time to relapse is progressively decreasing between each hospitalization **(left)**. Relapse time trajectory of a patient with five relapses where the relapse times are irregular and do not follow a trend **(right)**.

Relapse Trend Score

The application of a symbolic series-based approach to compute the relapse trend score is explained in detail in the **Supplementary Material**. Briefly, this method captures the relapse onset temporal pattern for each individual based on the time gap between relapses and converts it into a score between 0 to 1 that describes the trend observed in these time gaps. An individual with a history of n relapses ($n \geq 4$) has n-1 time units between the relapses (**Figure 1**). This series of time units, when iteratively grouped into three consecutive time units—with a shift in one time unit each iteration—will result in n-2 trend units. Each trend unit is a measure of direction of change between consecutive three time units and takes a score of 1 if consecutive time units form a predictable pattern, that is, are increasing or decreasing or remaining the same, and takes a score of 0 otherwise, in which case they are non-predictable or random patterns. The sum of these n-2 trend unit scores is then normalized within each cohort group based on the number of relapses. Thus, the temporal pattern score for an individual with n relapses is 1 when all n-2 trend scores are 1 and is 0 when none of the n-2 trend units scored is a 1, and anything in between will be a score between 0 to 1. A lower score suggests that relapse onset or time between hospitalizations is random, and a higher score suggests predictability over time. **Figure 2** illustrates trend scores for two individuals with 6 relapses, one with a predictable trend and the other with unpredictable relapses. The above methodology is explained in more detail in the **Supplementary Material**.

Statistical Analyses

Relapse trend score were computed using Matlab (The Mathworks Inc., Natick, MA) and data analyzed in SPSS statistical software v25. Patients were grouped into cohorts based on minimum number of relapses and regression models were developed within each cohort with age, gender and diagnosis as independent variables and the relapse trend score as the outcome variable. Linear and logistic regression were used, as appropriate, to explore the association between the outcome and the independent variables. Within the logistic model for association between relapse trend score and diagnostic information, the group that included patients with non-affective psychosis was used as the reference group and the affective disorder group was regressed against them. Odds ratios (ORs) and 95% CIs were obtained and used as a measure of effect size.

Role of the Funding Source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

RESULTS

The dataset included in the study had 8,658 subjects; 3,793 with non-affective psychosis and 4,865 with affective disorders. To analyze the relapse onset, a patient had to have at least four relapses so that the time gaps between them could be tested for a meaningful pattern. Hence, we retained only those patients from these two diagnostic groups, with more than four relapses and grouped them into cohorts based on their minimum number of relapses. The cohort size varied between 1,101 patients with four or more relapses and 43 patients who had 11 or more relapses. **Table 1** shows demographics, distribution

TABLE 1 | Distribution of age, gender, diagnostic information and relapse trend score among the patient cohorts.

Number of relapses	*N*	Age	Female	Psychotic disorders		Affective disorders	
		Mean (SD)	*N* (%)	*N* (%)	Relapse trend score mean (SD)	*N* (%)	Relapse trend score mean (SD)
4	1,101	40.1 (14.9)	480 (43.6)	652 (59.2)	0.36 (0.2)	449 (40.8)	0.43 (0.4)
5	728	39.7 (15.0)	319 (43.8)	437 (60.0)	0.38 (0.1)	291 (40.0)	0.51 (0.4)
6	490	39.2 (14.8)	212 (43.3)	299 (61.0)	0.39 (0.2)	191 (39.0)	0.60 (0.4)
7	335	39.4 (15.2)	138 (41.2)	194 (57.9)	0.41 (0.3)	141 (42.1)	0.66 (0.4)
8	231	40.0 (14.5)	92 (39.8)	142 (61.5)	0.42 (0.3)	89 (38.5)	0.72 (0.3)
9	153	39.6 (14.7)	70 (45.8)	87 (56.9)	0.41 (0.3)	66 (43.1)	0.74 (0.3)
10	89	41.1 (15.1)	32 (36.0)	46 (51.7)	0.38 (0.2)	43 (48.3)	0.78 (0.3)
11	43	40.74 (14.9)	12 (27.9)	23 (53.5)	0.44 (0.3)	20 (46.5)	0.77 (0.2)

of diagnostic information and mean relapse trend scores within each cohort group.

Age and Gender

Univariate regression models showed that age was significantly associated with relapse trend score in almost all patient cohorts analyzed, while gender had no effect on the relapse trend score in most patient cohorts (**Table 2**). Overall, the direction of associations indicated that the odds of older subjects exhibiting a pattern or trend in their gaps between hospitalizations were higher than younger adults.

Diagnosis

Within different cohort groups, diagnostic information was regressed against the relapse trend score. Logistic regression model fitting was statistically significant in all patient cohort groups (**Table 3**). Univariate analysis showed that patients with recurrent affective disorders were significantly more likely to exhibit a trend or pattern in time to relapse when compared to patients with non-affective psychosis (e.g., in cohorts with six or more relapses, OR 2.6, 95% CI 1.8–3.7, $p < 0.001$). This implies that the odds of a recurrent affective disorder patient exhibiting a trend or pattern in the time to relapse between his/her hospitalizations are 2.6 times the odds for a patient with recurrent psychosis. After adjusting for age, the recurrent affective disorder group still had a higher probability of having a trend or a pattern in time to relapse (**Table 3**).

TABLE 2 | Effect of age on the relapse trend score.

Number of relapses	Coefficient (B)	SE	*t*	*P*-value	95% CI of B
11	0.3	0.35	0.8	0.412	[-0.4–1.01]
10	0.7	0.24	2.9	0.004	[0.23–1.17]
9	0.6	0.19	3.1	0.002	[0.22–0.96]
8	0.4	0.16	2.8	0.006	[0.13–0.78]
7	0.5	0.13	4.0	0.000	[0.26–0.77]
6	0.5	0.12	4.2	0.000	[0.25–0.71]
5	0.3	0.10	3.2	0.002	[0.12–0.52]
4	0.3	0.09	3.3	0.001	[0.12–0.51]

TABLE 3 | Association of patient diagnosis (covaried with age) with relapse trend score.

Number of relapses	Coefficient (B)	SE	Wald χ^2	Odds ratio	95 % CI of OR
11	1.95	0.6	9.7	6.6	[2.0–21.5]*
10	2.13	0.4	22.5	9.9	[4.2–23.5]**
9	1.69	0.3	27.2	6.2	[3.3–11.7]**
8	1.55	0.3	33.5	5.0	[3.0–8.4]**
7	1.07	0.2	25.8	3.4	[2.2–5.0]**
6	0.87	0.2	24.9	2.6	[1.8–3.7]**
5	0.48	0.1	11.0	1.7	[1.3–2.2]*
4	0.19	0.1	2.3	1.3	[1.0–1.6]*

*$p < 0.05$; ** $p < 0.001$.

DISCUSSION

Our findings provide evidence that temporal patterns of relapse to hospitalization in non-affective psychotic disorders and affective disorders are significantly different. Longitudinal relapse trend analysis suggests that patients with affective disorders were more likely to have either increasing or decreasing times to relapse, while the time between relapse in non-affective psychotic disorders was more likely to be random. This distinction may reflect different underlying biological processes occurring in these conditions. For example, there may be more responsiveness to periodic environmental factors such as seasonal change in people with recurrent affective disorder (21). The biological processes behind non-affective psychosis may be less responsive to circadian rhythms and perhaps more likely to be affected by randomly occurring environmental, biological or social stress. Our findings are of note in view of concern that cross sectional categorical diagnostic systems such as DSM V define syndromes as heterogeneous in terms of etiology and treatment response (22). It is based on such concern the Research Domain operational Criteria (RoDC) (23) that considers classification of psychiatric disorders

on the basis of biological mechanisms, endophenotypes and biomarkers. However, in our study, patients in the broad cross-sectional diagnostic categories of affective disorder or non-affective psychosis derived from administrative hospital data, had significantly different patterns of changes in rate of relapse over time.

This is the first study to use an entropy-based approach to elucidate how temporal patterns of relapse differ between different psychiatric diagnoses. Entropy based approaches have been applied to analyze patient trajectories in finding treatment benefits, exploring relapse in placebo-controlled trials, and monitoring physical activity trends after rehabilitation. Haimovich et al. conducted an analysis of condition-specific hospital utilization rates using a clustering based computational approach and demonstrated that a substantial proportion of medical conditions exhibit seasonal variation in hospital utilization (19). This is a simple, coarse grained technique, that is robust in capturing predictability in longitudinal time-series. There are a limited number of studies in the literature where mathematical approaches have been applied to psychiatric readmission data to look for associations with clinical, environmental and health system characteristics. In particular, temporal trends or patterns in repeated admissions based longitudinal mental health patient trajectories have not been investigated previously.

Our study had several limitations. We used the relatively broad F-code categories of the ICD-10 diagnostic system to categorize patients. Unfortunately, the coding of the derived dataset that we received from the South Australian Mental Health Data Universe database system did not allow for further distinctions to be drawn between individual affective disorder categories within F30–F39 such as bipolar disorder and individual psychotic disorder categories within F20–F29 such as schizophrenia. A further subgroup analysis with the raw data presented to demonstrate heterogeneity if any within the F20 and F30 cohorts will be a future scope of this study. For example, the demonstration of a distinct set of affective diagnoses associated with a decreasing interval of time between relapse would be of interest. Another limitation relates to our definition of relapse which included only hospital and emergency department contacts. The inclusion of outpatient records of recurrence may have captured a greater number of less extreme relapses and thus affected our findings. A further limitation may relate to how we categorized patients into the affective and psychotic groups, particularly those who had different diagnoses when admitted at different times. Our decision to use the most common diagnosis, however, would conceivably have made the groups more heterogeneous and thus would have reduced the likelihood of us finding any distinction between the two categories in terms of illness trajectories.

CONCLUSION

To conclude, the presented approach is a proof of concept study toward longitudinal analysis of time trajectories in hospitalization data. We have woven together multiple hospitalizations of mental health patients and captured the trend in the time between each relapse. Temporal trends in relapse using time stamps along a mental health patient trajectory have not been investigated previously. This is a novel approach and a first step toward longitudinal trajectory-based approach to investigate relapse in mental health patients. Further investigation of patterns in mental health trajectories could provide insights into utilization of acute services over time and identify which individuals are at increased risk of readmissions.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

SI study design, literature search, figures, data collection, data analysis, data interpretation, writing and revision of MS. GS literature search, data interpretation, writing and revision of MS. NB study design, data analysis, data interpretation, writing and revision of MS. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank Prof. Tarun Bastiampillai for his support in facilitating data collection from the SA mental health data universe and Mr. Yang Yang for assisting with data retrieval processes. NB was co-funded by the Barossa Gawler Adelaide Hills Fleurieu Local Health Network.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.558056/full#supplementary-material

REFERENCES

1. Kraepelin E. Die Erscheinungsformen des Irreseins: (The manifestations of insanity). *Hist Psychiatry.* (1992) 3:509–29. doi: 10.1177/0957154X9200301208

2. Harrow M, Grossman LS, Jobe TH, Herbener ES. Do patients with schizophrenia ever show periods of recovery? A 15-year multi-follow-up study. *Schizophr Bull.* (2005) 31:723–34. doi: 10.1093/schbul/sbi026

3. Lang FU, Koesters M, Lang S, Becker T, Jaeger M. Psychopathological long-term outcome of schizophrenia–a review. *Acta Psychiatr Scand.* (2013) 127:173–82. doi: 10.1111/acps.12030
4. Cannon M, Caspi A, Moffitt TE, Harrington H, Taylor A, Murray RM, et al. Evidence for early-childhood, pan-developmental impairment specific to schizophreniform disorder: results from a longitudinal birth cohort. *Arch Gen Psychiatry.* (2002) 59:449–56. doi: 10.1001/archpsyc.59.5.449
5. Geddes JR, Verdoux H, Takei N, Lawrie SM, Bovet P, Eagles JM, et al. Schizophrenia and complications of pregnancy and labor: an individual patient data meta-analysis. *Schizophr Bull.* (1999) 25:413–23. doi: 10.1093/oxfordjournals.schbul.a033389
6. Ayesa-Arriola R, Terán JMP, Setién-Suero E, Neergaard K, Ochoa S, Ramírez-Bonilla M, et al. Patterns of recovery course in early intervention for FIRST episode non-affective psychosis patients: the role of timing. *Schizophr Res.* (2019) 209:245–54. doi: 10.1016/j.schres.2019.01.032
7. Velthorst E, Fett AKJ, Reichenberg A, Perlman G, van Os J, Bromet EJ, et al. The 20-year longitudinal trajectories of social functioning in individuals with psychotic disorders. *Am J Psychiatry.* (2017) 174:1075–85. doi: 10.1176/appi.ajp.2016.15111419
8. Gueorguieva R, Chekroud AM, Krystal JH. Trajectories of relapse in randomised, placebo-controlled trials of treatment discontinuation in major depressive disorder: an individual patient-level data meta-analysis. *Lancet Psychiatry.* (2017) 4:230–7. doi: 10.1016/S2215-0366(17)30038-X
9. Mirza SS, Wolters FJ, Swanson SA, Koudstaal PJ, Hofman A, Tiemeier H, et al. 10-year trajectories of depressive symptoms and risk of dementia: a population-based study. *Lancet Psychiatry.* (2016) 3:628–35. doi: 10.1016/S2215-0366(16)00097-3
10. Liberman AR, Halitjaha L, Ay A, Ingram KK. Modeling strengthens molecular link between circadian polymorphisms and major mood disorders. *J Biol Rhythms.* (2018) 33:318–36. doi: 10.1177/0748730418764540
11. Kessing LV, Andersen PK. Evidence for clinical progression of unipolar and bipolar disorders. *Acta Psychiatr Scand.* (2017) 135:51–64. doi: 10.1111/acps.12667
12. Olivares JM, Sermon J, Hemels M, Schreiner A. Definitions and drivers of relapse in patients with schizophrenia: a systematic literature review. *Ann Gen Psychiatry.* (2013) 12:1. doi: 10.1186/1744-859X-12-32
13. Lorine K, Goenjian H, Kim S, Steinberg AM, Schmidt K, Goenjian AK. Risk factors associated with psychiatric readmission. *J Nerv Ment Dis.* (2015) 203:425–30. doi: 10.1097/NMD.0000000000000305
14. Burns T. Hospitalisation as an outcome measure in schizophrenia. *Br J Psychiatry.* (2007) 191:s37–41. doi: 10.1192/bjp.191.50.s37
15. Cysarz D, Porta A, Montano N, Van Leeuwen P, Kurths J, Wessel N. Different approaches of symbolic dynamics to quantify heart rate complexity. *Ann Int Conf IEEE Eng Med Biol Soc.* (2013) 2013:5041–4. doi: 10.1109/EMBC.2013.6610681
16. Costa MD, Davis RB, Goldberger AL. Heart rate fragmentation: a new approach to the analysis of cardiac interbeat interval dynamics. *Front Physiol.* (2017) 8:255. doi: 10.3389/fphys.2017.00255
17. Schulz S, Voss A. Symbolic dynamics, poincaré plot analysis and compression entropy estimate complexity in biological time series. In: Barbieri R, Pasquale Scilingo E, Valenza G, editors. *Complexity and Nonlinearity in Cardiovascular Signals.* Cham: Springer (2017) p. 45–85.
18. Zivetz L. *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*, vol. 1. Geneva:World Health Organization (1992).
19. Castle DJ, Galletly CA, Dark F, Humberstone V, Morgan VA, Killackey E, et al. The 2016 Royal Australian and New Zealand College of Psychiatrists guidelines for the management of schizophrenia and related disorders. *Med J Australia.* (2017) 206:501–5. doi: 10.5694/mja16.01159
20. Ellis Australian P. New Zealand clinical practice guidelines for the treatment of depression. *Aust N Z J Psychiatry.* (2004) 38:389–407. doi: 10.1111/j.1440-1614.2004.01377.x
21. Haimovich JS, Venkatesh AK, Shojaee A, Coppi A, Warner F, Li SX, et al. Discovery of temporal and disease association patterns in condition-specific hospital utilization rates. *PloS ONE.* (2017) 12:e0172049. doi: 10.1371/journal.pone.0172049
22. Joyce DW, Kehagia AA, Tracy DK, Proctor J, Shergill SS. Realising stratified psychiatry using multidimensional signatures and trajectories. *J Transl Med.* (2017) 15:15. doi: 10.1186/s12967-016-1116-1
23. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. (2010) 167:748–51. doi: 10.1176/appi.ajp.2010.09091379

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 08 April 2021
doi: 10.3389/fpsyt.2021.610457

Digital Communication Biomarkers of Mood and Diagnosis in Borderline Personality Disorder, Bipolar Disorder, and Healthy Control Populations

George Gillett[1,2], *Niall M. McGowan*[2,3], *Niclas Palmius*[4], *Amy C. Bilderbeck*[2,5], *Guy M. Goodwin*[2] *and Kate E. A. Saunders*[2,3]*

[1] Oxford University Clinical Academic Graduate School, John Radcliffe Hospital, The Cairns Library IT Corridor Level 3, Oxford, United Kingdom, [2] Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, United Kingdom, [3] Oxford Health NHS Foundation Trust, Warneford Hospital, Oxford, United Kingdom, [4] Institute of Biomedical Engineering, University of Oxford, Oxford, United Kingdom, [5] P1vital Products, Manor House, Howbery Business Park, Wallingford, United Kingdom

OPEN ACCESS

Edited by:
Raz Gross,
Sheba Medical Center, Israel

Reviewed by:
Lior Carmi,
Sheba Medical Center, Israel
Renana Eitan,
Hebrew University Hadassah Medical School, Israel

***Correspondence:**
Kate E. A. Saunders
kate.saunders@psych.ox.ac.uk

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: 25 September 2020
Accepted: 10 March 2021
Published: 08 April 2021

Citation:
Gillett G, McGowan NM, Palmius N, Bilderbeck AC, Goodwin GM and Saunders KEA (2021) Digital Communication Biomarkers of Mood and Diagnosis in Borderline Personality Disorder, Bipolar Disorder, and Healthy Control Populations. Front. Psychiatry 12:610457. doi: 10.3389/fpsyt.2021.610457

Background: Remote monitoring and digital phenotyping harbor potential to aid clinical diagnosis, predict episode course and recognize early signs of mental health crises. Digital communication metrics, such as phone call and short message service (SMS) use may represent novel biomarkers of mood and diagnosis in Bipolar Disorder (BD) and Borderline Personality Disorder (BPD).

Materials and Methods: BD (n = 17), BPD (n = 17) and Healthy Control (HC, n = 21) participants used a smartphone application which monitored phone calls and SMS messaging, alongside self-reported mood. Linear mixed-effects regression models were used to assess the association between digital communications and mood symptoms, mood state, trait-impulsivity, diagnosis and the interaction effect between mood and diagnosis.

Results: Transdiagnostically, self-rated manic symptoms and manic state were positively associated with total and outgoing call frequency and cumulative total, incoming and outgoing call duration. Manic symptoms were also associated with total and outgoing SMS frequency. Transdiagnostic depressive symptoms were associated with increased mean incoming call duration. For the different diagnostic groups, BD was associated with increased total call frequency and BPD with increased total and outgoing SMS frequency and length compared to HC. Depression in BD, but not BPD, was associated with decreased total and outgoing call frequency, mean total and outgoing call duration and total and outgoing SMS frequency. Finally, trait-impulsivity was positively associated with total call frequency, total and outgoing SMS frequency and cumulative total and outgoing SMS length.

Conclusion: These results identify a general increase in phone call and SMS communications associated with self-reported manic symptoms and a diagnosis-moderated decrease in communications associated with depression in BD,

but not BPD, participants. These findings may inform the development of clinical tools to aid diagnosis and remote symptom monitoring, as well as informing understanding of differential psychopathologies in BD and BPD.

Keywords: bipolar disorder, borderline personality disorder, digital communications, smartphone, digital phenotyping, remote monitoring, depression, mania

INTRODUCTION

Bipolar Disorder (BD) and Borderline Personality Disorder (BPD) are psychiatric disorders with significant morbidity and associated mortality (1, 2). Both conditions share overlapping features, meaning they can be difficult to differentiate clinically and represent a diagnostic challenge in psychiatry (3–6). This is especially salient given misdiagnosis may lead to the selection of ineffective, or even harmful, treatments (7). Alongside core features of chronic mood instability and impulsivity, both conditions feature episodic exacerbation of symptoms. Individuals with BD experience episodes of depression and mania, while individuals with BPD experience acute crises often accompanied with suicidal thoughts or actions (8). The diagnostic overlap of these presentations and the fluctuant clinical course of the two disorders means that objective markers discerning diagnosis or mood may prove clinically useful in improving the accuracy of clinical diagnosis, predicting episode course and recognizing early signs of mental health crises.

Remote monitoring is concerned with the collection of clinically relevant data in ecologically-valid settings (9). Collecting time-stamped, longitudinal data in a patient's natural environment may provide a richer phenotype of mental distress than traditional forms of clinical assessment. Digital phenotyping represents a form of remote monitoring where personal digital devices, such as smartphones or wearables, are used to collect clinically relevant data (10, 11). This data may be used to identify new behavioral digital biomarkers, leading to the identification of novel phenotypes of psychiatric disorder and mental distress (9, 12). Previous research, for instance, has identified geolocation and actigraphy variables associated with clinical features in BD (13, 14). In BPD, similar passively-recorded digital markers are likely to provide insight into psychopathology and symptoms, given that BPD patients experience alexithymia and recall bias when reflecting on symptoms between clinical encounters (15, 16).

Digital phenotyping approaches are not constrained by our current classification of mental disorders and may inform more appropriate sub-grouping for diagnosis, prediction and treatment (10). This is particularly relevant in the management of depressive symptoms, where current diagnostic classification is highly heterogeneous (17). Smartphones may represent especially useful digital phenotyping tools given their relative low cost, high-frequency use and widespread ownership among the general population (11). It has been estimated that more than 90% of the world's adult population own a mobile phone (18).

Communications may represent an especially interesting subcategory of digital phenotyping in the context of BD and BPD. Observed changes in communication, such as increased talkativeness and pressured speech, are established features and predictive factors of (hypo)mania in BD (19–21). Meanwhile, symptoms, such as anhedonia, fatigue and reduced concentration may disrupt social communication in depression (22). Digital communications may also provide an empirical approach to assess psychological theories of BPD psychopathology and therapy which focus on interpersonal dysfunction as a core feature (23–25). Smartphone communication may be associated with emotional stability and mobile phone use has been hypothesized to be implicated in interpersonal attachment style (26, 27). Therefore, digital communications data may harbor clinically relevant digital biomarkers of both mood state and diagnosis in the clinically overlapping conditions of BD and BPD.

The development of any future clinical remote-monitoring tool is likely to integrate an array of variables when making predictions about diagnosis or mood (28). Therefore, it is first necessary to identify group-level associations between communications variables, diagnosis and mood symptoms in order to guide variable selection in model development (29). Previous work has investigated objective changes in phone call and short message service (SMS) use associated with both mood state and diagnosis in BD and healthy control (HC) cohorts (28, 30–32). However, findings are conflicting. Beiwinkel et al. (31) found that the frequency of outgoing SMS messages was negatively associated with depressive symptoms but not correlated with manic symptoms, while Faurholt-Jepsen et al. (30) found that call duration, but not call or text message frequency, was associated with depressive symptoms, and call frequency, incoming call duration and outgoing SMS message frequency were associated with manic symptoms. To our knowledge, patterns of digital communication are yet to be studied in BPD cohorts. Here, we present findings from an observational study of BD, BPD, and HC cohorts using self-report mood monitoring alongside passive monitoring of digital communications. We explore associations between mood, diagnosis, trait-impulsivity, and communications variables relating to phone call and SMS messaging.

METHODS

Participants

Data was collected as part of the Automated Monitoring of Symptom Severity (AMoSS) study, conducted between March 2014 and September 2018 (33, 34). Healthy volunteers were recruited from the community, BD and BPD participants were recruited from out-patient services or registration lists of ongoing studies. Participants were recruited for an initial 3-month study period, with an option to remain in the study for 12 months or longer. The study was observational in nature and independent of

TABLE 1 | Population characteristics and self-rated mood by diagnostic category.

		HC	BD	BPD	Total
Participant details					
	Participants, *n*	21	17	17	55
	Age, mean (SD)	42.38 (11.71)	42.24 (14.24)	38 (11.39)	40.98 (12.38)
	Male gender, % *(n)*	28.57% (6)	41.18% (7)	5.88% (1)	25.45% (14)
	BIS-11, mean (SD)	54.26 (6.40)	65.41 (9.03)	65.67 (11.58)	60.80 (10.49)
	Weeks in study, median (IQR)	23 (26)	19 (19)	21 (22)	21 (24.5)
Mood details					
	Aggregate weeks in study, *n*	642	456	401	1,499
	Euthymic weeks, *n* (%)	540 (84.11%)	279 (61.18%)	64 (15.96%)	883 (58.91%)
	Depressed weeks, *n* (%)	99 (15.42%)	90 (19.74%)	308 (76.81%)	497 (33.16%)
	Manic weeks, *n* (%)	3 (0.47%)	70 (15.35%)	11 (2.74%)	84 (5.60%)
	Mixed weeks, *n* (%)	0 (0%)	17 (3.73%)	18 (4.49%)	35 (2.33%)
	QIDS, median (IQR)	2 (5)	4 (7)	15 (8)	5 (10)
	ASRM, median (IQR)	0 (1)	1 (4)	1 (3)	0 (2)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

the clinical care participants received. Written informed consent was obtained from all participants. Approval was granted by the NRES Committee East of England–Norfolk (13/EE/0288) and Oxford Health NHS Foundation Trust.

Participant diagnoses were confirmed prior to study enrolment by an experienced psychiatrist (KEAS) using the Structured Clinical Interview for DSM-IV and the borderline items of the International Personality Disorder Examination (IPDE). HC status was confirmed by psychiatric assessment. Exclusion criteria for HC group were: any history of neurological disorder, head injury or major psychiatric illness, or having a first degree relative with a history of BD or BPD. Exclusion criteria for BD and BPD groups were a comorbid diagnosis of the other disorder. Due to a technical problem logging communications data, only a sub-set of the total AMoSS study population were included in this study. Our study population included a total of 55 participants; 21 HCs, 17 individuals with a diagnosis of BD and 17 individuals with a diagnosis of BPD. Demographic details by diagnostic group are reported in **Table 1**. The median number of weeks that participants provided digital communications and mood questionnaire data for was 21 weeks (**Table 1**).

Clinical Assessments

Participants completed a weekly remote mood assessment using the True Colours monitoring system (35). Depressive symptoms were assessed by the Quick Inventory of Depressive Symptomatology (QIDS), manic symptoms were assessed by the Altman Self-Rating Mania Scale (ASRM) (36, 37). For mood state, thresholds of QIDS $\geq$11 and ASRM <6 were used to define depressive state, QIDS <11 and ASRM $\geq$6 were used for manic state, QIDS $\geq$11 and ASRM $\geq$6 for mixed state and QIDS <11 and ASRM <6 for euthymic state. This is in-keeping with established thresholds for moderate or severe depressive and manic episodes (36, 38). Weeks where a participant did not complete QIDS or ASRM assessments were excluded from analysis. The Barratt Impulsiveness Scale (BIS-11) was recorded upon enrolment as a measure of trait-impulsivity (39). Baseline trait-impulsivity and summary statistics for self-reported mood measures by diagnostic group are reported in **Table 1**. Although HC participants reported symptoms of moderate depression in a number of weeks, no HC participant reported symptoms of severe depression (QIDS $\geq$16) at any point in the study (**Table 1**; **Figure 1**).

Communications Variables

Communications data were obtained from the AMoSS application that was installed on participants' smartphones at study entry. Participants without an Android device were given a smartphone and asked to use it as their primary means of communication throughout the study period. The time-stamp, length and directionality (incoming vs. outgoing) of communications (calls and SMS messages) were logged passively by the smartphone application. Weeks where a participant did not make or receive at least one phone-call or SMS were excluded from the analysis.

Selection of variables for regression analyses were guided by previous literature (28, 30). Primary communications variables were selected for their theoretical potential to directly reflect participant behavior; total and outgoing call frequency, mean total, incoming and outgoing call duration, total and outgoing SMS frequency and mean total and outgoing SMS length. Incoming call duration, but not frequency, was included due to a participant's agency to determine the length of incoming calls but not their frequency. Call and SMS frequency corresponds to the number of calls or SMS messages sent in the 6 days preceding, and day of, a completed mood assessment. Mean call duration and SMS length corresponds to the number of seconds and characters per phone-call and SMS message, respectively.

Secondary variables included cumulative duration and length, which correspond to the number of seconds or characters aggregated across all calls or text messages in the 6 days preceding, and day of, a completed mood assessment. These were selected in-line with previous methods of reporting call duration and SMS length in the literature (30) and to give a general measure of use of a communications modality (i.e., phone call or SMS messaging). Finally, variables summarizing the ratio of total and outgoing call frequency to SMS frequency and the ratio of total and outgoing call duration to SMS length were developed to

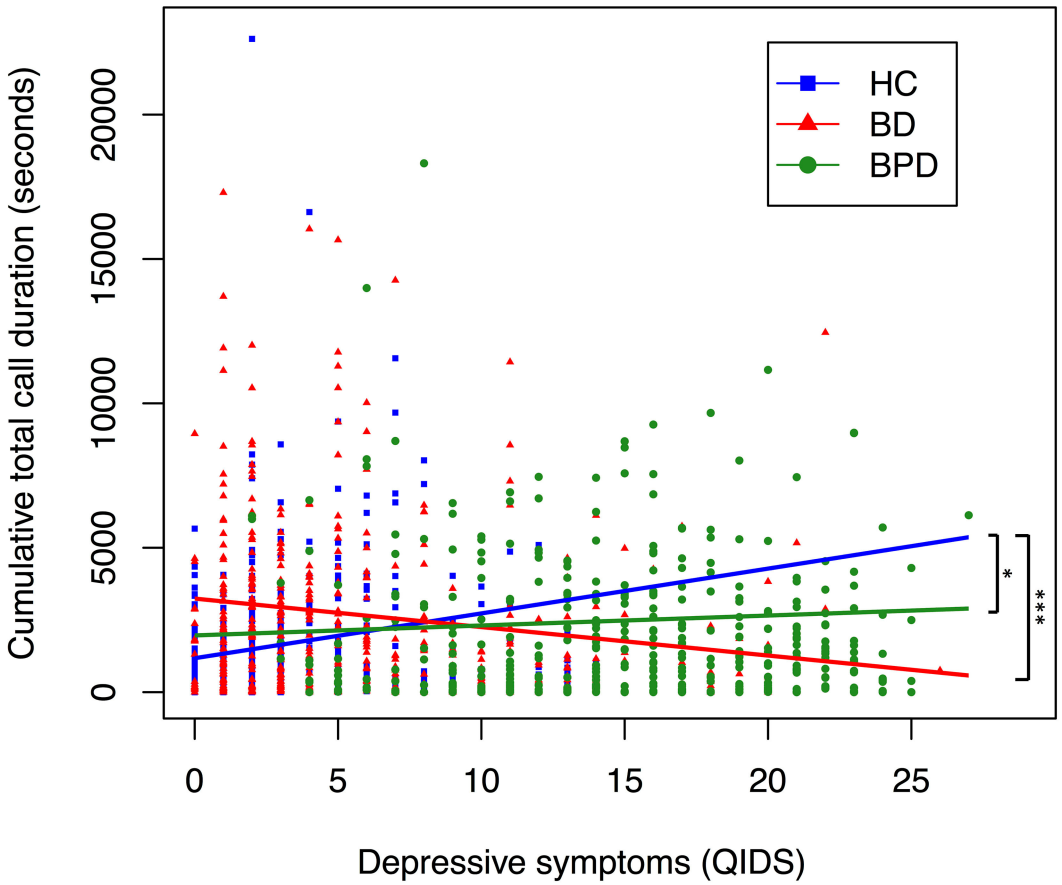


FIGURE 1 | Relationship between depressive symptoms and cumulative total call duration, by diagnostic group. A scatter plot displaying the relationship between depressive symptoms and cumulative total call duration. Each point corresponds to a participants' depressive symptoms (measured by QIDS) and their cumulative total call duration (measured in seconds) in the 6 days preceding, and day of, a completed mood assessment. Color coding corresponds to diagnosis. Trendline coefficients are taken from linear mixed-effects regression models adjusted for age. Significance testing performed with HC as reference; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (HC, $n = 642$; BD, $n = 456$; BPD, $n = 401$).

standardize participants' propensity for vocal communication to their propensity for written communication, where higher values indicate a preference for vocal communication.

Statistical Analysis

Linear mixed-effects regression models were performed with each communications variable of interest entered as a dependent variable, defined a priori. Random-effects models were used, with participant identification number entered as a random intercept. Age, diagnosis, mood state, mood symptoms, and trait-impulsivity were included as fixed effects to investigate the association between covariates and communications variables of interest. Interaction terms were inputted where relevant. Fixed effects and interaction terms are listed for each model. It was not possible to include gender as a fixed effect due to the high preponderance of female participants in our BPD sample, representative of the wider clinical population (40). Therefore, to investigate the possible effect of gender, models were replicated with gender included as a fixed effect for the BD and HC cohorts only (**Supplementary Material**). For mood state, euthymic state was used as a reference level in dummy coding. For diagnosis, HC was used as a reference level in dummy coding, apart from where stated otherwise. Regression analyses were performed with lmerTest (41) package in R (42), which performs *t*-tests using Satterthwaite's method for each covariate; *p*-values below 0.05 were considered statistically significant. Consistent with previous research, we report unstandardized coefficients (notated as B); which represent the amount (in frequency of calls/messages, seconds of call, or number of characters) by which the dependent communications variable changes for a change in the stated independent variable of one unit, keeping other independent variables constant (30). Where diagnosis was included as an independent variable, it was coded as 0 or 1 using dummy coding, and therefore in such cases the unstandardized coefficient represents the difference between the diagnostic groups, keeping other independent variables constant.

TABLE 2 | Phone call data by mood symptoms.

	Transdiagnostic model[a]			Adjusted by diagnosis[b]		
	Coefficient	*S.E.*	*p-value*	*Coefficient*	*S.E.*	*p-value*
Total call frequency[c]						
Depressive symptoms (QIDS)	−0.020	0.068	0.766	−0.049	0.071	0.488
Manic symptoms (ASRM)	0.265	0.103	0.010**	0.241	0.103	0.020*
Outgoing call frequency						
Depressive symptoms (QIDS)	0.011	0.046	0.808	−0.004	0.048	0.925
Manic symptoms (ASRM)	0.217	0.069	0.002**	0.203	0.070	0.004**
Mean total call duration						
Depressive symptoms (QIDS)	3.336	1.359	0.015*	3.485	1.514	0.022*
Manic symptoms (ASRM)	1.983	2.176	0.363	1.888	2.215	0.394
Mean incoming call duration						
Depressive symptoms (QIDS)	5.104	1.621	0.002**	5.714	1.891	0.003**
Manic symptoms (ASRM)	4.756	2.698	0.078	5.049	2.766	0.068
Mean outgoing call duration						
Depressive symptoms (QIDS)	0.106	1.622	0.948	−0.560	1.816	0.758
Manic symptoms (ASRM)	1.424	2.613	0.586	0.873	2.659	0.743

*[a] Transdiagnostic model adjusted by age only. All significant results remained significant when age removed from model. [b] Adjusted model adjusted for both age and diagnosis. [c] Analyses were performed separately for each variable in univariate analyses and with QIDS & ASRM variables together in multivariate analyses. Results remained significant when univariate analysis performed, multivariate analyses results presented here. ***p < 0.001, **p < 0.01, *p < 0.05.*

RESULTS

Mood

Mood Symptoms

Across the cohort, manic symptoms were positively associated with total call frequency (B = 0.27, SE = 0.10, $p = 0.01$) and outgoing call frequency (B = 0.22, SE = 0.07, $p < 0.01$) in the transdiagnostic model (**Table 2**). All results remained significant when adjusted for diagnosis. Manic symptoms were also positively associated with cumulative total call duration (seconds; B = 70.91, SE = 24.33, $p < 0.01$), cumulative incoming call duration (seconds; B = 33.50, SE = 13.69, $p = 0.02$) and cumulative outgoing call duration (seconds; B = 37.70, SE = 16.35, $p = 0.02$) (**Supplementary Table 1**) but not mean total, incoming or outgoing call duration (**Table 2**).

Depressive symptoms were positively associated with mean total call duration (seconds; B = 3.336, SE = 1.359, $p = 0.015$) and mean incoming call duration (seconds; B = 5.104, SE = 1.621, $p = 0.002$) and results remained significant when adjusted for diagnosis (**Table 2**). There was no strong evidence of a transdiagnostic association between depressive symptoms and other primary phone call variables (**Table 2**).

For SMS data, manic symptoms were positively associated with total SMS frequency (B = 1.62, SE = 0.40, $p \leq 0.01$) and outgoing SMS frequency (B = 0.72, SE = 0.20, $p < 0.01$) (**Table 3**). All results remained significant when adjusted for diagnosis.

There was no evidence of an association between transdiagnostic depressive symptoms and any SMS variable (**Table 3**, **Supplementary Table 1**).

Mood State

For phone call data, manic state was associated with increased total call frequency (B = 5.16, SE = 1.12, $p < 0.01$) and outgoing call frequency (B = 3.41, SE = 0.75, $p < 0.01$) compared to euthymia, but not mean total, incoming or outgoing call duration (**Supplementary Table 2**). All results remained significant when adjusted for diagnosis. Manic state was also associated with cumulative total call duration (seconds; B = 1,344.04, SE = 264.55, $p < 0.01$), cumulative incoming call duration (seconds; B = 581.58, SE = 150.39 $p < 0.01$) and cumulative outgoing call duration (seconds; B = 766.79, SE = 177.90, $p < 0.01$) (**Supplementary Table 3**).

Depressive state was not associated with any primary phone-call variable (frequency or mean duration) (**Supplementary Table 2**). However, depressive state was associated with increased cumulative incoming call duration (seconds; B = 217.23, SE = 102.74, $p = 0.04$) and a non-significant decrease in cumulative outgoing call duration (seconds; B = −164.56, SE = 122.57, $p = 0.18$) compared to euthymia (**Supplementary Table 3**).

For SMS data, there was no evidence of an association between depressive or manic states and SMS variables in either model (**Supplementary Table 2**). Mixed states were associated with increased total SMS frequency (B = 25.24, SE = 6.82, $p < 0.01$) and outgoing SMS frequency (B = 11.39, SE = 3.46, $p < 0.01$) compared to euthymia and results remained significant when adjusted for diagnosis (**Supplementary Table 2**).

To investigate whether the communications changes observed in mania are partially specific to vocal, rather than written, communication, we performed regression analyses for our secondary communications variables (**Supplementary Table 4**). There was no strong evidence of

TABLE 3 | SMS data by mood symptoms.

	Transdiagnostic model[a]			Adjusted by diagnosis[b]		
	Coefficient	*S.E.*	*p-value*	*Coefficient*	*S.E.*	*p-value*
Total SMS frequency[c]						
Depressive symptoms (QIDS)	−0.124	0.275	0.653	−0.211	0.278	0.447
Manic symptoms (ASRM)	1.618	0.402	< 0.001***	1.566	0.402	< 0.001***
Outgoing SMS frequency						
Depressive symptoms (QIDS)	−0.107	0.140	0.444	−0.151	0.141	0.283
Manic symptoms (ASRM)	0.718	0.204	< 0.001***	0.691	0.204	0.001***
Mean total SMS length						
Depressive symptoms (QIDS)	−0.645	0.408	0.115	−0.652	0.479	0.174
Manic symptoms (ASRM)	−0.306	0.707	0.665	−0.084	0.722	0.908
Mean outgoing SMS length						
Depressive symptoms (QIDS)	0.362	0.373	0.333	0.319	0.410	0.437
Manic symptoms (ASRM)	0.569	0.603	0.345	0.664	0.612	0.278

*[a]Transdiagnostic model adjusted by age only. All significant results remained significant when age removed from model. [b]Adjusted model adjusted for both age and diagnosis. [c]Analyses were performed separately for each variable in univariate analyses and with QIDS & ASRM variables together in multivariate analyses. Results remained significant when univariate analysis performed, multivariate analyses results presented here. ***p < 0.001, **p < 0.01, *p < 0.05.*

an association between manic symptoms or manic state and call frequency standardized to SMS frequency for either total or outgoing calls (**Supplementary Table 4**). However, increased manic symptoms were associated with both total call duration standardized to SMS length (B = 0.67, $p = 0.03$) and outgoing call duration standardized to SMS length (B = 0.49, $p = 0.02$), while manic state was associated with outgoing call duration standardized to SMS length (B = 5.94, SE = 2.57, $p = 0.02$) although did not reach significance ($p > 0.05$) for total call duration standardized to SMS length.

Diagnosis

For phone call data, BD diagnosis was associated with increased total call frequency (B = 5.91, SE = 2.79, $p = 0.04$) compared to HC (**Table 4**). Results remained significant when adjusted for mood symptoms, but not mood state (**Table 4**, **Supplementary Table 5**).

For SMS data, BPD diagnosis was associated with increased total SMS frequency (B = 54.52, SE = 23.79, $p = 0.03$) and outgoing SMS frequency (B = 29.05, SE = 12.32, $p = 0.02$), compared to HC (**Table 5**). BPD diagnosis was also associated with cumulative total SMS length (characters; B = 4,931.97, SE = 2,129.68, $p = 0.03$) and cumulative outgoing SMS length (characters; B = 2,927.61, SE = 1,384.30, $p = 0.04$) compared to HC (**Supplementary Table 6**), but not mean total or mean outgoing SMS length (**Table 5**). Results remained significant when adjusted for mood symptoms or mood state (**Supplementary Table 6**). BD diagnosis was associated with decreased mean total SMS length (B = −21.300, SE = 9.306, $p = 0.027$), but lost significance when adjusted for mood symptoms or state (**Table 5**, **Supplementary Table 5**).

All significant associations between BD or BPD diagnosis and communications variables were attenuated when adjusted for trait-impulsivity (**Supplementary Table 7**). In separate analyses, transdiagnostic trait-impulsivity adjusted for age was associated with all variables previously identified to be associated with BD or BPD diagnosis; increased total call frequency (B = 0.20, SE = 0.10, $p = 0.05$), total SMS frequency (B = 2.72, SE = 0.87, $p < 0.01$), outgoing SMS frequency (B = 1.36, SE = 0.45, $p < 0.01$), cumulative total SMS length (characters; B = 208.04, SE = 79.67, $p = 0.01$) and cumulative outgoing SMS length (characters; B = 106.20, SE = 52.76, $p = 0.05$) (**Supplementary Table 8**).

Interaction: Mood and Diagnosis

To assess whether diagnosis moderates the effect between mood and digital communications variables, we performed regression analyses for the interaction between diagnosis and depressive state (**Tables 6**, **7**, **Supplementary Table 9**). For phone call data, interaction between BD and depression was associated with decreased mean total call duration (seconds; B = −134.029, SE = 43.577, $p = 0.002$), mean incoming call duration (seconds; B = −126.671, SE = 53.030, $p = 0.017$), mean outgoing call duration (seconds; B = −126.342, SE = 53.620, $p = 0.019$), cumulative total call duration (seconds; B = −1598.62, SE = 464.92, $p < 0.01$), cumulative incoming call duration (seconds; B: −702.30, SE = 264.36, $p = 0.01$) and cumulative outgoing call duration (seconds; B = −875.95, SE = 320.467, $p = 0.01$) when HC was used as the reference dummy variable (**Table 6**, **Supplementary Table 9**). The interaction between BPD and depression was not significantly associated ($p > 0.05$) with any phone call variable other than mean incoming call duration (seconds; B = −106.436, SE = 49.903, $p = 0.033$) (**Table 6**). When BPD was used as the reference dummy variable, the interaction between BD and depression was associated with decreased total call frequency (B = −7.19, SE = 1.99, $p < 0.01$), outgoing call frequency (B = −3.12, SE = 1.36, $p = 0.02$), cumulative total call duration (seconds; B = −1,658.13, SE = 461.22, $p < 0.01$), cumulative incoming call duration (seconds; B = −594.73, SE = 261.93, $p = 0.02$) and cumulative outgoing

TABLE 4 | Phone call data by diagnosis (adjusted by mood symptoms).

	Unadjusted[a]			Adjusted by mood symptoms[b]		
	Coefficient	***S.E.***	***p-value***	***Coefficient***	***S.E.***	***p-value***
Total call frequency						
BD vs. HC	5.912	2.791	0.040*	5.678	2.787	0.047*
BPD vs. HC	3.561	2.835	0.216	3.962	2.945	0.184
Outgoing call frequency						
BD vs. HC	3.624	1.866	0.059	3.251	1.858	0.087
BPD vs. HC	2.185	1.896	0.256	2.054	1.964	0.300
Mean total call duration						
BD vs. HC	42.013	39.490	0.294	22.273	41.055	0.590
BPD vs. HC	37.596	40.345	0.357	−8.183	45.381	0.858
Mean incoming call duration						
BD vs. HC	20.257	40.202	0.618	−15.451	42.412	0.718
BPD vs. HC	46.534	41.316	0.268	−28.714	48.506	0.556
Mean outgoing call duration						
BD vs. HC	50.870	45.697	0.273	51.449	47.171	0.282
BPD vs. HC	38.332	46.948	0.419	44.658	52.793	0.401

[a]*Unadjusted model is adjusted by age only. All significant results remained significant when age removed from model.* [b]*Adjusted model adjusted for both age and mood symptoms (QIDS & ASRM).* ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

TABLE 5 | SMS data by diagnosis (adjusted by mood symptoms).

	Unadjusted[a]			Adjusted by mood symptoms[b]		
	Coefficient	***S.E.***	***p-value***	***Coefficient***	***S.E.***	***p-value***
Total SMS frequency						
BD vs. HC	36.311	23.501	0.129	34.506	23.354	0.146
BPD vs. HC	54.522	23.786	0.026*	55.645	23.841	0.023*
Outgoing SMS frequency						
BD vs. HC	18.164	12.176	0.142	17.661	12.172	0.153
BPD vs. HC	29.045	12.323	0.022*	30.265	12.419	0.018*
Mean total SMS length						
BD vs. HC	−21.300	9.306	0.027*	−18.306	9.799	0.067
BPD vs. HC	−8.674	9.541	0.368	−0.404	11.433	0.972
Mean outgoing SMS length						
BD vs. HC	−14.918	10.937	0.179	−17.776	11.269	0.121
BPD vs. HC	6.619	11.162	0.556	2.113	12.395	0.865

[a]*Unadjusted model is adjusted by age only. All significant results remained significant when age removed from model.* [b]*Adjusted model adjusted for both age and mood symptoms (QIDS & ASRM).* ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

call duration (seconds; B = −1,038.23, SE = 317.86, $p < 0.01$) (**Table 6, Supplementary Table 9**). Together these results suggest diagnosis may moderate the association between depression and phone call communications. Interaction trends between diagnostic groups and depressive symptoms for cumulative total, incoming and outgoing call duration are summarized in **Figures 1–3**, all other communications variables are summarized in **Supplementary Figures 1,2**.

For SMS data, the interaction between BD and depression was associated with decreased total SMS frequency (B = −28.78, SE = 7.18, $p < 0.01$), outgoing SMS frequency (B = −12.42, SE = 3.67, $p < 0.01$) and cumulative total SMS length (characters; B = −1,463.73, SE = 632.69, $p = 0.02$) when HC was used as the reference dummy variable (**Table 7**, **Supplementary Table 9**). The interaction between BPD and depression was not significantly associated ($p > 0.05$) with any SMS variable (**Table 7**). When BPD was used as the reference dummy variable, the interaction between BD and depression was associated with decreased total SMS frequency (B = −30.26, SE = 7.14, $p < 0.01$), outgoing SMS frequency (B = −12.35, SE = 3.65, $p < 0.01$) and cumulative total SMS length (characters; B = −1,749.79, SE = 628.92, $p = 0.01$), suggesting diagnosis

TABLE 6 | Phone call data by diagnosis & mood state interaction effects.

	Reference: HC group[a]			Reference: BPD group[b]		
	Coefficient	***S.E.***	***p-value***	***Coefficient***	***S.E.***	***p-value***
Total call frequency						
Depression	0.020	1.303	0.988	3.580	1.275	0.005**
BD vs. HC	6.941	2.715	0.014*	–	–	–
BPD vs. HC	0.972	2.910	0.740	–	–	–
BD vs. BPD	–	–	–	5.969	3.071	0.056
BD × Depression	−3.625	2.013	0.072	−7.185	1.994	< 0.001***
BPD × Depression	3.560	1.823	0.051	–	–	–
Outgoing call frequency						
Depression	−0.453	0.888	0.610	1.593	0.868	0.067
BD vs. HC	4.063	1.750	0.025*	–	–	–
BPD vs. HC	0.856	1.886	0.652	–	–	–
BD vs. BPD	–	–	–	3.207	1.993	0.113
BD × Depression	−1.076	1.368	0.432	−3.123	1.355	0.021*
BPD × Depression	2.047	1.242	0.100	–	–	–
Mean total call duration						
Depression	83.598	28.283	0.003**	18.205	28.628	0.525
BD vs. HC	62.671	43.466	0.157	–	–	–
BPD vs. HC	22.739	48.917	0.644	–	–	–
BD vs. BPD	–	–	–	39.932	51.836	0.444
BD × Depression	−134.029	43.577	0.002**	−68.636	43.784	0.117
BPD × Depression	−65.392	40.241	0.104	–	–	–
Mean incoming call duration						
Depression	86.272	35.785	0.016*	−20.164	34.789	0.562
BD vs. HC	27.733	44.997	0.542	–	–	–
BPD vs. HC	58.665	52.537	0.268	–	–	–
BD vs. BPD	–	–	–	−30.932	55.517	0.579
BD × Depression	−126.671	53.030	0.017*	−20.235	52.331	0.699
BPD × Depression	−106.436	49.903	0.033*	–	–	–
Mean outgoing call duration						
Depression	28.441	35.598	0.424	6.576	36.260	0.856
BD vs. HC	93.066	50.274	0.071	–	–	–
BPD vs. HC	34.878	57.968	0.549	–	–	–
BD vs. BPD	–	–	–	58.189	61.373	0.346
BD × Depression	−126.342	53.620	0.019*	−104.477	54.038	0.053
BPD × Depression	−21.865	50.811	0.667	–	–	–

*Data limited to depression & euthymia weeks (n = 1,380) to avoid rank deficiency. All analyses are adjusted for age. All significant results remained significant when age removed from model. [a]In dummy coding, HC group used as reference level, therefore diagnosis x depression represents the moderation effect compared to reference (HC). [b]In dummy coding, BPD group used as reference level, therefore diagnosis × depression represents the moderation effect compared to reference (BPD). ***p < 0.001, **p < 0.01, *p < 0.05.*

may also moderate the association between depression and SMS communications (**Table 7**, **Supplementary Table 9**).

DISCUSSION

Communications variables, incorporating phone call and SMS messaging, may represent digital biomarkers of mood symptoms, mood state and diagnosis in BD, BPD, and HC populations.

Specifically, we identified a positive association between both manic symptoms and manic state with total and outgoing phone call frequency and cumulative phone call duration, and a positive association between manic symptoms and increased total and outgoing SMS frequency. These results may reflect increased talkativeness and pressured speech, which are core features of ICD-10 and DSM-5 classification systems but are not currently operationally defined.

Furthermore, manic symptoms were associated with both increased total call duration and outgoing call duration standardized to SMS length and manic state was associated with increased outgoing call duration standardized to SMS length. This finding is novel and may refine the clinical phenotype of mania by suggesting that pressured speech and talkativeness may be objectively conceptualized as lengthening of oral relative to written communication. These results may also guide future

TABLE 7 | SMS data by diagnosis & mood state interaction effects.

	Reference: HC group[a]			Reference: BPD group[b]		
	Coefficient	*S.E.*	*p-value*	*Coefficient*	*S.E.*	*p-value*
Total SMS frequency						
Depression	0.646	4.562	0.887	2.122	4.495	0.637
BD vs. HC	41.816	30.723	0.180	–	–	–
BPD vs. HC	64.413	31.615	0.047*	–	–	–
BD vs. BPD	–	–	–	−22.597	32.840	0.495
BD × Depression	−28.781	7.180	< 0.001***	−30.257	7.138	< 0.001***
BPD × Depression	1.476	6.405	0.818	–	–	–
Outgoing SMS frequency						
Depression	−0.027	2.330	0.991	−0.101	2.296	0.965
BD vs. HC	20.689	15.558	0.190	–	–	–
BPD vs. HC	35.718	16.010	0.030*	–	–	–
BD vs. BPD	–	–	–	−15.029	16.631	0.371
BD × Depression	−12.424	3.667	0.001***	−12.350	3.646	0.001***
BPD × Depression	−0.074	3.271	0.982	–	–	–
Mean total SMS length						
Depression	3.506	9.930	0.724	8.771	9.553	0.359
BD vs. HC	−16.413	10.182	0.113	–	–	–
BPD vs. HC	−16.749	12.527	0.184	–	–	–
BD vs. BPD	–	–	–	0.336	13.395	0.980
BD × Depression	−18.988	14.442	0.189	−24.252	14.174	0.087
BPD × Depression	5.265	13.776	0.702	–	–	–
Mean outgoing SMS length						
Depression	−3.998	9.222	0.665	5.246	7.198	0.466
BD vs. HC	−13.666	11.384	0.236	–	–	–
BPD vs. HC	0.912	12.676	0.943	–	–	–
BD vs. BPD	–	–	–	−14.578	13.345	0.278
BD × Depression	−3.550	12.711	0.780	−12.794	11.322	0.259
BPD × Depression	9.244	11.697	0.430	–	–	–

*Data limited to depression & euthymia weeks (n = 1,380) to avoid rank deficiency. All analyses are adjusted for age. All significant results remained significant when age removed from model. [a]In dummy coding, HC group used as reference level, therefore diagnosis x depression represents the moderation effect compared to reference (HC). [b]In dummy coding, BPD group used as reference level, therefore diagnosis × depression represents the moderation effect compared to reference (BPD). ***p < 0.001, **p < 0.01, *p < 0.05.*

attempts to better understand the psychopathological and neurobiological basis of the increased drive to communicate observed in manic episodes. Writing SMS messages may require a more reflective capacity than oral conversation, and be less achievable to participants during a manic episode. Alternatively, it is possible that SMS communication is simply less immediately rewarding than oral communication. Known deficits in mentalization (the ability to understand other people's mental and emotional states) associated with manic episodes may also be relevant (43). Oral communication may decrease the amount of mentalization required, by providing immediate feedback, prosodic cues and potentially less ambiguous content, compared to written messaging.

Our findings are in-agreement with previous reports that manic symptoms, measured using the Young Mania Rating Scale, are associated with increased phone call and SMS communications in a separate cohort (30). We also identified a tentative association between mixed features and increased total and outgoing SMS frequency. Adjusting for diagnosis did not affect the relationship between manic symptoms and phone use.

For depression, across the whole sample, mean incoming call duration was correlated with depressive symptoms. It is plausible that this reflects increased concern from friends and family. Alternatively, increased incoming call duration may reflect features of the depressed clinical phenotype, such as psychomotor retardation and answer latency (44, 45). The latter explanation seems less likely since depressive state was weakly associated with non-significant reductions in mean outgoing call duration, although it is possible that incoming calls present greater cognitive challenges compared to outgoing calls, exacerbating the effect of psychomotor retardation.

For depressed mood there were important effects of diagnosis. BD exhibited decreased phone call and SMS communications when depressed. This might be expected from the behavioral impact of low mood via anhedonia, fatigue, reduced concentration and motor slowing (46). In contrast,

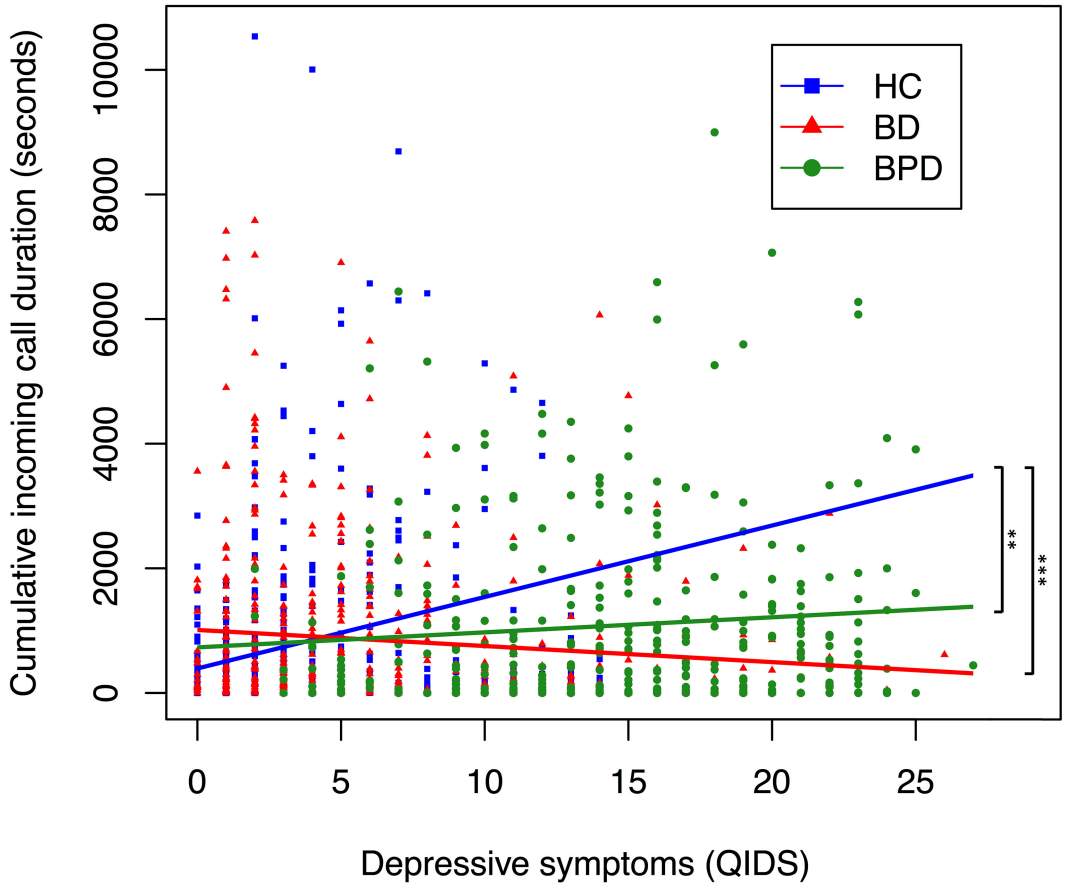


FIGURE 2 | Relationship between depressive symptoms and cumulative incoming call duration, by diagnostic group. A scatter plot displaying the relationship between depressive symptoms and cumulative total call duration. Each point corresponds to a participants' depressive symptoms (measured by QIDS) and their cumulative incoming call duration (measured in seconds) in the 6 days preceding, and day of, a completed mood assessment. Color coding corresponds to diagnosis. Trendline coefficients are taken from linear mixed-effects regression models adjusted for age. Significance testing performed with HC as reference; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (HC, $n = 642$; BD, $n = 456$; BPD, $n = 401$).

for BPD, depression was not strongly associated with any communications variable other than a reduction in mean incoming call duration.

These results add to an as yet inconsistent picture of digital communications in depressive states. While a pilot study identified decreased communications in depression, other work has identified increased phone call and SMS use in depressive states in the context of BD (28, 30, 31).

The apparent difference between the impact of depression in BD and BPD is of great interest. First, it suggests the practical possibility of a diagnostic biomarker, which would be welcome given the common clinical uncertainty in distinguishing the cause of mood instability (6). Second, while in BPD, distress is expressed in terms of depressive symptoms, they are notably more persistent than in BD (**Table 1**). The absence of decreased communications when depressed may be in-keeping with the clinical phenotype of BPD, where self-reported mental distress may not correlate well with the traditional depression phenotype (47). The absence of behavioral correlates of depression may reflect a different phenotype of depression in BPD with less core motor retardation and withdrawal. In particular, traditional clinical assessment tools may typically lack the resolution to discern these differential phenotypes, compared to the digital behavior metrics used in our study. Interestingly, our results add to a body of work suggesting that high QIDS scores in BPD individuals may not represent the same diagnostic entity of depression as in other diagnostic groups (33, 48). Furthermore, models developed to predict depressed mood in other diagnostic groups have translated poorly to BPD (13). If these results continue to be replicated in other domains, it may be that the mental distress reported by BPD individuals is more suitably conceptualized using a different diagnostic term other than depression, to reflect the different experiences, behavioral phenotypes and treatment outcomes associated with mental distress in BPD (49).

Our findings are especially interesting given that BPD is often defined as a clinical disorder of attachment, interpersonal dysfunction, perceived abandonment and the formation of

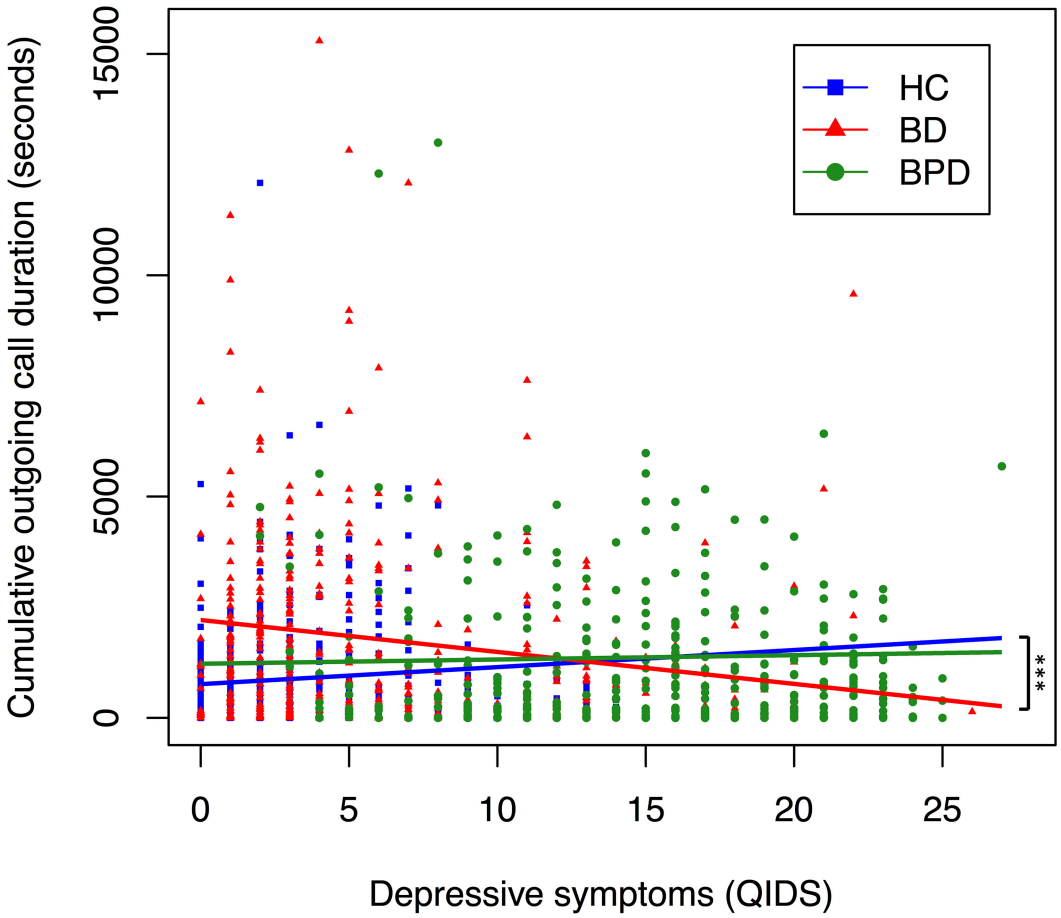


FIGURE 3 | Relationship between depressive symptoms and cumulative outgoing call duration, by diagnostic group. A scatter plot displaying the relationship between depressive symptoms and cumulative total call duration. Each point corresponds to a participants' depressive symptoms (measured by QIDS) and their cumulative outgoing call duration (measured in seconds) in the 6 days preceding, and day of, a completed mood assessment. Color coding corresponds to diagnosis. Trendline coefficients are taken from linear mixed-effects regression models adjusted for age. Significance testing performed with HC as reference; ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. (HC, $n = 642$; BD, $n = 456$; BPD, $n = 401$).

unstable relationships (7). It is possible that the persistence of social interaction during states of self-reported depression in BPD represents a type of mental distress closely associated with and possibly caused by such factors, which results in patients seeking to reaffirm their social relationships and allay perceived abandonment through persistent communication. Alternatively, it is possible that the chronic influence of interpersonal features of BPD simply over-ride any observable influence of mood on communications metrics. It is also possible that the traditional characterization of BPD as a disorder of interpersonal dysfunction results from the persistent seeking of social interaction during states of mental distress compared to other diagnostic cohorts (such as BD), who are deemed to internalize depressed mood and withdraw from social settings in line with social norms and expectations.

Regarding diagnosis, compared to HC, BD was associated with increased total call frequency, and BPD was associated with increased total and outgoing SMS frequency, even after adjustment for mood symptoms. These effects appear to have been largely driven by trait-impulsivity. This is in keeping with previous work in non-clinical populations which has identified associations between trait impulsivity and self-reported, often problematic mobile phone use in non-clinical populations (50–54). Phone-call variables were significantly associated with the motor component of impulsivity, whereas SMS use tended to be associated with the attentional and non-planning components of impulsivity (**Supplementary Table 8**). Although previous work has associated general mobile phone use with the urgency component of impulsivity (51), we believe this is the first finding of differential associations between components of impulsivity and phone call and SMS messaging. Self-reported trait impulsivity correlates poorly with laboratory assessments of impulsivity in BD (55) and digital communications may therefore represent a novel, ecologically-valid, objective marker of impulsivity if our findings are replicated in larger samples.

Limitations

Although phone call and SMS communication was frequent during the study period, it cannot be assumed that this represented a participants' complete engagement with digital communications. Social media and instant messaging applications are increasingly used in the general population and involve communication using both live and recorded written, vocal, photographic and video media. This fast-changing social ecosystem presents opportunities for future research, especially in light of previous work suggesting behaviors including propensity to send photographs may correlate with psychological traits and subjective well-being (56, 57). Social contacts may be sensitive to unique signs of illness relapse in individual participants, and incoming communications metrics beyond the scope of this study may therefore be required to detect change more reliably.

Participants were provided with a mobile phone upon study enrolment, and it is possible that they continued to use other phones during the study period. Equally, it is possible that the study phone was lent to others during the observation period. These are currently unavoidable drawbacks of ecological study designs which require trust that participants follow research instructions. Use of an Android device may also have caused a selection bias in our study population and skewed the digital behavior we observed; this has been discussed in the literature previously (28, 58).

Our results should also be interpreted in the context of the multiple analyses performed. Our study did not include adjustment for multiple testing and our results should therefore be considered to be exploratory in nature (59). Future research may focus on more specific and sophisticated measures of communication to further explore the general associations we have identified. Our results should also be interpreted in the context of our study's relatively small sample size. However, the sample size is comparable to previous analyses reported in the literature (28, 30) and our study included significant longitudinal follow-up, generating an extensive data-set.

Remote self-assessments are different from objective clinical assessments. However, it is impractical to achieve high-frequency longitudinal mood monitoring by clinical interview and the tools used in this study are clinically-validated self-report scales (60). It is possible that at extremes of mood states participants were less likely to engage in mood-monitoring, and mania in particular may not be as well-served by self-monitoring as depression. Likewise, the uneven contribution of data from different participants is an important limitation (**Table 1**), although the effect of this was mitigated in part by the use of random effects models.

Mobile phone communications have previously been associated with extraversion, agreeableness, openness and self-consciousness in non-clinical populations (27, 57). These traits were beyond the scope of this study and it is possible that they may partially explain differences in digital communications between diagnostic groups. Similarly, the unbalanced gender proportions between groups is a further limitation of our study, although the preponderance of female participants in our BPD sample is representative of the wider clinical population. To investigate the possible effect of gender, models were replicated with gender included as a fixed effect for the BD and HC cohorts (**Supplementary Tables 10–12**). This did not significantly alter the results, suggesting that gender is not a significant confounding factor for the associations we identify.

Conclusion

Our study highlights the potential to identify novel digital biomarkers of mood and diagnosis and demonstrates how such variables can identify behavioral phenotypes of mental distress specific to diagnostic categories. Future work could extend the associations between mood and a wider range of communications metrics in larger cohorts. The identification of such variables may inform the development of multivariate clinical prediction models for individual patients to support clinical diagnosis, prognosis and passive symptom monitoring.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

This study was reviewed and approved by NRES Committee East of England-Norfolk (13/EE/0288). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The AMoSS study was conceived and designed by KS and GMG. The Android application was developed by NP. Data was collected by KS and AB. The analyses presented here were conceived, designed and performed by GG. The manuscript was written by GG. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

This study was supported by the Wellcome Trust through a Centre Grant No. 98,461/Z/12/Z, The University of Oxford Sleep and Circadian Neuroscience Institute (SCNi). This work was also funded by a Wellcome Trust Strategic Award (CONBRIO: Collaborative Oxford Network for Bipolar Research to Improve Outcomes, Reference number 102,616/Z).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.610457/full#supplementary-material

REFERENCES

1. Vieta E, Berk M, Schulze TG, Carvalho AF, Suppes T, Calabrese JR, et al. Bipolar disorders. *Nat Rev Dis Primers*. (2018) 4:18008. doi: 10.1038/nrdp.2018.8
2. Gunderson JG, Herpertz SC, Skodol AE, Torgersen S, Zanarini MC. Borderline personality disorder. *Nature Rev Dis primers*. (2018) 4:18029. doi: 10.1038/nrdp.2018.29
3. Bayes A, Parker G, Fletcher K. Clinical differentiation of bipolar II disorder from borderline personality disorder. *Cur Opin Psychiatry*. (2014) 27:14–20. doi: 10.1097/YCO.0000000000000021
4. Ghaemi SN, Dalley S, Catania C, Barroilhet S. Bipolar or borderline: a clinical overview. *Acta Psychiatrica Scandinavica*. (2014) 130:99–108. doi: 10.1111/acps.12257
5. Bassett D, Mulder R, Outhred T, Hamilton A, Morris G, Das P, et al. Defining disorders with permeable borders: you say bipolar, I say borderline! *Bipolar Disord*. (2017) 19:320–3. doi: 10.1111/bdi.12528
6. Saunders KE, Bilderbeck AC, Price J, Goodwin GM. Distinguishing bipolar disorder from borderline personality disorder: a study of current clinical practice. *Eur Psychiatry*. (2015) 30:965–74. doi: 10.1016/j.eurpsy.2015.09.007
7. Leichsenring F, Leibing E, Kruse J, New AS, Leweke F. Borderline personality disorder. *Lancet*. (2011) 377:74–84. doi: 10.1016/S0140-6736(10)61422-5
8. Borschmann R, Henderson C, Hogg J, Phillips R, Moran P. Crisis interventions for people with borderline personality disorder. *Cochrane Database Syst Rev*. (2012) 6:Cd009353. doi: 10.1002/14651858.CD009353.pub2
9. Gillett G, Saunders KEA. Remote monitoring for understanding mechanisms and prediction in psychiatry. *Curr Behav Neurosci Rep*. (2019) 6:51–6. doi: 10.1007/s40473-019-00176-3
10. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*. (2016) 41:1691–6. doi: 10.1038/npp.2016.7
11. Insel TR. Digital phenotyping: technology for a new science of behavior. *JAMA*. (2017) 318:1215–6. doi: 10.1001/jama.2017.11295
12. Hidalgo-Mazzei D, Young AH, Vieta E, Colom F. Behavioural biomarkers and mobile mental health: a new paradigm. *Int J Bipolar Disord*. (2018) 6:9. doi: 10.1186/s40345-018-0119-7
13. Palmius N, Tsanas A, Saunders KEA, Bilderbeck AC, Geddes JR, Goodwin GM, et al. Detecting bipolar depression from geographic location data. *IEEE Trans BioMed Eng*. (2017) 64:1761–71. doi: 10.1109/TBME.2016.2611862
14. Shou H, Cui L, Hickie I, Lameira D, Lamers F, Zhang J, et al. Dysregulation of objectively assessed 24-hour motor activity patterns as a potential marker for bipolar I disorder: results of a community-based family study. *Transl Psychiatry*. (2017) 7:e1211. doi: 10.1038/tp.2017.136
15. Derks Y, Westerhof GJ, Bohlmeijer ET. A Meta-analysis on the association between emotional awareness and borderline personality pathology. *J Pers Disord*. (2017) 31:362–84. doi: 10.1521/pedi_2016_30_257
16. Winter D, Elzinga B, Schmahl C. Emotions and memory in borderline personality disorder. *Psychopathology*. (2014) 47:71–85. doi: 10.1159/000356360
17. Gillett G, Tomlinson A, Efthimiou O, Cipriani A. Predicting treatment effects in unipolar depression: a meta-review. *Pharmacol Ther*. (2020) 212:107557. doi: 10.1016/j.pharmthera.2020.107557
18. Topol EJ, Steinhubl SR, Torkamani A. Digital medical tools and sensors. *JAMA*. (2015) 313:353–4. doi: 10.1001/jama.2014.17125
19. Frye MA, Helleman G, McElroy SL, Altshuler LL, Black DO, Keck PE Jr, et al. Correlates of treatment-emergent mania associated with antidepressant treatment in bipolar depression. *Am J Psychiatry*. (2009) 166:164–72. doi: 10.1176/appi.ajp.2008.08030322
20. Guidi A, Salvi S, Ottaviano M, Gentili C, Bertschy G, de Rossi D, et al. Smartphone application for the analysis of prosodic features in running speech with a focus on bipolar disorders: system performance evaluation and case study. *Sensors*. (2015) 15:28070–87. doi: 10.3390/s151128070
21. Karam ZN, Provost EM, Singh S, Montgomery J, Archer C, Harrington G, et al. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. *Proceed IEEE Int Conf Acoust Speech Signal Process*. (2014) 2014:4858–62. doi: 10.1109/ICASSP.2014.6854525
22. Kupferberg A, Bicks L, Hasler G. Social functioning in major depressive disorder. *Neurosci Biobehav Rev*. (2016) 69:313–32. doi: 10.1016/j.neubiorev.2016.07.002
23. Lazarus SA, Cheavens JS, Festa F, Zachary Rosenthal M. Interpersonal functioning in borderline personality disorder: a systematic review of behavioral and laboratory-based assessments. *Clinical Psychol Rev*. (2014) 34:193–205. doi: 10.1016/j.cpr.2014.01.007
24. Bateman AW. Interpersonal psychotherapy for borderline personality disorder. *Clin Psychol Psychother*. (2012) 19:124–33. doi: 10.1002/cpp.1777
25. Ooi J, Michael J, Lemola S, Butterfill S, Siew CSQ, Walasek L. Interpersonal functioning in borderline personality disorder traits: a social media perspective. *Sci Rep*. (2020) 10:1068. doi: 10.1038/s41598-020-58001-x
26. Konok V, Gigler D, Bereczky BM, Miklósi Á. Humans' attachment to their mobile phones and its relationship with interpersonal attachment style. *Front Psychol*. (2016) 61:537–47. doi: 10.1016/j.chb.2016.03.062
27. Stachl C, Hilbert S, Au J-Q, Buschek D, De Luca A, Bischl B, et al. Personality traits predict smartphone usage. *Eur J Pers*. (2017) 31:701–22. doi: 10.1002/per.2113
28. Faurholt-Jepsen M, Busk J, Thornorarinsdottir H, Frost M, Bardram JE, Vinberg M, et al. Objective smartphone data as a potential diagnostic marker of bipolar disorder. *Aust N Z J Psychiatry*. (2019) 53:119–28 doi: 10.1177/0004867418808900
29. Lee YH, Bang H, Kim DJ. How to establish clinical prediction models. *Endocrinol Metabol*. (2016) 31:38–44. doi: 10.3803/EnM.2016.31.1.38
30. Faurholt-Jepsen M, Vinberg M, Frost M, Christensen EM, Bardram JE, Kessing LV. Smartphone data as an electronic biomarker of illness activity in bipolar disorder. *Bipolar Disord*. (2015) 17:715–28. doi: 10.1111/bdi.12332
31. Beiwinkel T, Kindermann S, Maier A, Kerl C, Moock J, Barbian G, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. *JMIR Ment health*. (2016) 3:e2. doi: 10.2196/mental.4560
32. Grünerbl A, Muaremi A, Osmani V, Bahle G, Ohler S, Tröster G, et al. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J Biomed Health Inform*. (2015) 19:140–8. doi: 10.1109/JBHI.2014.2343154
33. Tsanas A, Saunders KE, Bilderbeck AC, Palmius N, Osipov M, Clifford GD, et al. Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder. *J Affect Disord*. (2016) 205:225–33. doi: 10.1016/j.jad.2016.06.065
34. Saunders KE, Bilderbeck AC, Panchal P, Atkinson LZ, Geddes JR, Goodwin GM. Experiences of remote mood and activity monitoring in bipolar disorder: a qualitative study. *Eur Psychiatry*. (2017) 41:115–21. doi: 10.1016/j.eurpsy.2016.11.005
35. Goodday SM, Atkinson L, Goodwin G, Saunders K, South M, Mackay C, et al. The true colours remote symptom monitoring system: a decade of evolution. *J Med Internet Res*. (2020) 22:e15188. doi: 10.2196/15188
36. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry*. (2003) 54:573–83. doi: 10.1016/S0006-3223(02)01866-8
37. Altman EG, Hedeker D, Peterson JL, Davis JM. The altman self-rating mania scale. *Biol Psychiatry*. (1997) 42:948–55. doi: 10.1016/S0006-3223(96)00548-3
38. Miller CJ, Johnson SL, Eisner L. Assessment tools for adult bipolar disorder. *Clin Psychol*. (2009) 16(2):188–201. doi: 10.1111/j.1468-2850.2009.01158.x
39. Patton JH, Stanford MS, Barratt ES. Factor structure of the Barratt impulsiveness scale. *J Clin Psychol*. (1995) 51:768–74. doi: 10.1002/1097-4679(199511)51:6<768::aid-jclp2270510607>3.0.co;2-1
40. Ten Have M, Verheul R, Kaasenbrood A, van Dorsselaer S, Tuithof M, Kleinjan M, et al. Prevalence rates of borderline personality disorder symptoms: a study based on the Netherlands Mental Health Survey and Incidence Study-2. *BMC Psychiatry*. (2016) 16:249. doi: 10.1186/s12888-016-0939-x
41. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. *J Stat Softw*. (2017) 82:1–26. doi: 10.18637/jss.v082.i13

42. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing (2019). Available online at: https://www.R-project.org/ (accessed July 10, 2020).
43. Bodnar A, Rybakowski JK. Mentalization deficit in bipolar patients during an acute depressive and manic episode: association with cognitive functions. *Int J Bipolar Disord*. (2017) 5:38. doi: 10.1186/s40345-017-0107-3
44. Greden JF, Albala AA, Smokler IA, Gardner R, Carroll BJ. Speech pause time: a marker of psychomotor retardation among endogenous depressives. *Biol Psychiatry*. (1981) 16:851–9.
45. Alpert M, Kotsaftis A, Pouget ER. At issue: speech fluency and schizophrenic negative signs. *Schizophr Bull*. (1997) 23:171–7. doi: 10.1093/schbul/23.2.171
46. Benassi VA, Sweeney PD, Dufour CL. Is there a relation between locus of control orientation and depression? *J Abnorm Psychol*. (1988) 97:357–67. doi: 10.1037/0021-843X.97.3.357
47. Beatson JA, Rao S. Depression and borderline personality disorder. *Med J Aust*. (2013) 199:S24–7. doi: 10.5694/mja12.10474
48. Tsanas A, Saunders K, Bilderbeck A, Palmius N, Goodwin G, De Vos M. Clinical insight into latent variables of psychiatric questionnaires for mood symptom self-assessment. *JMIR Ment Health*. (2017) 4:e15. doi: 10.2196/mental.6917
49. Newton-Howes G, Tyrer P, Johnson T. Personality disorder and the outcome of depression: meta-analysis of published studies. *Br J Psychiatry*. (2006) 188:13–20. doi: 10.1192/bjp.188.1.13
50. Billieux J, Van der Linden M, d'Acremont M, Ceschi G, Zermatten A. Does impulsivity relate to perceived dependence on and actual use of the mobile phone? (2007). *Appl Cognit Psychol*. (2007) 21:527–37. doi: 10.1002/acp.1289
51. Billieux J, Van der Linden M, Rochat L. The role of impulsivity in actual and problematic use of the mobile phone. *Appl Cognit Psychol*. (2008) 22:1195–210. doi: 10.1002/acp.1429
52. Mitchell L, Hussain Z. Predictors of problematic smartphone use: an examination of the integrative pathways model and the role of age, gender, impulsiveness, excessive reassurance seeking, extraversion, and depression. *Behav Sci*. (2018) 8:74. doi: 10.3390/bs8080074
53. Mei S, Chai J, Wang SB, Ng CH, Ungvari GS, Xiang YT. Mobile phone dependence, social support and impulsivity in chinese university students. *Int J Environ Res Public Health*. (2018) 15:504. doi: 10.3390/ijerph15030504
54. Grant JE, Lust K, Chamberlain SR. Problematic smartphone use associated with greater alcohol consumption, mental health issues, poorer academic performance, and impulsivity. *J Behav Addict*. (2019) 8:335–42. doi: 10.1556/2006.8.2019.32
55. Newman AL, Meyer TD. Impulsivity: present during euthymia in bipolar disorder?-a systematic review. *Int J Bipolar Disord*. (2014) 2:2. doi: 10.1186/2194-7511-2-2
56. Gao Y, Li H, Zhu T. Predicting Subjective Well-Being by Smartphone Usage Behaviors. *Proc Int Conf Biomed Eng Syst Technol*. (2014) 5:317–22. doi: 10.5220/0004800203170322
57. Stachl C, Au Q, Schoedel R, Gosling SD, Harari GM, Buschek D, et al. Predicting personality from patterns of behavior collected with smartphones. *Proc Natl Acad Sci USA*. (2020) 117:17680–7. doi: 10.1073/pnas.1920484117
58. Götz FM, Stieger S, Reips U-D. Users of the main smartphone operating systems (iOS, Android) differ only little in personality. *PLoS ONE*. (2017) 12:e0176921. doi: 10.1371/journal.pone.0176921
59. Bender R, Lange S. Adjusting for multiple testing – when and how? *J Clin Epidemiol*. (2001) 54:343–9. doi: 10.1016/S0895-4356(00)00314-0
60. Simon J, Budge K, Price J, Goodwin GM, Geddes JR. Remote mood monitoring for adults with bipolar disorder: an explorative study of compliance and impact on mental health service use and costs. *Eur Psychiatry*. (2017) 45:14–9. doi: 10.1016/j.eurpsy.2017.06.007

Conflict of Interest: GMG is a NIHR Emeritus Senior Investigator, holds shares in P1vital and P1Vital products and has served as consultant, advisor or CME speaker in the last 3 years for Compass pathways, Evapharm, Janssen, Lundbeck, Medscape, P1Vital, Sage, Servier. KS is supported by the NIHR Oxford Health Biomedical Research Centre. AB has received salaries from P1vital Ltd. NP holds shares in SmartCare Analytics Ltd. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 16 April 2021
doi: 10.3389/fpsyt.2021.582745

Effects of 7.5% Carbon Dioxide and Nicotine Administration on Latent Inhibition

Kiri T. Granger[1,2,3]*, *Jennifer Ferrar*[1,4], *Sheryl Caswell*[1,3], *Mark Haselgrove*[2], *Paula M. Moran*[2], *Angela Attwood*[4] and *Jennifer H. Barnett*[1,3,5]

[1] Cambridge Cognition, Cambridge, United Kingdom, [2] School of Psychology, University of Nottingham, Nottingham, United Kingdom, [3] Monument Therapeutics, Cambridge, United Kingdom, [4] Alcohol & Tobacco Research Group, University of Bristol, Bristol, United Kingdom, [5] Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom

Stratified medicine approaches have potential to improve the efficacy of drug development for schizophrenia and other psychiatric conditions, as they have for oncology. Latent inhibition is a candidate biomarker as it demonstrates differential sensitivity to key symptoms and neurobiological abnormalities associated with schizophrenia. The aims of this research were to evaluate whether a novel latent inhibition task that is not confounded by alternative learning effects such as learned irrelevance, is sensitive to (1) an in-direct model relevant to psychosis [using 7.5% carbon dioxide (CO_2) inhalations to induce dopamine release *via* somatic anxiety] and (2) a pro-cognitive pharmacological manipulation (*via* nicotine administration) for the treatment of cognitive impairment associated with schizophrenia. Experiment 1 used a 7.5% CO_2 challenge as a model of anxiety-induced dopamine release to evaluate the sensitivity of latent inhibition during CO_2 gas inhalation, compared to the inhalation of medical air. Experiment 2 examined the effect of 2 mg nicotine administration vs. placebo on latent inhibition to evaluate its sensitivity to a potential pro-cognitive drug treatment. Inhalation of 7.5% CO_2 raised self-report and physiological measures of anxiety and impaired latent inhibition, relative to a medical air control; whereas administration of 2 mg nicotine, demonstrated increased latent inhibition relative to placebo control. Here, two complementary experimental studies suggest latent inhibition is modified by manipulations that are relevant to the detection and treatment of schizophrenia. These results suggest that this latent inhibition task merits further investigation in the context of neurobiological sub-groups suitable for novel treatment strategies.

Keywords: schizophrenia, biomarker, latent inhibition, carbon dioxide challenge, nicotine

OPEN ACCESS

Edited by:
Raz Gross,
Sheba Medical Center, Israel

Reviewed by:
Veena Kumari,
Brunel University London, United Kingdom
Christos Theleritis,
National and Kapodistrian University of Athens, Greece

***Correspondence:**
Kiri T. Granger
kgranger@monumenttx.com

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: *13 July 2020*
Accepted: *05 February 2021*
Published: *16 April 2021*

Citation:
Granger KT, Ferrar J, Caswell S, Haselgrove M, Moran PM, Attwood A and Barnett JH (2021) Effects of 7.5% Carbon Dioxide and Nicotine Administration on Latent Inhibition. Front. Psychiatry 12:582745. doi: 10.3389/fpsyt.2021.582745

INTRODUCTION

The biological heterogeneity of schizophrenia continues to be a major obstacle for clinical practice and the development of novel drug treatments. A non-invasive biomarker to define sub-groups of patients with common neurobiological underpinnings would improve detection, diagnosis and the efficacy of drug development. Abnormal attention is a core deficit of schizophrenia that is commonly modeled pre-clinically using a latent inhibition paradigm (1–4) which may have potential in this regard. In latent inhibition, a stimulus is rendered irrelevant by mere

exposure, before being established as a cue for an outcome. Latent inhibition is observed when participants learn more slowly about the preexposed cue than a non-preexposed control cue during a subsequent test of learning (5). Theoretical analyses of latent inhibition have focused upon an attentional explanation—proposing that during preexposure, attention diminishes to the preexposed stimulus so that, subsequently, participants take longer to learn the association between this stimulus and the outcome than the non-preexposed cue (6–8).

Disrupted latent inhibition is widely observed in schizophrenia [for a review see (9)] and can happen in two distinct ways: (1) *An attenuation of latent inhibition*, in which the difference in the rate of learning to the preexposed and non-preexposed stimuli is reduced (and we posit a disrupted ability to reduce attention to the preexposed/irrelevant stimulus). (2) *An enhancement of latent inhibition* in which the difference in the rate of learning to the preexposed and non-preexposed cues is increased (and we posit an enhanced ability to reduce attention to the preexposed/irrelevant stimulus). Latent inhibition thus provides a measure of the balance between these two extremes of attentional processing, which together, are thought to underpin the key symptoms of schizophrenia (4, 10). Attenuated latent inhibition is deemed particularly relevant to the positive symptoms (i.e., hyper-dopaminergic state) of the disorder; with an inability to reduce attention to irrelevant information driving a psychotic state. Whereas enhanced latent inhibition is related to the negative and cognitive symptoms [i.e., cholinergic and hypo-glutamatergic; see (11)]; where augmented reduction in attention to the preexposed stimulus is considered a reflection of an inability to switch attentional responding and learn that the preexposed stimulus is now a predictor of an outcome (9).

In line with the well-known dopaminergic contribution to psychosis (12, 13), rats treated with amphetamine show an attenuation of latent inhibition (14, 15) which is successfully reversed by dopamine-blocking antipsychotic drugs [for a review see (10)]. This has been replicated in humans [see (10, 11)], providing support for amphetamine-induced disrupted latent inhibition as a model of positive symptoms of schizophrenia. In contrast to dopaminergic effects, and in line with the idea that glutamatergic and cholinergic signaling drives the negative and cognitive symptoms of schizophrenia, NMDA antagonists (i.e., MK801) that inhibits glutamate as well as nicotinic acetylcholine receptors (nAChRs) (16) have demonstrated an opposing effect, producing enhancement (excess) of latent inhibition in humans and animals [(10); but see (17)].

The existence of dissociable forms of perturbation in latent inhibition is supported by observations of attenuated latent inhibition in acutely psychotic patients experiencing positive symptoms [e.g., (18–21)], and an enhancement of latent inhibition demonstrated in patients experiencing a predominance of negative and cognitive symptoms (9, 20, 22–24). As these attentional manifestations can be mapped onto underlying neural systems considered dysfunctional in schizophrenia, latent inhibition lends itself as a potential tool for detecting patients with different neurochemical states and symptomologies.

As anti-psychotic treatments are largely ineffective at treating the negative and cognitive symptoms of schizophrenia (25–27), many attempts have been made to develop non-dopaminergic treatments for cognitive impairment associated with schizophrenia. Several of these efforts have emphasized the α7 subtype of nAChRs due to the preponderance of patients with schizophrenia who self-medicate with nicotine to manage cognitive and negative symptoms and the side effects of anti-psychotic medications [(28), but see (29)]. This hypothesis is built on evidence that nicotinic receptor signaling is fundamentally decreased in individuals experiencing schizophrenia, and thus patients are using the most readily available method for pharmacologically targeting this system in an attempt to restore signaling to appropriate levels (30).

In humans, reports of the effects of nicotine on latent inhibition are however limited. Thornton et al. (31) reported that nicotine failed to affect latent inhibition in non-smokers who were tested following subcutaneous administration of nicotine, vs. a placebo-treated control group. Although, in a group of smokers vs. non-smokers, Della Casa and Feldon (32) reported that latent inhibition was enhanced. Pre-clinically however, a α7-nAChR partial agonist, SSR180711, has been shown to reinstate latent inhibition following administration of the NMDA receptor antagonist MK801 (33), as well as improve attention and memory performance. Furthermore, α7-nAChR agonists have been shown to improve P50 attentional gating deficits as well as cognitive performance on measures of sustained attention, measured by the Cambridge Neuropsychological Test Automated Battery (CANTAB) in patients with chronic schizophrenia (34). Additional evidence supports a moderate correlation between P50 and latent inhibition [$r > 0.6$ (35)]. With the pro-cognitive potential of nicotine-enhancing agents for the treatment of cognitive impairment associated with schizophrenia, the current study aimed to investigate the sensitivity of a novel latent inhibition task [see (36)] to nicotine exposure vs. placebo in non-smoking individuals. Treatment of improved attentional filtering (enhanced latent inhibition) following nicotine vs. placebo treatment could provide evidence to determine the future research and potential clinical validation of this latent inhibition task that may serve as a potential tool to identify patients with schizophrenia most likely to benefit cognitively from a nicotinic-based treatment.

This study aimed to evaluate the sensitivity of latent inhibition to both clinically-relevant (dopaminergic) and pro-cognitive pharmacological (nicotinic) manipulations. Experiment 1 explored the sensitivity of the latent inhibition task to a 7.5% carbon dioxide (CO_2) challenge as a model of anxiety-induced dopamine release. Given evidence that the 7.5% CO_2 challenge is accepted as a robust method to induce state anxiety (37, 38) and state anxiety increases dopamine release (39, 40), it was hypothesized that latent inhibition would be attenuated during the CO_2 gas inhalation, compared to inhalation of medical air, in a single-blind crossover design in healthy volunteers. Experiment 2 conversely explored the sensitivity of the latent inhibition task to a pro-cognitive model relevant to the treatment of cognitive impairment associated with schizophrenia by examining the effect of nicotine administration on latent

inhibition. It was hypothesized that latent inhibition would be increased (i.e., improved attentional filtering) following nicotine administration, compared to placebo, in a single-blind crossover design in non-smoking healthy volunteers.

EXPERIMENT 1: EFFECTS OF 7.5% CARBON DIOXIDE INHALATION ON LATENT INHIBITION

Materials and Methods

Design

In experiment 1, 30 healthy volunteers were administered either 7.5% CO_2 or medical air to induce dopamine release *via* induction of state anxiety, in a single-blind crossover design, with 30-min washout between gas inhalations. The gas orders were counterbalanced across participants.

Participants

Thirty non-smoking healthy volunteers were recruited from the University of Bristol and the local community *via* email lists, poster, and fllier advertisements and the Tobacco and Alcohol Research Group newsletter and website. The exclusion criteria were age under 18 or over 50 years, daily smoking, history of drug/alcohol dependency, pregnancy or breast feeding, recent use of prescribed or illicit drugs, uncorrected visual or hearing problems, diagnosed medical illness, and not being registered with a general practitioner. Pregnancy and recent drug use were assessed by urine screen, whereas all other criteria were confirmed by self-report. Participants were also excluded if they had high systolic or diastolic blood pressure (SBP/DBP) ($<$140/90 mmHg), bradycardia or tachycardia ($<$50 or $>$90 beats per min), or a body mass index (BMI) outside a healthy range ($<$18 or $>$30 kg/m^2) (all physically assessed). Psychiatric health was assessed using a truncated MINI International Neuropsychiatric Interview (41). Participants refrained from consuming alcohol for 36 h prior to the study day. Expired breath alcohol and carbon monoxide readings were taken, and participants were to be excluded if the readings were $>$0 or $\geq$10, respectively. No candidate participants had to be excluded from the research. The study was approved by the University of Bristol, Faculty of Science Research Ethics Committee. Sample size was determined based on a previous study of a similar nature (38).

Gases and Questionnaires

The gases were 7.5% CO_2 or medical air (21% oxygen; BOC Ltd.). These were administered using an oro-nasal mask (Hans Rudolph, Kansas City, MO, USA). Questionnaires included the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA) (42), Positive and Negative Affect Schedule (PANAS) (43), and the Oxford-Liverpool Inventory of Feelings and Experiences as a measure of schizotypy to ensure baseline schizotypy scores were within normative range [O-LIFE (44)].

Latent Inhibition Task

A modification of Granger et al.'s (36) latent inhibition task was used and delivered *via* the CANTAB Connect web-based software platform. Two equivalent versions of the task were used (one during each gas inhalation). Each participant completed the task on a 17-in. LCD monitor at a resolution of 1,280 $\times$ 1,024 with a 60-Hz refresh rate. The latent inhibition task was accessed *via* a web-based link that directed participants to the CANTAB Connect platform-hosting site for the task and data collection. Stimuli were white capital-letters in Arial-font (7 mm $\times$ 5 mm; h $\times$ w) presented for 1,000 ms each on the computer-screen with a black background. There were two versions of the task to enable repeat testing that were counterbalanced across participants. For version 1, the stimulus-letters were S and H; one of the letters served as the preexposed stimulus and the other was the non-preexposed stimulus, counterbalanced across participants. The target was the letter X, with filler-letters D, M, T, and V; see **Figure 1** for an example. For version 2, the stimulus letters were R and O, and again one of the letters served as the preexposed stimulus and the other was the non-preexposed stimulus, counterbalanced across participants. The target was the letter Z, with filler-letters F, N, K, and A.

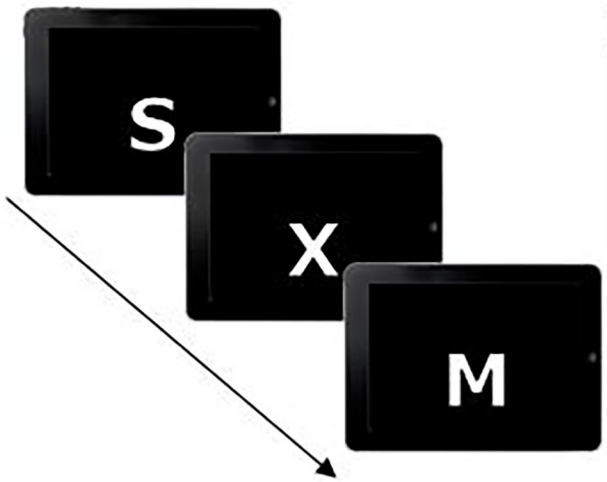
Each version of the task had two stages: Preexposure and Test. After reading an information sheet and signing a consent form, the following instructions were presented to participants on the computer monitor prior to the task:

> *"In this task you will see a sequence of letters appearing on the screen. Your task is to press the response button at the bottom of the screen each time the current letter is the same as the one that was presented before last, which is 2 positions back in the sequence. Otherwise, do not respond. When this task ends, you will be given a new set of instructions. Press the arrow below when you are ready to begin."*

During the preexposure stage the preexposed stimulus was presented 20 times, intermixed in a random order with presentations of filler letters each of which was presented 15 times; each stimulus was presented for 1,000 ms separated by a 150 ms inter-stimulus interval. The non-preexposed stimulus and target letter (X or Z) were not presented during the preexposure stage. Following completion of the pre-exposure phase, participants were presenting with a new set of instructions prior to the test phase:

> *"In this task you will see a sequence of letters appearing on the screen. Your task is to try and predict when a letter 'X' is going to appear. If you think you know when the 'X' will appear then you can press the response button early in the sequence, which is before the 'X' appears on screen. Alternatively, if you are unable to do this please press the response button as quickly as possible when you see the letter 'X'. There may be more than one rule that predicts the 'X'. Please try to be as accurate as you can, but do not worry about making the occasional error. If you understand the task, please press the arrow below when you are ready to begin."*

The test stage instructions were the same for the second version of the latent inhibition task but with the instruction to predict the letter "Z" rather than "X." For the test stage, the preexposed stimulus and the non-preexposed stimulus were each presented 20 times followed by a 1,000 ms presentation of the target stimulus. There were also 20 non-cued presentations of either "X"



Preexposure Stage	Test Stage
Preexposed Stimulus	*Training*
S (20)	S -> X (20)
	H -> X (20)
Filler Trials	*Filler Trials*
D (15)	D -> X (5) D (64)
M (15)	M -> X (5) M (64)
T (15)	T -> X (5) T (64)
V (15)	V -> (5) V (64)

FIGURE 1 | Experimental design and example stimuli for the test stage of the latent inhibition task. Each trial comprised a 1,000 ms presentation of a stimulus separated by an inter-stimulus interval (ISI) of 150 ms. Participants were required to use the computer mouse to click the button on-screen either when the target stimulus "X" appeared on screen, or before it appeared if they could predict it as the next letter in the sequence. The preexposed (PE) and non-preexposed (NPE) stimuli were counterbalanced across participants. Numbers in parentheses in the insert refer to trial frequencies.

or "Z" during which the target was preceded by one of the four filler letters, each of which preceding the target five times. In total there were 64 presentations of the filler letters throughout the test phase. The whole task lasted 7 min.

Reaction times (RTs) in the test stage were recorded from the onset of the preexposed and non-preexposed stimulus that preceded the target letter (X or Z) for each participant. Each stimulus was presented for 1,000 ms separated by a 150 ms inter-stimulus interval. Reaction times could range from 0 to 2,150 ms; reaction times <1,150 ms, indicated participants predicted the occurrence of the target as the next letter in the sequence. Whereas, reaction times between 1,150 and 2,150 ms, indicated participants responded to the target when it appeared on screen. Median reaction times for responses to the preexposed and non-preexposed stimuli were calculated for each participant as the median is less biased by extreme values compared to the mean. Correct responses were also calculated for each individual. If the participant had predicted the target (i.e., they had pressed the spacebar on the letter immediately preceding the target) it was deemed that this was a correct response. For each participant the number of correct responses to the preexposed and non-preexposed stimuli were counted separately for each stimulus type (preexposed and non-preexposed).

Procedure

Prior to the session, a telephone screen assessed basic eligibility. Eligible participants attended a single test session, at which full written informed consent was obtained and further screening assessments were conducted. If eligibility was met, baseline questionnaire (STICSA, PANAS, and O-LIFE) and cardiovascular [blood pressure (BP) and heart rate (HR)] measures were recorded. The inhalation began with 60 s of free breathing before the tasks were started (this allowed for the gas to start taking effect before data collection began). Inhalations then continued for the duration of the latent inhibition task (up to 20 min for each inhalation). Immediately after each inhalation, measures of BP, HR, STICSA, and PANAS were completed, and there was a 30-min washout period between gas inhalations. The second inhalation followed the same procedure as the first. After the inhalations were complete, participants remained in the room for a minimum of 20 min, to allow any effects to dissipate. Participants were then debriefed and reimbursed £20. A follow-up call was conducted 24 h later to assess whether any adverse events had occurred.

Results

Characteristics of Participants

The participants (n = 18; 60% female) were between 19 and 32 years of age (M = 23, SD = 3.4). STICSA state and trait baseline scores ranged between 21 and 50 (M = 28, SD = 7) and between 2 and 31 (M = 25, SD = 5), respectively. Baseline PANAS scores ranged between 21 and 43 (M = 25, SD = 5) and for the sub-dimensions of O-LIFE: Unusual Experiences (positive schizotypy); 0 and 19 (M = 4, SD = 5), Cognitive Disorganization; 0 and 18 (M = 7, SD = 6), Introvertive Anedonia (negative schizotypy); 1 and 11 (M = 4, SD = 3) Impulsive Non-conformity; 0 and 11 (M = 6, SD = 2). O-LIFE scores were relatively comparable to normative values and those reported in previous studies (44) demonstrating baseline schizotypy scores representative of a healthy sample.

Subjective and Cardiovascular Effects

State anxiety (STICSA), negative affect (PANAS-negative), SBP, DBP, and HR were higher, and positive affect (PANAS-positive) was lower, after CO_2 than after medical air inhalation (see

TABLE 1 | State anxiety, affect, and cardiovascular function show significant differences during CO_2 vs. air inhalation (paired *t*-test comparisons).

	Mean difference (SD): CO_2 vs. air	Effect size (Cohen's d)	*df*	95% CI	*p*-value
STICSA state	10.33 (11.11)	0.95	29	−6.18 to −14.48	0.001
PANAS-positive	−5.23 (4.92)	0.67	29	7.07–3.39	0.001
PANAS-negative	2.73 (3.76)	0.55	29	−1.33 to −4.13	0.001
Systolic BP	9.77 (10.67)	0.75	29	−5.79 to −13.75	0.001
Diastolic BP	1.60 (4.11)	0.18	29	−0.07 to −3.14	0.041
Heart rate	8.27 (10.57)	0.64	29	−4.32 to −12.21	0.001

STICSA, State-Trait Inventory for Cognitive and Somatic Anxiety; PANAS, Positive and Negative Affect Schedule; SBP, systolic blood pressure; DBP, diastolic blood pressure; HR, heart rate.

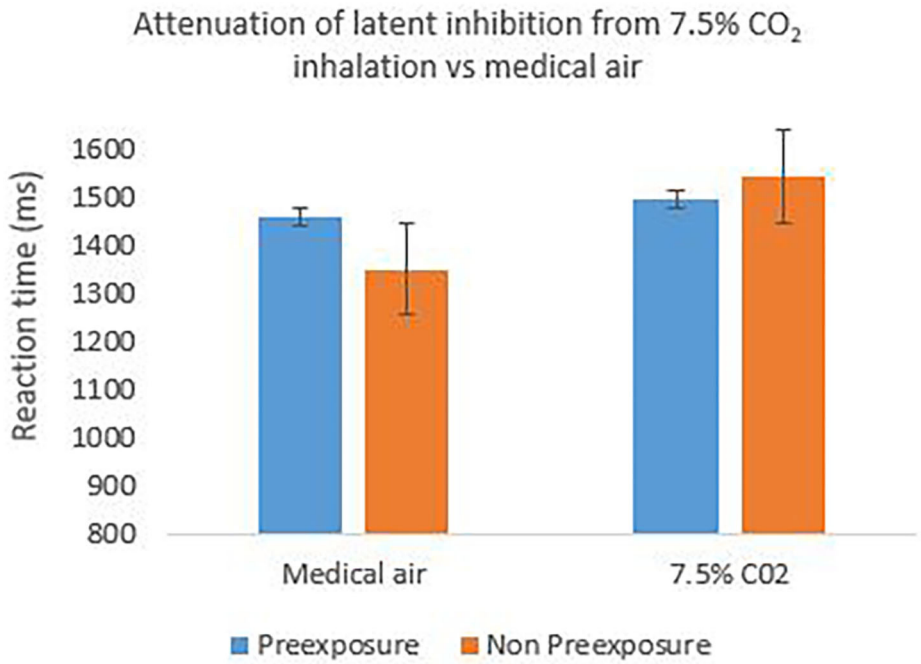


FIGURE 2 | The mean reaction time to the target cued by preexposed stimuli and non-preexposed stimuli in the test stage of the latent inhibition task. Successful effect of latent inhibition is seen in the medical air condition, but attenuated or reversed during CO_2 inhalation. Error bars are 1± within-subject standard error of the mean [see (45)].

Table 1), confirming the validity of the manipulation to induce state anxiety. Importantly, at baseline, there were no significant differences between conditions (CO_2 vs. medical air) for any subjective or cardiovascular event using independent sample *t*-tests (all $p > 0.45$).

Latent Inhibition: Reaction Time

Figure 2 shows the group mean of individual median reaction times to the target (X or Z) across the 20 test trials for the preexposed and non-preexposed stimuli. For the medical air condition, it can be seen that reaction times were slightly faster during the non-preexposed than the non-preexposed stimulus trials, indicating successful induction of the expected latent inhibition effect. In the CO_2 condition however, the effect is, if anything, in the reverse direction indicating slightly faster reaction times to the preexposed stimulus. This impression was explored using a 2 (stimulus: preexposed, non-preexposed) × 2 (gas: CO_2, medical air) repeated measures analysis of variance (ANOVA) on individual median reaction times, which revealed a significant main effect of stimulus $F_{(1, 29)} = 7.718$, $p = 0.009$, partial $\eta^2 = 0.210$ indicating an overall effect of latent inhibition; there was no significant main effect of gas ($F < 1$). Pre-planned comparisons revealed a significant effect of stimulus in the medical air $F_{(1, 29)} = 8.440$, $p = 0.0017$ partial $\eta^2 = 0.225$ but not the CO_2 condition $F_{(1, 29)} = 1.875$, $p = 0.181$; confirming an effect of latent inhibition observable in the anticipated direction in the medical air condition, and an absence of this effect, in the CO_2 condition, see **Figure 2**. The overall 2-way interaction (stimulus × gas) was however not significant ($F < 1$).

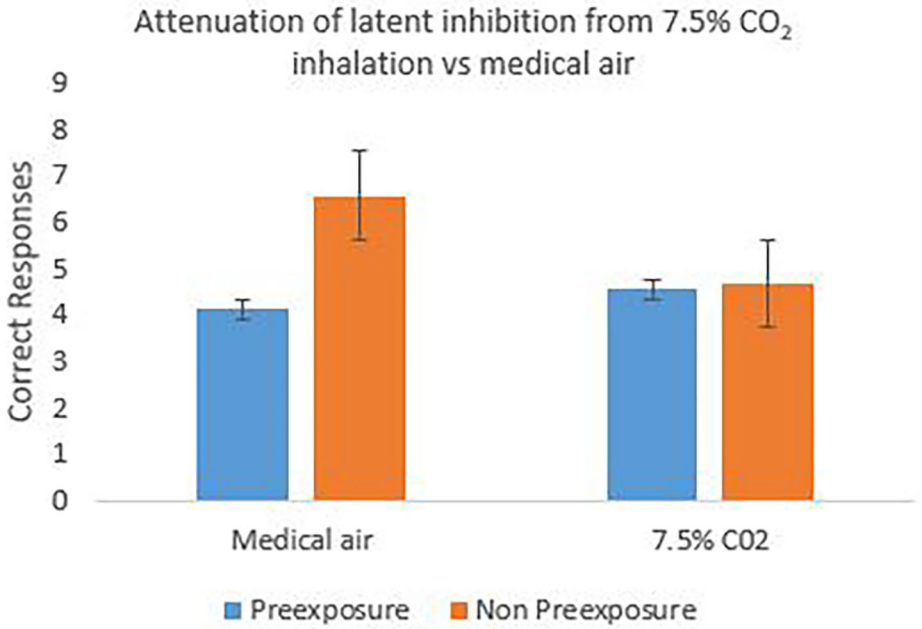


FIGURE 3 | The mean number of correct responses to the target cued by preexposed stimuli and non-preexposed stimuli in the test stage of the latent inhibition task. Successful effect of latent inhibition is seen in the medical air condition, but attenuated or reversed during CO_2 inhalation. Error bars are 1± within-subject standard error of the mean [see (45)].

Latent Inhibition: Correct Responses

Figure 3 shows the group mean of individual correct responses to the target (X or Z) across the 20 test trials with the preexposed and non-preexposed stimuli. For the medical air condition, it can be seen that correct responses were higher for the non-preexposed than the preexposed stimulus trials, illustrating a potential effect of latent inhibition. In the CO_2 condition, by contrast, the amount of correct responses to both preexposed and non-preexposed stimuli appear relatively equal, indicating an absence of latent inhibition. This impression was confirmed with pre-planned comparisons revealing a significant effect of stimulus (preexposed vs. non-preexposed) only in the medical air condition, $F_{(1, 29)} = 5.805$, $p = 0.023$, partial $\eta^2 = 0.167$, indicating the presence of latent inhibition. There was no significant effect of stimulus in the CO_2 condition $F_{(1, 29)} = 0.011$, $p = 0.919$, indicating the absence of this effect (see **Figure 3**) in this sample of participants. There was however no overall main effect of stimulus $F_{(1, 29)} = 2.690$, $p = 0.112$ or of gas using a 2 (stimulus: preexposed, non-preexposed) × 2 (gas: CO_2, medical air) repeated measures ANOVA but the overall 2-way interaction between stimulus × gas approached significance $F_{(1, 29)} = 3.554$, $p = 0.069$, partial $\eta^2 = 0.109$.

Discussion

Experiment 1 was successful in using 7.5% CO_2 inhalation vs. medical air inhalation to induce state anxiety with results were in the anticipated direction: state anxiety measured by the STICSA (42) was significantly higher following CO_2 inhalation with a large Cohen's d effect size. In addition, negative affect as measured by the PANAS (43), heart rate, systolic and diastolic blood pressure were all significantly higher following the inhalation of CO_2, with generally large effect sizes. The validity of this manipulation to induce state anxiety is in line with previous research findings [see (38)].

Using both reaction time and correct response data, the results indicated that an effect of latent inhibition (faster/better learning to the non-preexposed stimuli compared to the preexposed stimuli) was only observable during the inhalations of medical air. During the 7.5% CO_2 inhalations, the effect of latent inhibition was absent, which was particularly prominent when correct responses were used as the dependent variable. Interestingly, the absence of the latent inhibition effect in the CO_2 condition seems to be primarily driven by a reduction in learning to the non-preexposed stimulus, indicating an observation of an induced learning deficit by CO_2 exposure. The lack of overall interaction however between latent inhibition and gas condition is potentially due to a lack of power, as the sample size of the current study was relatively small. To increase the power of the study e.g., to 95%, we recommend the use of $N = 40$ in future studies to obtain a moderate effect size of $dz = 0.6$ at an alpha level of 5%. The direction of the current results nevertheless provide support for the Experiment 1 hypothesis and existing research that reports an absence and/or attenuation of latent inhibition under state anxiety, and by extension, augmented dopaminergic conditions relevant to schizophrenia [e.g., (10, 18–21)].

EXPERIMENT 2: EFFECTS OF NICOTINE ON LATENT INHIBITION IN NON-SMOKERS

Materials and Methods

Design

To assess the sensitivity of latent inhibition to a pro-cognitive pharmacological manipulation, Experiment 2 evaluated latent inhibition in healthy non-smoking volunteers who received a 2 mg dose of nicotine or placebo in a single-blind crossover design with 2-day washout between treatment administrations.

Participants

Twenty non-smoking healthy volunteers were recruited from among members of the University of Bristol and the local community *via* email lists, poster and fllier advertisements and the Tobacco and Alcohol Research Group newsletter and website. Non-smokers were defined as not having smoked in the past 12 months, and not smoked more than 100 cigarettes in their lifetime. The exclusion criteria were age under 18 or over 50 years, pregnancy or breast feeding, recent use of prescribed or illicit drugs, uncorrected visual or hearing problems. Participants refrained from consuming alcohol for 24 h prior to the study day and were required to refrain from caffeine consumption on test days prior to assessments. Expired breath alcohol and carbon monoxide readings were taken, and participants were to be excluded if the readings were >0 or ≥ 10, respectively. No candidate participants had to be excluded from the research. The study was approved by the University of Bristol Faculty Of Science Research Ethics Committee. Sample size was determined based on a previous study of a similar nature (46).

Questionnaires and Latent Inhibition Task

A 12-item visual analog scale (VAS) was used to assess aversive effects of nicotine (nausea, dizziness, sweatiness, light-headed, nervous, headache, heart racing, indigestion, tight-throat, increased saliva, change in taste, fatigue), which relate to the most common side effects associated with acute nicotine administration reported in previous studies (47, 48). Additional questionnaires included the STICSA as measure of state and trait anxiety (42) and the O-LIFE as measure of schizotypy (44) to ensure baseline schizotypy scores were within normative range. The modified version of the Granger et al. (36) latent inhibition task was used, as described in Experiment 1.

Procedure

Eligible participants attended two sessions (minimum 2 days apart) at approximately the same time of day. After providing informed consent at the first testing session, further screening assessments were conducted and an expired CO test using a piCO smokelyser (Bedfont Scientific Ltd.) was used to rule out recent smoking. Baseline questionnaire measures (VAS, O-LIFE and STICSA) were then completed, after which participants were administered either 2 mg nicotine mouth spray or placebo (peppermint mouth spray, Boots UK). Treatment administration was single-blind and order of administration was counterbalanced across participants. Following administration, participants were required to sit quietly for 30 min to allow peak plasma nicotine levels to be reached. After which, the latent inhibition task was completed, followed by the self-report questionnaires. Prior to the second session, there was a washout period for a minimum of 2 days. The second session followed the same procedure as the first but delivered the alternative treatment (i.e., nicotine or placebo). At the end of the second session participants were debriefed and reimbursed £30.

Results

Characteristics of Participants

The participants ($n = 12$; 60% female) were between 18 and 39 years of age (M = 23, SD = 4.6). STICSA trait baseline scores ranged between 22 and 42 (M = 31, SD = 6) and the sub-dimensions of O-LIFE between: Unusual Experiences (positive schizotypy); 0 and 15 (M = 5, SD = 4), Cognitive Disorganization; 0 and 18 (M = 8, SD = 6), Introvertive Anhedonia (negative schizotypy); 1 and 19 (M = 7, SD = 5) Impulsive Non-conformity; 1 and 14 (M = 6, SD = 3). O-LIFE scores were relatively comparable to normative values and those reported in previous studies (44) demonstrating baseline schizotypy scores representative of a healthy sample.

Subjective Effects (Nicotine vs. Placebo)

State anxiety (STICSA) and each of the VAS scores were higher after nicotine than after placebo (see **Table 2**), indicating that participants experienced the commonly experienced aversive effects of nicotine administration. At baseline, there were no significant differences between treatment groups (nicotine vs. placebo) for any of the subjective self-report measures (STICSA and VAS scores), derived using independent sample *t*-tests (all $p > 0.07$).

Latent Inhibition: Reaction Time

Figure 4 shows the group mean of individual median reaction times to the target (X or Z) across the 20 test trials with the preexposed and non-preexposed stimuli. For the nicotine condition, reaction times were faster during the non-preexposed than the preexposed stimulus trials, indicating an effect of latent inhibition compared to the placebo group. This impression was explored using a 2 (stimulus: preexposed, non-preexposed) × 2 (treatment: nicotine, placebo) repeated measures ANOVA on individual median reaction times, which revealed a significant main effect of stimulus $F_{(1,19)} = 6.246$, $p = 0.002$, partial $\eta^2 = 0.247$, indicating an overall effect of latent inhibition; there was no main effect of treatment ($F < 1$). Pre-planned comparisons revealed a significant effect of stimulus only in the nicotine treatment $F_{(1,19)} = 7.288$, $p = 0.014$, partial $\eta^2 = 0.277$, indicating an effect of latent inhibition. There was however no significant effect of stimulus in the placebo arm ($F < 1$), see **Figure 4**, indicating an absence/reduction of the effect compared to the nicotine treatment. The overall 2-way interaction (stimulus × treatment) was however not significant ($F < 1$).

Latent Inhibition: Correct Responses

Figure 5 shows the group mean of individual correct responses to the target (X or Z) across the 20 test trials with the preexposed and non-preexposed stimuli. For the nicotine

TABLE 2 | State anxiety and subjective measures demonstrate anticipated aversive effects of 2 mg nicotine vs. placebo in non-smokers (paired *t*-test comparisons).

	Mean difference (SD): nicotine vs. placebo	Effect Size (Cohen's d)	*df*	95% CI	*p*-value
STICSA state	3.25 (5.53)	0.54	19	0.66–5.84	0.017
Dizziness	23.60 (28.98)	1.11	19	10.04–37.16	0.002
Fatigue	1.85 (14.02)	0.08	19	−4.71–8.41	0.562
Headache	8.10 (20.26)	0.35	19	−1.38–17.58	0.090
Heart racing	11.00 (15.25)	0.69	19	3.86–18.14	0.004
Indigestion	2.45 (8.65)	2.44	19	−1.60–6.50	0.221
Nausea	13.90 (24.56)	0.74	19	2.41–25.39	0.020
Nervousness	7.90 (19.49)	0.54	19	−1.22–17.02	0.086
Salivation	7.65 (15.89)	0.38	19	0.21–15.09	0.044
Sweatiness	11.20 (19.47)	0.77	19	2.09–20.31	0.019
Taste	10.05 (20.75)	0.51	19	0.34–19.76	0.043
Throat-tightness	22.45 (31.41)	1.00	19	7.75–37.15	0.005

STICSA, State-Trait Inventory for Cognitive and Somatic Anxiety.

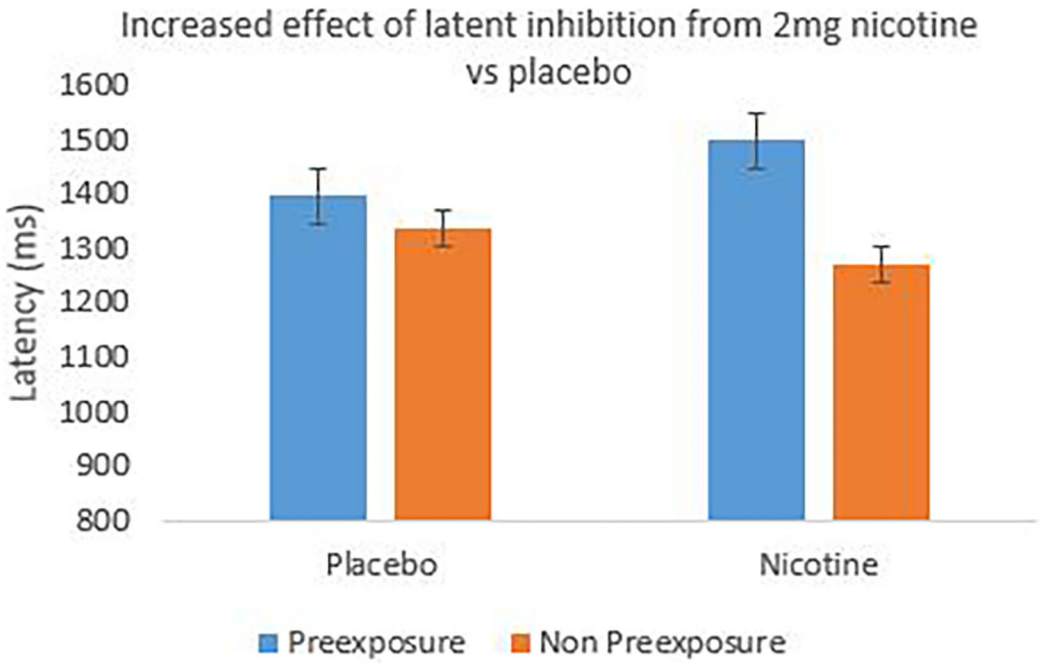


FIGURE 4 | The mean reaction time to the target cued by preexposed stimuli and non-preexposed stimuli in the test stage of the latent inhibition task. A potentiation of latent inhibition is seen in the nicotine treatment arm which is reduced for the placebo arm. Error bars are 1± within-subject standard error of the mean [see (45)].

treatment arm, correct responses were higher for the non-preexposed than the preexposed stimulus trials, indicating an effect of latent inhibition that appears increased relative to the placebo treatment arm. This impression was explored using a 2 (stimulus: preexposed, non-preexposed) × 2 (treatment: nicotine, placebo) repeated measures ANOVA on individual correct responses, which revealed a significant main effect of stimulus $F_{(1, 19)} = 7.563$, $p = 0.013$, partial $\eta^2 = 0.285$ indicating an overall effect of latent inhibition; there was no significant main effect of treatment ($F < 1$). Pre-planned comparisons revealed a significant effect of stimulus only in the nicotine treatment arm, $F_{(1, 19)} = 6.717$, $p = 0.018$, partial $\eta^2 = 0.261$, indicating an effect of latent inhibition. There was however no significant effect of stimulus in the placebo arm ($F < 1$), see **Figure 5**. The overall 2-way interaction (stimulus × treatment) was however not significant ($F < 1$).

Discussion

This experiment confirmed that nicotine (vs. placebo) induced the commonly experienced aversive effects in non-smokers [cf. (49)], in particular, state anxiety, racing heart, nervousness, sweatiness, and throat-tightness. Both reaction time and correct

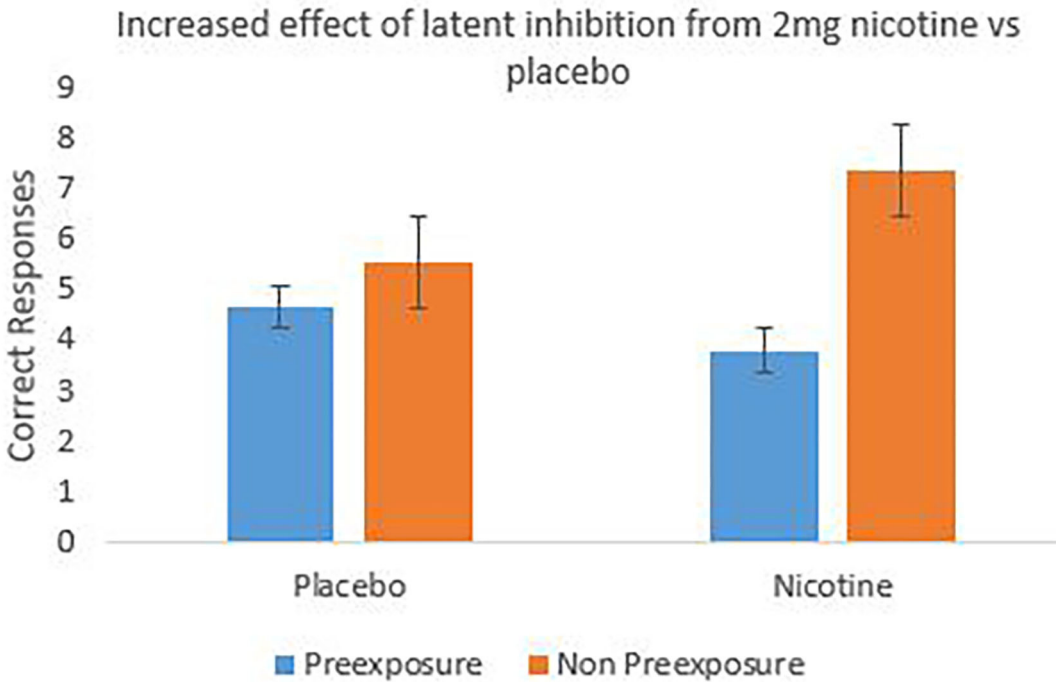


FIGURE 5 | The mean number of correct responses to the target cued by preexposed stimuli and non-preexposed stimuli in the test stage of the latent inhibition task. A potentiation of latent inhibition is seen in the nicotine treatment arm which is reduced for the placebo arm. Error bars are 1± within-subject standard error of the mean [see (45)].

response data confirmed an overall effect of latent inhibition. Nicotine treatment appeared to produce a greater degree of latent inhibition than the placebo arm (see **Figure 5** in particular) indicated by the significant effect of stimulus relative to nicotine administration but not placebo. Whilst the effect of stimulus in the placebo arm was not significant, the anticipated direction of effect for latent inhibition was observable (in particular more correct responses to the non-preexposed stimuli, compared to the preexposed stimuli; **Figure 5**). This lack of/small demonstration of latent inhibition in the placebo arm could potentially be a result of participant's anticipation of receiving nicotine, generating a compensatory response that is agonistic to the normal effect of nicotine [see e.g., (50)]. In line with this, the anticipatory effect would then also presumably be present in the nicotine treatment condition, but overcome by the pharmacological effect of nicotine itself, as illustrated by the presence of latent inhibition. The statistical exploration of this in the current study e.g., by exploring the order effects of treatment administration on latent inhibition is not however attainable due to restrictions on sample size. To increase the power of the study e.g., to 95%, we recommend the use of $N = 40$ in future studies to obtain a moderate effect size of $dz = 0.6$ at an alpha level of 5%. Nevertheless, the current finding which illustrates an observable effect of latent inhibition from nicotine administration, compared to placebo, provides support for existing research [e.g., (31)]. It would be of interest for future research to explore differences in latent inhibition to e.g., 2 vs. 4 mg of nicotine to establish dose sensitivity of the latent inhibition effect. In addition, to further understand the effects of nicotine on latent inhibition, a larger future research study could recruit smokers and non-smokers to evaluate whether a reduced effect of latent inhibition potentiation by nicotine is observed in those who already smoke cigarettes, compared to those who do not [cf. (51)].

General Discussion

Inhalation of 7.5% CO_2 raised self-report and physiological measures of anxiety and impaired latent inhibition compared to medical air control; whereas administration of nicotine demonstrated an increased effect of latent inhibition, compared to placebo control. Given supporting evidence that state anxiety increases dopamine (39, 40), the aim of Experiment 1 was to explore the sensitivity of the modified latent inhibition task (36) to an in-direct model relevant to psychosis (positive symptoms associated with schizophrenia) by using the 7.5% CO_2 challenge as a way to stimulate dopamine release *via* induction of state anxiety. In contrast, Experiment 2 aimed to explore the sensitivity of latent inhibition to a pro-cognitive model relevant to the treatment of cognitive impairment associated with schizophrenia by examining the effect of nicotine administration on latent inhibition vs. placebo. The results from these initial experiments suggest latent inhibition warrants further investigation as a potential biomarker for schizophrenia [see also (24)].

Given the sensitivity of latent inhibition to dopaminergic manipulations as seen from existing research [see (10)], and by corollary underlying dopaminergic perturbations observed

in psychosis patients (9), further studies should assess whether latent inhibition can be used as a tool to help identify patients and also accelerate or rationalize treatment strategies for patients with psychotic disorders to support decision making. With no biomarker currently available to identify, for example, individuals at ultra-high risk (UHR) for developing psychosis, a means to do so would allow anti-psychotic treatment to be initiated at an earlier stage to reduce the risk of conversion to a full-blown state of psychosis. Currently, treatment for psychosis is not initiated until the first full episode of the disorder emerges (52), and is thus rarely (if at all) provided to UHR individuals. Given existing research supporting the sensitivity of latent inhibition, it has the potential, with further clinical validation, to act as a surrogate marker to detect underlying neurotransmitter perturbations and provide a non-invasive proxy measure of e.g., hyper-dopaminergic state to identify which individuals would, along biological lines, be suited to receiving a dopamine blocker (the mainstream anti-psychotic treatment) to remediate psychosis, or a non-dopaminergic alternative. Considering around one third of patients are also classified as treatment resistant [see (53)], it is a major clinical need to identify ways for patients to be fast-tracked to an appropriate treatment, ideally at initial diagnosis depending upon their neurobiology. Experimental investigations should continue to focus on this in future research, particularly as specialist clinical services are well-placed to benefit from novel means for better identification and/or early treatment options for affected individuals.

The effect of latent inhibition by nicotine administration compared to placebo observed in Experiment 2, provides encouraging support for existing research demonstrating a potentiation of latent inhibition in smokers compared to non-smokers (32) and for preclinical findings that demonstrate demonstrating pro-cognitive effects of an α7-nAChR partial agonist, SSR180711 using latent inhibition as a model to demonstrate treatment efficacy [see (33)]. Given the sensitivity of latent inhibition to cholinergic manipulations and associated neurobiological disruptions, future research should investigate the sensitivity of latent inhibition as a stratification tool to identify the sub-population of patients with schizophrenia that could benefit from pro-cognitive treatment with a α7-nAChR agonist. Despite the biological complexity and heterogeneity of schizophrenia, inclusion criterion for previous clinical trials investigating these potentially pro-cognitive drugs have relied on subjective diagnoses and self-report measures (i.e., the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5). Since DSM-5 criteria neither determine the presence of cognitive impairments cognitive ability nor classify according to underlying neurobiological abnormalities, it is not surprising that these drugs have failed to universally improve cognition among such a heterogeneous group. To date, 87 novel agents have been unsuccessfully trialed for cognitive impairment associated with schizophrenia [see (54)]: a tool to enhance the prediction of treatment efficacy for a core area of schizophrenia where no treatments currently exist has the potential to greatly improve the chances of an effective drug becoming available.

Conclusions

The experiments reported here provide initial research findings that support the potential utility and sensitivity of latent inhibition to relevant manipulations which underpin key neurobiological dysfunctions and symptoms associated with schizophrenia; a tool that is sensitive to these neurobiological states and associated treatment-induced changes holds potential to advance schizophrenia research. Latent inhibition holds potential promise as a biomarker/stratification tool for use in both clinical practice and clinical development for patients that are in need of improved means of illness detection, and improved efficacy of treatment options and outcomes.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Bristol, Faculty of Science Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KG and JB designed the study protocol. KG drafted the manuscript. JF carried out recruitment and experimental testing with oversight and project management from AA. All authors critically reviewed and approved the manuscript prior to its submission for publication.

FUNDING

This research did not receive any specific grant from any funding agencies. The studies were supported by Cambridge Cognition Ltd.

REFERENCES

1. Lubow RE. Latent inhibition. *Psychol Bull.* (1973) 79:398–407. doi: 10.1037/h0034425
2. Hall G, Honey RC. Contextual effects in conditioning, latent inhibition, and habituation: Associative and retrieval functions of contextual cues. *J Exp Psychol Anim Behav Processes.* (1989) 15:232–41. doi: 10.1037/0097-7403.15.3.232
3. Moser PC, Hitchcock JM, Lister S, Moran PM. The pharmacology of latent inhibition as an animal model of schizophrenia. *Behav Brain Res.* (2000) 33:275–307. doi: 10.1016/S0165-0173(00)00026-6
4. Arad M, Weiner I. Disruption of latent inhibition induced by ovariectomy can be reversed by estradiol and clozapine as well as by co-administration of haloperidol with estradiol but not by haloperidol alone. *Psychopharmacology.* (2009) 206:731–40. doi: 10.1007/s00213-009-1464-0

5. Lubow RE, Moore AU. Latent inhibition: the effect of nonreinforced preexposure to the conditional stimulus. *J Comp Physiol Psychol.* (1959) 52:415–9. doi: 10.1037/h0046700
6. Lubow RE, Gewirtz JC. Latent inhibition in humans: data, theory, and implications for schizophrenia. *Psychol Bull.* (1995) 117:87–103. doi: 10.1037/0033-2909.117.1.87
7. Mackintosh NJ. Blocking of conditioned suppression: role of the first compound trial. *J Exp Psychol Anim Behav Processes.* (1975) 1:335–45. doi: 10.1037/0097-7403.1.4.335
8. Pearce JM, Hall G. A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol Rev.* (1980) 87:532–52. doi: 10.1037/0033-295X.87.6.532
9. Lubow RE, Weiner I. Issues in latent inhibition research and theory: an overview. In: Lubow RE, Weiner I, editors. *Latent Inhibition: Cognition, Neuroscience, and Applications to Schizophrenia.* New York, NY: Cambridge University Press (2010). p. 531–557.
10. Weiner I. The "two-headed" latent inhibition model of schizophrenia: modeling positive and negative symptoms and their treatment. *Psychopharmacology.* (2003) 169:257–97. doi: 10.1007/s00213-002-1313-x
11. Weiner I, Arad M. Using the pharmacology of latent inhibition to model domains of pathology in schizophrenia and their treatment. *Behav Brain Res.* (2009) 204:369–86. doi: 10.1016/j.bbr.2009.05.004
12. Abi-Dargham A, Gil R, Krystal J, Baldwin RM, Seibyl JP, Bowers M, et al. Increased striatal dopamine transmission in schizophrenia: confirmation in a second cohort. *Am J Psychiatry.* (1988) 155:761–7.
13. Bramness JG, Rognli EB. Psychosis induced by amphetamines. *Curr Opin Psychiatry.* (2016) 29:236–41. doi: 10.1097/YCO.0000000000000254
14. Solomon PR, Crider A, Winkelman JW, Turi A, Kramer RM, Kaplam LJ. Disrupted latent inhibition in the rat with chronic amphetamine or haloperidol-induced supersensitivity: relationship to schizophrenic attention disorder. *Biol Psychiatry.* (1981) 16:529–37.
15. Weiner I, Lubow RE, Feldon J. Disruption of latent inhibition by acute administration of low doses of amphetamine. *Pharmacol Biochem Behav.* (1988) 30:871–8. doi: 10.1016/0091-3057(88)90113-X
16. Amador M, Dani JA. MK-801 inhibition of nicotinic acetylcholine receptor channels. *Synapse.* (1991) 7:207–15. doi: 10.1002/syn.890070305
17. Lewis MC, Gould TJ. Latent inhibition of cued fear conditioning: an NMDA receptor-dependent process that can be established in the presence of anisomycin. *Eur J Neurosci.* (2004) 20:818–26. doi: 10.1111/j.1460-9568.2004.03531.x
18. Baruch I, Hemsley DR, Gray JA. Differential performance of acute and chronic schizophrenics in a latent inhibition task. *J Nerv Ment Dis.* (1988) 176:598–606. doi: 10.1097/00005053-198810000-00004
19. Gray NS, Hemsley DR, Gray JA. Abolition of latent inhibition in acute but not chronic schizophrenics. *Neurol Psychiatry Brain Res.* (1992) 1:83–9.
20. Rascle C, Mazas O, Vaiva G, Tournant M, Raybois O, Goudemand M, et al. Clinical features of latent inhibition in schizophrenia. *Schizophr Res.* (2001) 51:149–61. doi: 10.1016/S0920-9964(00)00162-6
21. Vaitl D, Lipp O, Bauer U, Schuler G, Stark R, Zimmermann M, et al. Latent inhibition and schizophrenia: Pavlovian conditioning of autonomic responses. *Schizophr Res.* (2002) 55:147–58. doi: 10.1016/S0920-9964(01)00250-X
22. Cohen E, Sereni N, Kaplan O, Weizman A, Kikinzon L, Weiner I, et al. The relation between latent inhibition and symptom-types in young schizophrenics. *Behav Brain Res.* (2004) 149:113–22. doi: 10.1016/S0166-4328(03)00221-3
23. Gal G, Barnea Y, Biran L, Mendlovic S, Gedi T, Halavy M, et al. Enhancement of latent inhibition in patients with chronic schizophrenia. *Behav Brain Res.* (2009) 197:1–8. doi: 10.1016/j.bbr.2008.08.023
24. Granger KT, Talwar A, Barnett JH. Latent inhibition and its potential as a biomarker for schizophrenia. *Biomark Neuropsychiatry.* (2020) 3:100025. doi: 10.1016/j.bionps.2020.100025
25. Marder SR. The NIMH-MATRICS project for developing cognition-enhancing agents for schizophrenia. *Dialogues Clin Neurosci.* (2006) 8:109–13. doi: 10.31887/DCNS.2006.8.1/smarder
26. Hill SK, Bishop JR, Palumbo D, Sweeney JA. Effect of second-generation antipsychotics on cognition: current issues and future challenges. *Expert Rev Neurother.* (2010) 10:43–57. doi: 10.1586/ern.09.143
27. MacKenzie NE. Antipsychotics, metabolic adverse effects, and cognitive function in schizophrenia. *Front Psychiatry.* (2018) 9:622. doi: 10.3389/fpsyt.2018.00622
28. Kumari V, Postma P. Nicotine use in schizophrenia: the self medication hypotheses. *Neurosci Biobehav Rev.* (2005) 29:21–34. doi: 10.1016/j.neubiorev.2005.02.006
29. Wootton R, Richmond R, Stuijfzand B, Lawn R, Sallis H, Taylor G, et al. Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. *Psychol Med.* (2019) 50:1−9. doi: 10.1017/S0033291719002678
30. Smucny J, Tregellas JR. Targeting neuronal dysfunction in schizophrenia with nicotine: evidence from neurophysiology to neuroimaging. *J Psychopharmacol.* (2017) 3:801–11. doi: 10.1177/0269881117705071
31. Thornton JC, Dawe S, Lee C, Capstick C, Corr PJ, Cotter P, et al. Effects of nicotine and amphetamine on latent inhibition in human subjects. *Psychopharmacology.* (1996) 127:164–73. doi: 10.1007/BF02805990
32. Della Casa V, Weiner I, Feldon J. Effects of smoking status and schizotypy on latent inhibition. *J Psychopharmacol.* (1999) 13:45–57. doi: 10.1177/026988119901300106
33. Barak S, Arad M, De Levie A, Black MD, Griebel G, Weiner I. Pro-cognitive and antipsychotic efficacy of the A7 nicotinic partial agonist SSR180711 in pharmacological and neurodevelopmental latent inhibition models of schizophrenia. *Neuropsychopharmacology.* (2009) 34:1753–63. doi: 10.1038/npp.2008.232
34. Shiina A, Shirayama Y, Niitsu T, Hashimoto T, Yoshida T, Hasegawa T, et al. A randomised, double-blind, placebo-controlled trial of tropisetron in patients with schizophrenia. *Ann Gen Psychiatry.* (2010) 9:1–10. doi: 10.1186/1744-859X-9-27
35. Jones LA, Hills PJ, Dick KM, Jones SP, Bright P. Cognitive mechanisms associated with auditory sensory gating. *Brain Cogn.* (2016) 102:33–45. doi: 10.1016/j.bandc.2015.12.005
36. Granger KT, Moran PM, Buckley MG, Haselgrove M. Enhanced latent inhibition in high schizotypy individuals. *Pers Individ Dif.* (2016) 91:31–9. doi: 10.1016/j.paid.2015.11.040
37. Garner M, Attwood A, Baldwin DS, James A, Munafo M. Inhalation of 7.5% carbon dioxide increases threat processing in humans. *Neuropsychopharmacology.* (2011) 36:1557–62. doi: 10.1038/npp.2011.15
38. Easey KE, Catling JC, Kent C, Crouch C, Jackson S, Munafò MR, et al. State anxiety and information processing: a 7.5% carbon dioxide challenge study. *Psychon Bull Rev.* (2018) 25:732–8. doi: 10.3758/s13423-017-1413-6
39. Mizrahi R, Addington R, Rusjan I, Ng A, Boileau I, Pruessner JC, et al. Increased stress-induced dopamine release in psychosis. *Biol Psychiatry.* (2012) 71:561–7. doi: 10.1016/j.biopsych.2011.10.009
40. Nagano-Saito A, Dagher A, Booij L, Gravel P, Welfeld K, Casey KF, et al. Stress-induced dopamine release in human medial prefrontal cortex−18F-fallypride/PET study in healthy volunteers. *Synapse.* (2013) 67:821–30. doi: 10.1002/syn.21700
41. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The mini-international neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry.* (1998) 59:22–33. doi: 10.1037/t18597-000
42. Ree MJ, French D, MacLeod C, Locke V. Distinguishing cognitive and somatic dimensions of state and trait anxiety: development and validation of the state-trait inventory for cognitive and somatic anxiety (STICSA). *Behav Cogn Psychother.* (2008) 36:313–32. doi: 10.1017/S1352465808004232
43. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol.* (1988) 54:1063–70. doi: 10.1037/0022-3514.54.6.1063
44. Mason O, Claridge G, Jackson M. New scales for the assessment of schizotypy. *Personal Individ Differ.* (1995) 18:7–13. doi: 10.1016/0191-8869(94)00132-C
45. Cousineau D. Confidence intervals in within-subject designs: a simpler solution to Loftus and Masson's method. *Tutorial Quantitative Methods Psychol.* (2005) 1:4–45. doi: 10.20982/tqmp.01.1.p042
46. Griesar WS, Zajdel DP, Oken BS. Nicotine effects on alertness and spatial attention in non-smokers. *Nicotine Tobacco Res.* (2002) 4:185–94. doi: 10.1080/14622200210123617

47. Blank M, Sams C, Weaver MF, Eissenberg T. Nicotine delivery, cardiovascular profile, and subjective effects of an oral tobacco product for smokers. *Nicotine Tob Res.* (2008) 10:417–21. doi: 10.1080/14622200801901880
48. Shahab L, McEwen A, West R. Acceptability and effectiveness for withdrawal symptom relief of a novel oral nicotine delivery device: a randomised cross-over trial. *Psychopharmacology.* (2011) 216:187–96. doi: 10.1007/s00213-011-2204-9
49. Adams S, Attwood A, Munafò M. Effects of nicotine and nicotine expectancy on attentional bias for emotional stimuli. *Nico. Tob Res.* (2015) 17:697–703. doi: 10.1093/ntr/ntu219
50. Siegel S, Baptista MAS, Kim JA, McDonald RV, Weise-Kelly L. Pavlovian psychopharmacology: the associative basis of tolerance. *Exp Clin Psychopharmacol.* (2000) 8:276-93. doi: 10.1037/1064-1297.8.3.276
51. Tregellas JR, Wylie KP. Alpha 7 nicotinic receptors as therapeutic targets in schizophrenia. *Nicotine Tob Res.* (2019) 21:349–56. doi: 10.1093/ntr/nty034
52. Fusar-Poli P, Bonoldi I, Yung AR, Borgwardt S, Kempton MJ, Valmaggia L, et al. Predicting psychosis: meta-analysis of transition outcomes in individuals at high clinical risk. *Arch Gen Psychiatry.* (2012) 69:220–9. doi: 10.1001/archgenpsychiatry.2011.1472
53. Lally J, MacCabe JH. Antipsychotic medication in schizophrenia: a review. *Br Med Bull.* (2015) 14:169–79. doi: 10.1093/bmb/ldv017
54. Cotter J, Barnett JH, Granger K. The use of cognitive screening in pharmacotherapy trials for cognitive impairment associated with schizophrenia. *Front Psychiatry*. (2019) 10:648. doi: 10.3389/fpsyt.2019.00648

Conflict of Interest: KG, SC, and JB are employees of Cambridge Cognition Ltd. KG, SC, and JB are also employees of Monument Therapeutics Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

STUDY PROTOCOL
published: 28 April 2021
doi: 10.3389/fpsyt.2021.611243

Unobtrusive Sensing Technology for Quantifying Stress and Well-Being Using Pulse, Speech, Body Motion, and Electrodermal Data in a Workplace Setting: Study Concept and Design

*Keisuke Izumi[1,2,3], Kazumichi Minato[4], Kiko Shiga[4], Tatsuki Sugio[4], Sayaka Hanashiro[4], Kelley Cortright[4], Shun Kudo[4], Takanori Fujita[3,5,6], Mitsuhiro Sado[4,7], Takashi Maeno[8], Toru Takebayashi[3,9], Masaru Mimura[4] and Taishiro Kishimoto[3,4,10]**

[1] Division of Rheumatology, Department of Internal Medicine, Keio University School of Medicine, Tokyo, Japan, [2] National Hospital Organization Tokyo Medical Center, Tokyo, Japan, [3] Medical AI Center, Keio University, Tokyo, Japan, [4] Department of Neuropsychiatry, Keio University School of Medicine, Tokyo, Japan, [5] Department of Health Policy and Management, Keio University School of Medicine, Tokyo, Japan, [6] World Economic Forum Centre for the Fourth Industrial Revolution Japan, Tokyo, Japan, [7] Center for Stress Research, Keio University, Tokyo, Japan, [8] Human System Design Laboratory, Graduate School of System Design and Management, Keio University, Tokyo, Japan, [9] Department of Preventive Medicine and Public Health, Keio University School of Medicine, Tokyo, Japan, [10] Department of Psychiatry, Donald and Barbara Zucker School of Medicine, New York, NY, United States

OPEN ACCESS

Edited by:
Qiang Luo,
Fudan University, China

Reviewed by:
Anke Maatz,
University of Zurich, Switzerland
Lizhu Luo,
University of Electronic Science and Technology of China, China

***Correspondence:**
Taishiro Kishimoto
taishiro-k@mti.biglobe.ne.jp

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: 28 September 2020
Accepted: 23 March 2021
Published: 28 April 2021

Citation:
Izumi K, Minato K, Shiga K, Sugio T, Hanashiro S, Cortright K, Kudo S, Fujita T, Sado M, Maeno T, Takebayashi T, Mimura M and Kishimoto T (2021) Unobtrusive Sensing Technology for Quantifying Stress and Well-Being Using Pulse, Speech, Body Motion, and Electrodermal Data in a Workplace Setting: Study Concept and Design. Front. Psychiatry 12:611243. doi: 10.3389/fpsyt.2021.611243

Introduction: Mental disorders are a leading cause of disability worldwide. Depression has a significant impact in the field of occupational health because it is particularly prevalent during working age. On the other hand, there are a growing number of studies on the relationship between "well-being" and employee productivity. To promote healthy and productive workplaces, this study aims to develop a technique to quantify stress and well-being in a way that does not disturb the workplace.

Methods and analysis: This is a single-arm prospective observational study. The target population is adult (>20 years old) workers at companies that often engage in desk work; specifically, a person who sits in front of a computer for at least half their work hours. The following data will be collected: (a) participants' background characteristics; (b) participants' biological data during the 4-week observation period using sensing devices such as a camera built into the computer (pulse wave data extracted from the facial video images), a microphone built into their work computer (voice data), and a wristband-type wearable device (electrodermal activity data, body motion data, and body temperature); (c) stress, well-being, and depression rating scale assessment data. The analysis workflow is as follows: (1) primary analysis, comprised of using software to digitalize participants' vital information; (2) secondary analysis, comprised of examining the relationship between the quantified vital data from (1), stress, well-being, and depression; (3) tertiary analysis, comprised of generating machine learning algorithms to estimate stress, well-being, and degree of depression in relation to each set of vital data as well as multimodal vital data.

Discussion: This study will evaluate digital phenotype regarding stress and well-being of white-collar workers over a 4-week period using persistently obtainable biomarkers such as heart rate, acoustic characteristics, body motion, and electrodermal activity. Eventually, this study will lead to the development of a machine learning algorithm to determine people's optimal levels of stress and well-being.

Ethics and dissemination: Collected data and study results will be disseminated widely through conference presentations, journal publications, and/or mass media. The summarized results of our overall analysis will be supplied to participants.

Registration: UMIN000036814

Keywords: adult psychiatry, mental health, occupational & industrial medicine, wearabe sensors, well-being, stress, protocols, depression

INTRODUCTION

Mental disorders are a leading cause of disability worldwide and, among mental disorders, major depressive disorder was ranked number 1 in years lived with disability in 2017 (1). The lifetime prevalence of depression in Japan is estimated at 6.2% (2002–2006 estimate), which makes it the country's most common mental illness (2). The disease costs are also enormous and are estimated to exceed 3.09 trillion Japanese Yen (∼30 billion U.S. dollars) per year (3). Depression has a significant impact in the field of occupational health because it is particularly prevalent during working age (20–65 years of age). It is estimated that more than half of the social loss due to depression is attributed to loss of labor productivity through absenteeism and presenteeism (2). The Japanese government has implemented measures against long working hours (Standards on limits of overtime work in 1998; the revision of the Industrial Safety and Health Act in 2006) and has introduced the stress check system (enforced as of December 2015), but there has not been a significant impact from those measures. In fact, the number of workers' compensation claims related to mental disorders are increasing each year.

On the other hand, there are a growing number of studies on the relationship between "well-being" and employee productivity. The happiness of employees has been reported to be associated with creativity and productivity (4). As the birthrate continues to decline and the aging population continues to increase in Japan, the working age population is also decreasing, requiring each employee to make the most of his/her abilities. Therefore, preventing negative factors such as depression and promoting well-being are major challenges for health management in the workplace and ensuring a stable economy.

Conventionally, it is known that heart rate variability (HRV) reflects autonomic nerve activity and serves as an index of psychological and physical stress. There are many suggested indicators for stress, including standard deviation of all normal-to-normal R-R intervals (SDNN) and the percentage of successive R-R intervals that differ by >50 ms (pNN50) with time domain variables and low frequency (LF), high frequency (HF), and their ratio (LF/HF) as frequency domain variables (5).

In addition, techniques for estimating emotions and depressive symptoms have been developed based on the analysis of speech including formant frequencies (6). The autonomic nervous system and voice characteristics are closely related to each other because most of the vocal fold movement is stimulated by the recurrent nerve, which branches off from the vagus nerve. Johannes et al. reported that speech fundamental frequency increased with psychological load while there was no significant difference with physical load, suggesting that speech fundamental frequency is a good indicator of psychological stress (7). Nakatsu et al. reported that combination of linear prediction cepstral coefficients and pitch-related characteristics predicted classification of 8 emotions using artificial neural network (8).

Furthermore, electrodermal activity measured by a wearable device can reflect the activity of eccrine sweat glands that are controlled only by sympathetic nerve activity, and is therefore expected to be a stress indicator (9). Previous studies have used the above mentioned approaches to measure subjects' degree of stress; however, this prior research comprises only feasibility studies that have simply verified device performance, and/or studies with only a small number of patients or healthy individuals (10, 11).

Moreover, such approaches are only used to evaluate the so-called "short-term stress" of the study period, and are not necessarily reflective of the effects of medium- to long-term stress. In a workplace environment, it is expected that regardless of whether employees experience temporary stress, there should also be situations where people feel a sense of freedom and accomplishment when a task is completed or a problem overcome. To date, very few studies have revealed the relationship between vital data and mid- to long-term stress and well-being in the workplace (12).

With the development of information and communication technology, similar approaches trying to utilize such biological and/or behavioral data to identify depression are reported recently. For example, in the case of HRV, Dell'Acqua et al. reported that HRV reduction can be the predictor for depression as HRV of individuals with dysphoria and in those with past depression was lower than controls (13). Kemp et al. in their meta-analysis on HRV and antidepressant treatment,

reported that depression was associated with reduced HRV, which decreased with increasing depression severity (14).

We have reported that the timing related speech features can reflect the severity of depression. Speech rate, pause time, and response time showed significant associations with the total score of Hamilton Depression Rating Scale (15, 16). We have also reported that body movement captured by infrared sensor can be reflective of depression severity (15, 17).

Not only using single modality but combining multimodal data and with machine learning approach it may be more realistic to screen depression or to predict severity of depression. Utilizing wrist band-type wearable device that record three-axis acceleration, heart rate, body temperature, and ultraviolet light exposure, we have reported that it was possible to identify patients with depression with an accuracy of 0.76, and to predict depression severity with a 0.61 correlation coefficient with Hamilton Depression Rating Scale score (15, 18).

This study, which is funded by the Japan Agency for Medical Research and Development (AMED), is an industry-academia collaborative research project that aims to develop new techniques for evaluating mid- to long-term stress and well-being using technologies that will not obstruct normal work environments. By doing so, we hope to promote healthy workplaces and, in the end, to prevent depression in the prime of life.

Research Objectives

The general aim of this study is to develop a technique to quantify stress and well-being in a way that does not disturb the workplace. Our specific objectives are: (1) To evaluate the relationship between the obtained questionnaire-based stress and well-being scores and the employees' vital data, which are collected using: a technique for extracting pulse waves from an image captured by a camera attached to the employee's computer, a technique for extracting emotional components from speech, and a technique for measuring electrodermal activity using a wristband-type wearable device; and (2) to gather information regarding how and when employees are coping to reduce stress or promote enhanced well-being by comparing questionnaire-based stress and well-being scores and the employees' vital data.

METHODS AND ANALYSIS

Study Design

This is a single-arm prospective observational study.

Participant Criteria

Inclusion Criteria

Adult (>20 years old) workers at companies that often engage in desk work; specifically, a person who sits in front of a computer for at least half their work hours (3.5 h a day or more).

Exclusion Criteria

People who correspond to any of the following groups are excluded from this study:

(1) People currently receiving treatment for mental illness, such as depression;
(2) People who suffer from diseases that may affect the acquisition of biometric information. For example, those who have a disease or disorder that affects pulse wave data measurement (persons who have paralysis or involuntary movements on their faces, or heart disease), those who have a disease or disorder that affects speech data measurement (speech difficulty caused by vocal cord extraction, etc.), or those who have a disease or disorder that affects measurement with wearable devices (persons with paralysis of the extremities or involuntary movement, etc.);
(3) People who have difficulty operating a computer, such as using email or the internet;
(4) People who cannot offer biometric information to researchers due to business/security reasons.

Participant and Public Involvement

This study was supported by AMED at the stage of developing proof of concept for quantification of stress and well-being using pulse, speech, body motion, and electrodermal data. The study design was made by industrial doctors who served as consultants for some of the companies for which the participants of this study work. These industrial doctors, who are members of our research team, conducted preliminary meetings with the participants, and based on those meetings, they arrived at the question this study hopes to answer: whether stress and well-being can be quantified by pulse, speech, and electrodermal data. The results of this study will be made available to participants through debriefing sessions at each participating company.

Data Collection

Data will be collected according to the observation period schedule in **Table 1**.

(A) Collection of background factors

After obtaining written consent, the following information will be obtained from each participant:

(1) Sex, age, job department, job content, duration of service, position/title, family composition, work commute, household income, etc.;
(2) Information on past medical checkups and stress check information (with consent of participant);
(3) Any current illnesses and prescriptions.

(B) Collection of biological data with sensing devices

Biological information will be recorded at participants' workplaces during the 4-week observation period using methods B-1 through B-3, as described below:

(B-1) Pulse wave data

Participants install software on their work computers that uses a camera built into or connected to the computer to record video images of the participant; the pulse wave data is extracted from the facial video images. The pulse wave data is automatically sent to cloud storage through the software. Participants are asked to start the software when they arrive for work; the software must also be restarted if the participant's computer is put into sleep

TABLE 1 | Schedule for data collection and evaluations during the study's observation period.

Data collection		At the beginning	At the mid-point (2 weeks)	At the end (4 weeks)
(A) Collection of background factors	Background characteristics (sex, age, department, work content, duration of service, etc.) Past stress check data, etc.	✓	If participant's environment changes, data will be updated.	
(B) Collection of data using external sensors	Pulse wave, speech, electrodermal activity, etc.	Acquired during business hours		
(C) Stress, well-being, and depression assessment using rating scale; self-reported daily condition	New occupational stress simple questionnaire (revised version): Estimated completion time, 5 min	✓	If participant's environment changes, data will be updated.	
	Perceived Stress Scale (PSS): Estimated completion time, 1 min			✓
	Satisfaction With Life Scale (SWLS): Estimated completion time, 1 min			✓
	Japanese version of Positive and Negative Affect Schedule (PANAS): estimated completion time, 1 min		✓	✓
	Japanese version of Flourishing Scale (FS-J): Estimated completion time, 1 min			✓
	Subjective well-being/ideal happiness: estimated completion time, 1 min			
	Japanese version of Patient Health Questionnaire-9 (PHQ-9): estimated completion time, 1 min			
	Self-reported daily condition: estimated completion time, 1 min	Daily at the end of work (not required)		

mode. This contactless pulse wave sensing system has a strong correlation in the R-R interval values compared to data obtained using ECG ($r^2 = 0.978$, $p < 0.00001$) (19).

(B-2) Voice data

Participants install software on their work computers that uses a microphone built into or connected to their work computer to record the emotional components (pitch, speed, etc.) of participants' speech data. The emotional component data is automatically sent to cloud storage through the software. Participants are asked to start the software when they arrive for work; the software must also be restarted if the participant's computer is put into sleep mode.

(B-3) Electrodermal activity data, body motion data, and body temperature

Participants are asked to wear the Embrace2 wristband-type wearable device, made by Empatica, Inc., continuously during work hours. The device is equipped with an electrodermameter, accelerometer, gyroscope, and thermometer (20, 21).

(C) Collection of stress, well-being, and depression rating scale assessment data; self-reported daily condition

Researchers will send participants an email with a unique URL link for a unique website where participants can answer questionnaires on stress and well-being online. The evaluation scales and their estimated completion times are as follows (see **Table 1** for the evaluation schedule):

- New Occupational Stress Questionnaire (modified version) (22)
- Perceived Stress Scale (PSS) (23)
- Satisfaction With Life Scale (SWLS) (24)
- Japanese version of Positive and Negative Affect Schedule (PANAS) (25)
- Japanese Flourishing Scale (FS-J) (26)
- Subjective Well-being/Ideal Happiness (27, 28)
- Japanese version of Patient Health Questionnaire-9 (PHQ-9) (29)
- Self-reported daily condition: an email with a unique URL will be sent to participants every business day; participants will input their condition (stress level, emotions, etc.) and sleep quality for the day in Likert scales, and include any special notes in 1–2 sentences (e.g., "I had a tough day," "I was praised by my boss today," etc.).

Data Storage

The data of background factors (Data A) will be recorded on paper by the researcher and then entered by the researcher into the password-locked computer in the laboratory of the researcher and stored.

Pulse wave and voice data (Data B-1 and B-2) will be quantified and automatically uploaded to the server used by the researcher team through a Secure Sockets Layer (SSL) connection. The research team will download this data via an

SSL connection to a research computer in the laboratory and analyzes it.

Data of electrodermal activity, body motion, and body temperature (Data B-3) will be captured by the wearable device and transferred to Empatica's cloud in encrypted form using proprietary software, where the raw data from the device will be analyzed and transformed into skin potential, heart rate, body movement, and temperature data. These sensing data will be eventually downloaded to the computers of the Keio University research team for analysis via an SSL connection.

The web-input data (Data C) such as stress, well-being, depression rating scale assessment data, and self-reported daily condition will be entered directly by participants by accessing an input format on a secure cloud computing service created by the research team. The entered data will be stored on the cloud, and the research team will access the cloud and downloads them to the laboratory's computer through an SSL connection.

All data above will be not accessible by the participants' employers.

Data Analysis

The analysis workflow is as follows: (1) primary analysis, comprised of using software to digitalize participants' vital information; (2) secondary analysis, comprised of examining the relationship between the quantified vital data from (1), stress, well-being, and depression; (3) tertiary analysis, comprised of generating machine learning algorithms to estimate stress, well-being, and degree of depression in relation to each set of vital data as well as multimodal vital data. The primary analysis is conducted with technology already established by Panasonic and NEC, who are industrial collaborators in this study. In this research, the results of the primary analysis are used to generate machine learning algorithms for the secondary and tertiary analyses.

Primary Analysis

Primary Analysis of Pulse Wave Data

Software from Panasonic installed on participants' work computers uses a camera to capture facial images of participants using facial detection. Based on skin color changes from blood flow in the images, the software extracts pulse wave data for the participant. The pulse wave data will then be clarified using filtering and noise removal techniques. SDNN, root mean square of successive differences in R-R intervals (RMSSD), Lorentz plot (Longitudinal axis/Transverse axis value), LH/HF ratio, and Tone-Entropy are calculated from the facial video images (1/30 s unit).

Primary Analysis of Voice

Software from Panasonic installed on participants' work computers uses a microphone to acquire speech data from participants. From this data, voice activity detection (VAD; presence/absence of voice), power (volume of speech), pitch, tension (strength of speech), and speech rate data are extracted (in units of 0.5 s), and an emotion estimate is calculated based on the results.

Primary Analysis of Electrodermal Activity

The Embrace2 wristband-type wearable device from Empatica, Inc., is equipped with an electrodermameter, accelerometer, gyroscope, and thermometer, which are used to record and analyze electrodermal activity, acceleration, angular velocity, and skin temperature.

Secondary Analysis

Each result from the primary analyses of pulse wave, speech, and electrodermal activity data will be compared with the self-assessment results for stress, well-being, and depression. Then, we will determine the relationships between the vital data, stress, well-being, and depression. For example, such an investigation could be done by comparing the stress and well-being score quartiles with the vital data, or comparing them among subjects grouped according to depression symptom severity. Multiple regression analysis will be performed in order to predict stress, well-being, or depression using various kinds of vital data. Moreover, cluster analysis will be performed to find a group of people that have similar digital phenotypes and to seek for the potential relationships with clinical phenotypes.

Tertiary Analysis

We will attempt to build a machine learning model to predict depression, stress, or well-being based on single modalities. Various methods, such as support vector machine, decision tree, and deep learning, are used for the machine learning analysis. Machine learning algorithms that estimate stress, well-being, and degree of depression are generated from each of these vital data sets. We will attempt to build a model not only for single modalities, but also one that can utilize all the modalities, namely pulse wave data, voice data, electrodermal activity data, body motion data, and body temperature data, together.

Sample Size

Based on the data obtained from the pilot study (28 cases) conducted prior to this study, we estimated the PSS scores using machine learning analysis of pulse wave and speech data. A gradient boosting decision tree was used for the machine learning algorithm, and hyper parameter was adapted by random search. Accuracy verification is based on 3-fold cross validation. In order to examine the accuracy for each data size, an arbitrary number of data points were extracted at random, and the accuracy for the data size was calculated. The pilot study sample size of 28 subjects was divided and incremented for prediction. The error in the PSS score range of 0–40 corresponds to a score of 2 when an error in the predicted value of PSS of up to around 5% is warranted. When aiming for a root mean square error of 2 or less, we found that 200 cases are required, as shown in **Figure 1**. Regarding dropouts, we considered that the 4-week-long observation as well as filling out multiple self-rating scales can be burdensome to participants, and the dropout rate can be high. Considering the dropout rate would be 30–35%, we will aim to recruit 300 cases.

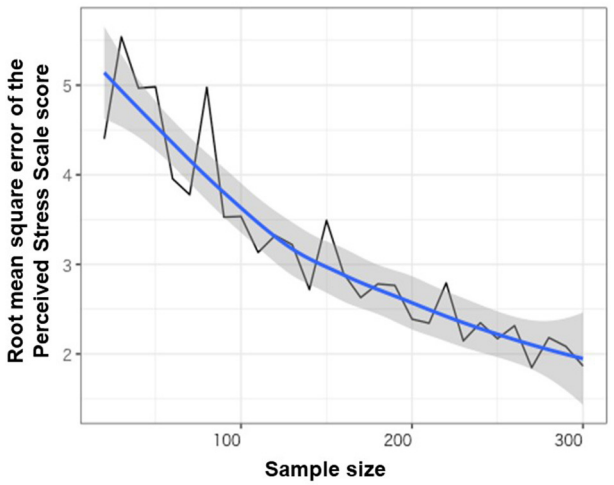


FIGURE 1 | Sample size and root mean square error of the PSS score in the pilot study.

DISCUSSION

The main aim of this study is to develop a technique to quantify stress and well-being in a way that does not disturb the workplace using vital data. Depression is a major problem in the field of occupational health, with its high incidence, especially among people of working age. The social significance of our project will be great because the occurrence of psychiatric disorders such as depression may be suppressed through self-management and improvement of working conditions in the companies if the quantitative measurement of stress and well-being, which is the ultimate goal of this study, can be achieved.

In this study, biometric information such as pulse wave, voice, and skin potential will be collected without any special intervention while the research collaborators are working. In addition, the research collaborators will be asked to cooperate in data entry through questionnaires, but as described below, the burden will be kept to a certain extent. We will use a web camera, a microphone, a wrist-band wearable device to obtain pulse, voice, and skin potential data, and a web-based questionnaire, which may cause psychological, physical, and time burdens. However, we believe that the burden will be small because we will not intervene and the questionnaire will be administered in a way that will not interfere with normal work and it takes a maximum of about 10 min per session.

The study will also include a rating scale for depression, on which significantly higher scores are presumed to indicate a higher risk of depression. If a recruited person will present any psychiatric disorder during the study, he or she will receive information that would not have been available to him or her if he or she had not participated in the study. The possibility that this may lead to a psychological burden on the individual cannot be denied. Only individuals who agreed that he or she will be informed of accidental findings are recruited. In addition, when the score of the rating scale for depression indicates he or she may have depression, an appropriate action should be taken, such as referring the participant to an appropriate medical institution if he or she wishes to do so. Moreover, the participant will be informed that sensing can be discontinued by shutting down the software or removing the wearable device, and that he or she may do so temporarily if it causes a psychological burden. Prior to the study, consent with the field companies will be obtained using a memorandum of understanding to ensure the privacy of research collaborators (e.g., not to disclose data without the participant's consent).

Challenges of the study are as follows. First, there will be a potential bias that those who are originally interested in the physical and mental health conditions may collaborate in the study. For this reason, participants who are currently being treated for psychiatric disorders, such as depression, will be excluded from the study, and efforts will be made to recruit a wide range of participants. Second, there will be also the issue of adherence. Failure to fill out the questionnaire or not obtaining biometric data may be considered. The researchers will try to reduce the burden on research participants as much as possible to minimize the time and effort required for inputting information and improve adherence. The researchers also consider rewarding each participant based on the response rate to the survey and the amount of time the participant will spend using the devices. Third, as the data collection is done in a natural setting in workplaces, we will lack the control data such as setting participants an experimental task where we can compare the data under strong stress.

ETHICS STATEMENT

Our study has received approval from the institutional review board at Keio University School of Medicine. Approval was granted on April 22, 2019. This study is registered in the University Hospital Medical Information Network (UMIN) (UMIN000036814). Participants' inclusion will be voluntary. Written consent will be obtained from every participant. Participants will be free to withdraw from the study at any time.

AUTHOR CONTRIBUTIONS

KI, KM, and TK conceived the original study concept. KI designed and managed the study, and wrote the initial draft of the manuscript. KM and KS designed and managed the study. TK designed and supervised the study, and assisted in drafting the manuscript. All authors contributed to the design of the study, protocol development, its implementation, critically reviewed the manuscript, approved the final version of the manuscript, agree to be responsible for the accuracy, and integrity of the work.

FUNDING

This work was supported by the Japan Agency for Medical Research and Development (AMED) (Grant No. 18le0110008h0001; Unobtrusive Sensing Technology for Quantifying Stress and Well-being to Promote a Healthy Workplace).

ACKNOWLEDGMENTS

The authors are grateful to Ms. Momoko Kitazawa, Mr. Shunya Kurokawa, Mr. Michitaka Yoshimura, Mr. Kuo-ching Liang, Mr. Asuka Koshi, Ms. Yuki Ishikawa, Ms. Yoko Usami, Ms. Kumiko Hiza, and Ms. Hiromi Mikami for supporting data collection and management in this study. The authors are grateful to Mr. Yuichiro Suzuki (Panasonic) and Mr. Yoshifumi Onishi (NEC) for technical support in this study. This manuscript has been released as a pre-print at medRxiv (30).

REFERENCES

1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. (2018) 392:1789–858. doi: 10.1016/S0140-6736(18)32279-7
2. Kawakami N. *A study on epidemiological survey on mental health: comprehensive research report: 2004-2006 Ministry of Health, Labour and Welfare Grant-in-Aid for Health Science Research Project*. (2007). Available online at: https://www.khj-h.com/wp/wp-content/uploads/2018/05/soukatuhoukoku19.pdf
3. Ministry of Health. *Labour and Welfare Disability Welfare Promotion Project Subsidy: 'Estimation of Social Cost of Mental Illness' Business Report*. (2011). Available online at: https://www.mhlw.go.jp/bunya/shougaihoken/cyousajigyou/dl/seikabutsu30-2.pdf
4. Lyubomirsky S, King L, Diener E. The benefits of frequent positive affect: does happiness lead to success? *Psychol Bull*. (2005) 131:803–55. doi: 10.1037/0033-2909.131.6.803
5. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. *Front Public Heal*. (2017) 5:1–17. doi: 10.3389/fpubh.2017.00258
6. Koolagudi SG, Rao KS. Emotion recognition from speech: a review. *Int J Speech Technol*. (2012) 15:99–117. doi: 10.1007/s10772-011-9125-1
7. Johannes B, Wittels P, Enne R, Eisinger G, Castro CA, Thomas JL, et al. Non-linear function model of voice pitch dependency on physical and mental load. *Eur J Appl Physiol*. (2007) 101:267–76. doi: 10.1007/s00421-007-0496-6
8. Nakatsu R, Nicholson J, Tosa N. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Knowledge Based Syst*. (2000) 13:497–504. doi: 10.1016/S0950-7051(00)00070-8
9. Sano A, Phillips AJ, Yu AZ, McHill AW, Taylor S, Jaques N, et al. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In: *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks 2015*. IEEE. (2015). p. 1–6. doi: 10.1109/BSN.2015.7299420
10. Dogan E, Sander C, Wagner X, Hegerl U, Kohls E. Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? Systematic review. *J Med Internet Res*. (2017) 19:e262. doi: 10.2196/jmir.7006
11. Dang M, Mielke C, Diehl A, Haux R. Accompanying depression with FINE - a smartphone-based approach. *Stud Health Technol Inform*. (2016) 228:195–9. doi: 10.3233/978-1-61499-678-1-195
12. Alberdi A, Aztiria A, Basarab A. Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review. *J Biomed Inform*. (2016) 59:49–75. doi: 10.1016/j.jbi.2015.11.007
13. Dell'Acqua C, Dal Bò E, Benvenuti SM, Palomba D. Reduced heart rate variability is associated with vulnerability to depression. *J Affect Disord Rep*. (2020) 1:100006. doi: 10.1016/j.jadr.2020.100006
14. Kemp AH, Quintana DS, Gray MA, Felmingham KL, Brown K, Gatt JM. Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis. *Biol Psychiatr*. (2010) 67:1067–74. doi: 10.1016/j.biopsych.2009.12.012
15. Kishimoto T, Takamiya A, Liang KC, Funaki K, Fujita T, Kitazawa M, et al. The project for objective measures using computational psychiatry technology. (PROMPT): Rationale, design, and methodology. *Contemp Clin Trials Commun*. (2020) 19:100649. doi: 10.1016/j.conctc.2020.100649
16. Yamamoto M, Takamiya A, Sawada K, Yoshimura M, Kitazawa M, Liang KC, et al. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PLoS ONE*. (2020) 15:e0238726. doi: 10.1371/journal.pone.0238726
17. Horigome T, Sumali B, Kitazawa M, Yoshimura M, Liang KC, Tazawa Y, et al. Evaluating the severity of depressive symptoms using upper body motion captured by RGB-depth sensors and machine learning in a clinical interview setting: a preliminary study. *Compr Psychiatr*. (2020) 98:152169. doi: 10.1016/j.comppsych.2020.152169
18. Tazawa Y, Liang KC, Yoshimura M, Kitazawa M, Kaise Y, Takamiya A, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. *Heliyon*. (2020) 6:e03274. doi: 10.1016/j.heliyon.2020.e03274
19. Suga K, Hori H, Katsuki A, Ohashi M, Tezuka T, Matsuo M, et al. The contactless vital sensing system precisely reflects R-R interval in electrocardiograms of healthy subjects. *PACE*. (2017) 40:514–5. doi: 10.1111/pace.13057
20. Nakashima Y, Umematsu T, Tsujikawa M, Onishi Y. An Effectiveness Comparison between the Use of Activity State Data and That of Activity Magnitude Data in Chronic Stress Recognition. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019*. IEEE (2019). p. 239–43. doi: 10.1109/ACIIW.2019.8925222
21. Empatica Embrace2 device website. Available online at: https://www.empatica.com/en-int/embrace2/
22. Tsutsumi A, Shimazu A, Eguchi H, Inoue A, Kawakami N. A Japanese stress check program screening tool predicts employee long-term sickness absence: a prospective study. *J Occup Health*. (2018) 60:55–63. doi: 10.1539/joh.17-0161-OA
23. Sumi K. Reliability and validity of the Japanese version of the perceived stress scale. *Jpn J Heal Psychol*. (2006) 19:44–53. doi: 10.11560/jahp.19.2_44
24. Diener E, Emmons RA, Larsen RJ GS. The satisfaction with life scale. *J Pers Assess*. (1985) 49:71–5. doi: 10.1207/s15327752jpa4901_13
25. Sato A. Development of the Japanese version of Positive and Negative Affect Schedule. (PANAS) scales. *Jpn J Personal*. (2001) 9:138–9. doi: 10.2132/jjpjspp.9.2_138
26. Sumi K. Temporal stability of the japanese versions of the flourishing scale and the scale of positive and negative experience. *J Psychol Psychother*. (2014) 4:1–5. doi: 10.4172/2161-0487.1000140
27. Huppert FA, Marks N, Michaelson J, Vázquez C, Vittersø J. *ESS6 - 2012/3 Question Module Design Final Template*. (2013). Available online at: https://www.europeansocialsurvey.org/docs/round6/questionnaire/ESS6_final_personal_and_social_well_being_module_template.pdf
28. Takahashi Y. The effect of culture on happiness in Japan: verification by ideal happiness. *J Behav Econ Financ*. (2018) 11:S9–12. doi: 10.11167/jbef.11.S13
29. Inagaki M, Ohtsuki T, Yonemoto N, Kawashima Y, Saitoh A, Oikawa Y, et al. Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: a cross-sectional study. *Gen Hosp Psychiatr*. (2013) 35:592–7. doi: 10.1016/j.genhosppsych.2013.08.001
30. Izumi K, Minato K, Shiga K, Sugio T, Hanashiro S, Cortright K, et al. Quantification of stress and well-being using pulse, speech, and electrodermal data: study concept and design. *medRxiv*. (2020) 2020.05.01.20082610. doi: 10.1101/2020.05.01.20082610

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 25 June 2021
doi: 10.3389/fpsyt.2021.574482

Feasibility of Repeated Assessment of Cognitive Function in Older Adults Using a Wireless, Mobile, Dry-EEG Headset and Tablet-Based Games

*Esther C. McWilliams[1], Florentine M. Barbey[2], John F. Dyer[1], Md Nurul Islam[2], Bernadette McGuinness[3], Brian Murphy[2,4], Hugh Nolan[2], Peter Passmore[3], Laura M. Rueda-Delgado[2,5] and Alison R. Buick[1,4]**

[1] Cumulus Neuroscience Ltd, Belfast, United Kingdom, [2] Cumulus Neuroscience Ltd, Dublin, Ireland, [3] Centre for Public Health, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom, [4] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, United Kingdom, [5] Trinity Centre for Biomedical Engineering, Trinity College, The University of Dublin, Dublin, Ireland

Access to affordable, objective and scalable biomarkers of brain function is needed to transform the healthcare burden of neuropsychiatric and neurodegenerative disease. Electroencephalography (EEG) recordings, both resting and in combination with targeted cognitive tasks, have demonstrated utility in tracking disease state and therapy response in a range of conditions from schizophrenia to Alzheimer's disease. But conventional methods of recording this data involve burdensome clinic visits, and behavioural tasks that are not effective in frequent repeated use. This paper aims to evaluate the technical and human-factors feasibility of gathering large-scale EEG using novel technology in the home environment with healthy adult users. In a large field study, 89 healthy adults aged 40–79 years volunteered to use the system at home for 12 weeks, 5 times/week, for 30 min/session. A 16-channel, dry-sensor, portable wireless headset recorded EEG while users played gamified cognitive and passive tasks through a tablet application, including tests of decision making, executive function and memory. Data was uploaded to cloud servers and remotely monitored via web-based dashboards. Seventy-eight participants completed the study, and high levels of adherence were maintained throughout across all age groups, with mean compliance over the 12-week period of 82% (4.1 sessions per week). Reported ease of use was also high with mean System Usability Scale scores of 78.7. Behavioural response measures (reaction time and accuracy) and EEG components elicited by gamified stimuli (P300, ERN, Pe and changes in power spectral density) were extracted from the data collected in home, across a wide range of ages, including older adult participants. Findings replicated well-known patterns of age-related change and demonstrated the feasibility of using low-burden, large-scale, longitudinal EEG measurement in community-based cohorts. This technology enables clinically relevant data to be recorded outside the lab/clinic, from which metrics underlying cognitive ageing could be extracted, opening the door to potential new ways of developing digital cognitive biomarkers for disorders affecting the brain.

Keywords: EEG, EEG biomarker, cognition, gamification, mobile EEG

OPEN ACCESS

Edited by:
Qiang Luo,
Fudan University, China

Reviewed by:
Xi Jiang,
University of Electronic Science and Technology of China, China
Albert Yang,
National Yang-Ming University, Taiwan

***Correspondence:**
Alison R. Buick
alison@cumulusneuro.com

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: *19 June 2020*
Accepted: *18 March 2021*
Published: *25 June 2021*

Citation:
McWilliams EC, Barbey FM, Dyer JF, Islam MN, McGuinness B, Murphy B, Nolan H, Passmore P, Rueda-Delgado LM and Buick AR (2021) Feasibility of Repeated Assessment of Cognitive Function in Older Adults Using a Wireless, Mobile, Dry-EEG Headset and Tablet-Based Games. Front. Psychiatry 12:574482. doi: 10.3389/fpsyt.2021.574482

INTRODUCTION

Recent advances in digital technologies provide a wealth of opportunity in the management of health conditions. In neurological disease the heterogeneity and complexity of conditions, along with continuing reliance on traditional subjective measurement tools, have presented a challenge for the development of data-driven biomarkers for diagnosis, monitoring and prediction of therapeutic response (1–7). The suite of tools described in this paper was designed to enable longitudinal, in-home data collection of brain electrophysiology and cognitive performance. The platform comprises (1) a dry sensor, wireless electroencephalography (EEG) headset that records brain activity, (2) gamified versions of cognitive tasks, and (3) cloud-based storage and automatic processing—with the aim of identifying potential digital biomarkers with utility in neuropsychiatric and neurodegenerative disorders.

EEG directly reflects neural synaptic function, with similar patterns from animal to human (8–10) and thus has substantial potential as a brain-based, translatable biomarker for diseases such as schizophrenia (11–18), depression (19–21) and Alzheimer's disease (AD) (22–29). However, traditional research EEG setups are effortful and time-consuming, requiring expensive equipment and the support of personnel with technical training. Single or infrequent lab-based EEG recording sessions may be affected by a range of factors including fluctuations in levels of participants' mental alertness, fatigue and task-induced mental workload (30, 31). Similarly, cognition as traditionally measured in therapeutic research and practise tends to take the form of clinician administered batteries of neuropsychological tests [e.g., (32, 33)] which, whilst low burden and relatively inexpensive, are subject to variability in scores on repeated testing occasions (34). Infrequent, "snapshot" assessments are subject to measurement error arising from multiple factors, such as practise effects (35–37), the "white-coat effect" related to anxiety about suspected cognitive impairment (38), and day-to-day fluctuations in context (39), in mood and in perceived stress (40–44).

The adoption of modern technology into medicine allows for more innovative forms of data collection (e.g., wearable devices), increasing objectivity and taking advantage of powerful analytical tools to probe complex diseases. Further, digital tools may allow for more frequent sampling and detection of subtle daily fluctuations, at minimal disruption to the patient since data may be collected both inside and outside of the clinic. Progress in modern electronics and dry sensor technology means that EEG is emerging from amongst standard brain imaging methods as a mobile technology, suitable for deployment to very large cohorts for convenient at-home use (45). Likewise, neuropsychological testing can now be completed outside of the clinic through the use of automated, web-based assessments (46–49).

Mobile EEG systems are advancing quickly. Several studies have shown that it is possible in principle to collect EEG recordings using consumer-grade hardware, and from the data extract potentially useful neuronal signals, including spectral band-power measures (45, 50, 51) and task-evoked event-related potentials (ERPs) (52–55). However, studies using such devices have typically required some specialist equipment (e.g., a computer running bespoke software to present stimuli and record EEG), and a specialist experimenter to set up and supervise the recording. In addition, most consumer-grade EEG platforms operate using low numbers of electrodes, leaving some research questions and certain types of analysis out of reach for researchers. To the authors' knowledge, there exists no prior example of large-scale, unsupervised in-home, repeated sampling ERP research using a dry-sensor, portable, user-friendly EEG platform.

Innovative solutions can be deployed to enable us to carry out unsupervised data collection without placing undue burden on the user, such as 'dry' sensors (i.e., eschewing the conductive gel used in the laboratory in favour of an easier electrode setup) and automated user-facing stepwise tutorials and notifications (to compensate for reduced environmental control outside the laboratory). Similarly, for use at home over repeated sessions, EEG/ERP tasks as used in research may not be particularly exciting or motivational for the user, but applying gamification can make these tasks more engaging and rewarding for participants (56) and gamified cognitive tasks can facilitate global data gathering on an unparalleled scale (57).

The study presented here was a first, proof-of-concept, field study to test the human-factors and technical feasibility of an early version of the Cumulus Neuroscience platform in a cohort of healthy adults spanning an age range up to 79 years old. In this paper we investigate the potential of this platform to capture in-home, frequent repeated measurement of EEG and behavioural metrics of cognitive ageing, metrics that also have broader appeal as potential cognitive biomarkers for the diagnosis and treatment of disorders affecting the brain. Use of the platform on a regular basis over 3 months assesses the acceptability of long-term use for future use cases where longitudinal progression tracking is required, avoids dependence on a single "snapshot" measurement, and allows for improved signal quality through aggregation of EEG data collected in the home (an unsupervised environment).

Analyses are presented that quantify reported ease-of-use, and resultant levels of weekly adherence over a period of 3 months of unsupervised at-home use. The gamified cognitive tasks are evaluated for face-validity, by comparing key known behavioural effects with data gathered in the home, and examining effects of age that have been reported in the literature. Similarly the EEG data is examined at grand-average level to confirm that it replicates the main features (waveform morphology and timing, frequency content, scalp topography) of the neural signatures that the gamified tasks are designed to elicit.

METHOD

Participants

89 healthy adult volunteers (67 female), aged between 40 and 79 years (mean 58.78, s.d. 8.86) with a Montreal Cognitive Assessment (MoCA) score $\geq$24 gave informed consent to take part in the procedures approved by Queen's University Belfast Ethics Committee. Recruitment channels included "Join

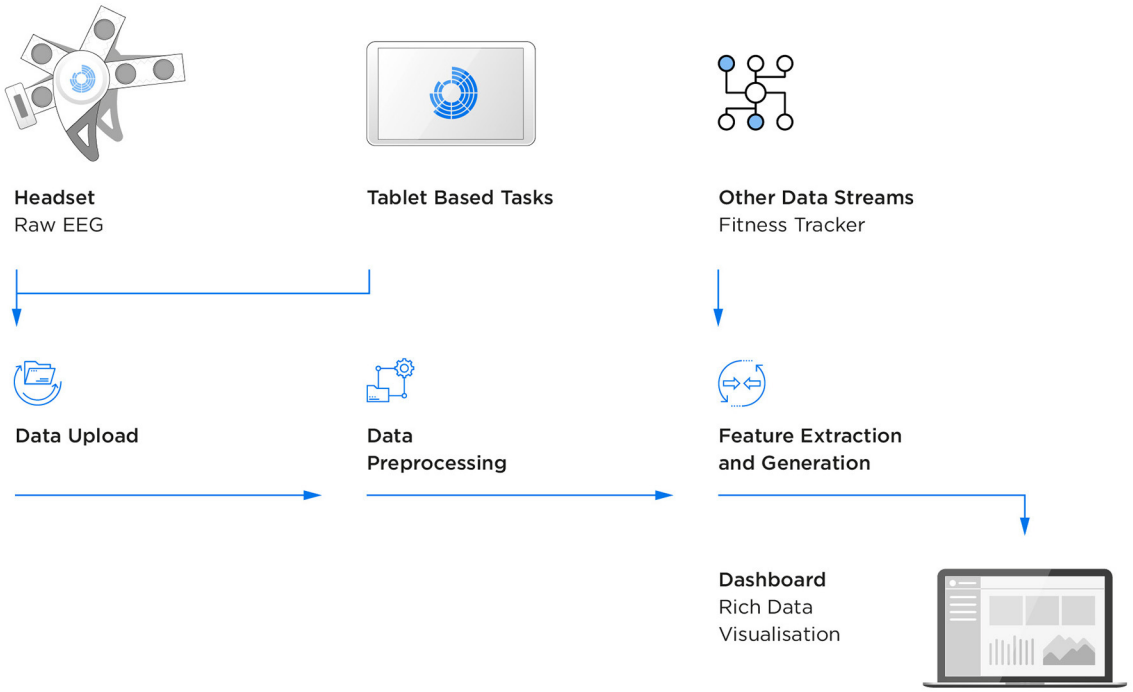


FIGURE 1 | Flow of EEG and behavioural data.

Dementia Research," local community groups and use of print media and social media.

The Platform

The platform was designed to enable frequent, objective sampling of brain-based markers of cognition inside and outside of the clinic/lab setting, using a dry sensor, wireless encephalography (EEG) headset that records brain activity, accompanied by gamified versions of cognitive tasks presented via a tablet-based app. Upon logging into the app, a stepwise tutorial guides the user through setup of the headset (covering placement on the head, positioning of the detachable mastoid sensors and feedback on sensor impedances) in preparation for recording data during the gamified tasks. Cloud-based secure methods are used for collection and automatic processing, as well as integration with other data streams (in this study participants wore a fitness tracker, the Withings Go) and web-based dashboards for monitoring and data visualisation on a daily, session-by-session basis (**Figure 1**).

EEG

The wireless EEG headset (**Figure 2**) consists of dry flexible Ag/AgCl coated polymer sensors at 16 channels (O1, O2, P3, Pz, P4, Cz, FT7,FC3, FCz, FC4, FT8, Fz, AF7, AF8, FPz). The left and right mastoids are used for reference and driven-bias, with single-use, snap-on electrodes attached to wires extending from the headset. The electronics and sensors are mounted on flexible neoprene, and the stretchable structure incorporates anatomical landmarks in the form-fit of the headset to encourage consistent placement by users in line with the 10–20 sensor system. The analogue headset has high input impedance of 1 GΩ, a configurable driven bias function for common-mode rejection, built-in impedance checking, and configurable gain and sampling rates. An onboard processor and Bluetooth module transmit 250 Hz EEG data to the tablet, from where it is transferred to a cloud server for storage and processing. EEG recording and behavioural events are timestamp synchronised to ±2 ms.

Cognitive Tasks

The gamified tasks (**Figure 3**) are based on well-known paradigms from experimental electrophysiology and cover a range of core cognitive functions. Cognitive/electrophysiological tasks were gamified with the aim of improving motivation, i.e., to enhance attentiveness during testing and to prevent boredom over repeated plays, while maintaining the core cognitive components of the original task. Feedback on gameplay performance was provided (e.g., points awarded for speed of responses where appropriate), along with personalised leaderboards to promote long-term adherence to the study schedule. Each daily session comprised a resting state plus two of the four other games (alternating between sessions). Participants also answered a daily health and lifestyle questionnaire to further contextualise the daily recordings. Daily sessions were designed to take <30 min total time from start to finish.

Two-Stimulus Visual Oddball

This gamified version of the classic 2-stimulus visual oddball paradigm (58), presents target stimuli (aliens—requiring the

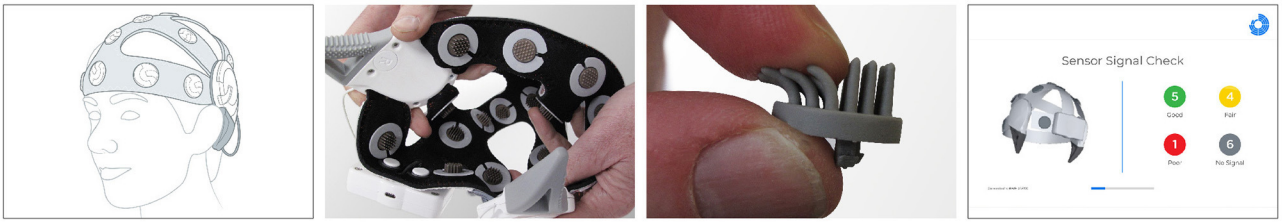


FIGURE 2 | Sixteen-channel wireless headset designed with pliable sensors and the sensor signal quality check.

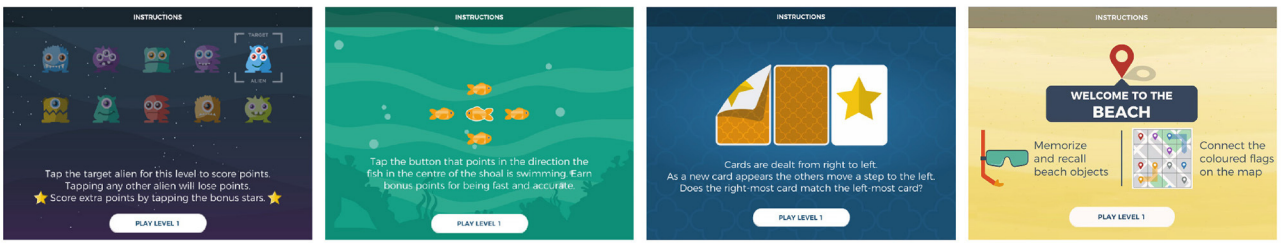


FIGURE 3 | Images of 2-stimulus visual oddball, flanker, n-back, and delayed match-to-sample gamified tasks.

participant to tap on the screen, $n = 30$) and non-target stimuli (visually different aliens—requiring no response from the participant, $n = 70$) across five levels of gameplay, as well as 15 "bonus" stars throughout the game (to enhance gameplay and not included for analysis). Behaviour (reaction time and response accuracy) and corresponding EEG correlates are indicative of neural dynamics of the decision-making process and the attention and working memory on which it relies (59–61). Using EEG, a positive voltage deflection can be observed over the parietal cortex starting ∼300 ms following presentation of the stimulus, known as the P300 event-related potential (ERP). With advancing age, the amplitude of the P300 is known to decrease, and its latency known to increase (62–64).

Flanker

Inhibition and error awareness were probed using a gamified version of the Erikson Flanker task (65). Fish served as directional stimuli and were presented across five levels with a shoal of fish (flanking stimuli) appearing first, followed by the central (target) fish. The participant was asked to tap on the side of the screen corresponding to the direction of the central fish, ignoring the flanking fish (either congruent or incongruent stimuli, split evenly between the 150 trials), presenting a cognitive challenge reflected in behavioural responses (accuracy and reaction time) and EEG. The relevant EEG metric extracted from this task is the Error-related negativity (ERN)—a negative voltage deflection observed on error trials most prominently over the fronto-central scalp, followed by a subsequent positive rebound in the signal (the Error Positivity, or Pe) (66). Previous studies have consistently reported a decrease in the negative amplitude of the ERN with progressing age (67).

N-Back

The visual n-back paradigm (68) taxes working memory and executive function with age-related differences in behavioural performance, according to recent meta-analysis (69). In the current study, this game had a continuous short-sequence memorisation of 4 different playing cards where the participant was asked whether the current card is a "match" or "no-match" to the card seen 2 trials before. This 2-back paradigm consisted of 100 trials presented across two levels, with a 33% match rate.

Delayed Match-to-Sample

A visual delayed match-to-sample task, this task targets recognition memory, a key cognitive function known to be affected by age (70), across 50 trials, presented in blocks of 5, with 50% overall match rate (71). Each level is set in a specific location (beach, jungle, etc) where the user is presented with a variety of objects which must be encoded into memory to be retrieved after a brief (10-s) distractor game which involves connecting dots. Points are earned by identifying previously presented items at retrieval and rejecting unseen items.

Resting State

In this passive task (72) participants selected a relaxing scene (forest, park or beach) for 1 min of restful eyes open followed by 1 min eyes closed. This task elicits resting electrocortical activity and seeks to produce an increase in the neural oscillatory power of the alpha frequency band (7–13 Hz) when eyes are closed relative to eyes open, a physiological measure sensitive to a range of neurocognitive and psychiatric disorders, ageing, as well as sleep quality and caffeine intake (22, 73–75).

Procedure

Participants attended two in-lab sessions, at baseline and following 12 weeks of at-home use of the platform.

Lab Sessions

Lab sessions consisted of neuropsychological testing followed by in-lab use of the platform. Neuropsychological assessment was carried out for screening and to facilitate potential longitudinal follow-up and/or comparison with other datasets, and is not analysed in this paper. The MoCA (76) was selected as a screening tool using a cut-off of ≥24 to be representative of normal cognitive function, in line with findings reported in the Irish older adult general population (77, 78). Participants then completed tasks from the Cambridge Neuropsychological Test Automated Battery (CANTAB) covering multiple domains including psychomotor skills, executive function, memory and attention domains before completing a session with the platform. Participants provided ratings of the usability of the platform at baseline, 6 and 12 weeks into the study on the System Usability Scale (SUS), a 10-item industry standard questionnaire designed specifically to assess use of technology (79).

Home Sessions

Participants were asked to use the platform as described in section The Platform, at home, over the course of 12 weeks. These sessions were ∼25–30 min and participants were asked to contribute 5 sessions per week (one session per day, days unspecified for participant convenience). Throughout the 12-week period, participants wore a fitness tracker to monitor their step count and sleep, and answered questions about daily well-being and lifestyle habits (not analysed in the current paper).

Analysis

To measure usability of the platform across age groups and feasibility of extracting features reflecting cognitive ageing, participants were assigned to three groups for analysis: those aged 40–54 ($n = 26$), 55–66 ($n = 35$), and 67–79 ($n = 17$) years. Usability measures used for analysis were adherence and participant-reported SUS scores, as well as technical measures of signal reliability. To investigate feasibility and explore effects of cognitive ageing, behavioural and EEG metrics were extracted across age groups. Additionally, event-related potential waveforms were plotted for comparison based on single game-play median epoch, single-participant averaged epochs and whole-sample grand averaged epochs. Validity of the approach can be established by confirming that behavioural and neural patterns observable in the literature (e.g., differences in timing between congruent and incongruent trials; the waveform and scalp topography of ERP components) are seen in the data recorded unsupervised in the home, and that age-related changes in these variables reflect the published consensus. Ninety five percentage confidence intervals are reported throughout using the upper and lower bounds.

Behavioural Analysis

Measures of accuracy and speed of response were extracted from the cognitive games played using the platform (2-stimulus oddball, flanker, n-back and delayed match-to-sample) to establish face validity against that which the literature leads us to expect. For this analysis, we averaged different behavioural measures across game-plays. To investigate reaction times, we chose the median reaction time per game-play, taking the median-average per participant to calculate age group mean comparisons and sample means. To compare rates of accuracy, we calculated percentage accuracy per game-play. We produced a mean accuracy rate per participant for age group and sample mean comparisons. Additionally, we calculated confidence intervals as an indication of variance. To visualise age group differences across game plays, the log-transformed game-play number was used as an explanatory variable of the different behavioural metrics per group in a linear model and a 95% confidence interval was calculated using 1,000 bootstrapping resamples.

EEG Analysis

The processing pipeline consists of filtering from 0.25 to 40 Hz, customised artefact removal, epoch extraction and baseline removal. Metric-based methods for removing invalid ERPs and PSDs were applied to outputs. Two event-related potential (ERP) components, the P300 (a positive-going waveform which peaks >300 ms after the presentation of an attended stimulus, associated with decision-making) and the Error-Related Negativity (a negative-going, response-locked waveform associated with error-awareness) were computed as the smoothed pointwise median of epochs within each session. Power spectral density (PSD) was computed using a 1,024 point Fast Fourier Transform (FFT) with Welch's method of averaging (using a 256 sample window) on the resting-state eyes-open and eyes-closed data. For this analysis, time-series data was converted to average reference to remove lateralised effects of the original single-mastoid reference. Mean and 95% confidence intervals were computed across all sessions from participants within each age group.

RESULTS

Usability

Of the 89 healthy adult participants that enrolled in the study (67 female, mean age = 58.78, mean MoCA score = 27.12), 11 participants withdrew and 78 (61 female) completed the study (mean age = 58.99 mean MoCA score = 27.06), yielding an attrition rate of 12.40%. Data from those who withdrew was excluded from the following analyses reported.

Reasons for withdrawal cited were work/caring/other commitments ($n = 3$) and/or illness/health-related issues ($n = 6$). Two participants cited both health and caring commitments. One participant mentioned a faulty headset as an additional factor in the decision to withdraw; this participant's headset had required repair. Four participants did not give any reason. The mean number of sessions contributed by participants who

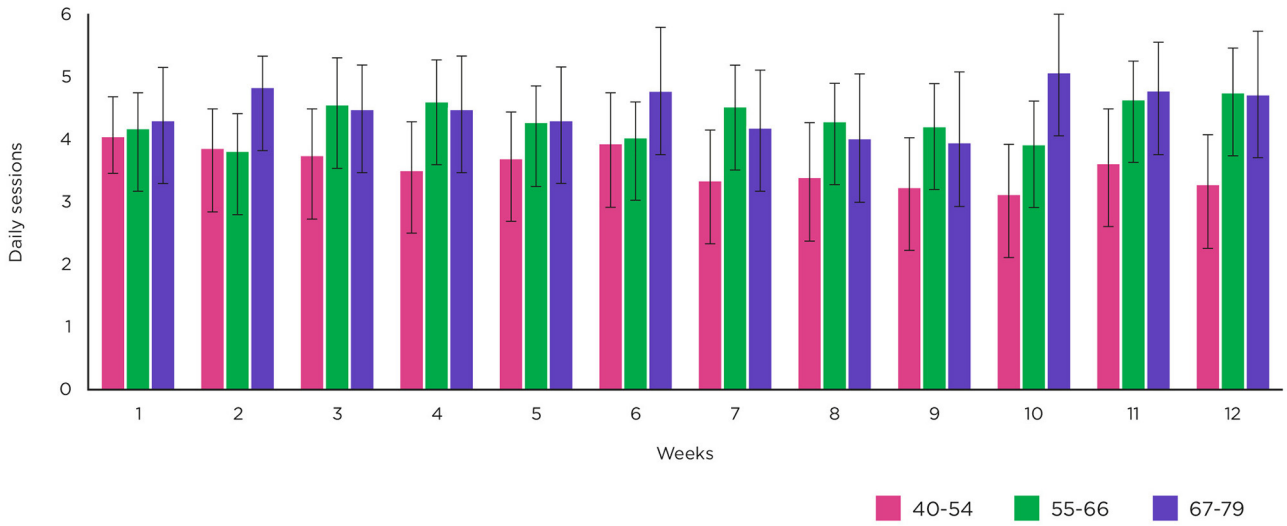


FIGURE 4 | Weekly adherence. Mean number of sessions per week across all participants and by age group. Error bars show upper and lower 95% confidence intervals.

TABLE 1 | System usability scale scores by age group at baseline, 6 and 12 weeks.

	40–54 years			55–66 years			67–79 years		
	Mean	**95% CI**	***n***	**Mean**	**95% CI**	***n***	**Mean**	**95% CI**	***n***
Baseline	81.35	±5.99	26	77.94	±4.60	34	66.62	±9.01	17
6 weeks	81.56	±5.81	24	85.44	±3.69	34	72.81	±7.08	16
12 weeks	80.50	±6.04	25	81.50	±4.62	35	68.38	±8.05	17

withdrew was 16.82 [8.41–25.23], ranging from 1 (completed in-lab) to 45 sessions. The mean duration of at-home involvement by those who withdrew was 5.55 weeks [3.22–7.88].

Figure 4 shows the rate of weekly adherence for those who completed the study ($n = 78$), including a breakdown of weekly adherence by age group [40–54 years ($n = 26$), 55–66 years ($n = 35$) and 67–79 years ($n = 17$)]. For those who completed, mean number of sessions contributed per week was 4.10 [3.97–4.23], out of a target of 5 per week and the mean total number of sessions contributed per participant was 49.14 [46.54–51.74]. By age group, mean number of sessions per week was 3.56 [3.33–3.78] for those aged 40–54, 4.31 [4.12–4.50] for those aged 55–66 and 4.48 [4.22–4.74] for those aged 67–79 years.

Participants were asked to evaluate usability via the System Usability Scale (SUS) at 3 timepoints. Mean SUS scores were 76.59 [72.94–80.24] at baseline, 81.45 [78.33–84.57] at 6 weeks and 78.28 [74.74–81.81] at 12 weeks (see **Table 1**). It is worth noting that the mean SUS score at baseline from those who subsequently withdrew was 75.23 [64.42–86.04] and of those who were still enrolled at 6 weeks, mean SUS score was 74.64 [62.62–86.67].

Ability to use the system to record usable EEG in the at-home setting, reported in **Figure 5**, was considered by measuring the percentage of time that individual sensors were connected to the scalp (i.e., recording non-saturated data) for the different age groups. Three thousand six hundred three sessions were successfully uploaded to the cloud server. Of these, 95.81% (3,452 sessions) contained portions of EEG data that could be used for analysis, even though certain sections of that session, or certain sensors, may be very noisy. One hundred and fifty-one sessions were rejected in their entirety, due to saturated data sections, high variance sections or gaps. By comparison, the behavioural data, where 99.03% (3,568 sessions) contained a complete set of response measures for at least one of the two games assigned per session. There were 116 sessions for which behavioural data, but not EEG data, was suitable for analysis.

Behaviour

To establish face validity of the gamified behavioural tasks, key metrics from each game were extracted to evaluate against what would be expected from traditional lab paradigms described in the literature. The temporal development of several illustrative metrics, stratified by age-group, is displayed in **Figure 6**.

Two-Stimulus Oddball Game

Similar to lab versions of simple decision-making task which do not challenge the accuracy of responses, the gamified version demonstrated high accuracy, mean 97.71 [97.08–98.35]%. Age

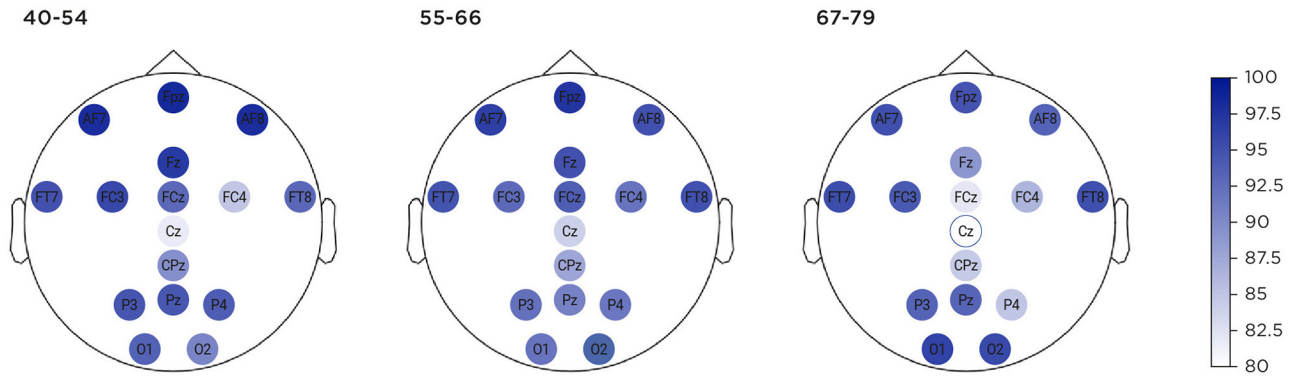


FIGURE 5 | Percentage of time each of 16 channels recorded non-saturated data, shown across age groups.

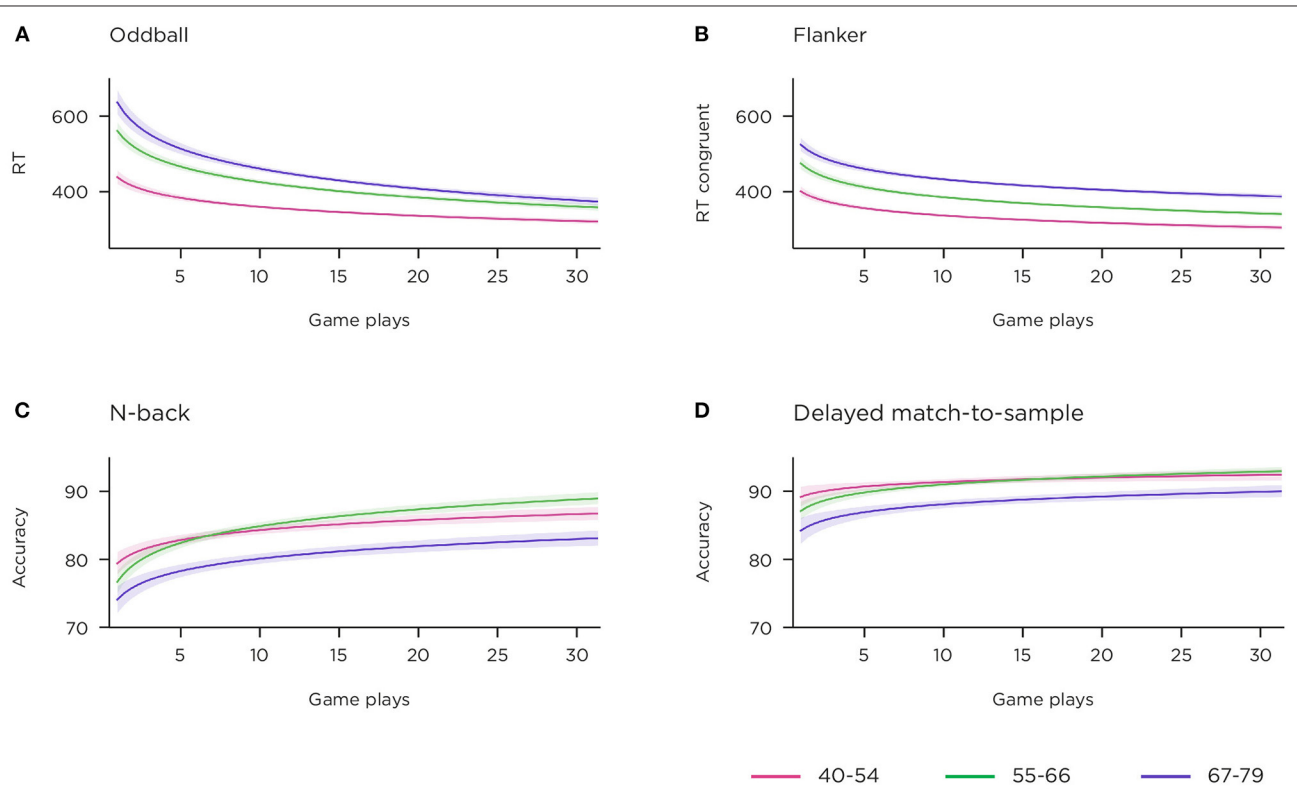


FIGURE 6 | Behavioural responses to gamified cognitive tasks over 12 weeks across age group. Shading indicates 95% confidence interval. **(A)** median correct RTs to targets in 2-stimulus oddball task; **(B)** median correct RTs to congruent trials in flanker task; **(C)** percentage accuracy, all trials, n-back task; **(D)** percentage accuracy, all trials, delayed match-to-sample task.

group means were 98.7 [98.32–99.08]% for 40–54 years, 98.05 [97.16–98.95]% for 55–66 years, and 95.5 [93.62–97.39]% for 67–79 years. As the game rewards speed of response (in its scoring procedure), and due to learning/strategizing effects, we expected reaction time (RT) to improve with repeated gameplays. **Figure 6A** shows the temporal trend of speed of response over consecutive sessions, and clear separation of the three age bands is visible. Reaction time per group, averaged over all sessions, was 366.76 [352.68–380.84] milliseconds (ms) for 40–54 years, 414.85 [393.17–436.54] ms for 55–66 years and 449.58 [423.45–475.70] ms for 67–79 years.

Flanker Game

This task is time-restricted, and encourages the player to make a trade-off between speed and accuracy of response (as erroneous response trials are those that yield the key EEG

metric). Incongruent trials require inhibition and possible motor-replanning relative to congruent trials, and this is seen in an inhibition cost of 62.16 [56.62–67.7] ms. The inhibition cost per age group was 55.87 [48.28–63.45] ms for 40–54 years, 64.67 [56.0–73.34] ms for 55–66 years and 66.89 [52.99–80.8] ms for 67–79 years.

Incongruent trials induced many more errors [8.45 (7.06–9.84)] than congruent trials [2.02 (1.66–2.68)]. Looking at the temporal development of the congruent condition reaction times alone (**Figure 6B**), a pattern of learning from session to session, and separation of age bands, is visible.

N-Back Game

This game requires cycling of information in and out of short-term memory. RTs were slower for non-match [1,075.83 (990.77–1,160.9) ms] than match trials [919.56 (857.76–981.36) ms], however accuracy rates were higher for non-match [87.71 (86.41–89.01)%] vs. match trials [74.76 (71.91–77.60)%]. Accuracy rate was 76.49 [72.96–80.02]% for 40–54 years, 75.56 [70.57–80.55]% for 55–66 years and 70.46 [64.6–76.32]% for 67–79 years on match trials and 87.51 [85.73–89.29]% for 40–54 years, 89.29 [87.41–91.16]% for 55–66 years and 84.77 [81.47–88.08]% for 67–79 years on non-match trials. **Figure 6C** shows accuracy rates on all trials, by age group, across repeated game-plays.

Delayed Match-to-Sample Game

This is not a speed challenge task, however, RTs to match trials were faster than non-match trials 886.99 [847.27–926.71] ms vs. 1,060.31 [1,013.4–1,107.22] ms, a pattern reflected in RTs by age group: 796.27 [754.02–838.52] ms, 890.05 [840.75–939.35] ms and 1,027.89 [914.95–1,140.83] ms for match trials compared to 944.11 [898.82–989.41] ms, 1,079.58 [1,007.6–1,151.56] ms and 1,208.17 [1,102.96–1,313.39] ms for non-match trials, for 40–54, 55–66, and 67–79 years. Memory performance is known to decrease with age, and divergence can be seen in **Figure 6D** for the oldest age-band, though again not between the younger and middle bands. Separate examination of the accuracy for match and non-match trials showed that this difference in performance was primarily driven by the non-match trials. Over the sample, accuracy was 93.16 [92.04–94.28] for match vs. 81.82 [79.35–84.30] for non-match trials, while accuracy rates across the age groups showed more difference for non-match compared to match trials: 85.67 [82.61–88.72], 80.98 [76.72–85.24], and 77.38 [72.51–82.24], compared to 93.85 [92.41–95.28], 93.12[91.24–94.99], and 92.12 [89.44–94.81] for non-match and match trials, respectively, for the age groups 40–54, 55–66, and 67–79 years. These results suggest that the non-match trials acted as effective lures.

EEG

For the resting-state task, power spectral density (PSD) was plotted at occipital sites to explore the effectiveness of the platform to measure change in alpha power between the eyes-open and eyes-closed conditions of resting-state task, across the three different age bands (**Figure 7**). Data from all 78 participants was included in the analysis. The number of sessions per comparison at electrode site O1 for eyes-open/eyes-closed were 904 for 40–54 years, 1,511 for 55–66 years, and 791 for 67–79 years. For electrode site O2, number of sessions were 903 (40–54 years), 1,513 (55–66 years) and 792 (67–79 years). The eyes-open data clearly shows the expected 1/f pattern of signal power falling with increasing frequency, and an alpha band peak around 10 Hz. As expected, the alpha peak amplitude is increased in the eyes-closed condition, as well as in the lower beta band (15–20 Hz). **Figure 7A** displays the effect of age group on absolute band power. There is a clear monotonic decrease in power with age in the difference condition with the largest eyes-open/eyes-closed difference for those aged 40–54 and the smallest difference for those aged 67–79 in the alpha and lower beta range. Furthermore, it can be seen that the average peak alpha frequency is highest for younger participants, and lowest for older participants. No consistent pattern is apparent in alpha power for the eyes-open and eyes-closed conditions alone, although there are clear distinctions between groups in the gamma range (30–35 Hz). This may indicate a difference in noise floor between the age groups. We applied a suitable normalisation by taking the relative power on this analysis (80). Relative power is shown in **Figure 7B**, again demonstrating a stratified pattern of age group on alpha power and peak frequency, most evident in the graph of the eyes-closed condition. It is noticeable that there is more fluctuation in the higher frequencies for the oldest age group (67–79 year olds).

Evoked and event-related potentials elicited in the gamified 2-stimulus oddball and flanker tasks were also extracted at a single-session, single-participant and grand average level (shown in **Figures 8**, **9**). **Figure 8C** illustrates the grand average ERP for target trials on the 2-stimulus oddball, at the centro-parietal location CPz, where the P300 is centred. This is a robust average over multiple sessions contributed by 77 participants, time-locked to the presentation of the stimulus (data from one participant, $n = 26$ sessions, did not meet quality thresholds for inclusion at this channel). Interpolated topographies (**Figure 8D**) across all 16 channels at ERP peaks are shown at 0, 200 and 420 ms post-stimulus onset. The principal waveform features of a P300 ERP are visible in the early sensory processing components ($\sim$0–250 ms with an occipital focus) and the P300 component ($\sim$300–500 ms, with a centro-parietal focus). A strong readiness potential can also be seen before stimulus presentation ($-$500–0 ms). The other two panels show the median stimulus-locked epoch from 29 correctly identified target trials from a single game-play session (**Figure 8A**), and the median-average across 18 out of a total of 21 game-play sessions (3 did not meet quality thresholds), contributed by a participant aged 44 years (**Figure 8B**). **Figure 8E** demonstrates examples of successful session-level ERPs evoked during a single game-play session, representing 6 users across the different age groups in the study (2 participants from each age group). Unsuccessful sessions yield waveforms that show a discernible ERP overlaid with noise, flat-line signals (e.g., due to an unconnected sensor), or noise of various heterogenous types.

Figure 9C illustrates the grand average difference ERP for error trials on the flanker task, at the central Cz location, where the ERN is observed (total 1,004 sessions). This is a robust average over multiple sessions contributed by 76

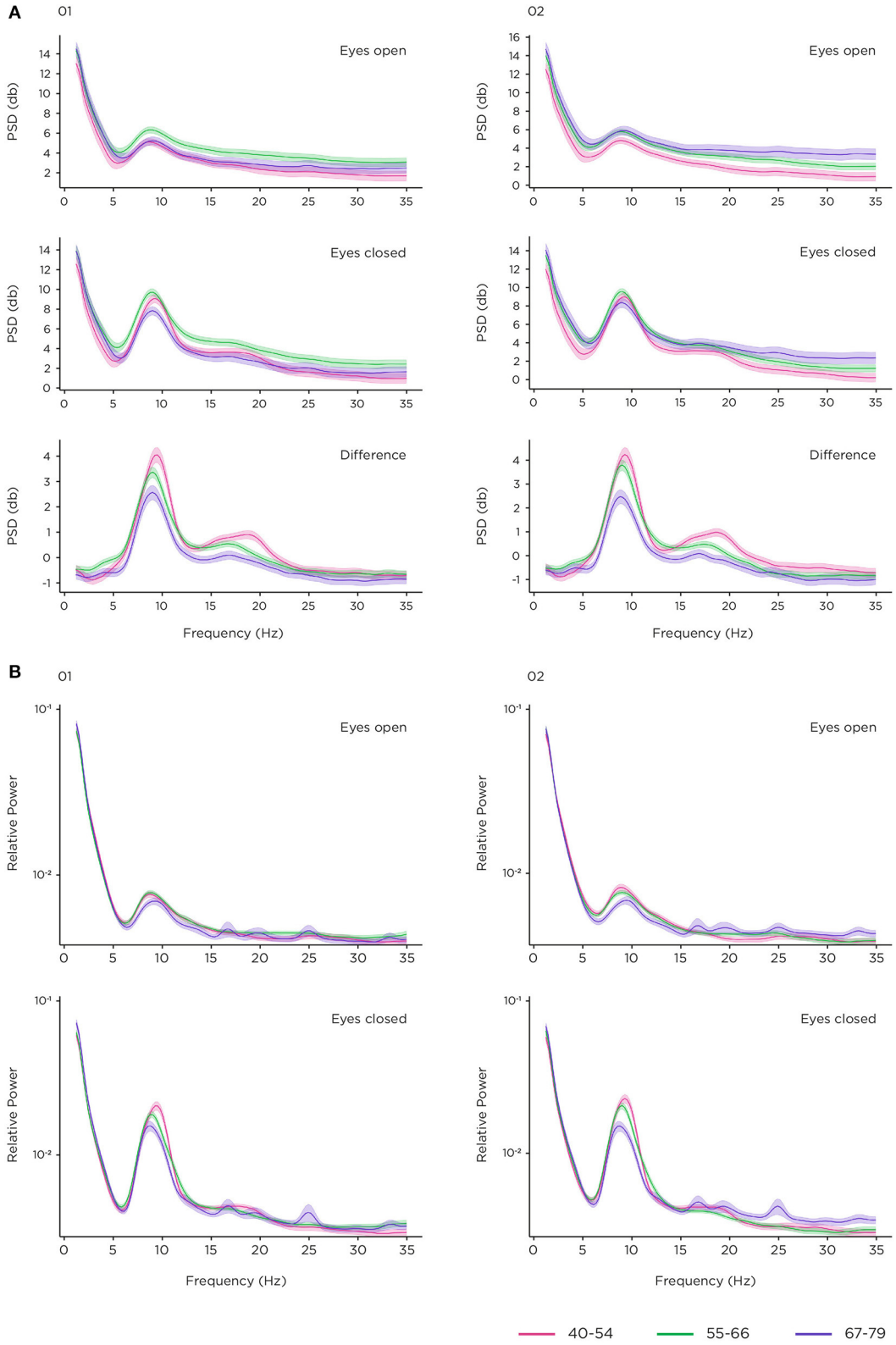


FIGURE 7 | Resting state task. **(A)** power spectral density (PSD) in decibels (dB) at O1 and O2 by age group in eyes-open and eyes-closed conditions, and the difference condition; **(B)** relative power at O1 and O2 by age group in eyes-open and eyes-closed conditions with logarithmic scaling for display only.

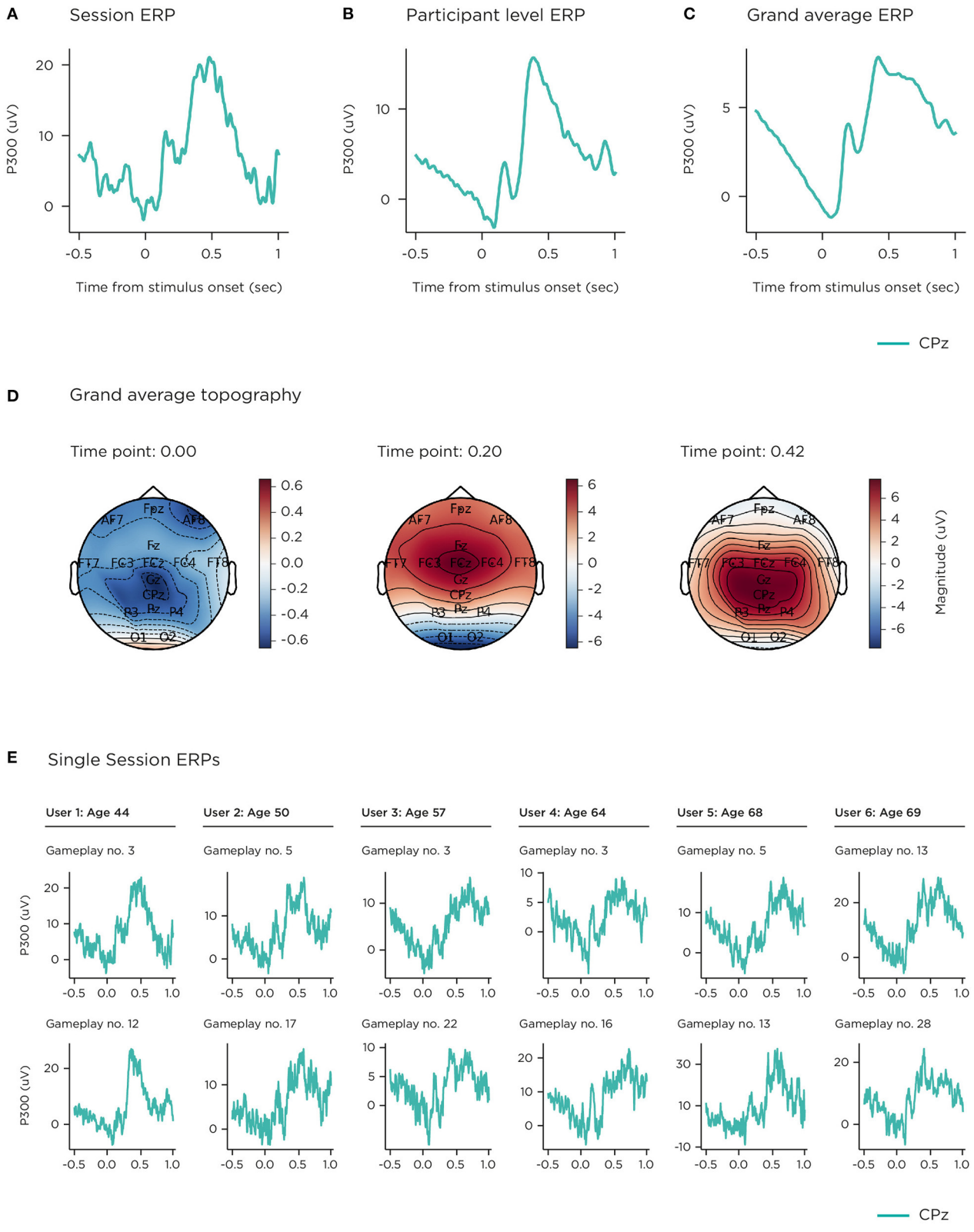


FIGURE 8 | P300. **(A)** single-session median; **(B)** single-participant mean; **(C)** grand mean; **(D)** grand mean topographies selected at ERP peak timepoints; **(E)** examples of single-session median ERPs successfully recorded from game plays from 6 different participants (2 participants per age group).

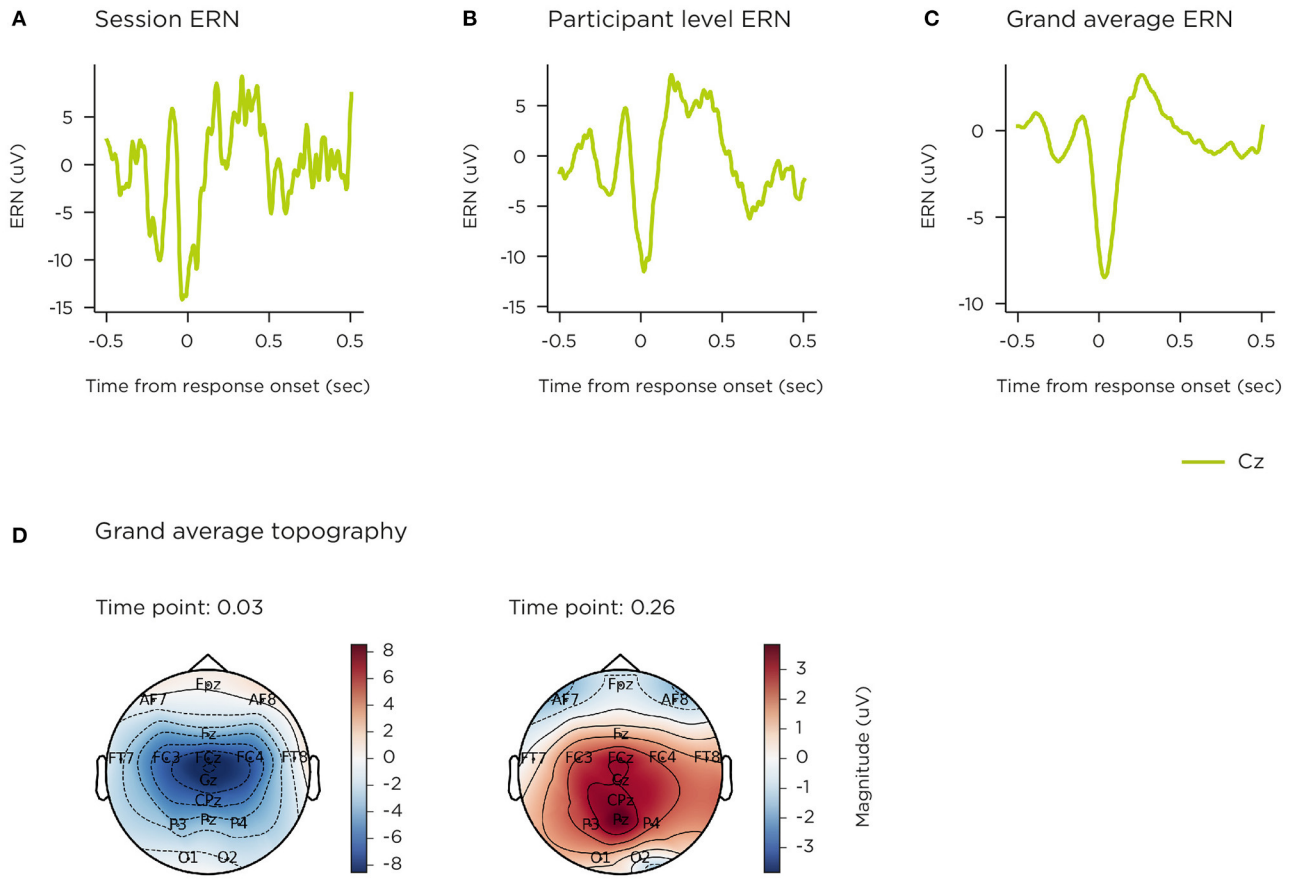


FIGURE 9 | ERN. **(A)** single-session median; **(B)** single-participant mean; **(C)** grand mean; **(D)** grand mean topographies selected at ERP peak timepoints.

participants, time-locked to touch-response (single-channel data was excluded from 2 participants, who contributed 6 and 8 sessions). Grand average topography (**Figure 9D**) across all 16 channels at ERP peaks are shown at 30 and 260 ms post-stimulus onset. The event-related negativity (ERN) waveform is clearly represented (peaking at ∼50 ms), with a central focus, as is the error positivity (Pe) with a centro-parietal focus (peak ∼250 ms). **Figure 9A** shows the corresponding median response-locked epoch across 12 error trials following a single game-play session, contributed by a participant aged 52 years. **Figure 9B** displays that participant's median-average across 14 out of 16 game-play sessions.

The P300 and ERN components were also compared by age group. **Figure 10A** shows grand-averaged epochs per age group on single channel CPz for the stimulus-locked ERP from the visual oddball, and **Figure 10B** shows these at Cz for the response-locked ERP of the flanker task. The number of participants and sessions per comparison at electrode site CPz were 26 and 397 for 40–54 years; 34 and 515 for 55–66 years; 17 and 240 for 67–79 years. For electrode site Cz, number of participants and sessions were 23 and 339 (40–54 years); 35 and 473 (55–66 years); 17 and 182 (67–79 years). The impact of signal variability from individual sessions (both noise, and genuine inter-individual differences) is quantified in the 95% confidence intervals illustrated.

The P300 shows early differences in latency in sensory processing, and separation in amplitude among the groups, where disruption increases with age. In the Flanker task ERP, the early ERN component is attenuated in both older groups, relative to the youngest group, and the later Pe component is reduced for the oldest group.

DISCUSSION

Usability

This proof-of-concept paper reports findings from the first time that this novel EEG platform was deployed in-field. After a single training session, participants, including older adults up to the age of 79 years, were able to use the technology at home to successfully perform EEG and behavioural recordings without the supervision of trained technicians. This study yielded 3,603 uploaded sessions, >95% of which contained usable data (i.e., EEG and behavioural metrics could be extracted from the submitted data), providing encouraging evidence supporting the feasibility of this technological approach to cognitive neuroscience research.

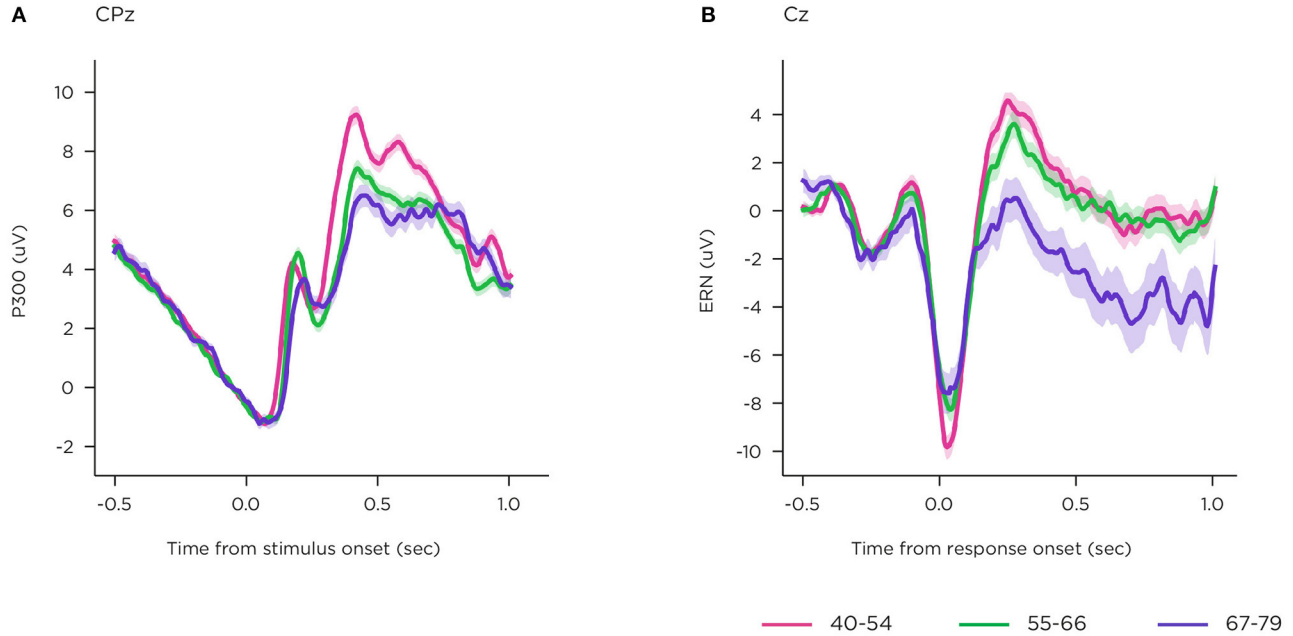


FIGURE 10 | Showing age-related differences in event-related components recorded using the platform. **(A)** P300; **(B)** ERN.

Users were asked to trial and evaluate this new system for monitoring brain health at home and to contribute five 30-min recording sessions per week, which was a considerable effort given that no extrinsic incentives or disincentives were applied to promote adherence to this schedule. Adherence to this schedule was remarkably high, relative to other reports of in-home monitoring devices in older populations over similar time courses [∼55% (81, 82)]. For the current study, attrition was low (12.4%) and the average contribution was >4 sessions per week (>80% adherence to schedule). Older adults had the highest rates of adherence, indicating that age was no impediment to using the system regularly. A high level of adherence was maintained throughout the 3-month period without substantial decline in the latter weeks of participation, testament to both the power of gamification and usability of the system motivating and facilitating repeated play, and the level of commitment from the study participants. Maintaining adherence in unsupervised environments is challenging and may be particularly so for psychiatric populations (83), however widespread evidence from other diseases, where there has been a broad uptake in new technologies, indicates that patient-centric digital monitoring provides more objective, frequent tracking with clear healthcare benefits (84–86) and have been shown to lead to better compliance relative to paper based assessments (86, 87).

Reported usability was somewhat lower for the oldest age-band, although it is worth noting that their lowest average score still falls within the range between "ok" and "good" (88). Contrary to our expectations, this did not result in reduced adherence, suggesting that many challenges of manual dexterity or familiarity with digital technologies had been successfully mitigated during the initial user-focused design process. However, signal reliability measures indicated that the oldest age band experienced the greatest difficulty in obtaining good sensor connectivity, particularly around the midline sensors (Cz) located at the top of the head. Younger participants achieved slightly better connectivity at Cz but also better connectivity on adjacent sensors. Head shape is variable at the crown, meaning that generic headset sizing options are not always optimal. That location may also require additional manual dexterity and adjustment, which is more difficult for older populations. In the light of these findings from the first deployment of this technology in-field, subsequent incremental improvements to the headset, app and enrolment training procedures have been deployed which have resulted in superior sensor connectivity and data quality (89, 90).

In addition to investigating the overall usability of the platform, empirical data collected in this study was used to assess the potential of collecting scientifically valid neurocognitive data from remote, fully autonomous participants.

Behaviour

All gamified cognitive tasks exhibited some degree of a learning effect. Reaction times generally decreased rapidly over the first five sessions of a given task (see **Figure 6**). This likely reflects the development of task-specific perceptual-motor skill, rather than a change in the underlying cognitive function probed by the task (i.e., "brain training"). Time spent developing task strategies, and learning the layout of the task environment, is likely to have enabled more effective allocation of visual attention and therefore faster responding (91). Age-related effects on speed of response were generally preserved throughout this learning phase and

into asymptote, with the oldest participants consistently making their responses more slowly than other age groups, consistent with the literature. Regarding accuracy over time, participants on average improved very slowly and consistently on the n-back task throughout the study, as expected since n-back variants are typically included with brain training suites due to their inherent learnability (92). Participants, regardless of age band, demonstrated less of a learning effect in the visual delayed match-to-sample task, potentially reflecting the simplicity of this basic old/new matching task for healthy adults however again, the oldest participants scored lowest on this task throughout the study, consistent with age-related decline in memory performance. These learning curves themselves (**Figure 6**), enabled through the ability to collect multiple assessments over time, may be a rich source of data and potentially informative measures of underlying cognitive function (93), with recent research showing that rate of learning, in the context of a cognitive task conducted over multiple days, can differentiate groups by age (94) and neuropathology (95).

EEG

EEG signals described in this paper (resting-state spectral activity and the P300 and ERN ERPs) demonstrate morphology consistent with those elicited by non-gamified, laboratory paradigms described in the literature. Furthermore, grand average visualisations of P300 and ERN ERPs across age bands replicate classic electrophysiological patterns of age-related change.

The study design included repeated use to permit aggregation of EEG data collected in the home, as a means of improving reliability and signal to noise ratio. The focus in this paper is on group level grand-average analyses, common in cognitive neuroscience literature. Although, as can be seen in the data presented in this paper, it is feasible to obtain cognitive ERP components from users (across different age groups) based on single, home-based sessions. As might be expected, not all sessions were available for analysis with factors such as saturated signal or high variance rendering the data unusable. However, over 95% of sessions contained EEG from which at least a portion of the data was usable, even though certain sections of that session, or certain sensors, may have been noisy. In order to support participants to achieve good signal quality, the system included a sensor signal check step at the beginning of every session to give feedback on impedance levels to encourage self-adjustment for a good connection.

EEG devices that offer miniaturisation of the EEG amplifier, use of Bluetooth technology to transmit EEG data, precise stimulus event-marking, and a choice of wet or dry sensor set-ups, have been extensively evaluated in the literature [e.g., (54, 96–99)]. These demonstrate reduced set-up times and greater portability while generally maintaining good signal-to-noise ratio [e.g., (52, 53, 100, 101); but see Duvinage et al. (102), Maskeliunas et al. (103)]. However, the authors are not aware of reports of any other mobile EEG system for which repeated ERP data collection has been demonstrated in participants' home environment without a researcher present, as was the case in this study.

The P300 elicited from the 2-stimulus oddball task exhibited reduced amplitude and latency for the older age groups, consistent with previous studies (62–64). Whilst the underlying mechanisms are yet not fully elucidated, recent evidence points toward the P300 reflecting the accumulation of information leading to a decision (60), the ability to do this effectively being impaired by ageing and cognitive decline (104). ERP components evoked from the Flanker task also demonstrated sensitivity to ageing with a smaller ERN for the older age groups and a weak Pe for the oldest age band, reproducing known effects in the literature, reflecting a general weakening of the processes underpinning cognitive control in ageing populations (67, 105). Resting state EEG PSD demonstrated alpha band increase in the eyes closed condition relative to eyes open as expected. In the absolute power analyses differences in noise levels were observed. Grummet et al. (106) discusses variability in noise floor in dry EEG, which in this case may potentially be driven by factors such as systematic variation in the use of the headset, and age-related changes in skin condition (107). However, having controlled for differences in noise floor levels, we observed alpha band power and peak frequency stratification across that age groups that align with the effects of ageing on brain activity during resting state typically reported in the literature (74, 108, 109).

CONCLUSION

In this paper we have described the first large scale field trial of a new suite of tools to collect clinically relevant domain-specific markers of brain function and cognitive performance unsupervised in the home. Human-factors feasibility was demonstrated by high reported usability, low levels of withdrawal, and adherence of >80% over a 5-day-per-week, 3-month long, uncompensated participation. Newly gamified versions of established tasks were trialled and were successful in replicating key aspects of behaviour from their lab-based counterparts. Widespread learning effects were observed, as would be expected on repeated plays, but age-related differences were preserved over many weeks of repeated play. Grand average EEG data from the resting state, visual oddball and flanker tasks all illustrated core features of frequency content, waveform morphology and timing, and scalp topographies to confirm that they faithfully replicate the lab-based tasks on which they were modelled.

Challenges of data quality were encountered. On an average session, 14 or 15 sensors (of 16) provided EEG signals that could be analysed, the remainder lost due to issues with contact reliability with particular scalp locations and age cohorts. Certain sessions were evaluated as too noisy for inclusion in grand average analyses and early behavioural sessions proved more variable than later ones. Since this study was completed, incremental improvements to the headset, tablet-based app and participant familiarisation procedures have been made that have increased signal quality (89, 90).

While the focus of this paper is on ageing, and cognitive functions of relevance to Alzheimer's disease and other pathologies underlying dementia, this suite of tools can also

include additional tasks (e.g., emotional face processing, passive auditory oddball) suitable for use in mood disorders, psychosis (110), and measurement of treatment response in psychiatry (89, 111).

Advances in wearable electronics, dry sensors and user-facing interactive technologies enable EEG as an easy-to-use affordable biomarker of cognition, grounded directly in brain function. Decades of scientific literature support EEG as an emerging translational biomarker for disease cases in neuropsychiatric (schizophrenia, depression) (20, 112, 113) and neurodegenerative (e.g., Alzheimer's) disease (114–116). Sampling a broad suite of cognitive functions (including memory, attention, and executive function) offers coverage of multiple cognitive domains, which has greater predictive accuracy for disease progression (117, 118). Cloud computing can securely collect data from distributed locations, automatically evaluate quality, and use machine learning techniques to derive composite markers based on neural activity and behavioural performance from single and multiple cognitive domains (119). These innovations in technology, supported by scientific literature, make it possible to use large-scale longitudinal sampling of real-world data, to support potential future use-cases in early detection, personalised medicine, progression tracking, and measurement of treatment response for neuropsychiatric disorders.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the data collected using the Cumulus Neuroscience platform is commercially sensitive and contains proprietary information. Requests to access supporting data will be considered from bona fide researchers upon reasonable request. Requests to access the datasets should be directed to alison@cumulusneuro.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Queen's University Belfast Faculty Research Ethics Committee (EPS). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

EM, BMu, BMc, PP, and AB contributed to the conception and design of the study and paper. AB performed data collection. EM, LR-D, HN, MI, JD, FB, BMu, and AB performed analysis and interpretation. EM, JD, BMu, and AB wrote the manuscript. LR-D, HN, and MI wrote sections of the manuscript. All authors critically revised the manuscript and approve the submitted version.

FUNDING

The study on which this paper is based was funded by the Innovate UK Biomedical Catalyst programme (Grant Number: IUK 102862). LR-D was supported by Science Foundation Ireland (18/IF/6272) and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 893823.

ACKNOWLEDGMENTS

The authors wish to thank Liam Watson and Aoife Sweeney for their part in data collection and Weronika Jaworowska-Bialko for her contributions to usability analysis. We also extend our gratitude to all our participants who gave up their time so generously to take part in this 12-week study. We would like to thank Xi Jiang and Albert Yang for their helpful and constructive feedback during review.

REFERENCES

1. Brietzke E, Hawken ER, Idzikowski M, Pong J, Kennedy SH, Soares CN. Integrating digital phenotyping in clinical characterization of individuals with mood disorders. *Neurosci Biobehav Rev.* (2019) 104:223–30. doi: 10.1016/j.neubiorev.2019.07.009
2. Carpenter G, Harbin HT, Smith RL, Hornberger J, Nash DB. A promising new strategy to improve treatment outcomes for patients with depression. *Popul Health Manage.* (2019) 22:223–8. doi: 10.1089/pop.2018.0101
3. Ermers NJ, Hagoort K, Scheepers FE. The predictive validity of machine learning models in the classification and treatment of major depressive disorder: state of the art and future directions. *Front Psychiatry.* (2020) 11:472. doi: 10.3389/fpsyt.2020.00472
4. Gillan CM, Whelan R. What big data can do for treatment in psychiatry. *Curr Opin Behav Sci.* (2017) 18:34–42. doi: 10.1016/j.cobeha.2017.07.003
5. Jollans L, Whelan R. Neuromarkers for mental disorders: harnessing population neuroscience. *Front Psychiatry.* (2018) 9:242. doi: 10.3389/fpsyt.2018.00242
6. Newson JJ, Thiagarajan TC. EEG frequency bands in psychiatric disorders : a review of resting state studies. *Front Hum Neurosci.* (2019) 12:521. doi: 10.3389/fnhum.2018.00521
7. Russ TC, Woelbert E, Davis KAS, Hafferty JD, Ibrahim Z, Inkster B, et al. How data science can advance mental health research. *Nat Hum Behav.* (2019) 3:24–32. doi: 10.1038/s41562-018-0470-9
8. Clapp WC, Hamm JP, Kirk IJ, Teyler TJ. Translating long-term potentiation from animals to humans: a novel method for noninvasive assessment of cortical plasticity. *Biol Psychiatry.* (2012) 71:496–502. doi: 10.1016/j.biopsych.2011.08.021
9. Drinkenburg WHIM, Ruigt GSF, Ahnaou A. Pharmaco-EEG studies in animals: an overview of contemporary translational applications. *Neuropsychobiology.* (2015) 72:151–64. doi: 10.1159/000442210
10. Leiser SC, Dunlop J, Bowlby MR, Devilbiss DM. Aligning strategies for using EEG as a surrogate biomarker: a review of preclinical and clinical research. *Biochem Pharmacol.* (2011) 81:1408–21. doi: 10.1016/j.bcp.2010.10.002
11. Atkinson RJ, Michie PT, Schall U. Duration mismatch negativity and P3a in first-episode psychosis and individuals at ultra-high risk of psychosis. *Biol Psychiatry.* (2012) 71:98–104. doi: 10.1016/j.biopsych.2011.08.023
12. Erickson MA, Ruffle A, Gold JM. A meta-analysis of mismatch negativity in schizophrenia: from clinical risk to disease specificity and progression. *Biol Psychiatry.* (2016) 79:980–7. doi: 10.1016/j.biopsych.2015.08.025
13. Hamilton HK, Boos AK, Mathalon DH. Electroencephalography and event-related potential biomarkers in individuals at clinical high risk for psychosis. *Biol Psychiatry.* (2020) 88:294–303. doi: 10.1016/j.biopsych.2020.04.002

14. Light GA, Naatanen R. Mismatch negativity is a breakthrough biomarker for understanding and treating psychotic disorders. *Proc Natl Acad Sci USA*. (2013) 110:15175–6. doi: 10.1073/pnas.1313287110
15. Näätänen R, Todd J, Schall U. Mismatch negativity (MMN) as biomarker predicting psychosis in clinically at-risk individuals. *Biol Psychol.* (2016) 116:36–40. doi: 10.1016/j.biopsycho.2015.10.010
16. Nagai T, Tada M, Kirihara K, Araki T, Jinde S, Kasai K. Mismatch negativity as a "translatable" brain marker toward early intervention for psychosis: a review. *Front Psychiatry.* (2013) 4:115. doi: 10.3389/fpsyt.2013.00115
17. Perez VB, Roach BJ, Woods SW, Srihari VH, McGlashan TH, Ford JM, et al. Chapter 10 - Early auditory gamma-band responses in patients at clinical high risk for schizophrenia. In: Başar E, Başar-Eroğlu C, Özerdem A, Rossini PM, Yener GG, editors. *Application of Brain Oscillations in Neuropsychiatric Diseases: Selected Papers from "Brain Oscillations in Cognitive Impairment and Neurotransmitters"*, Vol. 62. Istanbul: Elsevier (2013). p. 1–375.
18. Perez VB, Woods SW, Roach BJ, Ford JM, McGlashan TH, Srihari VH, et al. Automatic auditory processing deficits in schizophrenia and clinical high-risk patients: forecasting psychosis risk with mismatch negativity. *Biol Psychiatry.* (2014) 75:459–69. doi: 10.1016/j.biopsych.2013.07.038
19. Normann C, Schmitz D, Fürmaier A, Döing C, Bach M. Long-Term plasticity of visually evoked potentials in humans is altered in major depression. *Biol Psychiatry.* (2007) 62:373–80. doi: 10.1016/j.biopsych.2006.10.006
20. Olbrich S, van Dinteren R, Arns M. Personalized medicine: review and perspectives of promising baseline EEG biomarkers in major depressive disorder and attention deficit hyperactivity disorder. *Neuropsychobiology.* (2015) 72:229–40. doi: 10.1159/000437435
21. Stewart JL, Bismark AW, Towers DN, Coan JA, Allen JJB. Resting frontal EEG asymmetry as an endophenotype for depression risk: sex-specific patterns of frontal brain asymmetry. *J Abnorm Psychol.* (2010) 119:502–12. doi: 10.1037/a0019196
22. Babiloni C, Lizio R, Del Percio C, Marzano N, Soricelli A, Salvatore E, et al. Cortical sources of resting state EEG rhythms are sensitive to the progression of early stage Alzheimer's disease. *J Alzheimers Dis.* (2013) 34:1015–35. doi: 10.3233/JAD-121750
23. Choi J, Ku B, You YG, Jo M, Kwon M, Choi Y, et al. Resting-state prefrontal EEG biomarkers in correlation with MMSE scores in elderly individuals. *Sci Rep.* (2019) 9:10468. doi: 10.1038/s41598-019-46789-2
24. Ghorbanian P, Devilbiss DM, Verma A, Bernstein A, Hess T, Simon AJ, et al. Identification of resting and active state EEG features of alzheimer's disease using discrete wavelet transform. *Ann Biomed Eng.* (2013) 41:1243–57. doi: 10.1007/s10439-013-0795-5
25. Lai CL, Lin RT, Liou LM, Liu CK. The role of event-related potentials in cognitive decline in Alzheimer's disease. *Clin Neurophysiol.* (2010) 121:194–9. doi: 10.1016/j.clinph.2009.11.001
26. Musaeus CS, Nielsen MS, Østerbye NN, Høgh P. Decreased parietal beta power as a sign of disease progression in patients with mild cognitive impairment. *J Alzheimers Dis.* (2018) 65:475–87. doi: 10.3233/JAD-180384
27. Parra MA, Ascencio LL, Urquina HF, Manes F, Ibáñez AM. P300 and neuropsychological assessment in mild cognitive impairment and alzheimer dementia. *Front Neurol.* (2012) 3:172. doi: 10.3389/fneur.2012.00172
28. Smailovic U, Jelic V. Neurophysiological markers of alzheimer's disease: quantitative EEG approach. *Neurol Ther.* (2019) 8:37–55. doi: 10.1007/s40120-019-00169-0
29. Tóth B, File B, Boha R, Kardos Z, Hidasi Z, Gaál ZA, et al. EEG network connectivity changes in mild cognitive impairment - Preliminary results. *Int J Psychophysiol.* (2014) 92:1–7. doi: 10.1016/j.ijpsycho.2014.02.001
30. Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci Biobehav Rev.* (2014) 44:58–75. doi: 10.1016/j.neubiorev.2012.10.003
31. Wilson GF, Russell CA, Monnin JW, Estepp JR, Christensen JC. How does day-to-day variability in psychophysiological data affect classifier accuracy? *Proc Hum Fact Ergonom Soc.* (2010) 1:264–8. doi: 10.1177/154193121005400317
32. Duke Han S, Nguyen CP, Stricker NH, Nation DA. Detectable neuropsychological differences in early preclinical alzheimer's disease: a meta-analysis. *Neuropsychol Rev.* (2017) 27:305–25. doi: 10.1007/s11065-017-9366-0
33. Hassenstab J, Ruvolo D, Jasielec M, Xiong C, Grant E, Morris JC. Absence of practice effects in preclinical Alzheimer's disease. *Neuropsychology.* (2015) 29:940–8. doi: 10.1037/neu0000208
34. Binder LM, Iverson GL, Brooks BL. To err is human: "abnormal" neuropsychological scores and variability are common in healthy adults. *Arch Clin Neuropsychol.* (2009) 24:31–46. doi: 10.1093/arclin/acn001
35. Cooley SA, Heaps JM, Bolzenius JD, Salminen LE, Baker LM, Scott SE, et al. Longitudinal change in performance on the montreal cognitive assessment in older adults. *Clin Neuropsychol.* (2015) 29:824–35. doi: 10.1080/13854046.2015.1087596
36. Falleti M, Maruff P, Collie A, Darby D. Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *J Clin Exp Neuropsychol.* (2006) 28:1095–112. doi: 10.1080/13803390500205718
37. Goldberg TE, Harvey PD, Wesnes KA, Snyder PJ, Schneider LS. Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimers Dement Diag Assess Dis Monitor.* (2015) 1:103–11. doi: 10.1016/j.dadm.2014.11.003
38. Mario B, Massimiliano M, Chiara M, Alessandro S, Antonella C, Gianfranco F. White-coat effect among older patients with suspected cognitive impairment: prevalence and clinical implications. *Int J Geriatr Psychiatry.* (2009) 24:509–17. doi: 10.1002/gps.2145
39. Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. *Assessment.* (2018) 25:14–30. doi: 10.1177/1073191116643164
40. Brose A, Schmiedek F, Lövdén M, Lindenberger U. Daily variability in working memory is coupled with negative affect: the role of attention and motivation. *Emotion.* (2012) 12:605–17. doi: 10.1037/a0024436
41. Brose A, Lövdén M, Schmiedek F. Daily fluctuations in positive affect positively co-vary with working memory performance. *Emotion.* (2014) 14:1–6. doi: 10.1037/a0035210
42. Hess TM, Ennis GE. Age differences in the effort and costs associated with cognitive activity. *J Gerontol Ser B Psychol Sci Soc Sci.* (2012) 67:447–55. doi: 10.1093/geronb/gbr129
43. Metternich B, Schmidtke K, Hüll M. How are memory complaints in functional memory disorder related to measures of affect, metamemory and cognition? *J Psychosomat Res.* (2009) 66:435–44. doi: 10.1016/j.jpsychores.2008.07.005
44. Stawski RS, Sliwinski MJ, Smyth JM. Stress-related cognitive interference predicts cognitive function in old age. *Psychol Aging.* (2006) 21:535–44. doi: 10.1037/0882-7974.21.3.535
45. Hashemi A, Pino LJ, Moffat G, Mathewson KJ, Aimone C, Bennett PJ, et al. Characterizing population EEG dynamics throughout adulthood. *ENeuro.* (2016) 3:1–13. doi: 10.1523/ENEURO.0275-16.2016
46. Cormack FK, Taptiklis N, Abbott RA, Anatürk M, Cartland I, Coppieters L, et al. Changes to validity of online cognitive assessment in young and older adults: a comparison to supervised testing using the cantab battery. *Alzheimers Dement.* (2016) 12:P286–7. doi: 10.1016/j.jalz.2016.06.520
47. Lancaster C, Koychev I, Blane J, Chinner A, Chatham C, Taylor K, et al. Gallery game: smartphone-based assessment of long-term memory in adults at risk of Alzheimer's disease. *J Clin Exp Neuropsychol.* (2020) 42:329–43. doi: 10.1080/13803395.2020.1714551
48. Moore RC, Swendsen J, Depp CA. Applications for self-administered mobile cognitive assessments in clinical research: a systematic review. *Int J Methods Psychiatr Res.* (2017) 26:e1562. doi: 10.1002/mpr.1562
49. Resnick HE, Lathan CE. From battlefield to home: a mobile platform for assessing brain health. *MHealth.* (2016) 2:30. doi: 10.21037/mhealth.2016.07.02
50. Rogers JM, Johnstone SJ, Aminov A, Donnelly J, Wilson PH. Test-retest reliability of a single-channel, wireless EEG system. *Int J Psychophysiol.* (2016) 106:87–96. doi: 10.1016/j.ijpsycho.2016.06.006
51. Wong SWH, Chan RHM, Mak JN. Spectral modulation of frontal EEG during motor skill acquisition: a mobile EEG study. *Int J Psychophysiol.* (2014) 91:16–21. doi: 10.1016/j.ijpsycho.2013.09.004
52. Badcock NA, Mousikou P, Mahajan Y, De Lissa P, Thie J, McArthur G. Validation of the emotiv EPOC® EEG gaming systemfor measuring research quality auditory ERPs. *PeerJ.* (2013) 2013:e38. doi: 10.7717/peerj.38

53. Badcock NA, Preece KA, de Wit B, Glenn K, Fieder N, Thie J, et al. Validation of the emotiv EPOC EEG system for research quality auditory event-related potentials in children. *PeerJ.* (2015) 3:e907. doi: 10.7717/peerj.907
54. Krigolson OE, Williams CC, Norton A, Hassall CD, Colino FL. Choosing MUSE: validation of a low-cost, portable EEG system for ERP research. *Front Neurosci.* (2017) 11:109. doi: 10.3389/fnins.2017.00109
55. Kuziek JWP, Shienh A, Mathewson KE. Transitioning EEG experiments away from the laboratory using a Raspberry Pi 2. *J Neurosci Methods.* (2017) 277:75–82. doi: 10.1016/j.jneumeth.2016.11.013
56. Lumsden J, Edwards EA, Lawrence NS, Coyle D, Munaf,ò MR. Gamification of cognitive assessment and cognitive training: a systematic review of applications and efficacy. *JMIR Ser Games.* (2016) 4:e11. doi: 10.2196/games.5888
57. Coughlan G, Coutrot A, Khondoker M, Minihane AM, Spiers H, Hornberger M. Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. *Proc Natl Acad Sci USA.* (2019) 116:9285–92. doi: 10.1073/pnas.1901600116
58. Herrmann CS, Knight RT. Mechanisms of human attention: event-related potentials and oscillations. *Neurosci Biobehav Rev.* (2001) 25:465–76. doi: 10.1016/S0149-7634(01)00027-6
59. Polich J, Kok A. Cognitive and biological determinants of P300: an integrative review. *Biol Psychol.* (1995) 41:103–46. doi: 10.1016/0301-0511(95)05130-9
60. Twomey DM, Murphy PR, Kelly SP, O'Connell RG. The classic P300 encodes a build-to-threshold decision variable. *Euro J Neurosci.* (2015) 42:1636–43. doi: 10.1111/ejn.12936
61. Woods DL, Wyma JM, Yund EW, Herron TJ, Reed B. Age-related slowing of response selection and production in a visual choice reaction time task. *Front Hum Neurosci.* (2015) 9:193. doi: 10.3389/fnhum.2015.00193
62. Fjell AM, Walhovd KB. P300 and neuropsychological tests as measures of aging: scalp topography and cognitive changes. *Brain Topogr.* (2001) 14:25–40. doi: 10.1023/A:1012563605837
63. Pavarini SCI, Brigola AG, Luchesi BM, Souza ÉN, Rossetti ES, Fraga FJ, et al. On the use of the P300 as a tool for cognitive processing assessment in healthy aging: a review. *Dement Neuropsychol.* (2018) 12:1–11. doi: 10.1590/1980-57642018dn12-010001
64. van Dinteren R, Arns M, Jongsma MLA, Kessels RPC. P300 development across the lifespan: a systematic review and meta-analysis. *PLoS ONE.* (2014) 9:e0087347. doi: 10.1371/journal.pone.0087347
65. Yeung N, Botvinick MM, Cohen JD. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol Rev.* (2004) 111:931–59. doi: 10.1037/0033-295X.111.4.931
66. Nieuwenhuis S, Ridderinkhof KR, Blom J, Band GPH, Kok A. Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology.* (2001) 38:752–60. doi: 10.1111/1469-8986.3850752
67. Hoffmann S, Falkenstein M. Aging and error processing: age related increase in the variability of the error-negativity is not accompanied by increase in response variability. *PLoS ONE.* (2011) 6:e0017482. doi: 10.1371/journal.pone.0017482
68. Dong S, Reder LM, Yao Y, Liu Y, Chen F. Individual differences in working memory capacity are reflected in different ERP and EEG patterns to task difficulty. *Brain Res.* (2015) 1616:146–56. doi: 10.1016/j.brainres.2015.05.003
69. Bopp KL, Verhaeghen P. Aging and n-back performance: a meta-analysis. *J Gerontol Ser B.* (2018) 75:229–40. doi: 10.1093/geronb/gby024
70. Grady C. The cognitive neuroscience of ageing. *Nat Rev Neurosci.* (2012) 13:491–505. doi: 10.1038/nrn3256
71. Mecklinger A. On the modularity of recognition memory for object form and spatial location: a topographic ERP analysis. *Neuropsychologia.* (1998) 36:441–60. doi: 10.1016/S0028-3932(97)00128-0
72. Alloway CED, Ogilvie RD, Shapiro CM. The alpha attenuation test: assessing excessive daytime sleepiness in narcolepsy-cataplexy. *Sleep.* (1997) 20:258–66. doi: 10.1093/sleep/20.4.258
73. Barry RJ, Rushby J, Wallace M, Clarke A, Johnstone S, Zlojutro I. Caffeine effects on resting-state arousal. *Clin Neurophysiol.* (2005) 116:2693–700. doi: 10.1016/j.clinph.2005.08.008
74. Rossini PM, Rossi S, Babiloni C, Polich J. Clinical neurophysiology of aging brain: from normal aging to neurodegeneration. *Progr Neurobiol.* (2007) 83:375–400. doi: 10.1016/j.pneurobio.2007.07.010
75. Stampi C, Stone P, Michimori A. A new quantitative method for assessing sleepiness: the alpha attenuation test. *Work Stress.* (1995) 9:368–76. doi: 10.1080/02678379508256574
76. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc.* (2005) 53:695–9. doi: 10.1111/j.1532-5415.2005.53221.x
77. Coen RF, Cahill R, Lawlor BA. Things to watch out for when using the montreal cognitive assessment (MoCA). *Int J Geriatr Psychiatry.* (2011) 26:107–8. doi: 10.1002/gps.2471
78. Kenny RA, Coen RF, Frewen J, Donoghue OA, Cronin H, Savva GM. Normative values of cognitive and physical function in older adults: findings from the irish longitudinal study on ageing. *J Am Geriatr Soc.* (2013) 61:S279–90. doi: 10.1111/jgs.12195
79. Brooke J. SUS: a "quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. *Usability Evaluation in Industry.* London: Taylor and Francis (1996). p. 189–94.
80. McEvoy K, Hasenstab K, Senturk D, Sanders A, Jeste SS. Physiologic artifacts in resting state oscillations in young children: methodological considerations for noisy data. *Brain Imag Behav.* (2015) 9:104–14. doi: 10.1007/s11682-014-9343-7
81. Chaudhry SI, Mattera JA, Curtis JP, Spertus JA, Herrin J, Lin Z, et al. Telemonitoring in patients with heart failure. *N Engl J Med.* (2010) 363:2301–9. doi: 10.1056/NEJMoa1010029
82. Ong MK, Romano PS, Edgington S, Aronow HU, Auerbach AD, Black JT, et al. Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure. *JAMA Intern Med.* (2016) 176:310. doi: 10.1001/jamainternmed.2015.7712
83. Sajatovic M, Velligan DI, Weiden PJ, Valenstein MA, Ogedegbe G. Measurement of psychiatric treatment adherence. *J Psychosomat Res.* (2010) 69:591–9. doi: 10.1016/j.jpsychores.2009.05.007
84. Argent R, Daly A, Caulfield B. Patient involvement with home-based exercise programs: can connected health interventions influence adherence? *JMIR MHealth UHealth.* (2018) 6:e47. doi: 10.2196/mhealth.8518
85. Maguire R, Fox PA, McCann L, Miaskowski C, Kotronoulas G, Miller M, et al. The eSMART study protocol: a randomised controlled trial to evaluate electronic symptom management using the advanced symptom management system (ASyMS) remote technology for patients with cancer. *BMJ Open.* (2017) 7:e015016. doi: 10.1136/bmjopen-2016-015016
86. Masterson Creber RM, Hickey KT, Maurer MS. Gerontechnologies for older patients with heart failure: what is the role of smartphones, tablets, and remote monitoring devices in improving symptom monitoring and self-care management? *Current Cardiovasc Risk Rep.* (2016) 10:30. doi: 10.1007/s12170-016-0511-8
87. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient compliance with paper and electronic diaries. *Control Clin Trials.* (2003) 24:182–99. doi: 10.1016/S0197-2456(02)00320-3
88. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. *Int J Hum Comput Interact.* (2008) 24:574–94. doi: 10.1080/10447310802205776
89. Barbey F, Dyer JF, McWilliams EC, Nolan H, Murphy B. Conventional wet eeg vs dry-sensor wireless eeg: comparing signal reliability through measures of neuronal integrity. In: Advances in Alzheimer's and Parkinson's Therapies: An AAT-AD/PD Focus Meeting (2020). Available online at: https://aat-adpd.kenes.com/
90. Murphy B, Barbey F, Buick AR, Dyer JF, Farina F, McGuinness B, et al. Replicating lab electrophysiology with older users in the home, using gamified dry EEG. *Alzheimers Dement.* (2019) 15:P867. doi: 10.1016/j.jalz.2019.06.4606
91. Reingold EM, Charness N, Pomplun M, Stampe DM. Visual span in expert chess players: evidence from eye movements. *Psychol Sci.* (2001) 12:48–55. doi: 10.1111/1467-9280.00309
92. Colom R, Ramon FJ, Abad FJ, Shih PC, Privado J, Froufe M, et al. Adaptive n-back training does not improve fluid intelligence at the construct level: gains on individual tests suggest that training may enhance visuospatial

processing. *Intelligence.* (2013) 41:712–27. doi: 10.1016/j.intell.2013.09.002
93. Hassenstab J, Monsell SE, Mock C, Roe CM, Cairns NJ, Morris JC, et al. Neuropsychological markers of cognitive decline in persons with Alzheimer disease neuropathology. *J Neuropathol Exp Neurol.* (2015) 74:1086–92. doi: 10.1097/NEN.0000000000000254
94. Baker JE, Bruns L, Hassenstab J, Masters CL, Maruff P, Lim YY. Use of an experimental language acquisition paradigm for standardized neuropsychological assessment of learning: a pilot study in young and older adults. *J Clin Exp Neuropsychol.* (2019) 42:55–65. doi: 10.1080/13803395.2019.1665626
95. Lim YY, Baker JE, Bruns L, Mills A, Fowler C, Fripp J, et al. Association of deficits in short-term learning and Aβ and hippocampal volume in cognitively normal adults. *Neurology.* (2020) 95:e2577–85. doi: 10.1212/WNL.0000000000010728
96. De Vos M, Gandras K, Debener S. Towards a truly mobile auditory brain-computer interface: exploring the P300 to take away. *Int J Psychophysiol.* (2014) 91:46–53. doi: 10.1016/j.ijpsycho.2013.08.010
97. Hairston WD, Whitaker KW, Ries AJ, Vettel JM, Bradford JC, Kerick SE, et al. Usability of four commercially-oriented EEG systems. *J Neural Eng.* (2014) 11:046018. doi: 10.1088/1741-2560/11/4/046018
98. Kam JWY, Griffin S, Shen A, Patel S, Hinrichs H, Heinze HJ, et al. Systematic comparison between a wireless EEG system with dry electrodes and a wired EEG system with wet electrodes. *NeuroImage.* (2019) 184:119–29. doi: 10.1016/j.neuroimage.2018.09.012
99. Radüntz T, Meffert B. User experience of 7 mobile electroencephalography devices: comparative study. *JMIR MHealth UHealth.* (2019) 7:1–18. doi: 10.2196/14474
100. Debener S, Minow F, Emkes R, Gandras K, de Vos M. How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology.* (2012) 49:1617–21. doi: 10.1111/j.1469-8986.2012.01471.x
101. Stopczynski A, Stahlhut C, Larsen JE, Petersen MK, Hansen LK. The smartphone brain scanner: a portable real-time neuroimaging system. *PLoS ONE.* (2014) 9:e0086733. doi: 10.1371/journal.pone.0086733
102. Duvinage M, Castermans T, Petieau M, Hoellinger T, Cheron G, Dutoit T. Performance of the emotiv Epoc headset for P300-based applications. *BioMed Eng Online.* (2013) 12:1–15. doi: 10.1186/1475-925X-12-56
103. Maskeliunas R, Damasevicius R, Martisius I, Vasiljevas M. Consumer grade EEG devices: are they usable for control tasks? *PeerJ.* (2016) 4:e1746. doi: 10.7717/peerj.1746
104. Porcaro C, Balsters JH, Mantini D, Robertson IH, Wenderoth N. P3b amplitude as a signature of cognitive decline in the older population: an EEG study enhanced by functional source separation. *NeuroImage.* (2019) 184:535–46. doi: 10.1016/j.neuroimage.2018.09.057
105. Larson MJ, Clayson PE, Keith CM, Hunt IJ, Hedges DW, Nielsen BL, et al. Cognitive control adjustments in healthy older and younger adults: conflict adaptation, the error-related negativity (ERN), and evidence of generalized decline with age. *Biol Psychol.* (2016) 115:50–63. doi: 10.1016/j.biopsycho.2016.01.008
106. Grummett TS, Leibbrandt RE, Lewis TW, DeLosAngeles D, Powers DMW, Willoughby JO, et al. Measurement of neural signals from inexpensive, wireless and dry EEG systems. *Physiol Measure.* (2015) 36:1469. doi: 10.1088/0967-3334/36/7/1469
107. Hurlow J, Bliss DZ. Dry skin in older adults. *Geriatr Nurs.* (2011) 32:257–62. doi: 10.1016/j.gerinurse.2011.03.003
108. Ishii R, Canuet L, Aoki Y, Hata M, Iwase M, Ikeda S, et al. Healthy and pathological brain aging: from the perspective of oscillations, functional, connectivity, signal complexity. *Neuropsychobiology.* (2017) 75:151–61. doi: 10.1159/000486870
109. Klimesch W. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res Rev.* (1999) 29:169–95. doi: 10.1016/S0165-0173(98)00056-3
110. Dyer JF, Barbey F, Barrett SL, Pickering EC, Buick AR, Mulholland C, et al. Gamified mobile EEG for early detection of psychotic disorders: identifying needs from clinicians and end-users. In: Nutt D, and Blier P, editors. *British Association for Psychopharmacology Summer Meeting.* Manchester (2019). p. A53.
111. Murphy B, Barbey F, Bianchi M, Buhl DL, Buick AR, Danyeli L, et al. *Demonstration of a Novel Wireless EEG Platform to Detect the Acute and Long-Term Effects of Ketamine, in the Lab and in the Home.* Glasgow: FENS (2020).
112. Maran M, Grent-'t-Jong T, Uhlhaas PJ. Electrophysiological insights into connectivity anomalies in schizophrenia: a systematic review. *Neuropsychiatr Electrophysiol.* (2016) 2:1–9. doi: 10.1186/s40810-016-0020-5
113. Randeniya R, Oestreich LKL, Garrido MI. Sensory prediction errors in the continuum of psychosis. *Schizophren Res.* (2018) 191:109–22. doi: 10.1016/j.schres.2017.04.019
114. Al-Nuaimi AHH, Jammeh E, Sun L, Ifeachor E. Complexity measures for quantifying changes in electroencephalogram in alzheimer's disease. *Complexity.* (2018) 2018:22–4. doi: 10.1155/2018/8915079
115. Ferreira D, Jelic V, Cavallin L, Oeksengaard AR, Snaedal J, Høgh P, et al. Electroencephalography is a good complement to currently established dementia biomarkers. *Dement Geriatr Cogn Disord.* (2016) 42:80–92. doi: 10.1159/000448394
116. Horvath A, Szucs A, Csukly G, Sakovics A, Stefanics G, Kamondi A. EEG and ERP biomarkers of Alzheimer's disease: a critical review. *Front. Biosci.* (2018) 23:4587. doi: 10.2741/4587
117. Belleville S, Fouquet C, Hudon C, Zomahoun HTV, Croteau J. Neuropsychological measures that predict progression from mild cognitive impairment to alzheimer's type dementia in older adults: a systematic review and meta-analysis. *Neuropsychol. Rev.* (2017) 27:328–353. doi: 10.1007/s11065-017-9361-5
118. Chehrehnegar N, Nejati V, Shati M, Rashedi V, Lotfi M, Adelirad F, et al. Early detection of cognitive disturbances in mild cognitive impairment: a systematic review of observational studies. *Psychogeriatrics.* (2020) 20:212–28. doi: 10.1111/psyg.12484
119. Murphy B, Aleni A, Belaoucha B, Dyer JF, Nolan H. Quantifying cognitive aging and performance with at-home gamified mobile EEG. In: *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI).* Singapore (2018). p. 1–4. doi: 10.1109/PRNI.2018.8423954

Conflict of Interest: LR-D, AB, EM, JD, HN, MI, FB, and BMu are employees of the company Cumulus Neuroscience Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

STUDY PROTOCOL
published: 01 July 2021
doi: 10.3389/fpsyt.2021.554811

SIMON: A Digital Protocol to Monitor and Predict Suicidal Ideation

*Laura Sels[1,2,3], Stephanie Homan[2], Anja Ries[1,2], Prabhakaran Santhanam[4], Hanne Scheerer[2], Michael Colla[2], Stefan Vetter[2], Erich Seifritz[2], Isaac Galatzer-Levy[5], Tobias Kowatsch[4,6], Urte Scholz[7] and Birgit Kleim[1,2]**

[1] Experimental Psychopathology and Psychotherapy, Department of Psychology, University of Zurich, Zurich, Switzerland, [2] Department of Psychiatry, Psychotherapy and Psychosomatics, University of Zurich, Zurich, Switzerland, [3] Experimental Clinical and Health Psychology, Faculty Psychology and Educational Sciences, Ghent University, East Flanders, Belgium, [4] Centre for Digital Health Interventions, Department of Management, Technology, and Economics, Swiss Federal Institute of Technology, Zurich, Switzerland, [5] Psychiatry, New York University School of Medicine, New York, NY, United States, [6] Department of Management, Technology, and Economics at ETH Zurich, Centre for Digital Health Interventions, Institute of Technology Management, University of St. Gallen, St. Gallen, Switzerland, [7] Applied Social and Health Psychology, Department of Psychology, University of Zurich, Zurich, Switzerland

Each year, more than 800,000 persons die by suicide, making it a leading cause of death worldwide. Recent innovations in information and communication technology may offer new opportunities in suicide prevention in individuals, hereby potentially reducing this number. In our project, we design digital indices based on both self-reports and passive mobile sensing and test their ability to predict suicidal ideation, a major predictor for suicide, and psychiatric hospital readmission in high-risk individuals: psychiatric patients after discharge who were admitted in the context of suicidal ideation or a suicidal attempt, or expressed suicidal ideations during their intake. Specifically, two smartphone applications -one for self-reports (SIMON-SELF) and one for passive mobile sensing (SIMON-SENSE)- are installed on participants' smartphones. SIMON-SELF uses a text-based chatbot, called Simon, to guide participants along the study protocol and to ask participants questions about suicidal ideation and relevant other psychological variables five times a day. These self-report data are collected for four consecutive weeks after study participants are discharged from the hospital. SIMON-SENSE collects behavioral variables -such as physical activity, location, and social connectedness- parallel to the first application. We aim to include 100 patients over 12 months to test whether (1) implementation of the digital protocol in such a high-risk population is feasible, and (2) if suicidal ideation and psychiatric hospital readmission can be predicted using a combination of psychological indices and passive sensor information. To this end, a predictive algorithm for suicidal ideation and psychiatric hospital readmission using various learning algorithms (e.g., random forest and support vector machines) and multilevel models will be constructed. Data collected on the basis of psychological theory and digital phenotyping may, in the future and based on our results, help reach vulnerable individuals early and provide links to just-in-time and cost-effective interventions or establish prompt mental health service contact. The current effort may thus lead to saving lives and significantly reduce economic impact by decreasing inpatient treatment and days lost to inability.

Keywords: suicidal ideation, digital monitoring, inpatient, ecological momentary assessment, passive mobile sensing

OPEN ACCESS

Edited by:
Raz Gross,
Sheba Medical Center, Israel

Reviewed by:
Shira Barzilay,
Icahn School of Medicine at Mount Sinai, United States
Maria Luisa Barrigón,
Autonomous University of Madrid, Spain
Enrique Baca-Garcia,
University Hospital Fundación Jiménez Díaz, Spain

***Correspondence:**
Birgit Kleim
birgit.kleim@uzh.ch

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: *17 July 2020*
Accepted: *12 May 2021*
Published: *01 July 2021*

Citation:
Sels L, Homan S, Ries A, Santhanam P, Scheerer H, Colla M, Vetter S, Seifritz E, Galatzer-Levy I, Kowatsch T, Scholz U and Kleim B (2021) SIMON: A Digital Protocol to Monitor and Predict Suicidal Ideation. Front. Psychiatry 12:554811. doi: 10.3389/fpsyt.2021.554811

INTRODUCTION

Digitalization has captured much of human society and is omnipresent in individuals' everyday lives. People carry their smartphone with them most of the time, even in times of crisis (1). This innovation provides new opportunities to help reach vulnerable individuals in critical moments [e.g., (2, 3)]. One group that could particularly benefit from this are individuals at risk for suicide. Suicide is one of the leading causes of deaths, and the numbers continue to rise. As a consequence, a better understanding, prediction, and prevention has been made one of the top priorities on international research agendas including the World Health Organization (4).

One of the greatest challenges to understand, predict, and prevent suicide has long been that it has to be intervened upon as it occurs and evolves in real life. Recent studies show that suicidal thoughts vary considerably throughout daily life, and can escalate quickly [for overviews, see (5, 6)]. Mobile technology can help address this challenge. For example, smartphones can be leveraged to perform real time collection of relevant self-report data and behavior, which can lead to just-in-time interventions (7). For instance, iHealth or intelligent Health has been proposed, in which the incorporation of new technologies into clinical practice helps shifting mental health care from a reactive to a proactive, participatory, and personalized domain, by for instance enhancing real-time self-monitoring and supporting medical decision making (8). With regards to suicide specifically, there has been a rapid increase in the use of mobile technology to help prevent suicide, but a major problem is that existing suicide prevention smartphone applications are not evidence-based or clinically validated (9, 10).

Before just-in-time interventions are possible, proximal risk factors of suicidal behavior have to be identified. Proximal risk factors are factors that predict the short-term occurrence of suicidal behaviors (11). Recently, there has been an increase in research that investigated proximal risk factors of suicidal ideation in daily life, of which most are based on Joiner's interpersonal theory of suicide (12–18). A key concept of Joiner's interpersonal theory of suicide, and a development beyond earlier suicide theories, is its *ideation-to action* framework, which explains why many individuals that think about suicide do not actually commit an attempt.

Joiner's interpersonal theory of suicide is one of the most rigourously researched and empirically supported theories of suicide (19, 20). The theory assumes a range of proximal suicide risk factors, and provides testable predictions of who will most likely develop suicidal ideations and who will most likely attempt suicide. It thus holds much promise to further our understanding of how certain suicide risk factors interact, and provides concrete targets for prevention and intervention efforts. In essence, it proposes that an individual will not die by suicide unless he or she has both the desire to die by suicide and the ability to do so. According to the theory, suicidal desire is caused by the simultaneous presence of two causal risk factors: (1) thwarted belongingness and (2) perceived burdensomeness, and hopelessness about these states (21, 22). Thwarted belongingness describes the experience of alienation from friends, family, or other subjectively important social circles. These comprise loneliness (i.e., feeling disconnected from others) and the absence of reciprocal care (i.e., having no one to turn to). Perceived burdensomeness refers to the view that one's existence is a burden on friends, family members, and/or society. It comprises two facets: self-hate (i.e., hating oneself) and feelings of liability (i.e., viewing one's death as more valuable than personal worth to others). Importantly, these cognitive-affective states are seen as dynamic and influenced by inter- and intra- personal factors such as experiencing family conflict, living alone, lacking social support, and readiness to interpret others behavior as rejection (22).

Specifically relevant for clinical practice are new clinical concepts that, building further on the research above, explicitly focus on imminent, acute risk factors, such as the Suicide Crisis Syndrome [SCS; (23)] and Acute Suicidal Affective Disturbance [ASAD; (24)]. For instance, ASAD is theorized to be characterized by: (1) a geometric increase in suicidal intent over the course of hours or days; (2) one or both of the following: marked social alienation (i.e., perceptions of being a liability on others) and/or marked self-alienation (i.e., perceptions of one's self being a burden); and (3) perceptions that these are hopelessly intractable; and (4) two or more manifestations of overarousal (i.e., insomnia).

Advances in real-time monitoring technology, also called ecological momentary assessment (EMA) or experience sampling (25), in which people's current behaviors and experiences are repeatedly sampled in real time in their natural environments (26), have thus recently made it possible to investigate such proximal and imminent factors as they occur and arise in daily life. Also here, the need and potential for individualized medicine is advocated, in which smartphone-based ecological momentary assessment and passive collection of information from sensors can provide a digital phenotype to develop tailored therapeutic and preventive approaches for suicide (10, 27). The big advantage of including the use of passive mobile sensing, is that it leverages the data people generate every day through their normal phone use without placing any additional burden to them. Emerging studies in this regard indeed suggest the potential utility of passive mobile sensing in predicting mental health [for a review, see (28)], mental health crises [e.g., see the EARS-project; (29)], and suicide risk (30).

Although existing research has now shown the potential short-term predictive value of some of these factors for suicidal ideation, the available evidence is inconclusive and cannot provide clear recommendations for clinical routine care yet (5). For instance, in past studies increases in hopelessness and loneliness went together with momentary suicidal ideation but were limited in predicting short-term change in suicidal ideation (16). To move the field forward, there has been a call for (1) larger, longer studies, (2) studies conducted during critical high-risk periods, and (3) the use of passive mobile sensing information (e.g., via smartphones or wearables that can deliver behavioral data without placing additional burden on participants) to improve predictability of suicidal ideation (5). Indeed, in this

regard, projects are rapidly arising that exactly tailor to these needs, such as MAPS (Mobile Assessment for the Prediction of Suicide; https://grantome.com/grant/NIH/U01-MH116923-01), the Emma app [Ecological Momentary Mental Assessment; (31)], or the Smartcrisis Study [Smartphone Survey of Suicidal Risk; (32, 33)]. Preliminary results from this research indeed suggests the feasibility (33, 34) and potential utility of combining EMA with passive mobile sensing in predicting and intervening in suicidal crises (34).

In our study, we aim to build further on this rapidly increasing research by designing and implementing a digital mental health protocol based on psychological theory – the interpersonal theory of suicide – and passive mobile sensing information. We focus on a high-risk population: psychiatric patients after discharge from an inpatient stay who were admitted in the context of suicidal ideation or a suicidal attempt, or expressed suicidal ideations at their intake interview after admission. Especially the month after discharge is a critical period associated with high rates of suicidality and mood deterioration and readmission (35). The objective of this study is to test in a sample of 100 participants whether (1) implementation of a digital mental health protocol or smartphone applications, based on self-reports and behavioral measures, is feasible and accepted and whether (2) suicidal ideation and psychiatric hospital readmission can be predicted from variables derived from these applications.

METHODS AND ANALYSES

Selection of Participants

One-hundred participants will be recruited from the Psychiatric University Hospital, Zurich, Switzerland. This number was determined based on a power analysis for multilevel data of a longitudinal study design (36). We considered a three-level nested structure of the longitudinal data with repeated EMAs (Level 1), collected across subjects (Level 2), and nested within different days (Level 3) and the simplest model, an unconditional three-level model (37) with

$$Y_{tij} = \gamma_{000} + u_{00j} + r_{oij} + \epsilon_{tij} \quad (1)$$

where Y is the suicidal ideation at hour t for participant i at day j as modeled by a linear combination of a grand mean suicidality score (γ_{000}) averaged across all repeated measures for all participants during all days. In addition, we added three random effect estimates at Level 3 (u_{00j}), Level 2 (r_{oij}), and Level 1 (ϵ_{tij}).

Consequently, we computed the intra class correlation (ICC), the design effect, and finally the power. First, the ICC is defined as the proportion of outcome variation on Level 2 and the expected correlation on Level 1t, and calculated with

$$ICC = \frac{\tau_{000}}{(\tau_{000} + \sigma^2)} \quad (2)$$

where τ_{000} is the random intercept and σ^2 the unexplained variability in outcomes. We chose an approximation using the ICC of a previous, similar study (18) with ICC = 0.52. Next, we computed the design effect, a parameter that quantifies the violation of independence on the estimates of the standard error (38), with

$$Design\ Effect = 1 + (m - 1) \times ICC \quad (3)$$

where m is the number of assessments per subject (m = 5 × 28). This results in a design effect of 73.8 which indicates the need for multilevel modeling (38). Finally, the power can be calculated with

$$Power = \frac{n \times m}{1 + (m - 1) \times ICC} \quad (4)$$

where n is the number of participants, m the number of assessments per participant, adjusted for the ICC. This can be rewritten as

$$n = \frac{(Power \times (1 + m - 1) \times ICC)}{m} \quad (5)$$

Assuming no missing data with $Power = 80\%$, $m = 140$, and $ICC = 0.52$, we would need 42 subjects. Yet, missing data especially when dealing with EMA should be taken into account. Thus, we calculated the sample size for different percentages of missing data points (50, 60, 70, and 80%). Results do not suggest a sample size larger than $n = 42$. Last, due to the imputed ICC, we also computed the sample size with different values for the ICC (**Figure 1**). Based on this, a sample size of 80 would be sufficient even in the case of an ICC of 1. Considering also the likely dropout rate, we aim at recruiting 100 participants which will allow us to detect the true effect with 80% probability at an alpha level of 0.05.

Besides multilevel modeling, we aim to apply machine leaning models to predict suicidality. The goal of the ML models will be to model the relationships between predictors and outcome (suicidal ideation, suicide attempts during follow up, hospital readmission), which requires equivalent power to detect any given univariate relationship between a dependent and independent variable. Model fit is estimated by permitting high dimensionality while penalizing model fit for increased complexity through the use of a loss function. While power is less of a concern in ML models, reproducibility and over-fitting is a significant risk, requiring strategies such as cross-validation to guard against this risk. Given the relatively small set of theory driven features included in the model, we anticipate $n = 100$ will allow for model estimation ($n = 60$) and hold-out cross-validation ($n = 40$) will be sufficient to train and test an ML model using the proposed predictors to forecast primary outcomes [see also (39, 40)].

Patients are included if they meet the following criteria: (a) admission to the hospital after a suicide attempt or in the context of suicidal ideation, and/or suicidal ideation were identified in the first diagnostic intake interview, (b) sufficient knowledge of

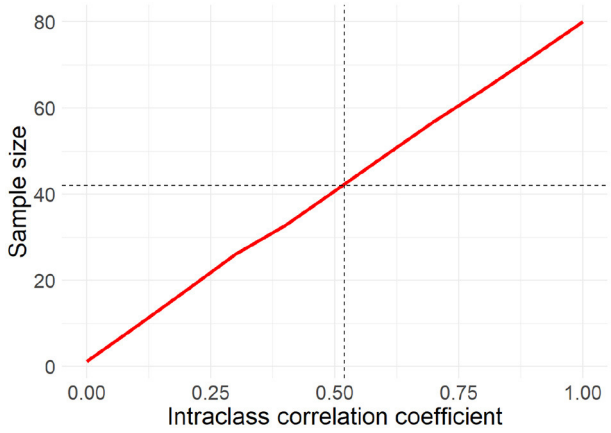


FIGURE 1 | Relationship between the intra class correlation coefficient and the sample size. With increasing correlation a larger sample size is needed to still detect the true effect with a probability of 80% at an alpha level of 0.05.

the German language, (c) having a smartphone, (d) discharge in accord with a clinician, with established outpatient care contact to the physician or psychologist. Patients are excluded if they meet the following criteria: (a) having plans to leave the greater Zurich area within the study period, (b) sharing a smartphone with another person, (c) being active military personnel (as passive sensing and EMA assessments would be challenging in active duty). There are no age restrictions. Researchers will keep track of all incoming patients in the hospital and contact the treating psychologist or physician in case of eligibility. When a patient meets the inclusion criteria and the treating psychologist or physician approves, the patient will be approached by the researcher and informed about the study.

Based on the Psychiatric University Hospital's report from 2019, patients have an average inpatient stay of 24.6 days. The average patient is 40.2 years old, with females (47.2%) and males (52.8%) almost equally distributed, and admitted mainly because of substance use disorders (27%), schizophrenia spectrum disorders (24%), affective disorders (26%), anxiety disorders (11%), and personality disorders (7.5%).

Procedure and Materials

The study will consist of different parts: a baseline assessment, a 4-week period of ecological momentary assessment in which the smartphone applications run, and a follow-up. Participants will be reimbursed with up to 120 CHF if they answer the smartphone applications' questions in more than 60 % of the time.

Baseline Assessment

The baseline assessment entails (1) detailed information about the study and informed consent, (2) assessment of the current mental disorders with the Mini International Neuropsychiatric Interview [MINI version (14)], (3) a short video-taped semi-structured qualitative interview, (4) electronic questionnaires that evaluate relevant psychological variables, and (5) the installation of the smartphone applications on participants' phones. During the baseline assessment, participants will also get a booklet that contains additional information on the aims of the study, crisis information in case of emergency, and the smartphone applications. The baseline assessment will thus occur within the hospital stay, after patients are able to and have provided informed consent to participate in the study. The exact timing of this assessment is expected to vary, as it depends on patients' acute symptom severity and their capacity to perform an interview, practical constraints and the schedule of the patient.

Table 1 lists the questionnaires and other assessments that will be used at baseline and/or at follow-up. These measures are thus a combination of self- and clinician-reports (MINI), and a video-taped qualitative interview for which participants provide separate consent. During the qualitative video interview, participants answer questions about experiences with different valences (i.e., positive, negative, neutral) and temporal dimensions (i.e., past, present, future). The videos will be used to derive markers for psychopathology using physiology, facial activity, language use, and vocal characteristics.

Ecological Momentary Assessment

Two smartphone applications will be installed on participants' smartphones. The first application (SIMON-SELF) is used for collecting self-report data according to a pre-defined ecological momentary assessment protocol. The second application collects smartphone sensor data (SIMON-SENSE). The two applications are made available for Android and iOS and described in more detail in the following paragraphs.

SIMON-SELF

MobileCoach (www.mobile-coach.eu) (53, 54), an open source software platform for delivering ecological momentary assessments and digital health interventions, was used to develop the SIMON-SELF application. The configuration of the ecological momentary assessments, i.e., timing and the self-report items, is defined via a graphical user interface by the co-authors of this paper on the MobileCoach server. The server then sends this content to SIMON-SELF, a mobile application that uses a conversational agent (named Simon in this study) to administer the self-reports to the study participants. A conversational agent is a computer program that imitates a human being, and which has the potential to establish a working alliance with participants (55) and thus, to increase involvement with the application (56, 57). Exemplary screenshots of SIMON-SELF are depicted in **Figure 2**.

After the application is installed on the participants' smartphone via Google's Play Store or Apple's App Store, conversational agent Simon provides details about the mobile application. This includes a demonstration on how to fill out the self-reports. Then, Simon asks participants to indicate whether they are about to leave the clinic, so that the experience sampling protocol covering the period post-discharge will be promptly initiated for the day following their discharge.

From then on, participants will be asked to answer the experience sampling surveys 5 times per day during a defined period of 12 h, according to a stratified random interval scheme with the time frame being divided into five equal intervals. This

TABLE 1 | Questionnaires and assessments conducted at baseline and/or follow-up.

Questionnaires and assessments	Administration	Baseline	Follow-up
Demographic and personal information	Self-report	x	
Mini International neuropsychiatric interview (41)	Semi-structured interview, Clinician report	x	
Video-taped qualitative interview	Semi-structured interview	x	
Beck depression inventory-II (42)	Self-report	x	x
Positive and negative affect scale (PANAS) (43)	Self-report	x	x
Patient health questionnaire (PHQ) (44)	Self-report	x	x
Suicide attempts (45)	Self-report	x	
Childhood trauma questionnaire (46)	Self-report	x	
Life events questionnaire (47)	Self-report	x	
Interpersonal needs questionnaire [INQ-15; (48)]	Self-report	x	x
Beck scale for suicide ideation (BSS; German validated version; (49))	Self-report	x	x
Beck hopelessness scale (BHS; German validated version; (49))	Self-report	x	x
Acquired capability for suicide scale (ACSS-20; German validated, revised version from (50))	Self-report	x	
Generalized self-efficacy scale (51)	Self-report	x	x
The trait hope scale (52)	Self-report	x	x
Suicidal crisis information	Information hospital and self-report		x
Research experience questionnaire	Self-report		x
App questionnaire			x

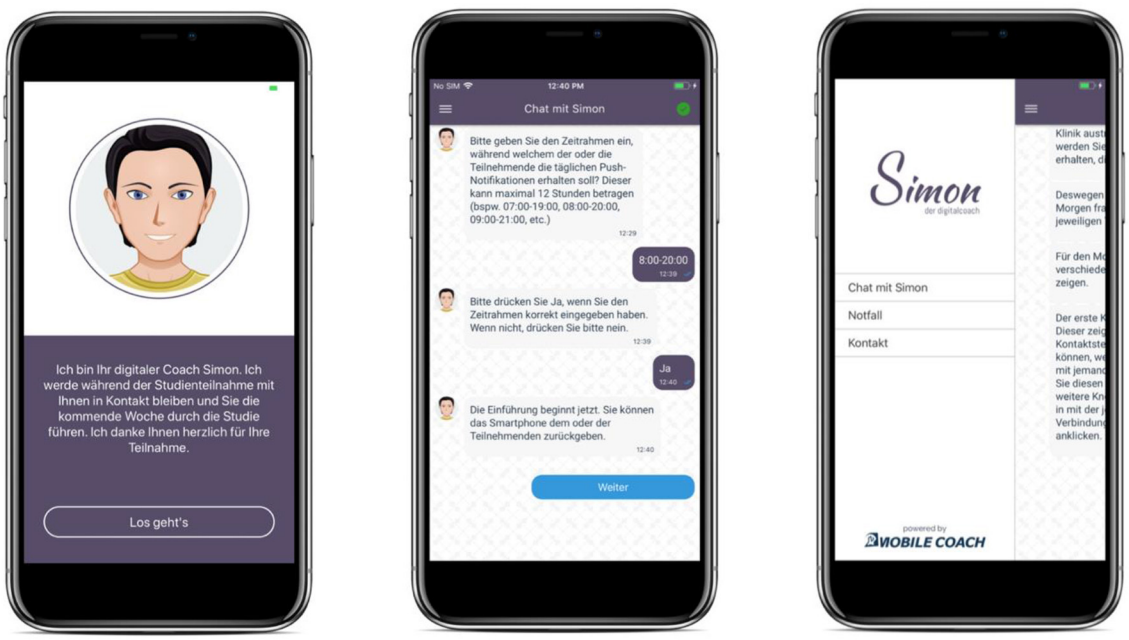


FIGURE 2 | The SIMON-SELF application: welcome screen **(left)**, the chat screen **(middle)** and the menu **(right)**.

means that individualization of the time frame is possible, but only with a fixed range of 12 h (e.g., from 9 AM to 9 PM, from 10 AM to 10 PM, et cetera). Participants will be asked upon installation of the application what timeframe they prefer. Every day, Simon will greet participants in the morning, wish them good night in the evening, and will prompt them to answer the questions. The specific questions that will be asked, can be found in **Table 2**. One block of questions is only asked in the morning (e.g., about sleep), one block only in the evening (e.g., about mood), and there is one block that shows up with every survey.

Compliance to the protocol is promoted through multiple strategies. Every second day, Simon gives participants an update on their compliance, and the feedback depends on more/less than 60% compliance. Simultaneously, participants will be sent automatically generated text messages with the same feedback via SMS, i.e., an additional communication channel compared to in-app chat messages. Finally, Simon tells participants every week that the researchers are very grateful for their participation in the study, and that they are helping to improve future suicide prevention methods. To further increase compliance, researchers will contact participants in case of non-compliance. Finally, participants will receive a personalized summary of their collected self-reports after successful study completion. Specifically, they receive visual feedback, containing a series of charts that summarize their changes in key variables, e.g., sleep, suicidality, other psychological characteristics and feelings, over the time of the study.

In addition, an emergency button is made available in the side menu of SIMON-SELF (see screenshot three, **Figure 2**). It provides three different helpline numbers, according to urgency and specific need. Participants receive information about these helplines and their services upon installation of the application. There is also a number available on the application that participants can reach in case of technical issues.

SIMON-SENSE

To assess relevant context variables such as physical activity, sleep, and social connectedness, the mobile sensing application SIMON-SENSE records sensor data commonly available via smartphones. We use the open source framework AWARE (67) for this purpose. SIMON-SENSE records and sends the data in a secure way to a server located at the university of the corresponding author (University of Zurich). The specific data sources, data types and collection frequencies are listed in **Table 3**. The application runs in the background, and thus requires no interaction with study participants. Because this application might drain the battery of participants' phones, Simon reminds participants every evening to charge their smartphone.

Follow-Up

After 4 weeks of ecological momentary assessment, a follow-up assessment takes place at the Psychiatric University Hospital, Zurich. Participants will fill out questionnaires (see **Table 1**) of which most are validated and have been assessed at baseline already. To gain insight into user experience of the apps, participants will fill in the Research Experience Questionnaire and the App Questionnaire. In addition to these quantitative measures, research assistants will be instructed to encourage participants to give also qualitative feedback on the app usage. Both sources of feedback will be valuable for the further development, particularly of the in-house developed SIMON-SELF app and the design of subsequent studies. Finally, participants receive payment for their participation in the study.

Data Management

The experience sampling and the passive mobile sensing data will be transmitted via a Secure Sockets Layer (SSL) connection to a study server. This server can only be accessed by a password. Data from the baseline and follow-up questionnaires will also be saved on this study server. The study server is provided by the University of Zurich, Switzerland. To match different datasources, a unique user number is generated for each participant. The only file containing participant's full personal information and respective unique user number, is kept in a separate document and stored in a locked file cabinet.

Data Analyses

The main research aim of this study is to investigate short-term predictors of suicidal ideation and psyciatric hospital readmission in a high risk-population. To this end, two kinds of analyses are planned.

First, prediction models using various learning algorithms will be developed. The development of such models involves several steps. In a first step, the raw sensor data has to be preprocessed involving feature extraction, scaling, selection, and dimensionality reduction. Smartphone data "features" derived from sensor data and the experience sampling indices as well as data from the baseline questionnaires will then be fed into machine learning models to identify the variables, and combinations thereof, that predict suicidal ideation and psychiatric hospital readmission. In a second step, the data will be split into a training and test data set to assess how the derived algorithms generalize to new data (68). The training dataset will also be split into subsets and k-fold cross-validations will be applied. The performance of the resulting model will then be evaluated using the test data set. This procedure will be repeated for various learning algorithms (e.g., random forest, support vector machines). The learning algorithms will also discard irrelevant information that does not help to improve the predictive value of the model using partitioning for categorical states (is suicidal ideation high/low, were participants readmitted to the hospital). After comparing the performance across algorithms, the best overall model will be selected.

We expect to construct a model that efficiently predicts suicidal ideation and psychiatric hospital readmission using a combination of sensory data and psychological data.

Second, longitudinal multilevel models will be applied to predict suicidal ideation and psychiatric hospital readmission from a combination of predictors based on theory. This will allow to compare between patient differences (between-person level) and to make predictions on an individual level (within-person level) by fitting individual symptom trajectories. Considering the

TABLE 2 | Self-report items of the ecological momentary assessment administered through SIMON-SELF.

	Response scale	Construct & reference
Only shown during first beep of the day		
1. How long did it take you to fall asleep yesterday?	Min	Sleep, derived from the Sleep Condition Indicator (58)
2. If you woke up during the night: how long were you awake for in total?	Slider scale from *0 min* to *≥ 61 min*	Sleep, derived from the Sleep Condition Indicator (58)
3. How would you rate your sleep quality?	Slider scale from *Very good* to *very poor*	Sleep, derived from the Sleep Condition Indicator (58)
4. Did you have nightmares?	Binary: yes/no	Nightmares
5. (conditional upon item 4) How distressing were they?	Slider scale from *Not at all* to *extremely*	Nightmares
Shown during every beep of the day		
6. At this moment, I feel little interest or pleasure in doing things.	Slider scale from Not at all to extremely	Depression (59)
7. At this moment, I feel down or depressed.	Slider scale from *Not at all* to *extremely*	Depression (59)
8. At this moment, I feel useless.	Slider scale from *Not at all* to *extremely*	Perceived burdensomness [Hallensleben et al., 2018; (60)]
9. At this moment, I feel like a burden for others.	Slider scale from *Not at all* to *extremely*	Perceived burdensomness [Hallensleben et al., 2018; (60)]
10. At this moment, I feel lonely.	Slider scale from *Not at all* to *extremely*	Thwarted belongingness [Hallensleben et al., 2018; (60)]
11. At this moment, I feel like I do not belong.	Slider scale from *Not at all* to *extremely*	Thwarted belongingness [Hallensleben et al., 2018; (10)]
12. At this moment, I feel hopeless.	Slider scale from *Not at all* to *extremely*	Hopelessness (16)
13. At this moment, the future seems hopeful to me and things are changing for the better.	Slider scale from *Not at all* to *extremely*	Hope (61)
14. At this moment, I feel that life is not worth living for me.	Slider scale from *Not at all* to *extremely*	Passive suicidal ideation [Hallensleben et al., 2018; (60)]
15. At this moment, I feel there are more reasons to die than to live for me.	Slider scale from *Not at all* to *extremely*	Passive suicidal ideation [Hallensleben et al., 2018; (60)]
16. At this moment, I feel that I want to die by suicide.	Slider scale from *Not at all* to *extremely*	Active suicidal ideation [Hallensleben et al., 2018; (60)]
17. At this moment, I think about taking my life.	Slider scale from *Not at all* to *extremely*	Active suicidal ideation [Hallensleben et al., 2018; (60)]
18. At this moment, I feel that I have control over the things that happen to me.	Slider scale from *Not at all* to *very confident*	Daily locus of control/ self-efficacy (62)
19. Move the sliders to express how you actually feel while watching the picture. Move the slider to rate your level of pleasure.	Slider scale with pleasure	Affect: The Affective Slider (63)
20. Move the sliders to express how you actually feel while watching the picture. Move the slider to rate your level of arousal.	Slider scale with arousal	Affect: The Affective Slider (63)
Only shown during the last beep of the day		
1. At this moment, I feel	Slider scale going from *tired* to *awake*	Awake-Affect (64)
2. At this moment, I feel	Slider scale going from *content* to *discontent*	Content-Affect (64)
3. At this moment, I feel	Slider scale going from *agitated* to *calm*	Agitated-Affect (64)
4. At this moment, I feel	Slider scale going from *full of energy* to *without energy*	Full of energy-Affect (64)
5. At this moment, I feel	Slider scale going from *unwell* to *well*	Unwell-Affect (64)
6. At this moment, I feel	Slider scale going from *relaxed* to *tense*	Relaxed-Affect (64)
7. Please indicate the persons you spent time with today (indicate none or as many as applicable)	☐ romantic partner ☐ parent(s) ☐ sibling(s) ☐ friend(s)	

(Continued)

TABLE 2 | Continued

	Response scale	Construct & reference
	☐ housemate(s) (not friend or family) ☐ coworkers or classmates ☐ other, specify:	
8. Choose the person you interacted with most today (Indicate only one)	☐ romantic partner ☐ parent(s) ☐ sibling(s) ☐ friend(s) ☐ housemate(s) (not friend or family) ☐ coworkers or classmates ☐ other, specify:	
9. To what degree have you disclosed your feelings to this person during the day?	Slider scale *not at all* to *fully*	(65)
10. To what degree have you suppressed your feelings to this person during the day?	Slider scale *not at all* to *fully*	
11. To what extent did you feel that this person understood you?	Slider scale *not at all* to *fully*	(66)
12. To what degree did you feel that this person expressed liking and encouragement for you?	Slider scale *not at all* to *fully*	(66)
13. To what degree did you feel that this person valued your abilities and opinions?	Slider scale *not at all* to *fully*	(66)
14. Have you experienced a conflict with this person throughout the day?	☐ Yes ☐ No	

TABLE 3 | Data sources, data types and collection frequency of the SIMON-SENSE application.

Sensor	Variable	Data type	Frequency[a]
Accelerometer	Physical activity	3D Float	Every 60 milliseconds
Gyroscope	Physical activity	3D Float	Every 60 milliseconds
Ambient light	Ambient Light	Float	Every 60 milliseconds
GPS	Location	Float (Multidimensional)	Every 180 s or 150 meters location change
Triangulation (Cell/Wi-Fi)	Location	Float (Multidimensional)	Every 300 s or 1,500 meters location change
Screen usage	Screen on/off	Binary (on/off)	Continuous
Bluetooth	Social connectedness	Categorical/string	Every 5 min
Wi-fi	Social connectedness	Categorical/string	Every min
Network	Network events	Categorical/string	Continuous
Application logs[b]	Application logs	Strings (Usage, Notifications, crashes)	Every 30 s
Ambient noise	Noise level	Categorical/float	Every 5 min

[a] *Estimated frequencies only. Actual frequencies may vary depending on device and operating system.*
[b] *Application log data is only collected for Android devices due to restrictions of iOS.*

dynamic nature of suicidal ideation, it is cruicial to identify what predicts within-person changes.

Ethics

This study follows ethical and safety guidelines, such as those put forward by Nock and others (69). In accordance with these guidelines, participants will not be excluded on the basis of elevated risk of suicide, participants will be elaborately informed before participation on all suggested elements (e.g., whether responses will trigger intervention actions; providing participants with information about who will have access to their data), and recommended technical and safety procedures are in place (e.g., figuring out what to do when technology fails, and providing participants with standardized informations on items of data-collection). Regarding safety specifically, all participants will receive detailed information about local help lines in case of crisis, and emergency. This information will be presented with the mobile application multiple times throughout the 4-weeks assessment. A standard operation procedure is established in cases of emergencies according to which the researchers are going to act.

We decided not to monitor and pro-actively respond to various levels of risk in real-time with interventions (although suggested by Nock and others), due to several reasons. First, there is currently no agreement on how to determine a

participant's current level of risk and criteria for acute level of risk. Further, data cannot be monitored continuously due to several practical reasons (specifically, data are only uploaded when a Wi-Fi connection is available and our study does not provide 24-7 tracking of the data overnight). This study is not an intervention study, but rather a naturalistic study that monitors potentially powerful predictors of suicidal ideation and hospital readmission, as well as suicidal ideation itself. It is stressed to participants that this is not an intervention study, but that the information collected as part of this study will inform and help develop such efforts. Participants can thus only be enrolled in the study if they have a physician and/or psychotherapist attending to them following discharge. Informed consents are obtained after patients received elaborate information about the study procedures and the fact that they can exit the study at any time.

The setup of the study has been discussed with clinicians, psychologists, and patients and piloted to minimize any potential risks or problems. Treating physicians and psychotherapists are involved when patients are approached and enrolled into the study.

The study was reviewed and approved by the Ethics Committee of the Faculty of Arts and Social Sciences of the University of Zurich, Switzerland. All collected data will be anonymised. Results will be published in medical and technical peer-reviewed journals.

DISCUSSION

This study builds further on an emerging line of research by testing in a large sample of high-risk individuals whether (1) a digital mental health protocol with self-reports and behavioral measures can be implemented and whether (2) suicidal ideation and psychiatric hospital readmission can be predicted from variables derived from this protocol. The results from this study will build on and extend the growing body of research on prediction markers of suicidal ideation by mobile health technology [for an overview see (5)]. Identifying reliable prediction markers of suicidal ideation is crucial to help develop just-in-time and cost-effective interventions. For instance, information about these predictors could then be fed back to clinicians and mental health services in real time to provide the support and interventions needed by each individual patient.

A better treatment of suicidal ideation is of vital importance as suicide is a major public health concern. As a consequence, it has been placed high on many national and international research agendas. In addition to being one of the most dramatic intrapersonal consequences of mental health problems, its interpersonal and economic costs are also enormous [e.g., (70)]. Digital technologies provide exciting opportunities to help reduce the number of suicide by accounting for particular challenges associated with its prevention.

Limitations

To optimize continued participation in this population, in which drop-out and low compliance are common problems, and because there is no intervention aspect to the study for participants, we decided to reimburse them. However, this decision may limit the ecological validity of the study in the sense of being comparable with real-world usage of smartphone applications for high-risk suicidal individuals (who are not reimbursed). Further, we decide to conduct a follow-up after four weeks, immediately after the EMA-part of the study, because of multiple reasons. First, we aim to diminish participant drop-out. Second, the first weeks after psychiatric discharge contain a much higher risk for suicide than any period thereafter or other treatment events (71–73). However, this choice has as a disadvantage that given the rarity of suicide, a low incidence of suicidal crises is expected to occur in such a short period. Finally, we determined sample size on power considerations for multilevel data of a longitudinal study design, and acknowledge that this is on the lower side for machine learning models.

DATA AVAILABILITY STATEMENT

Data will be available on OSF.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by PhF UZH Ethics comittee (IRB). The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individuals for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

BK and US conceptualized and initiated the project. HS, MC, SV, and ES provided input on psychiatric background, and recruitment and feasibility. PS and TK provided input on the technical background and wrote the paragraphs about conversational agents, sensor data collection, and MobileCoach for ecological momentary assessments. LS wrote the first draft of the manuscript under supervision of BK. AR and SH contributed to further versions of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The project is co-funded by the Swiss National Foundation's Digital Lives Funding Scheme, Award number 10DL12_183251 to BK and US.

ACKNOWLEDGMENTS

The authors thank Nina Klee, Carlo Berther, and Sarina Blaser for their invaluable help with the study setup, patient recruitment, and data collection. We would also like to thank all physicians and psychologists who helped recruiting patients, as well as the patients themselves for their efforts invested in this study.

REFERENCES

1. Larsen ME, Nicholas J, Christensen H. A systematic assessment of smartphone tools for suicide prevention. *PLoS ONE.* (2016) 11:e0152285. doi: 10.1371/journal.pone.0152285
2. Müller AM, Blandford A, Yardley L. The conceptualization of a just-in-time adaptive intervention (JITAI) for the reduction of sedentary behavior in older adults. *Mhealth.* (2017) 37. doi: 10.21037/mhealth.2017.08.05
3. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobiel sensing and support for people with depression: a pilot trial in the wild. *JMIR Mhealth Uhealth.* (2016) 4:e111. doi: 10.2196/mhealth.5960
4. World Health Organization. *Preventing Suicide: A Global Imperative*. World Health Organization (2014). Available online at: https://www.who.int/health-topics/suicide
5. Kleiman EM, Nock MK. Real-time assessment of suicidal thoughts and behaviors. *Curr Opin Psychol.* (2018) 22:33–7. doi: 10.1016/j.copsyc.2017.07.026
6. Spangenberg L, Forkmann T, Glaesmer H. Investigating dynamics and predictors of suicidal behaviors using ambulatory assessment. *Neuropsychiatrie.* (2015) 29:139–43. doi: 10.1007/s40211-015-0142-1
7. Ballard ED, Gilbert JR, Wusinich C, Zarate Jr CA. New methods for assessing rapid changes in suicide risk. *Front Psychiatry.* (2021) 12. doi: 10.3389/fpsyt.2021.598434
8. Berrouiguet S, Perez-Rodriguez MM, Larsen M, Baca-Garcia E, Courtet P, Oquendo M. From eHealth to iHealth: transition to participatory and personalized medicine in mental health. *J Med Int Res.* (2018) 20:e2. doi: 10.2196/jmir.7412
9. Torous J, Larsen ME, Depp C, Cosco TD, Barnett I, Nock MK, et al. Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps. *Curr Psychiatry Rep.* (2018) 20:51. doi: 10.1007/s11920-018-0914-y
10. Melia R, Hickey FK, Bogue J, Duggan J, O' Sullivan M, Young K. Mobile health technology interventions for suicide prevention: systematic review. *JMIR Mhealth Uhealth*. (2020) 8:e12516. doi: 10.2196/12516
11. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull.* (2017) 143:187–232. doi: 10.1037/bul0000084
12. Armey MF, Brick L, Schatten HT, Nugent NR, Miller IW. Ecologically Assessed Affect and Suicidal Ideation Following Psychiatric Inpatient Hospitalization. *Gen Hosp Psychiatry*. (2018) 63:89–96. doi: 10.1016/j.genhosppsych.2018.09.008
13. Ben-Zeev D, Young MA, Depp CA. Real-time predictors of suicidal ideation: mobile assessment of hospitalized depressed patients. *Psychiatry Res.* (2012) 197:55–9. doi: 10.1016/j.psychres.2011.11.025
14. Czyz EK, Horwitz AG, Yeguez CE, Ewell F, King CA. Parental self-efficacy to support teens during a suicidal crisis and future adolescent emergency department visits and suicide attempts. *J Clin Child Adol Psychol.* (2018) 47:S384–S396. doi: 10.1080/15374416.2017.1342546
15. Hallensleben N, Glaesmer H, Forkmann T, Rath D, Strauss M, Kersting A, et al. Predicting suicidal ideation by interpersonal variables, hopelessness and depression in real-time. An ecological momentary assessment study in psychiatric inpatients with depression. *Eur Psychiatry.* (2019) 65:43–50. doi: 10.1016/j.eurpsy.2018.11.003
16. Kleiman EM, Turner BJ, Fedor S, Beale EE, Huffman JC, Nock MK. Examination of real-time fluctuations in suicidal ideation and its risk factors: results from two ecological momentary assessment studies. *J Abnor Psychol.* (2017) 126:726–38. doi: 10.1037/abn0000273
17. Kyron MJ, Hooke GR, Page AC. Daily assessment of interpersonal factors to predict suicidal ideation and non-suicidal self-injury in psychiatric inpatients. *J Cons Clin Psychol.* (2018) 86:556. doi: 10.1037/ccp0000305
18. Littlewood DL, Kyle SD, Carter LA, Peters S, Pratt D, Gooding P. Short sleep duration and poor sleep quality predict next-day suicidal ideation: an ecological momentary assessment study. *Psychol Med.* (2019) 49:403–11. doi: 10.1017/S0033291718001009
19. Joiner TE Jr, Van Orden KA, Witte TK, Selby EA, Ribeiro JD, Lewis R, et al. Main predictions of the interpersonal-psychologica theory of suicidal behavior: empirical tests in two samples of young adults. *J Abno Psychol.* (2009) 118:634–46. doi: 10.1037/a0016500
20. Van Orden K, Conwell Y. Suicides in late life. *Curr Psychiatry Rep.* (2011) 13:234–41. doi: 10.1007/s11920-011-0193-3
21. Joiner TE Jr. *Why People die by Suicide*. Cambridge, MA: Harvard University Press (2005).
22. Van Orden KA, Witte TK, Cukrowicz KC, Braithwaite SR, Selby EA, Joiner TE Jr. The interpersonal theory of suicide. *Psychol. Rev.* (2010) 117:575–600. doi: 10.1037/a0018697
23. Schuck A, Calati R, Barzilay S, Bloch-Elkouby S, Galynker I. Suicide crisis syndrome: a review of supporting evidence for a new suicide-specific diagnosis. *Behav Sci Law.* (2019) 37:223–39. doi: 10.1002/bsl.2397
24. Tucker RP, Michaels MS, Rogers ML, Wingate LR, Joiner TE. Construct validity of a proposed new diagnostic entity: acute suicidal affective disturbance (ASAD). *J Affect Dis.* (2016) 189:365–78. doi: 10.1016/j.jad.2015.07.049
25. Larson R, Csikszentmihalyi M. The experience sampling method. *N Direc Methodol Soc Behav Sci.* (1983) 15:41–56.
26. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. *Ann Rev Clin Psychol.* (2008) 4:1–32. doi: 10.1146/annurev.clinpsy.3.022806.091415
27. Barrigon ML, Courtet P, Oquendo M, Baca-Garcia E. Precision medicine and suicide: an opportunity for digital health. *Curr Psychiatry Rep.* (2019) 12:1–8. doi: 10.1007/s11920-019-1119-8
28. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and well-being. *J Biomed Infor.* (2018) 77:120–32. doi: 10.1016/j.jbi.2017.12.008
29. Lind MN, Byrne ML, Wicks G, Smidt AM, Allen NB. The effortless assessment of risk states (EARS). *JMIR Mental Health.* (2018) 5:e10334. doi: 10.2196/10334
30. Haines-Delmont A, Chahal G, Bruen AJ, Wall AK, Khan CT, Sadashiv R, et al. Testing suicide risk prediction algorithms using phone measurements with patients in acute mental health settings: feasibility study. *JMIR Mhealth Uhealth.* (2020) 8:e15901. doi: 10.2196/15901
31. Morgieve M, Genty C, Azé J, Dubois J, Leboyer M, Vaiva G, et al. A digital companion: the EMMA app, for ecological momentary assessment and prevention of suicide: quantitative case series study. *JMIR Mhealth Uhealth.* (2020) 8:e15741. doi: 10.2196/15741
32. Berrouiget S, Barrigon ML, Castroman JL, Courtet P, Artés-Rodriguez A, Baca-Garcia E. Combining mobile-health (mHealth) and articial intelligence (AI) methods to avoid suicide attempts: the smartcrises study protocol. *BMC Psychiatry*. (2019) 19:1–9. doi: 10.1186/s12888-019-2260-y
33. Moreno-Munoz P, Romero-Medrano L, Moreno A, Herrera-Lopez J, Baca-Garcia E, Artés-Rodriguez A. Passive detection of behavioral shifts for suicide attempt prevention. *Arxiv* (2020). Retrieved from: https://arxiv.org/abs/2011.09848
34. Porras-Segovia A, Molina-Madueno RM, Berrouiguet S, Lopez-Castroman J, Barrigon ML, Pérez-Rodriguez MS, et al. Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: a real-world feasibility study. *J Affect Dis.* (2020) 274:733–41. doi: 10.1016/j.jad.2020.05.067
35. Chung DT, James Ryan C, Hadzi-Pavlovic D, Preet Singh S, Stanton C, Large MM. Suicide rates after discharge from psychiatric facilities: a systematic review and meta-analysis. *JAMA Psychiatry*. (2017) 74:694–702. doi: 10.1001/jamapsychiatry.2017.1044
36. Diggle P, Diggle PJ, Heagerty P, Liang KY, Heagerty PJ, Zeger S. *Analysis of Longitudinal Data.* Oxford: Oxford University Press (2002).
37. Peugh JL, Heck RH. Conducting three-level longitudinal analyses. *J Early Adol.* (2017) 37:7–58. doi: 10.1177/0272431616642329
38. Peugh JL. A practical guide to multilevel modeling. *J School Psychol.* (2010) 48:85–112. doi: 10.1016/j.jsp.2009.09.002
39. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*. (2005) 21:1509–15. doi: 10.1093/bioinformatics/bti171
40. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS ONE.* (2019) 14:e0224365. doi: 10.1371/journal.pone.0224365
41. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The mini-international neuropsychiatric interview (MINI): the development and validation of a structured diagnostic psychiatric interview

for DSM-IV and ICD-10. *J Clin Psychiatry*. (1998) 59:22–33. doi: 10.1037/t18597-000
42. Beck AT, Steer RA, Brown GK. Beck depression. Inventory-II. (1996) 78:490–8. doi: 10.1037/t00742-000
43. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Person Soc Psychol.* (1988) 54:1063–70. doi: 10.1037/0022-3514.54.6.1063
44. Spitzer RL, Kroenke K, Williams JB, Patient Health Questionnaire Primary Care Study Gr. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA*. (1999) 282:1737–44. doi: 10.1001/jama.282.18.1737
45. Chu C, Hom MA, Stanley IH, Gai AR, Nock MK, Gutierrez PM, et al. Non-suicidal self-injury and suicidal thoughts and behaviors: a study of the explanatory roles of the interpersonal theory variables among military service members and veterans. *J Consu Clin Psychol.* (2018) 56–68. doi: 10.1037/ccp0000262
46. Bernstein DP, Fink L, Handelsman L, Lovejoy M, Wenzel K, Sapareto E, et al. Initial reliability and validity of a new retrospective measure of child abuse and neglect. *Am J Psychiatry*. (1994) 1132–6. doi: 10.1176/ajp.151.8.1132
47. Weathers FZ, Blake DD, Schnurr PP, Kaloupek DG, Marx BP, Keane TM. *The Life Events Checklist For DSM-5 (LEC-5).* (2013). Available online at: www.ptsd.va.gov. Retrieved from: (March, 13, 2019).
48. Van Orden KA, Cukrowicz KC, Witte TK, Joiner JT. Thwarted belongingness and perceived burdensomness: construct validity and psychometric properties of the interpersonal needs questionnaire. *Psychol Assess*. (2012) 24:197–215. doi: 10.1037/a0025358
49. Kliem S, Lohmann A, Möble T, Brähler E. Psychometric properties and measurement invariance of the Beck hopelessness scale (BHS): results from a German representative population sample. *BMC Psychiatry.* (2018) 1:110. doi: 10.1186/s12888-018-1646-6
50. Spangenberg L, Glaesmer H, Scherer A, Gecht J, Barke A, Mainz V, et al. Fearlessness about death and suicidal behavior: psychometric properties of the german version of the revised acquired capability for suicide scale (ACSS-FAD). *Psychiatr Praxis*. (2016) 43:95–100. doi: 10.1055/s-0034-1387375
51. Schwarzer R, Jerusalem M. Generalized self-efficacy scale. In: Weinmann J, Wright S, Johnston M, editors. *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs*. Windsor: Nfer-Nelson (1995). p. 35–37.
52. Snyder CR, Harris C, Anderson JR, Holleran SA, Irving LM, Sigmon ST. The will and the ways: development and validation of an individual-differences measure of hope. *J Person Soc Psychol*. (1991) 60:570–85. doi: 10.1037/0022-3514.60.4.570
53. Filler A, Kowatsch T, Haug S, Wahle F, Staake T, Fleisch E. MobileCoach: A Novel Open-Source Platform For the Design of Evidence-Based, Scalable, and Low-Cost Behavioral Health Interventions -Overview and Preliminary Evaluation of the Public Health Context. 14th ed. St. New York, NY: Annual Wireless Telecommunications Symposium (2015).
54. Kowatsch T, Volland D, Shih I, Rüegger D, Künzler F, Barata F, et al. *Design and Evaluation of a Mobile Chat App For the Open Source Behavioral Health Intervention Platform MobileCoach. Nternational Conference on Design Science Research in Information System and Technology*. Cham: Springer (2017). p. 485–489. Available online at: https://link.springer.com/chapter/10.1007/978-3-319-59144-5_36#citeas
55. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working lliance in automated health behavior change interventions. *Pat Educ Couns*. (2005) 59:21–30. doi: 10.1016/j.pec.2004.09.008
56. Tinschert P, Barata F, Kramer J, Rassouli F, Steurer-Stey C, Puhan M, et al. *Don't Lose Heart: Preliminary Engagement Results in an Ecological Momentary Assessment (EMA) Study Evaluating Digital Biomarkers For Asthma*. Auckland: Abstract presented at the International Society for Research on Internet Interventions (ISRII) (2019).
57. Hauser-Ulrich S, Künzli H, Meier-Peterhans D, Kowatsch T. A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): pilot randomized controlled trial. *JMIR mHealth and uHealth*. (2020) 8:e15806. doi: 10.2196/15806
58. Espie CA, Kyle SD, Hames P, Gardani M, Fleming L, Cape J. The sleep condition indicator: a clinical screening tool to evaluate insomnia disorder. *BMJ Open.* (2014) 4:e004183. doi: 10.1136/bmjopen-2013-004183
59. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care.* (2003) 11:1248–92. doi: 10.1097/01.MLR.0000093487.78664.3C
60. Forkmann T, Spangenberg L, Rath D, Hallensleben N, Hegerl U, Kersting A, et al. Assessing suicidality in real time: a psychometric evaluation of self-report items for the assessment of suicidal ideation and its proximal risk factors using ecological momentary assessments. *J Abno Psychol*. (2018) 127:758–69. doi: 10.1037/abn0000381
61. Fraser C, Keating M. The effect of a creative art program on self-esteem, hope, perceived social support, and self-efficacy in individuals with multiple sclerosis: a pilot study. *J Neurosci Nurs*. (2014) 46:330–6. doi: 10.1097/JNN.0000000000000094
62. Ryon HS, Gleason ME. The locus of control in daily life. *Person Soc Psychol Bull.* (2014) 40:121–31. doi: 10.1177/0146167213507087
63. Betella A, Verschure PF. The affective slider: a digital self-assessment scale for the measurement of human emotions. *PLoS ONE.* (2016) 11:e0148037. doi: 10.1371/journal.pone.0148037
64. Wilhelm P, Schoebi D. Assessing mood in daily life. *Eur J Pyschol Assess.* (2007) 23:258–67. doi: 10.1027/1015-5759.23.4.258
65. Laurenceau J-P, Barrett LF, Pietromonaco PR. Intimacy as an interpersonal process: the importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *J Person Soc Psychol.* (1998) 74:1238. doi: 10.1037/0022-3514.74.5.1238
66. Gadassi R, Bar-Nahum LE, Newhouse S, Anderson R, Heiman JR, Rafaeli E, et al. Perceived partner responsiveness mediates the association between sexual and marital: a daily diary study in newlywed couples. *Arch Sex Behav*. (2016) 45:109–20. doi: 10.1007/s10508-014-0448-2
67. Ferreira D, Vassilis K, Dey AK. AWARE: mobile context instrumentation framework. *Front ICT*. (2015) 6. doi: 10.3389/fict.2015.00006
68. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Med Genom.* (2011) 4. doi: 10.1186/1755-8794-4-31
69. Nock MK, Kleiman EM, Abraham M, Bentley KH, Brent DA, Buonopane RJ, et al. Consensus statement on ethical & safety practices for conducting digital monitoring studies with people at risk of suicide and related behaviors. *Psychiatr Res Clin Prac.* (2020) 3:57–66. doi: 10.1176/appi.prcp.20200029
70. Yang B, Lester D. Recalculating the economic cost of suicide. *Death Stud.* (2007) 31:351–61. doi: 10.1080/07481180601187209
71. Goldacre M, Seagroatt V, Hawton K. Suicide after discharge from psychiatric inpatient care. *Lancet.* (1993) 342:283–6. doi: 10.1016/0140-6736(93)91822-4
72. Qin P, Nordentoft M. Suicide risk in relation to psychiatric hospitalization: evidence based on longitudinal registers. *Arch Gen Psychiatry*. (2005) 62:427–32. doi: 10.1001/archpsyc.62.4.427
73. Valenstein M, Kim HM, Ganoczy D, McCarthy JF, Zivin K, Austin KL, et al. Higher-risk periods for suicide among VA patients receiving depression treatment: prioritizing suicide prevention efforts. *J Affect Dis.* (2009) 112:50–58. doi: 10.1016/j.jad.2008.08.020

Conflict of Interest: The authors PS and TK are affiliated with the Center for Digital Health Interventions, a joint initiative of the Department of Management, Technology, and Economics at ETH Zurich and the Institute of Technology Management at the University of St. Gallen, which is funded in part by the Swiss health insurer CSS. TK is also co-founder of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways and has used the open source MobileCoach platform for that purpose, too. However, Pathmate Technologies is not involved in the intervention described in this paper.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

METHODS
published: 23 September 2021
doi: 10.3389/fpsyt.2021.574440

An Iterative and Collaborative End-to-End Methodology Applied to Digital Mental Health

Laura Joy Boulos[1]*, *Alexandre Mendes*[2]*, *Alexandra Delmas*[2]* and *Ikram Chraibi Kaadoud*[2]*

[1] Saint-Joseph University, Beirut, Lebanon, [2] Groupe onepoint, Paris, France

Artificial intelligence (AI) algorithms together with advances in data storage have recently made it possible to better characterize, predict, prevent, and treat a range of psychiatric illnesses. Amid the rapidly growing number of biological devices and the exponential accumulation of data in the mental health sector, the upcoming years are facing a need to homogenize research and development processes in academia as well as in the private sector and to centralize data into federalizing platforms. This has become even more important in light of the current global pandemic. Here, we propose an end-to-end methodology that optimizes and homogenizes digital research processes. Each step of the process is elaborated from project conception to knowledge extraction, with a focus on data analysis. The methodology is based on iterative processes, thus allowing an adaptation to the rate at which digital technologies evolve. The methodology also advocates for interdisciplinary (from mathematics to psychology) and intersectoral (from academia to the industry) collaborations to merge the gap between fundamental and applied research. We also pinpoint the ethical challenges and technical and human biases (from data recorded to the end user) associated with digital mental health. In conclusion, our work provides guidelines for upcoming digital mental health studies, which will accompany the translation of fundamental mental health research to digital technologies.

Keywords: digital mental health, an end-to-end methodology, human factors, cognitive biases, machine learning, knowledge discovery data base (KDD), interdisciplinar intersectoral collaborations, ethics

OPEN ACCESS

Edited by:
Jennifer H. Barnett, Cambridge Cognition, United Kingdom

Reviewed by:
Erping Long, National Institutes of Health (NIH), United States
Ellen E. Lee, University of California, San Diego, United States

***Correspondence:**
Laura Joy Boulos
laurajoyboulos@gmail.com
Alexandre Mendes
a.mendes@groupeonepoint.com
Alexandra Delmas
a.delmas@groupeonepoint.com
Ikram Chraibi Kaadoud
ichraibik@outlook.fr

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: *19 June 2020*
Accepted: *12 August 2021*
Published: *23 September 2021*

Citation:
Boulos LJ, Mendes A, Delmas A and Chraibi Kaadoud I (2021) An Iterative and Collaborative End-to-End Methodology Applied to Digital Mental Health. Front. Psychiatry 12:574440. doi: 10.3389/fpsyt.2021.574440

INTRODUCTION

Digital Health Definition

Digital health can be defined as the concept of healthcare meeting the Internet (1). It ranges from telehealth and telecare systems (2) to patient portals and personal health records (3, 4), mobile applications (5), and other online platforms and devices. However, and as opposed to digitized versions of traditional health approaches, digital health interventions (DHIs) (6) utilize artificial intelligence (AI) algorithms and other machine learning (ML) systems to monitor and predict symptoms of patients in an adaptive feedback loop (7). Improvements in ML over recent years have demonstrated potential within a variety of diseases and medical fields including neurological and mental health disorders (8) both at an individual-patient level and applied to larger populations for scalable understanding, management, and intervention of mental health conditions in different cohorts and various settings (7). In addition, and because to our knowledge, effective coverage does not exceed 50% in any country and is much lower in low- and middle-income

countries, DHIs also address social problems in the healthcare system such as poor access, uncoordinated care, and increasingly heavy costs (9). Digital mental health interventions could thus give much needed attention to underresearched and undertreated populations (10).

Digital Mental Health Technology Advances

The keywords "digital mental health" in PubMed's search engine (accessed April 2020) show that 2019 has the largest number of published articles compared to any prior year. The trend is also rising for the keywords "mental health mobile apps," providing evidence that interest in both (i) publication of articles about digital health and (ii) technical advances is rising. Advances in digital health technologies in mental health are occurring at a rapid pace in research laboratories both in academic institutions and in the industry (11). The rapidly growing number of biological devices and the exponential accumulation of data in the mental health sector aim at facilitating the four purposes of healthcare: diagnosis, monitoring, treatment, and prevention (1).

For Diagnosis

Important digital health interventions for characterization or diagnosis include algorithms for illness detection and classification (11). One digital tool that is further revolutionizing mental healthcare is conversational AI (12). Although the clinician–AI collaborations have yet to be specified and the cognitive biases considered (see *Designing digital health systems with human factors approach*), a blended approach (in an AI-delivered human-supervised model) (12, 13) is alluring.

For Monitoring

The use of data generated by personal electronic devices to monitor mental health parameters may result in useful biobehavioral markers that could in turn optimize diagnosis, treatment, and prevention and a global clinical improvement (14). This has led to the conception of all sorts of wearable devices and connected objects such as smart watches to collect data in healthy and pathological populations in a scalable unobtrusive way (15, 16), smart textiles to collect and monitor physiological outcome measure such as in athletes (17), or smart homes to monitor biophysiological measures of older people (18). This has also led to the development of various mobile applications (linked or not to a wearable device) that monitor given behaviors or cognitions in specific populations. This is the case of *eMoods*, a mood tracking app conceived for patients with bipolar disorders to follow their fluctuations. This is also the case of *PROMIS*, a mobile application to self-report different cognitive, emotional, and mood measures (19).

For Treatment

Beyond diagnosis and monitoring allowed mainly by data interpretation, some digital mental health interventions include assisting and treatment options (1). This is particularly timely as the Food & Drug Administration (FDA) has just approved its first prescription video game in mental health for kids with ADHD: *EndeavorRx* (20).

While digitized versions of classical clinical approaches propose digital conversational agents such as chatbots that provide coaching and cognitive behavioral therapies in a conceptually similar value than a human healthcare provider (7, 21), AI-based algorithms and data-driven digital health initiatives further aim at implementing more adaptive algorithms and flexible, personalized treatments *via* AI and ML (8, 21). Such is the case of *Open Book*, an assistive technology tool for adaptive, personalized text simplification for people with autism spectrum disorder (22). It is also the case of *Entourage*, a novel digital intervention that improves social connection for people with social anxiety symptoms (23), or *Doppel*, a device that helps people manage their daily stress by modulating physiological and emotional states through a heartbeat-like rhythm tactile sensation (24). Other digital mental health interventions for treatment purposes include virtual reality-based exposure [in the treatment of anxiety disorders for instance (25)] as well as the use of robotic technology [to improve social interactions in people with dementia for instance (26)].

For Prevention

By opening new modes of real-time assessment [through longitudinal data collection or through the presence of sensors in smartphones for instance, to track sleep, movement, speech… (27, 28)], digital mental health interventions enable catching new episodes of a given disorder at a very early stage. It is especially the case for suicide preventions (29).

The Need to Homogenize R&D Processes

In contrast, there is only scarce clinically significant outcomes of digitalized solutions. Although both advances in fundamental research and technical innovations are occurring rapidly, translation from one to the other has been slower (11). This can be explained by the lack of better-designed clinical trials and the loss of interest at the patient level in digital health products over time, both of which lead to poor long-term data and scarce information on whether new behavior facilitated by a digital health tool is long-lasting (30).

Another major problem at the time is the disparity of research and development processes across fields and sectors. One way of accelerating the potential benefits of digital mental health interventions and optimizing the transformation of fundamental discoveries into innovative digital technologies applied to routine clinical practice would be to propose a methodology that could be used across disciplines and sectors in the field of mental health. This would include homogenizing research and development processes in academia as well as in the private sector; improving technical methods that standardize, aggregate, and exchange data; centralizing data into federalizing platforms focusing on scalability; and establishing data repositories, common data standards, and collaborations (14, 31).

The Global Pandemic Context

In March 2020, the WHO declared the novel coronavirus disease of 2019 also known as COVID-19 as a global pandemic. Today, a year later, the WHO counts 185,038,214 confirmed cases of COVID-19 globally, including 3,250,648 deaths. Amid this

rapidly evolving sanitary crisis, digital innovation is being used to respond to the urgent needs of the pandemic. Actions in the field have been involving multiple stakeholders, from frontline healthcare to public health and governmental entities. They have also raised new challenges regarding the link between academia and the industry, the different velocities at which the two sectors evolve, the ethical questions of data collection, and the various geographical and socioeconomic inequalities due to limitations in capacity or resources (32).

Apart from the direct risks of COVID-19 on health and the healthcare system, the uncertainty of the context and the high death rate due to the virus also exacerbate the risk of mental health problems and worsen existing psychiatric symptoms, further impairing the daily functioning and cognition of patients (33).

While these illnesses do not all represent an immediate threat to life, they will have long-lasting serious effects on individuals and large populations. Emerging mental health issues should thus be addressed promptly. In addition, the multiple logistic changes imposed on us by the pandemic pose a unique challenge in mental health service delivery. For example, the restriction in freedom of movement and face-to-face therapies increases psychological distress (32). The limited knowledge on the virus and the overwhelming news that surround it also increase anxiety and fear in the public (33, 34). In addition, long quarantine durations are generating frustration, boredom, stigma, and stress, as well as financial loss that also affects mental health. This is without mentioning highly vulnerable populations such as healthcare providers (32), university students (35), children (36), and naturally anxious individuals (37) who are more prone to developing mental illnesses such as posttraumatic stress disorder or anxiety and mood disorders during this pandemic crisis. This is also without mentioning the already.

In this context and with the advent of AI, a digital methodology that optimizes and homogenizes research processes in an intersectoral and transdisciplinary approach makes more sense than ever, specifically in the field of mental health. Implementing such approaches could help detect and monitor mental health symptoms and their correlation to COVID-19 parameters (whether individuals are affected by the virus or know people affected by it, how political decisions impact mood and anxiety of general populations, etc.). Early detection and close monitoring would in turn allow adequate in-time treatment in the short term and prediction as well as prevention in the longer term.

Introduction to Our Work

Here, we propose an end-to-end methodology that highlights key priorities for optimal translational digital mental health research. Each step of the process is elaborated from brainstorming to product creation, with a focus on data analysis. Based on iterative processes, the methodology aims at being cross-sectorial, at the intersection between academia and the private industry. By formalizing the methodology around a mental health use case, the methodology also aims at being interdisciplinary, encompassing different fields (from computational neuroscience to psychology and well-being) all while stressing on the importance of human factors in the digitalization of health. An important goal of the methodology is thus to allow robust collaborations between experts from different fields and sectors (practicing clinicians, AI researchers in academic institutions, and R&D researchers in private industries) to pinpoint then advance foundational and translational research relevant to digital mental health and to create ultimately digital tools that satisfy various stakeholders (usability, clinical benefit, economic benefit, security, and safety). All in all, our methodology has the short-term ambition to propose guidelines for upcoming digital mental health studies and the ultimate ambition to transform the gap between fundamental and applied research into a federalizing platform.

CONCEPTUALIZATION AND PROJECT LEARNING

Project Idea and Concept

Evaluating the Feasibility of an Idea

All research begins with a question. Not all questions are testable though, and the scientific method only includes questions that can be empirically tested (observable/detectable/measurable) (38). Similarly, not all questions lead to the development of solutions. As a matter of fact, only few research projects directly reach practical solutions. However, in the digital health sector, research tends to have (or at least ought to have) a very pragmatic, concrete, and measurable outcome (39). The selection of ideas is therefore one of the most complex steps of the research process in the digital health sector since, in addition to verifying whether their idea can be transformed into a project, researchers must also evaluate whether the project can lead to practical often technical solutions, and when. As the world of technologies moves fast (11), by the time that an idea leads to a solution, the solution might lose some or all its value. It is thus crucial to assess whether the idea is feasible and realistic early in the process.

As a result of the COVID-19 pandemic context for instance, there has been an increase in the usage of telehealth medicine and alternative digital mental health options such as mobile applications and web-based platforms (40). Although the need is real and measurable, research projects must be cost-effective, and depending on the investment needed, they ought to be useful not only within a short-term period (i.e., to treat current psychiatric illnesses) but also in a longer time frame, to treat for instance the expected rise in symptoms of trauma among the general population (40).

In addition, according to Gartner (1), digital health research follows a hype cycle divided into five stages illustrated in **Figure 1**:

1. the innovation trigger,
2. a peak of inflated expectations,
3. a trough of disillusionment,
4. a slope of enlightenment, and
5. a plateau of productivity.

An ideal digital health research project predicts the failures that will occur at the third stage and the plateau that will be reached

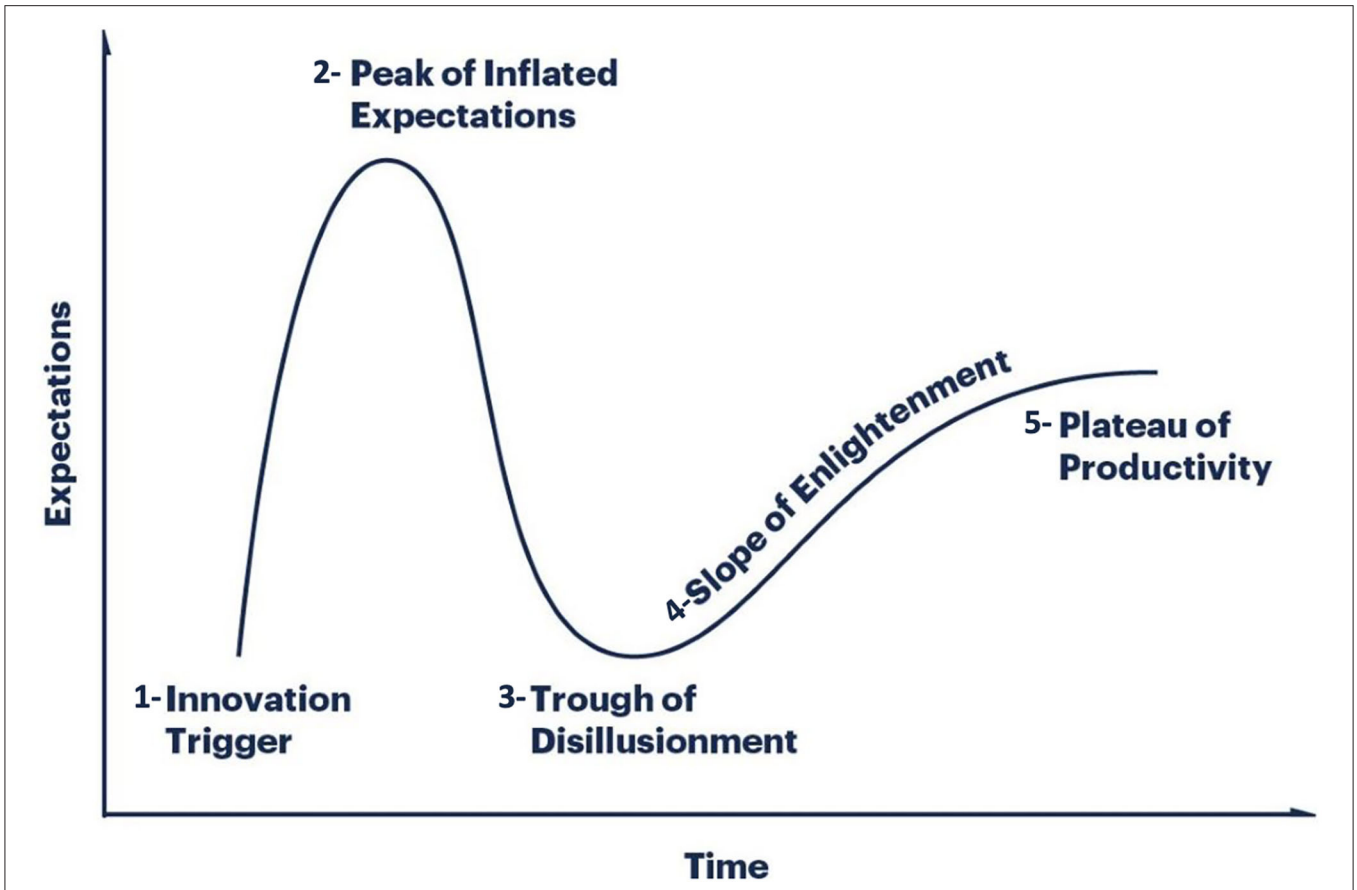


FIGURE 1 | The five stages of the hype cycle of digital health research: (2) the innovation trigger, (3) a peak of inflated expectations, (4) a trough of disillusionment, (5) a slope of enlightenment, and (6) a plateau of productivity. Adapted from Gartner (41).

at the fifth stage in order to prepare for them and increase the chances of overcoming them. For instance, treating traumas through mobile applications might 1 day be the old-fashioned way of approaching such disorders. This is one major challenge as there are no generic methods describing a digital health project from research inception to solution development (1).

Evaluating an initial idea further faces more classical challenges such as finding the good mix between focused enough to be interesting yet broad enough to build on existing knowledge (42). A digital health research project should balance more than any other research project between ambitious but not overambitious as the competitive landscape is both wide and niche. Going back to our COVID-19 example, this would mean developing digital health technologies that are precise enough to treat specifically traumas in a pandemic context, but broad enough to be adaptable when traumas would not be the main mental health issue anymore, in a near enough future.

Defining the Goal and the Approach

What is it that we want to put in light? Defining the goals and objectives of a digital health research project is essential as it keeps the project focused (11). The process of goal definition usually begins by writing down the broad and general goals of the study. As the process continues, the goals become more clearly defined and the research issues are narrowed to an extent that depends on the adopted approach. For instance, the general goal of a mental health mobile application could be to improve mental health conditions; this is the case of the 1,009 psychosocial wellness mobile apps that were found in a study looking to differentiate scientifically evidenced apps from the success stories due to a media buzz (43). A more palpable goal could be to promote behavioral change; this is the case of *notOK*, a suicide prevention application that alerts the support system of a patient when negative thoughts are too close to an acting out. This is also the case of *Twenty-Four Hours A Day*, an addiction app that offers 366 meditations (one per day) to help abstinent patients focus on sobriety. The goal ought, however, to be further narrowed as the design of the application might consider eliciting not only more engagement on the mobile app overall, but perhaps effective engagement defined by specific patterns (44). *Twenty-Four Hours A Day* could, for example, be used effectively during a year at the end of which users could lose interest, potentially resulting in a relapse. Narrowing the number of users could allow a deeper

engagement of actual users; more is not always better in digital health (45).

Given the multistakeholder nature of healthcare and their varying incentives, the best approach to impactful and useful digital health research may differ depending on the project. The main challenge is to find the right balance that maximizes clinical impact all while utilizing efficient resources and at a rate that corresponds to the needs of the market in a globally very dynamic and rapidly changing digital health landscape. This brings us back to creating a requirement set broad enough to encapsulate concepts important to all products, but not too inclusive that the requirements are not relevant anymore (39).

This also allows us to emphasize on the importance of staying flexible and ready to change strategies depending on the number and the rate at which new technical solutions are deployed with time. *Headspace* for instance had started as an events company organizing mindfulness trainings and workshops; as they stayed open to opportunities, they later developed their mobile application that is currently being used in several clinical trials (46). In the case of this app, adopting a ready-to-change pivot strategy allowed them to seize an opportunity and scale drastically.

One way to stay flexible is to inject some agility in the research processes. Agility uses iterations (also called sprints) to create short loops of work (1–4 weeks) that start with planning and end with retrospection, favoring more frequent deliverables (such as quick posters or abstract publications, proof of concepts or minimum viable products) (47). If the concept of agility springs from the software development field, it has been more broadly applied in different fields and sectors recently, such as in mobile health technology (47). A clear step-by-step example applied to our use case, i.e., digital mental health, is the text-based coaching practical guidance provided by Lattie et al. (48).

All in all, it is crucial to define then narrow the goal progressively while balancing between clinical requirements and market realities by staying agile and considering the potential conceptions and misconceptions of all stakeholders.

Clarifying Digital Health Research Conceptions and Misconceptions

Everyone is susceptible to the misconceptions of research, development, and innovation, including researchers and any other individual in academia or in the private industry (see *Identifying cognitive biases in digital health to improve health outcomes*). The what of research is challenging in itself and even more so in the digital health context that often includes translational application at the end of the process as well as the need to confront the views and requirements of academia and the industry. It is therefore critical to identify these misconceptions early in the research project to reduce them and promote alternative conceptions where necessary. Most common misconceptions include the following (49):

- Good research procedures necessarily yield positive results.
- Research becomes true when published.
- Properly conducted research never yields contradictory findings.
- It is acceptable to modify research data to make them look perfect.
- There is only one way of interpreting results.

Discussing conceptions and misconceptions of research can reduce cognitive biases (see *Identifying cognitive biases in digital health to improve health outcomes*) and improve research outcomes all while favoring a holistic approach to research (42):

- How would you describe research to your grandmother?
- What is the difference between academic (moving knowledge further, contributing to the development of the discipline, explaining, arguing, conceptualizing, theorizing, developing insights, being rigorous and methodical, situated within a theoretical or conceptual tradition) and industrial research (fact-finding, collecting and reporting, producing and developing)?
- How to combine different views and different approaches and methods of research into an R&D model that serves research and innovation in the digital health sector?

This "awakening" step is of particular importance in DHI as interdisciplinary and intersectoral collaborations increase by the day (see *Identifying the team and potential partners or collaborators*).

Extending the Literature to a Market Research

Reviewing the literature is an inevitable step of a research project (for further details, see **Appendix 1**). Nonetheless, it cannot factor in major advances in health technology if relying only on peer-reviewed sources (50). Given both the size (valued at 75 bn in 2017 by Technavio's Global Digital Health Market research report) and the evolution rate (projected to reach 223 bn in 2023 as predicted by Global Market Insights) of the digital health market, it seems crucial to complement the literature review with adequate market research also called gray literature.

Given the complexity that is characteristic of the digital health landscape of technologies, market research cannot be straightforward. For it to be as thorough as possible, it should include project reports, market research foresees, policy documents, and industry white papers (39). For instance, in the oversaturated market of mobile apps advocating for wellness and self-care, one approach would be to conduct a systematic review of publicly available apps on the stores using key words related to the topic (43).

In the context of digital mental health research, the market research would allow researchers not only to compare the potential outcome of their research to the state of current technology (51) but also to predict or at least speculate whether their solution will still have the same value by the time it reaches the market. Such market research could also provide researchers with an overview of the general landscape, i.e., of the unexplored new market areas (blue ocean strategy; 47).

Identifying the Team and Potential Partners or Collaborators

Common benefits to collaboration including brainstorming, division of labor, and speed of execution are challenged by

the difficulty of developing a shared vision and defining roles and responsibilities for the different collaborators (52). These challenges are exacerbated in the context of digital health as the field is essentially both interdisciplinary and intersectoral (53), bringing together academic researchers, private industries and their R&D departments, clinicians, patients, and other healthcare consumer groups (54). Indeed, while collaborations in the field are facilitated by complementary roles, authentic communication between partners, and clearly outlined goals or expectations prior to the collaboration, they can also be jeopardized by misaligned expectations, differences in productivity timelines, and balancing business outcomes vs. the generation of scientific evidence (53). It is thus crucial not only to identify the right fit for a collaboration but also to outline and communicate openly about goals, expectations, and timelines. This was done by *X2AI*, a US-based digital health company that developed in collaboration with experts (including clinical, ethical, technical, and research collaborators) an ethical code for startups, labs, and other entities delivering emotional AI services for mental health support (55). Once the project is developed, it moves to the commitment phase or project planning.

Project Execution

Sampling

The rapid advancement of digital health technologies has produced a research and development approach characterized by rapid iteration, often at the expense of medical design, large cohort testing, and clinical trials (39, 43). According to the WHO's guidance for digital health research (56), digital research measures are too often evaluated in studies with varying samples and lack of or poor validation. Additional challenges with digital health research include a potentially unrepresentative sample (57). Consequently, insufficient sample sizes may make it difficult for these data to be interpreted through ML techniques (58) (see *Data postprocessing: visualization and evaluation*). Underestimation occurs when a learning algorithm is trained on insufficient data and fails to provide estimates for interesting or important cases, instead approximating mean trends to avoid overfitting (59).

It is, however, necessary to pursue the adequate amount of evaluation and verification to avoid dubious quality and ensure usefulness and adequacy of the solution (60). To do so, it is crucial to improve sampling strategies by including underrepresented groups in the recruitment, collecting and analyzing reasons for declining, analyzing the profiles of recurrent participants (61), and creating ultimately novel smart sampling approaches (62).

Choosing the Appropriate Material and Method

Research projects in the digital health sector can take the form of cohort studies, randomized trials, surveys, or secondary data analysis such as decision analyses, cost-effectiveness analyses, or meta-analyses. To sum things up, there are three basic methods of research:

1. *Surveys* by e-mail, *via* a web platform or *via* a mobile application. They usually involve a lengthy questionnaire that is either more in-depth (usually by email) or more cost-effective (web- and app-based surveys) (63).
2. *Observation* monitors subjects without directly interacting with them. This can be done either in the environment of the subject with different monitoring devices (ecological environment) or in a lab setting using one-way mirrors, sensors, and cameras to study biophysiological markers or behavior (controlled environment) (64). Faster digital tools now allow monitoring patients *via* their health insurance or *via* different health apps.
3. *Experiments* allow researchers to modify variables and explain changes observed in a dependent variable by a change observed in the independent variable. Experiments were mostly restricted to laboratory contexts as it is very difficult to control all the variables in an environment. This contextual limitation is, however, blurred with digital health research and the use of technologies in less controlled environments. In addition, and even within a laboratory, attention should be given to hardware and software variability between devices as it can affect stimulus presentation and perception of a stimulus as well as human–machine interaction (64).

Although there is no one best method for all digital health research projects, a well-defined problem usually hints at the most appropriate method of research. There also often are cost/quality trade-offs that urge the researcher to consider budget and time as part of the general design process.

PROJECT DESIGN

Designing Digital Health Systems With Human Factors Approach

What Is User-Centered Design?

A user-centered design (UCD) is an iterative design process in which designers focus on users and their needs in each phase of the design process. Design teams may include professionals from multiple disciplines (ethnographers, psychologists, engineers), as well as domain experts, stakeholders, and the users themselves. They also involve users throughout the design process *via* a variety of research and design techniques (surveys, interviews, brainstorming), to create highly usable and accessible products. Each iteration of the UCD approach involves four distinct phases illustrated in **Figure 2** (65) [see norm ISO (9241-210, 2010)]:

1. understanding the context of use,
2. identifying and specifying user requirements,
3. designing solutions, and
4. evaluating the outcomes of the design to assess its performance.

Iterations are repeated until the evaluation phase is satisfactory.

The term "user-centered method" was first used in 1986 by Don Norman (66), who argued the "*importance of design in our everyday lives, and the consequences of errors caused by bad designs*." Ambler later highlighted the efficiency of agility (47) by demonstrating that UCD reduces computing costs (67). UCD approaches further provide advantages in a digital change context (68), all of which can be distinguished in four ways (69):

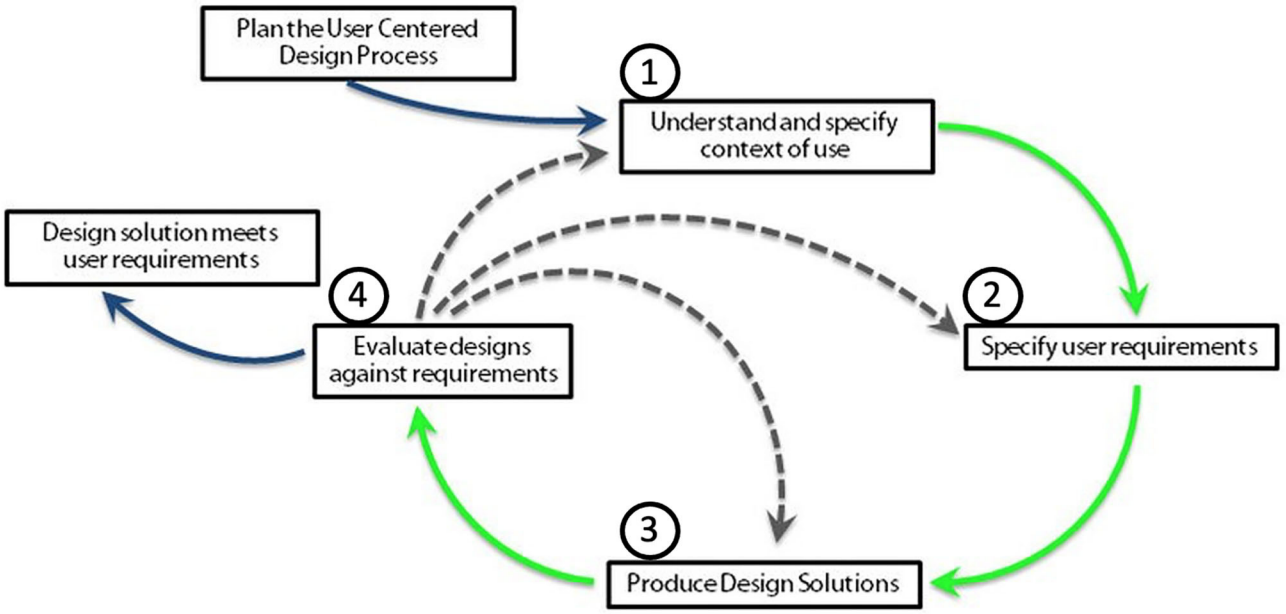


FIGURE 2 | The four phases of the user-centered design: (2) understanding the context of use, (3) identifying and specifying user requirements, (4) designing solutions, and (5) evaluating the outcomes of the design to assess its performance. Adapted from Nielsen (65).

- User involvement increases the likelihood for a product to meet expectations which in turn increases sales and reduces customer services costs.
- Tailored products reduce the risk of human error (29, 70).
- Designer–user reconciliation increases empathy and creates ethical designs that respect privacy.
- By focusing on specific product users, designers recognize the diversity of cultures and human values through UCD—a step in the right direction to create sustainable businesses.

User-Centered Design in Digital Health

Digital health asserts a translational vision of changed practices and care systems [new modes of assessment through virtual reality (71) and the presence of sensors in smartphones for instance (27, 28)] to drive better health outcomes. However, the human–technology interaction was only put in light recently (72): it took a decade to first develop and then apply a theoretical understanding of the scope for a substantial, human-centered "design-reality" gap in healthcare (73).

In terms of functionalities, the focus is on usability of parameters such as appearance, appeal, and ease of navigation, as well as various interventions that include quizzes, games, self-monitoring tools, progress reports, downloadable documents, and other similar features [e.g., for social anxiety disorder (74)]. On the other hand, numerous barriers potentially prevent people from participating in evaluations of DHIs such as being too busy, feeling incapable of using the technology, or disliking its impersonal nature (75, 76).

Increasing interest in human factors has underpinned key developments in digital health, spanning intervention development, implementation, and the quest for patient-centered care (77). The emergence of ML chatbots and other patient-centered designs within Internet-based cognitive behavioral therapy has proven to facilitate access and improve tailored treatments (78). This is mainly due to the digital removal of several barriers such as reduced perceptions of stigma (very present in face-to-face services) and a rapid response to the need of "in the moment" support for mental distress. All these reasons increase the demand for digital mental healthcare in formal healthcare settings (79).

Benefits, Facilitators, and Barriers of UCD in DHI

To truly benefit from DHI, privacy and data governance, clinical safety (handling crisis in mental health apps for instance), and evidence for effectiveness must be at the core of the design (80, 81). This is unfortunately not always the case as shown by a smartphone app review revealing that, out of all health apps, only 11 were identified as "prescriptible" [meaning that they included randomized controlled trials (RCTs) reporting of effectiveness without clinical intervention] (82).

The UCD of digital health systems enables greater engagement and long-term use of digital tools (83). However, little attention is given to human factors such as ethnography of users or usability testing (77), or to the real-world difficulties that individuals face (84, 85) such as technology cost and privacy or security issues (86). These barriers reduce health outcomes with poor user engagement despite mobile health interventions (87–89). The decision-making power toward consumers is in turn insufficient (80), raising questions of access [namely in low- and

middle-income countries (90)], equity, health literacy, privacy, and care continuity (14).

In their review of all barriers and facilitators for DHI engagement and recruitment, O'Connor et al. (91) distinguished four themes:

1. personal agency and motivation,
2. personal life and values,
3. the engagement and recruitment approach, and
4. DHI quality.

Education (91) and age (92, 93) were given particular attention as poor computer skills in both low-education individuals and old adults added to the enrollment struggle. In the same vein, literacy skills (94, 95) and the ability to pay for the technology (96) have impact on people's ability to interact with and use DHIs. All these factors ought to be further explored.

In summary, adopting a UCD of DHI would optimize long-term tool acceptance (6). Interdisciplinary collaborations could provide knowledge about "the context of use" (97), but it is crucial to further identify the technological and economic feasibility of the design (98, 99). In addition to the central role of human factors in DHI, attention should also be given to cognitive biases that come with ML strategy implementation and data interpretation.

Identifying Cognitive Biases in Digital Health to Improve Health Outcomes

Studies from the past decades point at the vulnerability of the human mind to cognitive biases, logical fallacies, false assumptions, and other reasoning failures (100). In the health system context, cognitive biases can be defined as faulty beliefs that affect decision-making and can result in the use of heuristics in the diagnostic process (101, 102). Kahneman and Tversky introduced a dual-system theoretical framework to explain judgments, decision under uncertainty, and *cognitive biases* (103, 104). In this model, illustrated in **Figure 3**, system 1 refers to an automatic, intuitive, unconscious, fast, and effortless decision process. Conversely, system 2 makes deliberate, non-programmed, conscious, slow, and effortful decisions. Most cognitive biases are likely due to the overuse of system 1 vs. system 2 (100, 105–107).

Cognitive Biases Included in Diagnostic Reasoning and Healthcare Strategies

"*Diagnostic reasoning is the complex cognitive process used by clinicians to ascertain a correct diagnosis and therefore prescribe appropriate treatment for patients*" (108): the ultimate consequences of diagnostic errors include unnecessary hospitalizations, medication underuse and overuse, and wasted resources (109, 110).

Diagnostic reasoning and risk of errors can be explained by adapting the dual-system model to the health system context. For instance (99—see **Appendix 2**), system 2 overrides system 1 when physicians take a time-out to reflect on their thinking. System 1 also often irrationally overrides system 2 when physicians ignore evidence-based clinical decision rules that outperform them. Depending on what system overrides the other, the calibration (the degree to which the perceived and actual diagnostic accuracy corresponds) will differ.

The main cognitive biases affecting medical performance and diagnosis are the following (111, 112):

- *Premature closure* (113–117): an automatic process that occurs when the provider closes the diagnostic reasoning process by clinging to an early distractor/diagnosis without fully considering all the salient cues (106).
- *Search satisficing* (112, 118): a subtype of premature closure in which searches for further evidence are terminated after a diagnosis is reached. This is the case for medical students that do not initiate a search for a secondary diagnosis (118).
- *Availability* (106, 113, 118): falsely enhancing the probability of a diagnosis following the recent exposure of the physician to that diagnosis (106).
- *Anchoring* (112, 113, 115): a subtype of premature closure in which a provider stakes their claim on a diagnosis, minimizing information that do not support the diagnosis with which they have attached their proverbial anchor (115).
- *Base rate neglect* (119): predicting the diagnosis occurrence probability when two independent probabilities are erroneously combined, ignoring the base rate and leading to under- or overestimating the diagnosis possibility (120).
- *Diagnostic momentum* (112, 119, 121): a subtype of anchoring and premature closure in which the suggestion power of colleagues is taken at face value. For example, the diagnosis of anxiety disorder of the patient established from her family doctor through to the emergency department (ED), and although she might well have had hyperventilation due to anxiety, other possibilities were not ruled out earlier on in her care (112).
- *Overconfidence and lower tolerance to risk/ambiguity* (122, 123). Because of these two biases, misdiagnosis, mismanagement, and mistreatment are frequently associated with poorer outcomes, leading to patient dissatisfaction and medical complaints and eventually to a dropout of the digital health system (79, 124–126).

In the specific context of digital mental health, it is important to identify potential cognitive biases in patients as well in order to avoid misinterpretation and treatment misusage. In addition to the eight cognitive biases mentioned above, other cognitive factors such as coping strategies (127, 128) and the role of emotional stimuli (e.g., in depression, there is a lack of such a bias) (129) require particular attention in order to design tailored digital treatments and to drive ultimately an effective digital health strategy.

Early recognition of the cognitive biases of physicians is crucial to optimize medical decisions, prevent medical errors, provide realistic patient expectations, and decrease healthcare costs (107, 126, 130). Some debiasing strategies include the following:

1. Advocating for a view in which clinicians can change thinking patterns through awareness of bias and feedback (100). It consists of theories of reasoning and medical decision-making, bias inoculation, simulation training,

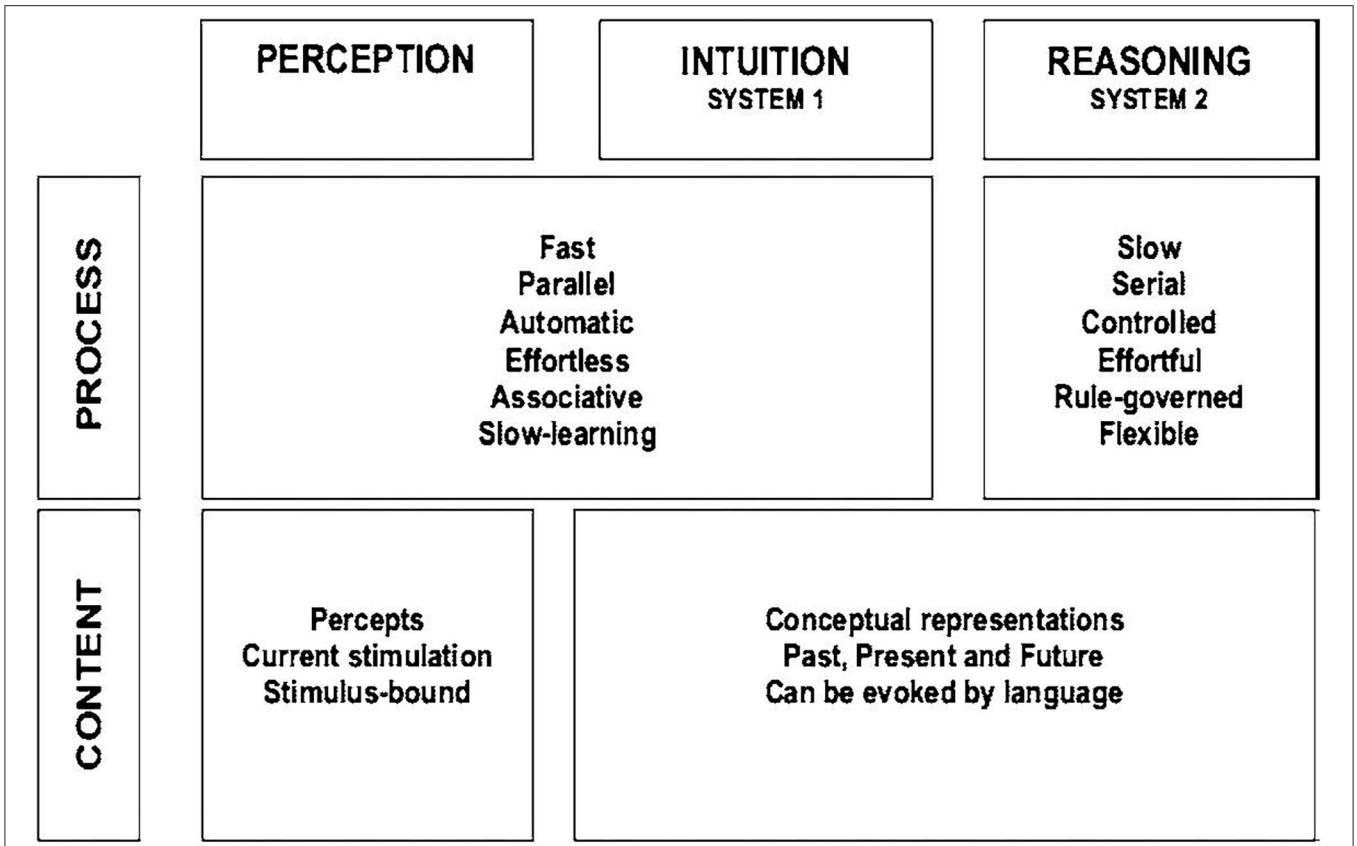


FIGURE 3 | Properties of the dual-system theoretical framework of Kahneman (103) to explain judgments, decision under uncertainty, and cognitive biases: (2) system 1 refers to an automatic, intuitive, unconscious, fast, and effortless decision process; (3) system 2 makes deliberate, non-programmed, conscious, slow, and effortful decisions. Schema taken from Kahneman (103).

computerized cognitive tutoring, metacognition, slow-down strategies, group decision strategy, and clinical decision support systems to force diagnostic reasoning out of bias-prone thought analytic processes.

2. Digital cognitive behavioral therapy [see, for review, 125] through which positive cognitive bias modification could be used as a potential treatment for depression (131), for anxiety disorders (25), for persecutory delusions (132), for improvement of social interaction in autism spectrum disorders and dementia (26), and for people with suicidal thoughts (133).

There is, however, no consensus regarding the efficacy of such debiasing approaches (118). In addition, other biases such as *aggregation bias* (the assumption that aggregated data from clinical guidelines do not apply to their patients) or *hindsight bias* (the tendency to view events as more predictable than they really are) also compromise a realistic clinical appraisal and could lead to medical errors (134, 135). This brings us to the urgent need for transparent and explicit data and strategy.

Biases in Defining Machine Learning Strategies

Cognitive biases exposed previously mainly concern physicians and their ability to analyze a digital diagnosis. Data scientists are also prone to specific cognitive biases given the strong interpretative component of data science and ML (136). Biases affecting data scientists in the digital mental health setting include but are not limited to the following (136):

- *survivorship*: a selection bias in which data scientists implicitly filter data based on some arbitrary criteria and then try to make sense out of it without realizing or acknowledging that they are working with incomplete data;
- *retrospective cost*: the tendency to make decisions based on how much of an investment they have already made, which leads to even more investment but no returns whatsoever;
- *illusion of causality*: the belief that there is a causal connection between two events that are unrelated;
- *availability*: the natural tendency to base decisions on information that is already available without looking at potentially useful alternatives that might be useful; and
- *confirmation*: the interpretation of new information in a way that makes it compatible with prior beliefs.

Despite these data science biases, a promise of ML in healthcare is precisely to avoid biases. The biases of scientists and clinicians would be circumvented by an algorithm that would objectively synthesize and interpret the data in the medical record and/or

offer clinical decision support to guide diagnosis and treatment (58). In the digital health context, integration of ML to clinical decision support tools such as computerized alerts or diagnostic support could offer targeted and timely information that would in turn improve clinical decisions (58, 137–140). With the rise of ML in the DHI, data sources and data collection methods should be further examined to better understand their potential impact (141–143). Biases that could be introduced through reliance on data derived from the electronic health record include but are not limited to the following:

- *Missing data*: If communicated sources such as patient-reported data are incomplete (missing or inaccessible), algorithms (that only use available data) may correctly misinterpret available data (144). Algorithms could thus be a bad choice for people with missing data (145) [people with low socioeconomic status (146) or those with psychosocial issues (147) for instance].
- *Misclassification and measurement errors*: Misclassification of diseases and measurement errors are common sources of bias in observational studies and analyses based on electronic health record data. Care quality may be affected by implicit biases related to patient factors, such as sex and race, or practitioner factors [e.g., patient with low socioeconomic status (148) or women (149)]. If patients receive differential care or are differentially incorrectly diagnosed based on sociodemographic factors, algorithms may reflect practitioner biases and misclassify patients based on those factors (58).

We mostly identified and described biases that interfere once the data are already collected. It is important to note that biases can also interfere earlier in the process, at every step of it, from brainstorming to literature reviewing (143). The main recommendation is to stay alert to all different biases, whether they are mentioned in this paper or not.

DATA COLLECTION–ANALYSIS

From Data to Information

As seen above, decision-making in the medical field often has far-reaching consequences. To better measure these consequences, it is essential to build certainties: certainties on the data used, their source, their format, and their update; certainties on the information put forward and their implications; and certainties on the tools exploiting these data as well as on the reliability of the algorithms and visual representations made available. These questions concern data in a broad way.

It thus seems important to start this technical part of the paper by defining the notions of data, information, and knowledge, as all three are involved in decision-making processes.

We will then focus our approach on the data and the different steps to structure, exploit, and enhance them.

Definitions: Data, Information, and Knowledge

Grazzini and Pantisano (150) defined each concept as follows:

Data can be considered as raw material given as input to an algorithm. Since it cannot be reproduced when lost, it must be carefully preserved and harvested. It can be of different forms: a continuous signal as in the recording of an electroencephalogram (EEG), an image representing a magnetic resonance imaging (MRI), a textual data, or a sequence of numerical values representing a series of physiological measurements or decisions taken *via* an application. Data can be complete, partial, or noisy. For example, if only a portion of an EEG recording is available, then the data are partial. Conversely, an EEG recording that has been completed but that has some parts unusable is said to be noisy because the noise alters the completeness of the recording. Two types of data can be distinguished: unstructured data, i.e., data directly after their collection or generation, and structured data, i.e., data that have been analyzed, worked on, and put in relation to each other to put them in a format suitable for the analysis considered afterwards. In the second case, it is considered information. Importantly, data by themselves are worthless.

Information is dependent on the original data and the context. If it is lost, it can be reproduced by analyzing the data. Depending on the data processed at time *t*, information must be accurate, relevant, complete, and available. Information is intelligible by a human operator and can be used in a decision-making process. It is therefore significant and valuable since it provides an answer to a question. It can take various forms such as a text message, a table of numerical values, graphs of all kinds, or even in the shape of a sound signal.

When semantics are added to a set of information, it becomes *knowledge*. Information, depending on the context, will not have the same impact. It is the context and the semantics brought by it and the human operator involved that will determine the value of that knowledge.

To illustrate these definitions in a mental health setting, in the case of a patient undergoing a follow-up with a psychiatrist: the psychiatrist can make his patient pass numerous tests in order to collect data: MRI, EEG, and textual answers to questionnaires. These data, once processed, formatted, and analyzed together, will represent a set of information on the condition of the patient. It is the combination of the knowledge and experience of the doctor, combined with his knowledge of the patient, his family context, and the current socioeconomic context, that will enable him to have a global knowledge of his patient and to provide him with the best possible support.

The passage from data to information thus requires a majority of digital processing to highlight correlations according to a given context. However, the passage to the knowledge stage requires considering individuals involved (see *Project design*). **Figure 4** illustrates data transformation into information through digital processing and into knowledge through human evaluation.

There is thus an increasing complexity in this process of transforming data into information and then into knowledge which make it difficult to identify and extract. We will present a process dedicated to these tasks in the following section.

Knowledge Data Extraction in the Literature

The process of Knowledge Discovery of Data (KDD) is defined as the process of discovering useful knowledge from data (151). As a three-step process, the KDD includes (2) a preprocessing step which consists of data preparation and selection, (3) a

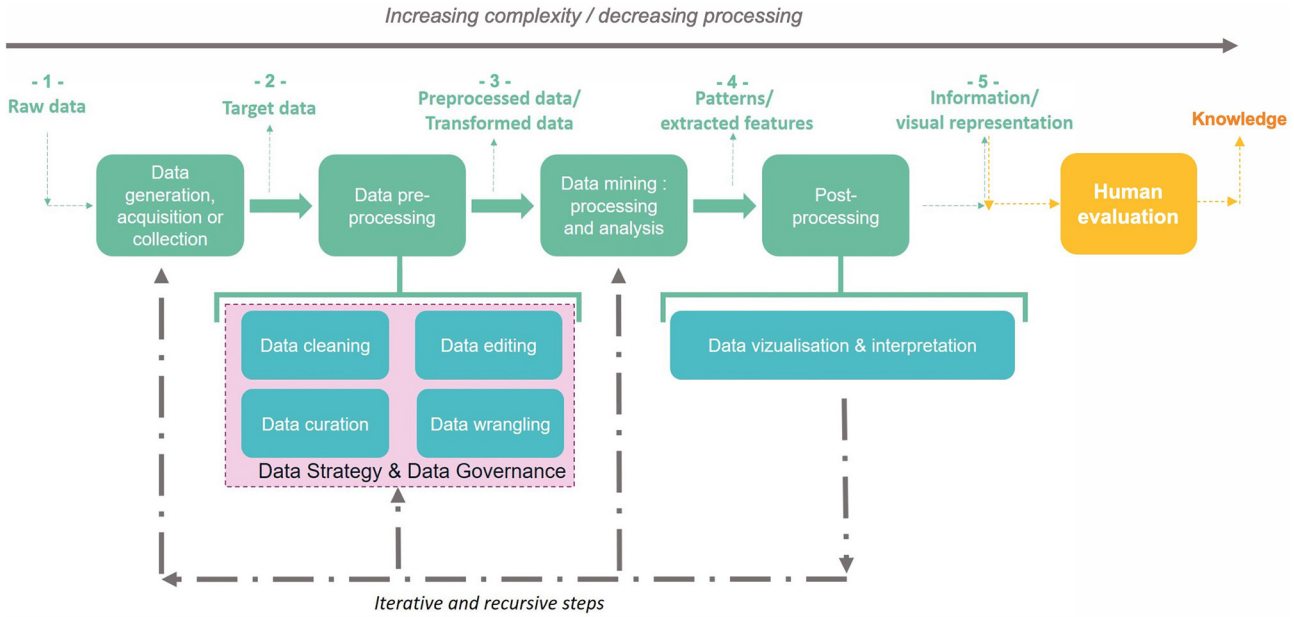


FIGURE 4 | Data transformation into information (from 1 to 5) then knowledge through KDD steps: (a) raw data can be collected to become target data, which can also be acquired or generated by an external process: it is the data generation, acquisition, or collection step. (b) Target data are submitted to a data preprocessing step according to the data mining techniques targeted: target data can thus be cleaned, filtered, completed, and anonymized if necessary. This second step allows to obtain preprocessed data. (c) Preprocessed data are submitted to data mining techniques to detect, identify, and extract patterns and relevant features. (d) Discovered patterns and features can thus be rearranged (e.g., with visual tools) for or within the phase of interpretation during the postprocessing step. The result is the generation of information. (e) Information evaluated by a human becomes knowledge: this step is an external one (i.e., not technical) and takes into account knowledge of the situation (context, issues, stakeholders, etc.).

data mining step involving the application of one or many algorithms in order to extract information (i.e., patterns), and (4) a postprocessing step to analyze extracted information manually by a human operator and lead to knowledge discovery.

As an iterative and interactive process, KDD involves many steps and decisions of the users. Iterations can continue as long as extracted information does not satisfy the decision-maker (see *Identifying cognitive biases in digital health to improve health outcomes*).

Concretely, as illustrated in **Figure 5**, the KDD stages encompass the following: (2) understanding the scope of the application field; (3) creation of the target dataset; (4) data cleaning and preprocessing; (5) data reduction and projection: reducing the number of variables to be analyzed by reducing the dimensionality of the data, extracting invariant representations, or searching for relevant characteristics; (6) matching the goals of the KDD process with the right method(s) in data mining; (7) exploratory analysis and selection model and hypothesis: selection of the data mining algorithm and method that will be used for the pattern search; (8) data mining: searching for interesting patterns in a particular form of representation, which includes rule and tree classification, regression, and clustering; (9) data postprocessing and visualization: interpretation of the patterns found with possible return to any step from 1 to 7 for a new cycle; and (10) action on discovered knowledge.

Here, we mainly focus our approach on the technical aspect, i.e., data and their transformation into information. We aim to present a complete and global approach by covering the KDD stages in the life cycle of a digital health product from the definition of the scientific question to data collection and analysis (for further details, see **Appendix 3**). We will reveal our approach in the following section.

Information Data Extraction Applied to Technology

We are aligned with the three-step approach of Fayyad (151) for information extraction:

- data preprocessing (data cleaning, data editing, data curation, and data wrangling),
- data mining with a special focus on biostatistics and AI and ML algorithms, and
- postprocessing focusing on data visualization.

Figure 4 proposes a representation of the global approach for information extraction for a specific question or product design.

Upstream of these activities, we would like to highlight two areas that are essential to good data management and that allow an optimization of the research of a team: data strategy, which aims at standardizing data management, and data governance, or the implementation of solutions to respond to the strategic issues defined beforehand.

Global Approach: Data Strategy and Governance

Data Strategy

Within a digital research project, technical and operational tasks are either managed by the same person (it is

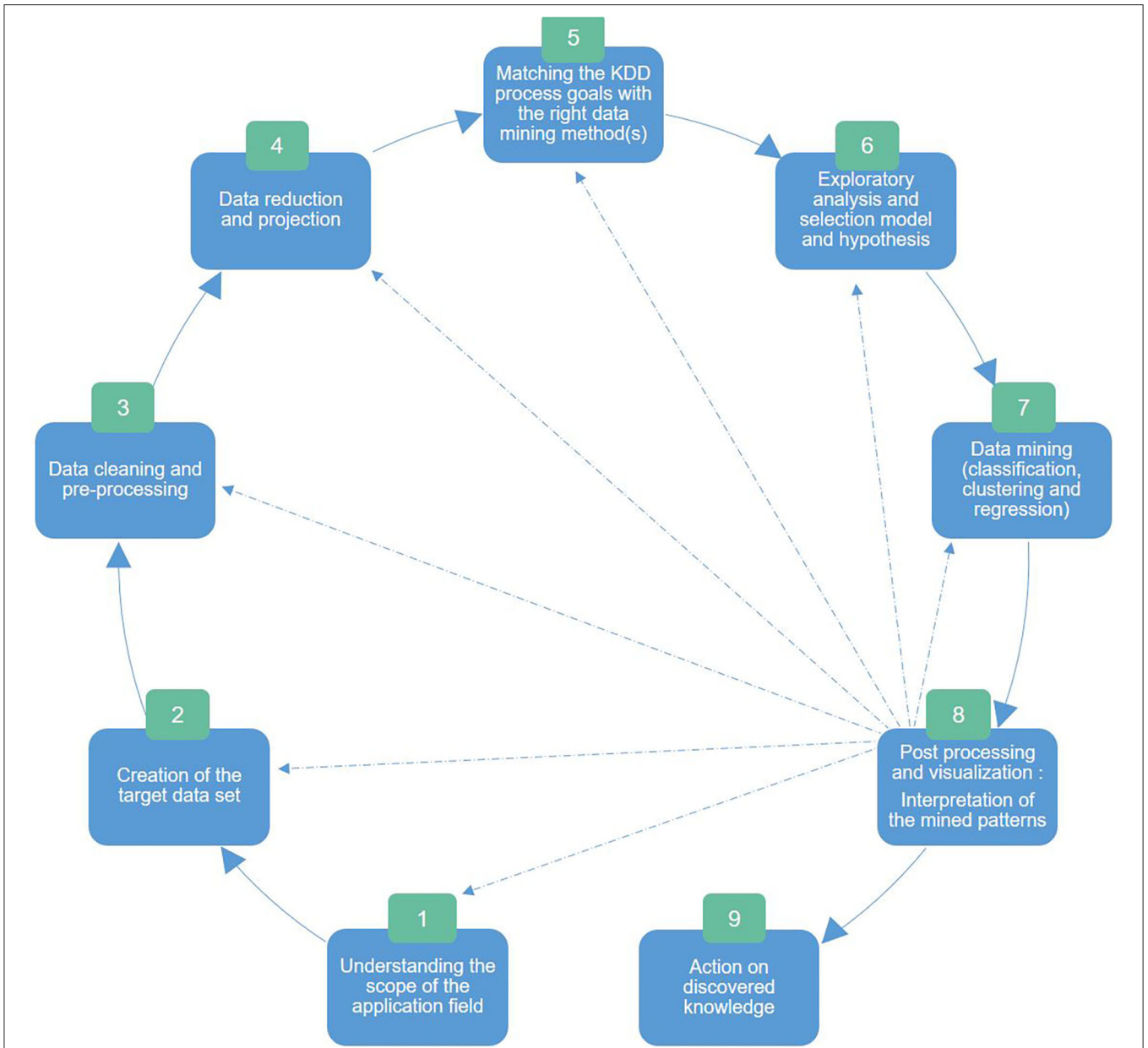


FIGURE 5 | The nine Knowledge Discovery of Data stages adapted from Fayyad (151): (2) understanding the scope of the application field, (3) creation of the target dataset, (4) data cleaning and preprocessing, (5) data reduction and projection: dimensionality reduction and extraction of invariant representations and relevant characteristics, (6) matching the KDD goals with the right data mining method(s), (7) exploratory analysis and selection model and hypothesis: selection of the data mining algorithm and method that will be used for the pattern search, (8) data mining: searching for interesting patterns in a particular form of representation (e.g., tree classification, regression, and clustering), (9) data postprocessing and visualization: interpretation of the patterns found with possible return to any step from 1 to 7 for a new cycle, and (10) action on discovered knowledge.

generally the case in a fundamental research team) or by distinct groups (it is the case for R&D groups in which technical teams focus on system architecture, development, quality, and testing, while operational teams handle experimental requirements and process definition). These concepts are classically and poorly applied to data (in fundamental and R&D teams), thus slowing down the improvement of accuracy, access, sharing, and reuse of data (152).

Data strategy applied to research aims at using, sharing, and moving data resources efficiently (adapted from 147) in order to manage projects easily, facilitate scientific collaboration, and accelerate decision-making regarding new project ideas.

Data strategy contains five core components that work together to comprehensively support an optimal data management:

- *Identification*: to set common data definition shared with the team, collaborators, and more broadly with the scientific community. In the mental health context for instance, it is crucial to define all biomarkers (whether they are genetic, molecular, anatomical, or environmental) to encompass the complexity of psychiatric disorders.
- *Storage*: to maintain data in information technology (IT) systems that allow easy access, sharing, and data processing.
- *Provision*: to anticipate and to prepare data in order to directly share or reuse it with adequate documentation explaining rules and definition.
- *Processing*: to aggregate data from different IT systems and obtain a centralized 360° data vision. In a mental health project, it could be pertinent to aggregate clinical, biological, and imaging data for instance.
- *Governance*: see dedicated chapter below (*Data governance*).

Data Governance

According to the Data Governance Institute, data governance is "*a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods*." Data governance is generally informal in fundamental research labs due to reduced (<15 people) and homogeneous (with the same background) teams in which processes, information, and tools are shared by everyone. This informal approach is less applicable with team expansion, different profile recruitment, scientific collaboration, or any operation that implies cross-functional activities. Initially designed for private industries, formal data governance approaches allow to frame cross-functional activities with a set of objectives adapted from the Data Governance Institute:

- to optimize decision-making: with a deeper knowledge of data assets and related documentation. This is helpful for instance when a choice has to be made between several scientific projects or strategies and the formal data governance approach estimates the ratio between investment and expected scientific value;
- to reduce operational friction: with defined and transparent roles and accountabilities regarding data and data use;
- to protect the needs of teams within a scientific collaboration framework;
- to train teams and collaborators to build common standards for approaching data issues;
- to reduce costs and increase effectiveness through effort coordination;
- to ensure transparency of processes;
- to accelerate and facilitate scientific collaboration;
- to allow scientific audit; and
- to respect compliance with the required documentation.

Data governance should not be applied as a theoretical concept but should rather be considered for its potential added value when it comes to pain points and the definition of use cases (e.g., to ensure data quality of a mental health digital project focused on schizophrenia). Good practices could anticipate value creation and changes triggered by the data governance framework (harmonious collaborations and their impact on data-related decisions for instance).

Regarding enterprise systems, research teams and/or scientific collaboration will require only a restricted number of rules and thus do not need a large organization assigned to data governance but more likely some clear and identified accountabilities and documentation for all scientific members (for further details, see **Appendix 4**).

In summary, data strategy and governance give a starting framework to structure the data management policy and strategy of a research team. Depending on the size of the team, the issues at stake, and the collaborations, these steps can have a real added value. As research teams do not rely on nor need large organizations for their data governance, it is also important to include other operational steps in a digital health data-centered project.

Operational Approach: From Preparation and Mining to Visualization

Data Preprocessing: Cleaning and Making Data Available

The preprocessing step consists in preparing the dataset to be mined (see **Figure 4**). This implies the following: (2) *data cleaning*, which consists in removing noise, corrupted data, and inaccurate records (3, 153, 154) *data editing* to control data quality by reviewing and adjusting it (155) and to anonymize data when needed with respect to data privacy standards (4, 156, 157) *data curation* to manage data maintainability over time for reuse and preservation (158); and 4) *data wrangling* or the process of mapping data from one type to another to fit the selected mining technique (e.g., from natural language to numerical vectors) (159). It is an important step in the KDD process (160) since the quality of the analysis of a data mining algorithm relies on the data available for the analysis. This step is inevitable as each dataset must be preprocessed before being mined. Alternatives (161) to preprocessing data exist but depend on the objective and the nature of available data, which makes it overwhelming to unexperienced users (162). It is thus essential to fix an explicit objective (i.e., a question to answer or a hypothesis to study) before preprocessing to choose the appropriate techniques.

Data Mining: From Biostatistics to Machine Learning

Biostatistics

Unlike ML, biostatistics are not used to establish predictions; hence, they do not require a large amount of data. Biostatistics study inferences between different populations by establishing a quantitative measure of confidence on a given sample of the population (163).

The frontiers between statistics and ML can be blurry as data analyses are often common to both [it is the case for the bootstrap method used for statistical inference and for the random forest (RF) algorithm]. It is thus important to differentiate statistics (that require us to choose a model incorporating our knowledge

of the system) from ML (that requires us to choose a predictive algorithm by relying on its empirical capabilities) (163).

Data Mining

Data mining is characterized by the willingness to find any possible means in order to be able to answer the research question. It can thus be defined as the process of analyzing large amounts of data to uncover patterns, associations, anomalies, commonalities, and statically significant structures in data (164). The two main goals of *data mining* are thus *prediction* of future behavior according to discovered patterns and *description* or the presentation in human-understandable shapes of the patterns found. To do so, data mining focuses on the analysis and extraction of features (extractable measurements or attributes) and patterns (arrangements or ordering with an underlying structure). Subfields of data mining include pattern recognition domain (or the characterization of patterns) (165) and *pattern detection* and *matching* (mining data to characterize patterns).

Data mining also includes subsets of popular algorithms:

- *Classification* consists in learning a function that classifies data into one or more predefined classes. For example, to predict generalized anxiety disorder among women, it is possible to either use RF to implement featured selection of the data mining classifier on the mental health data (166), or to use decision tree-based classification (167) or Shapley value algorithm (168).
- *Regression* consists in learning a function that matches data with a real predictor variable. The purpose of these algorithms is to analyze the relationship of variables with respect to the others, one by one, and to make predictions according to these relationships. It can be a statistical method or a ML algorithm. For example, Yengil et al. (169) used regression algorithms to study depression and anxiety in patients with beta thalassemia major and to further evaluate the impact of the disorder on quality of life.
- Another type of algorithm is *clustering* that consists in detecting a finite set of categories to describe data. Categories can be mutually exclusive and exhaustive or consist of a richer representation such as hierarchical or overlapping categories. The k-mean algorithm for instance can describe a population of patients as a finite set of clusters, each one grouping individuals sharing same features (e.g., children vs. adults).
- *Summarization* methods are used to find a compact description for a subset of data.
- *Dependency modeling* consists in finding a model that describes significant dependencies between variables. This can be done at the structural level (specifying dependent variables) or the quantitative level (specifying the strength of a dependency using numerical scales).

All these methods aim at extracting features or patterns following the search method as previously discussed in *Defining the goal and the approach*.

Data Postprocessing: Visualization and Evaluation

To efficiently communicate scientific information, data visualization (or graphic representation) should be specifically designed for the targeted audience. This can involve exploratory and/or explanatory objectives (170):

- *Pure exploratory*: addressed to teammates and collaborators to highlight main results in order to make data memorable and to identify the next strategic steps of the project.
- *Explanatory/exploratory mixed*: addressed to the scientific community, to share information and provide reliable (accessible and intelligible) data that can be analyzed and challenged by others. It can also support the scientific story telling in a grant application.
- *Pure explanatory*: addressed to patients, to quickly and efficiently explain scientific information with an appropriate and tailored content.

As seen in **Figures 4**, **5**, evaluating and interpreting mined patterns or extracted data through visualization can possibly induce returning to any previous step from preprocessing to data mining until discovered knowledge answers the fixed goal.

An Example of Our Method Applied to Mental Health

There is a growing number of mobile apps dedicated to mental health. Among them, "Moodfit" shapes up the mood, "Mood mission" teaches coping skills, "Talkspace" provides a virtual space for therapy, "Sanvello" acts as a stress relief, "Headspace" opens a virtual door to meditation, and "Shine" answers the specific mental health needs of BIPOC communities. However, there is no single guide for the development of evidence-based MHapps (171). An analysis of all apps dedicated to depression on the major marketplaces (Apple App and Google Play stores) shortlisted 293 apps that self-advertised as research-based (172). Among these apps, only 3.41% had published research that supports their claims of effectiveness, among which 20.48% were affiliated with an academic institution or medical facility. This analysis strongly indicates the need for mental health applications to be more rigorous (172), i.e., by following a strict method.

We have thus applied our end-to-end methodology to build a mobile application called i-decide (www.i-decide.fr) that facilitates decision processes under uncertainty. The application aims at complementing existing neuropsychological testing that take places punctually in a controlled setting by collecting longitudinal data on a daily basis. The data collected concern decision processes and all cognitive and emotional functions that impact decision-making (Boulos et al., in revision). All data are used to feed an algorithm that learns optimal choices (that reduce long hesitations and associated anxiety as well as the percentage of regret postdecision) under uncertain conditions. We tested the application on a population of 200 adult users with no diagnosed mental illness. Results revealed time slots during which decision-making was optimal as well as clusters of decision profiles according to stress, motivation, daily goals, support system, and the ratio of minor vs. major decisions (Boulos et al., in revision). More information can be found on the mobile application's website www.i-decide.fr.

DISCUSSION

Summary

AI algorithms together with advances in data storage have recently made it possible to better characterize, predict, prevent, and eventually cure a range of psychiatric illnesses. Amid the rapidly growing number of biological devices and the exponential accumulation of data in the mental health sector, the upcoming years are facing a need to homogenize research and development processes in academia as well as in the private sector and to centralize data into federalizing platforms. In this work, we describe an end-to-end methodology that optimizes and homogenizes digital biophysiological and behavioral monitoring with the ultimate ambition to bridge the gap between fundamental and applied research.

Methodology and Recommendations

The first step described project conception and planning stages. We proposed approaches to evaluate the feasibility of a digital mental health project, to define its goal, and to design the research approach accordingly. We clarified digital mental health research conceptions and misconceptions and described the difficulties of combining academic literature and market research. We further underlined the importance of collaborations in the interdisciplinary and intersectoral field to better understand what digital mental health is. We finally focused on the concrete planning of such methodology, that is, how to inject agility every step of the way to create ultimately platforms that reconcile different stakeholders to provide the best assistance possible to patients with mental health issues.

The second step zoomed in on the specificities of project design in mental health. We explained the importance of digital health interventions, the necessity to have clear goals, and the importance of human factors in defining them (introducing the user-centered design). We finally described cognitive biases and their impact on both physicians and data scientists in digital mental health.

The third, last, and more technical step described the stages from data collection to data analysis and visualization. We differentiated the notions of data (raw element), information (transformed data), and knowledge (transformed data with semantic contextual value) to then focus on the key steps of data in a digital mental health research. We provided recommendations for data management, strategy, and governance depending on the size and type of research structure and further elaborated a KDD-based operational approach that can be especially useful for small research teams that wish to work from collection to processing.

Issues at Stake: Ethics and Biases

Exploring the literature around digital mental health interventions leads us to question existing practices, that is, both their strengths and their issues. There are so many questions the scientific community and other stakeholders should consider when developing digital mental health solutions, and these include ethics and biases.

For a trial to be ethical, the assumption of equipoise (i.e., equilibrium) should be included in the design. While general designing and conducting RCT principles (173) are applicable to DHIs, specific DHI features deserve consideration when a trial is expected to provide evidence for rational decision-making: (2) the trial context, (3) the trade-off between external validity (the extent to which the results apply to a definable group of patients in a particular setting) and internal validity (how the design and conduct of the trial minimizes potential for bias) (174, 175) (e.g., of poor trade-off: recruiting highly motivated participants because of missing follow-up data) (174), (4) the specification of the intervention and delivery platform, (5) the choice of the comparator, and (6) establishing methods for separate data collection from the DHI itself.

Detailed specification of DHI is important, because it is required for the replication of trial results, the comparison between DHIs, and synthesizing data across trials in systematic reviews and meta-analyses (176). The relevant data to collect would then focus on usage, adherence, demographic access parameters, and user preferences (6, 177), even if participants are biased because they have access to a myriad of other DHIs. Indeed, someone who has sought help for a problem, entered a trial, and been randomized to the comparator arm, only to find the intervention unhelpful, may well search online until they find a better resource (178).

Finally, a well-designed RCT, especially for its ethical part, highlights the need to create interdisciplinarity. Researchers in digital mental health could learn from the multicycled iterative approach adopted in the industry for optimized development. Researchers from an engineering or computer science background may be surprised by the reliance on RCTs, whereas those from a biomedical or behavioral sciences background may consider that there is too much emphasis on methods other than RCTs. By enhancing critical thinking, interdisciplinarity in a team also tends to reduce cognitive biases. Although we have dedicated an entire part of this paper to cognitive biases (*Identifying cognitive biases in digital health to improve health outcomes*), there are several important points yet to be discussed. This includes the impact of biases in the decision-making process in digital mental health, the repercussion of the biases of practitioners on the data, and the biases of algorithms. One important message is that there are numerous cognitive biases across multiple domains (such as perception, statistics, logic, causality, social relations…) and that these biases are generally unconscious and effortless, making them hardly detectable and even less so controllable (179). Another important point is how AI and ML acceptability by the community on a social level can in turn affect the cognitive biases of physicians, researchers, and patients on digital mental health. In **Appendix 5**, we discuss these different issues and propose recommendations to better control the impact of cognitive biases in digital mental health research with the ultimate ambition to improve diagnostic reasoning and health outcomes.

Technical Challenges

In addition to the ethical considerations, working with data comes with technical challenges, three of which we wish to

highlight: (2) interoperability that is defined as the property that facilitates rapid and unrestricted sharing and use of data or resources between disparate systems *via* networks (180), (3) the trade-off between anonymization (to respect data privacy standards) and anonymization willingness, and (4). ML interpretability and explainability issues in digital mental health and digital health in general.

The multiplicity of tools that needs to be functional all while operating easily with other tools rushed the need for a "plug-and-play" interoperability. This is particularly the case for the medical field and its daily clinical use of various medical devices (MRI, computed tomography, ultrasound…). Beyond the traditional interoperability between different healthcare infrastructures, the will of patients to consult and understand their own data is imposing a new infrastructure-to-individual interoperability (181). In the light of this context, we believe that interoperability should be considered by a research team for their data strategy, especially when the research involves collaborations that are wanted or already in place. Beyond optimizing the collaboration and facilitating patient contribution, this could avoid data manipulation mistakes, as well as security or confidentiality failure.

Beyond confidentiality, one of the most sensitive points is privacy. In the context of digital mental health, and given the fact that it is a relatively young field with little information regarding clinically relevant variables (157), the bigger the data volume, the easier it is to identify relevant variables. The need for large data volumes is, however, challenged by the difficulty to collect these data all while respecting the strict health ethics and laws. It is thus crucial to set up the right privacy strategy. We would like also to highlight the existence of other technical challenges such as anonymization of data and explainable AI that are growing research fields (for further details, see **Appendix 6**).

CONCLUSION

In conclusion, our interdisciplinary collaboration to provide an end-to-end methodology for digital mental health research using interpretable techniques and a human-centered design with a special attention to data management and respecting privacy is therefore (2) a moral subject because it is linked to the transparency of the algorithms and, by extension, to the deriving decisions; (3) an ethical subject because it requires taking into account all people involved, their cognitive biases, and their impact on trials, experiments, and algorithms; and (4) a lever of trust for the end user specially in the mental health field where personal privacy is a critical but essential part that has to be respected.

Beyond this work, we find through our review of the literature that the various approaches taken to address different facets of product conception and design from research to market are siloed. Advances are often made separately and little attention is given to interdisciplinary and intersectoral centralizing approaches like ours in an attempt to provide a complete end-to-end methodology. We cannot stress enough on the timely importance of collaborations in digital mental health to reduce the disciplinary and sectoral gap and create platforms that deliver solutions trusted both by scientists and end users.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct and intellectual contribution to the work and approved it for publication.

FUNDING

We express our gratitude to onepoint, especially Erwan Le Bronec, for the financial support to the R&D department, which permitted us to carry out this work. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication. The only contribution of the funder is to pay the preliminary publishing fees.

ACKNOWLEDGMENTS

First and foremost, we acknowledge all the contributing authors who participated equally in this work. We would also like to thank Coralie Vennin for her referral to the appropriate literature on the psychological impact of the health crisis due to COVID-19.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.574440/full#supplementary-material

REFERENCES

1. Klonoff DC, King F, Kerr D. New opportunities for digital health to thrive. *J Diabetes Sci Technol.* (2019) 13:159–63. doi: 10.1177/1932296818822215
2. May C, Mort M, Williams T, Mair F, Gask L. Health technology assessment in its local contexts: studies of telehealthcare. *Soc Sci Med.* (2003) 57:697–710. doi: 10.1016/S0277-9536(02)00419-7
3. Devlin AM, McGee-Lennon M, O'Donnell CA, Bouamrane, M.-M., Agbakoba R, et al. Delivering digital health and well-being at scale: lessons learned during the implementation of the dallas program in the United Kingdom. *J Am Med Informat Assoc.* (2016) 23:48–59. doi: 10.1093/jamia/ocv097
4. Pagliari C, Detmer D, Singleton P. Potential of electronic personal health records. *BMJ.* (2007) 335:330–3. doi: 10.1136/bmj.39279.482963.AD

5. Bailey SC, Belter LT, Pandit AU, Carpenter DM, Carlos E, Wolf MS. The availability, functionality, and quality of mobile applications supporting medication self-management. *J Am Med Informat Assoc.* (2014) 21:542–6. doi: 10.1136/amiajnl-2013-002232
6. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. *Am J Prev Med.* (2016) 51:843–51. doi: 10.1016/j.amepre.2016.06.008
7. Palanica A, Docktor MJ, Lieberman M, Fossat Y. The need for artificial intelligence in digital therapeutics. *Digital Biomarkers.* (2020) 4:21–5. doi: 10.1159/000506861
8. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. *Per Med.* (2018) 15:429–48. doi: 10.2217/pme-2018-0044
9. Reti SR, Feldman HJ, Ross SE, Safran C. Improving personal health records for patient-centered care. *J Am Med Informat Assoc.* (2010) 17:192–5. doi: 10.1136/jamia.2009.000927
10. Aboujaoude E, Gega L, Parish MB, Hilty DM. Editorial: digital interventions in mental health: current status and future directions. *Front Psychiatry.* (2020) 11:111. doi: 10.3389/fpsyt.2020.00111
11. Allen B, Seltzer SE, Langlotz CP, Dreyer KP, Summers RM, Petrick N, et al. A road map for translational research on artificial intelligence in medical imaging: from the 2018 national institutes of Health/RSNA/ACR/The academy workshop. *J Am College Radiol.* (2019) 16:1179–89. doi: 10.1016/j.jacr.2019.04.014
12. Miner AS, Shah N, Bullock KD, Arnow BA, Bailenson J, Hancock J. Key considerations for incorporating conversational AI in psychotherapy. *Front Psychiatry.* (2019) 10:746. doi: 10.3389/fpsyt.2019.00746
13. Wentzel J, van der Vaart R, Bohlmeijer ET, van Gemert-Pijnen JEWC. Mixing online and face-to-face therapy: how to benefit from blended care in mental health care. *JMIR Mental Health.* (2016) 3:e9. doi: 10.2196/mental.4534
14. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digital Med.* (2019) 2:88. doi: 10.1038/s41746-019-0166-1
15. Kamdar MR, Wu MJ. PRISM: a data-driven platform for monitoring mental health. *Pac Symp Biocomput.* (2016) 21:333–44. doi: 10.1142/9789814749411_0031
16. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. *JMIR MHealth UHealth.* (2019) 7:e10828. doi: 10.2196/10828
17. Khundaqji H, Hing W, Furness J, Climstein M. Smart shirts for monitoring physiological parameters: scoping review. *JMIR MHealth UHealth.* (2020) 8:e18092. doi: 10.2196/18092
18. Chadborn NH, Blair K, Creswick H, Hughes N, Dowthwaite L, Adenekan O, et al. Citizens' juries: when older adults deliberate on the benefits and risks of smart health and smart homes. *Healthcare.* (2019) 7:54. doi: 10.3390/healthcare7020054
19. Reading Turchioe M, Grossman LV, Baik D, Lee CS, Maurer MS, Goyal P, et al. Older adults can successfully monitor symptoms using an inclusively designed mobile application. *J Am Geriat Soc.* (2020) 68:1313–8. doi: 10.1111/jgs.16403
20. Kollins SH, DeLoss DJ, Cañadas E, Lutz J, Findling RL, Keefe RSE, et al. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial. *Lancet Digital Health.* (2020) 2:e168–78. doi: 10.1016/S2589-7500(20)30017-0
21. Ginsburg GS, Phillips KA. Precision medicine: from science to value. *Health Affairs.* (2018) 37:694–701. doi: 10.1377/hlthaff.2017.1624
22. Cerga-Pashoja A, Gaete J, Shishkova A, Jordanova V. Improving reading in adolescents and adults with high-functioning autism through an assistive technology tool: a cross-over multinational study. *Front Psychiatry.* (2019) 10:546. doi: 10.3389/fpsyt.2019.00546
23. Rice S, O'Bree B, Wilson M, McEnery C, Lim MH, Hamilton M, et al. Leveraging the social network for treatment of social anxiety: pilot study of a youth-specific digital intervention with a focus on engagement of young men. *Internet Interventions.* (2020) 20:100323. doi: 10.1016/j.invent.2020.100323
24. Azevedo R, Bennett N, Bilicki A, Hooper J, Markopoulou F, et al. The calming effect of a new wearable device during the anticipation of public speech. *Sci Rep.* (2017) 7:2285. doi: 10.1038/s41598-017-02274-2
25. Valmaggia LR, Latif L, Kempton MJ, Rus-Calafell M. Virtual reality in the psychological treatment for mental health problems: an systematic review of recent evidence. *Psychiatry Res.* (2016) 236:189–195. doi: 10.1016/j.psychres.2016.01.015
26. Riek LD. Chapter 8—robotics technology in mental health care. In: DD Luxton, editor. *Artificial Intelligence in Behavioral and Mental Health Care*. Notre Dame, IN: Department of Computer Science and Engineering, University of Notre Dame. (2016). p. 185–203. doi: 10.1016/B978-0-12-420248-1.00008-8
27. Abdullah S, Matthews M, Frank E, Doherty G, Gay G, Choudhury T. Automatic detection of social rhythms in bipolar disorder. *J Am Med Informat Assoc.* (2016) 23:538–43. doi: 10.1093/jamia/ocv200
28. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res.* (2015) 17:e175. doi: 10.2196/jmir.4273
29. Christensen H, Cuijpers P, Reynolds CF. Changing the direction of suicide prevention research: a necessity for true population impact. *JAMA Psychiatry.* (2016) 73:435–6. doi: 10.1001/jamapsychiatry.2016.0001
30. Lo B, Shi J, Hollenberg E, Abi-Jaoude A, Johnson A, Wiljer D. Surveying the role of analytics in evaluating digital mental health interventions for transition-aged youth: a scoping review. *JMIR Mental Health.* (2020) 7:e15942. doi: 10.2196/15942
31. Blumenthal D. Realizing the value (and Profitability) of digital health data. *Ann Intern Med.* (2017) 166:842–3. doi: 10.7326/M17-0511
32. Ping NPT, Shoesmith WD, James S, Hadi NMN, Yau EKB, Lin LJ. Ultra brief psychological interventions for covid-19 pandemic: introduction of a locally-adapted brief intervention for mental health and psychosocial support service. *Malaysian J Med Sci.* (2020) 27:51. doi: 10.21315/mjms2020.27.2.6
33. Yang Y, Li W, Zhang Q, Zhang L, Cheung T, Xiang YT. Mental health services for older adults in China during the COVID-19 outbreak. *Lancet Psychiatry.* (2020) 7:e19. doi: 10.1016/S2215-0366(20)30079-1
34. Shigemura J, Ursano RJ, Morganstein JC, Kurosawa M, Benedek DM. Public responses to the novel 2019 coronavirus (2019-nCoV) in Japan: Mental health consequences and target populations. *Psychiatry Clin Neurosci.* (2020) 74:281–2. doi: 10.1111/pcn.12988
35. Wang CJ, Car J, Zuckerman BS. The power of telehealth has been unleashed. *Pediatr Clin North Am.* (2020) 67:xvii–xviii. doi: 10.1016/j.pcl.2020.05.001
36. Ravens-Sieberer U, Kaman A, Erhart M, Devine J, Schlack R, Otto C. Impact of the COVID-19 pandemic on quality of life and mental health in children and adolescents in Germany. *Euro Child Adolesc Psychiatry.* (2021). doi: 10.1007/s00787-021-01726-5. [Epub ahead of print].
37. Huremović D. *Psychiatry of Pandemics.* Cham: Springer (2019). doi: 10.1007/978-3-030-15346-5
38. Eston RG, Rowlands AV. Stages in the development of a research project: putting the idea together. *British J Sports Med.* (2000) 34:59–64. doi: 10.1136/bjsm.34.1.59
39. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. *NPJ Digital Med.* (2019) 2:38. doi: 10.1038/s41746-019-0111-3
40. Marshall JM, Dunstan DA, Bartik W. The role of digital mental health resources to treat trauma symptoms in australia during COVID-19. *Psychol Trauma.* (2020) 12:S269–71. doi: 10.1037/tra0000627
41. Hype Cycle Methodology. Available online at: https://www.gartner.com/en/research/methodologies/gartner-hype-cycle (accessed August 26, 2021).
42. Trigwell K, Dunbar-Goddet H. *The Research Experience of Postgraduate Research Students at the University of Oxford*. Oxford: Institute for the Advancement of University Learning, University of Oxford (2005).
43. Lau N, O'Daffer A, Colt S, Yi-Frazier JP, Palermo TM, McCauley E, et al. Android and iPhone mobile apps for psychosocial wellness and stress management: systematic search in app stores and literature review. *JMIR MHealth UHealth.* (2020) 8:e17798. doi: 10.2196/17798
44. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. *J Med Internet Res.* (2017) 19:e232. doi: 10.2196/jmir.7126

45. Alshurafa N, Jain J, Alharbi R, Iakovlev G, Spring B, Pfammatter A. Is more always better?: Discovering incentivized mhealth intervention engagement related to health behavior trends. *Proc ACM Interact Mob Wearable Ubiquitous Technol.* (2018) 2:153. doi: 10.1145/3287031
46. Champion L, Economides M, Chandler C. The efficacy of a brief app-based mindfulness intervention on psychosocial outcomes in healthy adults: a pilot randomised controlled trial. *PLoS ONE.* (2018) 13:e0209482. doi: 10.1371/journal.pone.0209482
47. Wilson K, Bell C, Wilson L, Witteman H. Agile research to complement agile development: a proposal for an mHealth research lifecycle. *NPJ Digital Med.* (2018) 1:46. doi: 10.1038/s41746-018-0053-1
48. Lattie EG, Graham AK, Hadjistavropoulos HD, Dear BF, Titov N, Mohr DC. Guidance on defining the scope and development of text-based coaching protocols for digital mental health interventions. *Digital Health.* (2019) 5:2055207619896145. doi: 10.1177/2055207619896145
49. Meyer JHF, Shanahan MP, Laugksch RC. Students' conceptions of research. I: A qualitative and quantitative analysis. *Scand J Educ Res.* (2005) 49:225–44. doi: 10.1080/00313830500109535
50. Loncar-Turukalo T, Zdravevski E, Machado da Silva J, Chouvarda I, Trajkovik V. Literature on wearable technology for connected health: scoping review of research trends, advances, and barriers. *J Med Internet Res.* (2019) 21:e14017. doi: 10.2196/14017
51. Urquhart C, Currell R. Systematic reviews and meta-analysis of health IT. *Stud Health Technol Inform.* (2016) 222:262–74. Available online at: https://www.nobelprize.org/prizes/economic-sciences/2002/kahneman/lecture
52. Murray E, Burns J, May C, Finch T, O'Donnell C, Wallace P, et al. Why is it difficult to implement e-health initiatives? A qualitative study. *Implement Sci.* (2011) 6:6. doi: 10.1186/1748-5908-6-6
53. Liu C, Shao S, Liu C, Bennett GG, Prvu Bettger J, Yan LL. Academia-industry digital health collaborations: a cross-cultural analysis of barriers and facilitators. *Digital Health.* (2019) 5:2055207619878627. doi: 10.1177/2055207619878627
54. Lupton D. Digital health now and in the future: findings from a participatory design stakeholder workshop. *Digital Health.* (2017) 3:2055207617740018. doi: 10.4324/9781315648835
55. Joerin A, Rauws M, Fulmer R, Black V. Ethical artificial intelligence for digital health organizations. *Cureus.* (2020) 12:e7202. doi: 10.7759/cureus.7202
56. Jandoo T. WHO guidance for digital health: What it means for researchers. *Digit Health.* (2020) 6:2055207619898984. doi: 10.1177/2055207619898984
57. Murray E, Khadjesari Z, White IR, Kalaitzaki E, Godfrey C, McCambridge J, et al. Methodological challenges in online trials. *J Med Internet Res.* (2009) 11:e9. doi: 10.2196/jmir.1052
58. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med.* (2018) 178:1544–7. doi: 10.1001/jamainternmed.2018.3763
59. d'Alessandro B, O'Neil C, LaGatta T. Conscientious classification: a data scientist's guide to discrimination-aware classification. *Big Data.* (2017) 5:120–34. doi: 10.1089/big.2016.0048
60. Plante TB, Urrea B, MacFarlane ZT, Blumenthal RS, Miller ER, Appel LJ, et al. Validation of the instant blood pressure smartphone app. *JAMA Intern Med.* (2016) 176:700–2. doi: 10.1001/jamainternmed.2016.0157
61. Poli A, Kelfve S, Motel-Klingebiel A. A research tool for measuring non-participation of older people in research on digital health. *BMC Public Health.* (2019) 19:1487. doi: 10.1186/s12889-019-7830-x
62. Dockendorf MF, Murthy G, Bateman KP, Kothare PA, Anderson M, Xie I, et al. Leveraging digital health technologies and outpatient sampling in clinical drug development: a phase i exploratory study. *Clin Pharmacol Ther.* (2019) 105:168–76. doi: 10.1002/cpt.1142
63. Liang J, He X, Jia Y, Zhu W, Lei J. Chinese mobile health apps for hypertension management: a systematic evaluation of usefulness. *J Healthc Eng.* (2018) 2018:7328274. doi: 10.1155/2018/7328274
64. Germine L, Reinecke K, Chaytor NS. Digital neuropsychology: challenges and opportunities at the intersection of science and software. *Clin Neuropsychol.* (2019) 33:271–86. doi: 10.1080/13854046.2018.1535662
65. Nielsen J. *Usability Engineering (New edition)*. Morgan Kaufmann Publishers In. (1994).
66. Norman DA, Draper SW. *User Centered System Design: New Perspectives on Human-computer Interaction.* Boca Raton, FL: CRC Press (1986). doi: 10.1201/b15703
67. Ambler SW. Lessons in agility from Internet-based development. *IEEE Software.* (2002) 19:66–73. doi: 10.1109/52.991334
68. Morville P, Callender J. *Search Patterns (1st ed.).* O'Reilly (2010).
69. Benyon D. *Designing Interactive Systems: A Comprehensive Guide to HCI, UX and Interaction Design.* Pearson (2014). Available online at: https://www.pearson.com/uk/educators/higher-education-educators/program/Benyon-Designing-Interactive-Systems-A-comprehensive-guide-to-HCI-UX-and-interaction-design-3rd-Edition/PGM1047688.html (accessed August 26, 2021).
70. Faurholt-Jepsen M, Vinberg M, Frost M, Christensen EM, Bardram J, Kessing LV. Daily electronic monitoring of subjective and objective measures of illness activity in bipolar disorder using smartphones–the MONARCA II trial protocol: a randomized controlled single-blind parallel-group trial. *BMC Psychiatry.* (2014) 14:309. doi: 10.1186/s12888-014-0309-5
71. Freeman D. Studying and treating schizophrenia using virtual reality: a new paradigm. *Schizophr Bull.* (2008) 34:605–10. doi: 10.1093/schbul/sbn020
72. Granja C, Janssen W, Johansen MA. Factors determining the success and failure of ehealth interventions: systematic review of the literature. *J Med Internet Res.* (2018) 20:e10235. doi: 10.2196/10235
73. Heeks R. Health information systems: failure, success and improvisation. *Int J Med Inform.* (2006) 75:125–37. doi: 10.1016/j.ijmedinf.2005.07.024
74. Stott R, Wild J, Grey N, Liness S, Warnock-Parkes E, Commins S, et al. Internet-delivered cognitive therapy for social anxiety disorder: a development pilot series. *Behav Cogn Psychother.* (2013) 41:383–97. doi: 10.1017/S1352465813000404
75. Gorst SL, Armitage CJ, Brownsell S, Hawley MS. Home telehealth uptake and continued use among heart failure and chronic obstructive pulmonary disease patients: a systematic review. *Ann Behav Med.* (2014) 48:323–36. doi: 10.1007/s12160-014-9607-x
76. Sanders C, Rogers A, Bowen R, Bower P, Hirani S, Cartwright M, et al. Exploring barriers to participation and adoption of telehealth and telecare within the Whole System Demonstrator trial: a qualitative study. *BMC Health Serv Res.* (2012) 12:220. doi: 10.1186/1472-6963-12-220
77. Huckvale K, Wang CJ, Majeed A, Car J. Digital health at fifteen: More human (more needed). *BMC Med.* (2019) 17:62. doi: 10.1186/s12916-019-1302-0
78. Nutt AE. *'The Woebot will see you now'—The Rise of Chatbot Therapy.* Washington Post (2017). Available online at: https://www.washingtonpost.com/news/to-your-health/wp/2017/12/03/the-woebot-will-see-you-now-the-rise-of-chatbot-therapy/ (accessed August 26, 2021).
79. Fairburn CG, Patel V. The impact of digital technology on psychological treatments and their dissemination. *Behav Res Ther.* (2017) 88:19–5. doi: 10.1016/j.brat.2016.08.012
80. Berwick DM. What "patient-centered" should mean: confessions of an extremist. *Health Affairs.* (2009) 28:w555–65. doi: 10.1377/hlthaff.28.4.w555
81. Torous J. Mobile telephone apps first need data security and efficacy. *BJPsych Bulletin.* (2016) 40:106–7. doi: 10.1192/pb.40.2.106b
82. Byambasuren O, Sanders S, Beller E, Glasziou P. Prescribable mHealth apps identified from an overview of systematic reviews. *Npj Digital Med.* (2018) 1:1–12. doi: 10.1038/s41746-018-0021-9
83. Miyamoto S, Henderson S, Young HM, Ward D, Santillan V. Recruiting rural participants for a telehealth intervention on diabetes self-management. *J Rural Health.* (2013) 29:69–77. doi: 10.1111/j.1748-0361.2012.00443.x
84. Goel MS, Brown TL, Williams A, Cooper AJ, Hasnain-Wynia R, Baker DW. Patient reported barriers to enrolling in a patient portal. *J Am Med Informat Assoc.* (2011) 18(Suppl. 1):i8–12. doi: 10.1136/amiajnl-2011-000473
85. Lakerveld J, Ijzelenberg W, van Tulder MW, Hellemans IM, Rauwerda JA, van Rossum AC, et al. Motives for (not) participating in a lifestyle intervention trial. *BMC Med Res Methodol.* (2008) 8:17. doi: 10.1186/1471-2288-8-17
86. O'Connor S, Mair FS, McGee-Lennon M, Bouamrane M, O'Donnell K. Engaging in large-scale digital health technologies and services. What factors hinder recruitment? *Stud Health Technol Inform.* (2015) 210:306–10.
87. Dyrbye LN, Shanafelt TD, Sinsky CA, Cipriano PF, Bhatt J, Ommaya A, et al. *Burnout Among Health Care Professionals: A Call to Explore and Address this Underrecognized Threat to Safe, High-Quality Care.* NAM Perspectives.

Discussion Paper, National Academy of Medicine, Washington, DC (2017). doi: 10.31478/201707b
88. Greenhalgh T, Procter R, Wherton J, Sugarhood P, Hinder S, Rouncefield M. What is quality in assisted living technology? The ARCHIE framework for effective telehealth and telecare services. *BMC Med.* (2015) 13:91. doi: 10.1186/s12916-015-0279-6
89. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evid Based Ment Health.* (2018) 21:116–9. doi: 10.1136/eb-2018-102891
90. Opoku D, Stephani V, Quentin W. A realist review of mobile phone-based health interventions for non-communicable disease management in sub-Saharan Africa. *BMC Med.* (2017) 15:24. doi: 10.1186/s12916-017-0782-z
91. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. *BMC Med Inform Decis Mak.* (2016) 16:120. doi: 10.1186/s12911-016-0359-3
92. Choi NG, DiNitto DM. The digital divide among low-income homebound older adults: internet use patterns, ehealth literacy, and attitudes toward computer/internet use. *J Med Internet Res.* (2013) 15:e93. doi: 10.2196/jmir.2645
93. Selwyn N, Gorard S, Furlong J, Madden L. Older adults' use of information and communications technology in everyday life. *Ageing Soc.* (2003) 23:561–82. doi: 10.1017/S0144686X03001302
94. Cashen MS, Dykes P, Gerber B. eHealth technology and Internet resources: Barriers for vulnerable populations. *J Cardiovasc Nursing.* (2004) 19:209–14; quiz 215–216. doi: 10.1097/00005082-200405000-00010
95. Kontos E, Blake KD, Chou WYS, Prestin A. Predictors of eHealth usage: Insights on the digital divide from the Health Information National Trends Survey 2012. *J Med Internet Res.* (2014) 16:e172. doi: 10.2196/jmir.3117
96. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. *J Med Internet Res.* (2012) 14:e19. doi: 10.2196/jmir.1619
97. de Ruyter B. User centred design. In Aarts E, Marzano S, editors. In: *The New Everyday: Views on Ambient Intelligence*. 010 Publishers (2003).
98. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. *Am J Prev Med.* (2009) 36:452–7. doi: 10.1016/j.amepre.2009.02.002
99. Steen M. Human-centered design as a fragile encounter. *Design Issues.* (2012) 28:72–80. JSTOR. doi: 10.1162/DESI_a_00125
100. Croskerry P. From mindless to mindful practice—Cognitive bias and clinical decision making. *N Engl J Med.* (2013) 368:2445–8. doi: 10.1056/NEJMp1303712
101. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ Qual Safety.* (2013) 22(Suppl. 2):ii58–64. doi: 10.1136/bmjqs-2012-001712
102. Zwaan L, Thijs A, Wagner C, Timmermans DRM. Does inappropriate selectivity in information use relate to diagnostic errors and patient harm? The diagnosis of patients with dyspnea. *Soc Sci Med.* (2013) 91:32–8. doi: 10.1016/j.socscimed.2013.05.001
103. Kahneman D. 'Maps of bounded rationality: a perspective on intuitive judgment and choice'. *Nobel Prize Lecture*. (2002) 8, 351–401. Available online at: https://www.nobelprize.org/prizes/economic-sciences/2002/kahneman/lecture/
104. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science.* (1974) 185:1124–31. doi: 10.1126/science.185.4157.1124
105. Ely JW, Graber ML, Croskerry P. Checklists to reduce diagnostic errors. *Acad Med.* (2011) 86:307–13. doi: 10.1097/ACM.0b013e31820824cd
106. Mamede S, van Gog T, van den Berge K, van Saase, JLCM, Schmidt HG. Why do doctors make mistakes? A study of the role of salient distracting clinical features. *Acad Med.* (2014) 89:114–20. doi: 10.1097/ACM.0000000000000077
107. van den Berge K, Mamede S. Cognitive diagnostic error in internal medicine. *Eur J Intern Med.* (2013) 24:525–9. doi: 10.1016/j.ejim.2013.03.006
108. Appel S, Wadas T, Talley M, Williams A. Teaching diagnostic reasoning: transitioning from a live to a distance accessible online classroom in an Adult Acute Care Nurse Practitioner Program. *J Nursing Educ Practice*. (2014) 3. doi: 10.5430/jnep.v3n1 2p125
109. Ioannidis JP, Lau J. Evidence on interventions to reduce medical errors: an overview and recommendations for future research. *J Gen Intern Med.* (2001) 16:325–34. doi: 10.1046/j.1525-1497.2001.00714.x
110. OECD. *Health at a Glance 2013*. (2013). Available online at: https://www.oecd-ilibrary.org/content/publication/health_glance-2013-en (accessed August 26, 2021).
111. Lawson TN. Diagnostic reasoning and cognitive biases of nurse practitioners. *J Nurs Educ.* (2018) 57:203–8. doi: 10.3928/01484834-20180322-03
112. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. *BMC Med Inform Decision Making.* (2016) 16:138. doi: 10.1186/s,12911-016-0377-1
113. Brosinski CM. Implementing diagnostic reasoning to differentiate Todd's paralysis from acute ischemic stroke. *Adv Emerg Nurs J.* (2014) 36:78–86. doi: 10.1097/TME.0000000000000007
114. Elstein AS. Thinking about diagnostic thinking: a 30-year perspective. *Adv Health Sci Educ.* (2009) 14(Suppl).1:7–18. doi: 10.1007/s10459-009-9184-0
115. Ilgen JS, Eva KW, Regehr G. What's in a label? Is diagnosis the start or the end of clinical reasoning? *J General Internal Med.* (2016) 31:435–7. doi: 10.1007/s11606-016-3592-7
116. Monteiro SM, Norman G. Diagnostic reasoning: where we've been, where we're going. *Teach Learn Med*. (2013) 25(Suppl. 1):S26–32. doi: 10.1080/10401334.2013.842911
117. Pirret AM, Neville SJ, La Grow SJ. Nurse practitioners versus doctors diagnostic reasoning in a complex case presentation to an acute tertiary hospital: A comparative study. *Int J Nurs Stud.* (2015) 52:716–26. doi: 10.1016/j.ijnurstu.2014.08.009
118. Sherbino J, Kulasegaram K, Howey E, Norman G. Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: a controlled trial. *CJEM.* (2014) 16:34–40. doi: 10.2310/8000.2013.130860
119. Thammasitboon S, Cutrer WB. Diagnostic decision-making and strategies to improve diagnosis. *Curr Probl Pediatr Adolesc Health Care.* (2013) 43:232–41. doi: 10.1016/j.cppeds.2013.07.003
120. Thompson C. Clinical experience as evidence in evidence-based practice. *J Adv Nurs.* (2003) 43:230–7. doi: 10.1046/j.1365-2648.2003.02705.x
121. Thammasitboon S, Thammasitboon S, Singhal G. Diagnosing diagnostic error. *Curr Probl Pediatr Adolesc Health Care.* (2013) 43:227–31. doi: 10.1016/j.cppeds.2013.07.002
122. Baldwin RL, Green JW, Shaw JL, Simpson DD, Bird TM, Cleves MA, et al. Physician risk attitudes and hospitalization of infants with bronchiolitis. *Acad Emergency Med.* (2005) 12:142–6. doi: 10.1197/j.aem.2004.10.002
123. Yee LM, Liu LY, Grobman WA. The relationship between obstetricians' cognitive and affective traits and their patients' delivery outcomes. *Am J Obstetr Gynecol.*(2014) 211:692.e1–6. doi: 10.1016/j.ajog.2014.06.003
124. Källberg, A.-S., Göransson KE, Östergren J, Florin J, Ehrenberg A. Medical errors and complaints in emergency department care in Sweden as reported by care providers, healthcare staff, and patients—a national review. *Europ J Emergency Med.* (2013) 20:33–8. doi: 10.1097/MEJ.0b013e32834fe917
125. Studdert DM, Mello MM, Gawande AA, Gandhi TK, Kachalia A, Yoon C, et al. Claims, errors, and compensation payments in medical malpractice litigation. *New Engl J Med.* (2006) 354:2024–33. doi: 10.1056/NEJMsa054479
126. Zwaan L, Thijs A, Wagner C, van der Wal G, Timmermans DRM. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. *Acad Med.* (2012) 87:149–56. doi: 10.1097/ACM.0b013e31823f71e6
127. Lazarus RS, Folkman S. Stress, *Appraisal, and Coping*. San Francisco, CA: Department of Medicine, School of Medicine University of California (1984).
128. Miller SM. Monitoring and blunting: validation of a questionnaire to assess styles of information seeking under threat. *J Pers Soc Psychol.* (1987) 52:345–53. doi: 10.1037/0022-3514.52.2.345
129. Deldin PJ, Keller J, Gergen JA, Miller GA. Cognitive bias and emotion in neuropsychological models of depression. *Cogn Emot.* (2001) 15:787–802. doi: 10.1080/02699930143000248
130. Andel C, Davidow SL, Hollander M, Moreno DA. The economics of health care quality and medical errors. *J Health Care Finance.* (2012) 39:39–50.
131. Blackwell SE, Browning M, Mathews A, Pictet A, Welch J, Davies J, et al. Positive imagery-based cognitive bias modification as a web-based treatment

tool for depressed adults: a randomized controlled trial. *Clin Psychol Sci.* (2015) 3:91–111. doi: 10.1177/2167702614560746
132. Freeman D, Bradley J, Antley A, Bourke E, DeWeever N, Evans N, et al. Virtual reality in the treatment of persecutory delusions: randomised controlled experimental study testing how to reduce delusional conviction. *Br J Psychiatry.* (2016) 209:62–7. doi: 10.1192/bjp.bp.115.176438
133. van Spijker BAJ, van Straten A, Kerkhof AJFM. Effectiveness of online self-help for suicidal thoughts: results of a randomised controlled trial. *PLoS ONE.* (2014) 9:e90118. doi: 10.1371/journal.pone.0090118
134. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med.* (2003) 78:775–80. doi: 10.1097/00001888-200308000-00003
135. Michaels AD, Spinler SA, Leeper B, Ohman EM, Alexander KP, Newby LK, et al. Medication errors in acute cardiovascular and stroke patients: a scientific statement from the American Heart Association. *Circulation.* (2010) 121:1664–82. doi: 10.1161/CIR.0b013e3181d4b43e
136. Agarwal R. *Five Cognitive Biases in Data Science (and How to Avoid Them)*. Built In (2020). Available online at: https://builtin.com/data-science/cognitive-biases-data-science (accessed August 26, 2021).
137. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science.* (2017) 356:183–6. doi: 10.1126/science.aal4230
138. Eubanks V. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. St Martin's Press (2018).
139. Noble SU. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: New York University Press (2018). doi: 10.2307/j.ctt1pwt9w5
140. O'Neil C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown (2016).
141. Çalikli G, Bener AB. Influence of confirmation biases of developers on software quality: an empirical study. *Software Qual J.* (2013) 21:377–416. doi: 10.1007/s11219-012-9180-0
142. Nelson GS. Bias in artificial intelligence. *N C Med J.* (2019) 80:220–2. doi: 10.18043/ncm.80.4.220
143. Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal ME, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplin Rev.* (2020) 10:e1356. doi: 10.1002/widm.1356
144. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA.* (2017) 318:517–8. doi: 10.1001/jama.2017.7797
145. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med.* (2018) 378:981–3. doi: 10.1056/NEJMp1714229
146. Arpey NC, Gaglioti AH, Rosenbaum ME. How socioeconomic status affects patient perceptions of health care: a qualitative study. *J Prim Care Community Health.* (2017) 8:169–75. doi: 10.1177/2150131917697439
147. Ng JH, Ye F, Ward LM, Haffer SCC, Scholle SH. Data on race, ethnicity, and language largely incomplete for managed care plan members. *Health Affairs.* (2017) 36:548–52. doi: 10.1377/hlthaff.2016.1044
148. Rauscher GH, Khan JA, Berbaum ML, Conant EF. Potentially missed detection with screening mammography: does the quality of radiologist's interpretation vary by patient socioeconomic advantage/disadvantage? *Ann Epidemiol.* (2013) 23:210–4. doi: 10.1016/j.annepidem.2013.01.006
149. Li S, Fonarow GC, Mukamal KJ, Liang L, Schulte PJ, Smith EE, et al. Sex and race/ethnicity-related disparities in care and outcomes after hospitalization for coronary artery disease among older adults. *Circul Cardiovasc Qual Outcomes.* (2016) *9(2 Suppl.* 1):S36–44. doi: 10.1161/CIRCOUTCOMES.115.002621
150. Grazzini J, Pantisano F. *Guidelines for Scientific Evidence Provision for Policy Support Based on Big Data and Open Technologies* (2015).
151. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Magazine.* (1996) 17:37.
152. SAS Institute Inc. *SAS Institute Whitepaper. "The 5 Essential Components of a Data Strategy.* (2016)." Available online at: https://www.sas.com/content/dam/SAS/en_au/doc/whitepaper1/five-essential-components-data-strategy.pdf (accessed August 26, 2021).
153. Rahm E, Do HH. Data cleaning: problems and current approaches. *IEEE Data Eng Bull.* (2000) 23:3–13.
154. Müller H, Freytag JC. Problems, methods, and challenges in comprehensive data cleansing. *Humboldt-Univ. zu Berlin.* (2005).
155. Famili A, Shen WM, Weber R, Simoudis E. Data preprocessing and intelligent data analysis. *Intelligent Data Analy.* (1997) 1:3–23. doi: 10.3233/IDA-1997-1102
156. Mittelstadt B, Floridi L. *The Ethics of Biomedical Big Data.* Vol. 29. Springer(2016). doi: 10.1007/978-3-319-33525-4
157. Aledavood T, Triana Hoyos AM, Alakörkk,ö T, Kaski K, Saramäki J, Isometsä E, et al. Data collection for mental health studies through digital platforms: requirements and design of a prototype. *JMIR Res Protoc.* (2017) 6:e110. doi: 10.2196/resprot.6919
158. Gopal K. *Data Curation: The Processing of Data.* IndianJournal.com. (2016).
159. Furche T, Gottlob G, Libkin L, Orsi G, Paton NW. Data wrangling for big data: challenges and opportunities. *EDBT.* (2016) 16:473–8.
160. Pyle D. *Data Preparation for Data Mining.* Morgan Kaufmann (1999).
161. Alasadi SA, Bhaya WS. Review of data preprocessing techniques in data mining. *J Eng Appl Sci.* (2017) 12:4102–7. doi: 10.36478/jeasci.2017.4102.4107
162. Bilalli B, Abelló A, Aluja-Banet T, Wrembel R. Intelligent assistance for data pre-processing. *Computer Standards Interfaces.* (2018) 57:101–9. doi: 10.1016/j.csi.2017.05.004
163. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods.* (2018) 15:233–4. doi: 10.1038/nmeth.4642
164. Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare. *Int J Bio-Sci Bio-Technol.* (2013) 5:241–66. doi: 10.14257/ijbsbt.2013.5.5.25
165. Bishop C. *Pattern Recognition and Machine Learning.* Springer-Verlag (2006). Available online at: https://www.springer.com/gp/book/9780387310732 (accessed August 26, 2021).
166. Husain W, Xin LK, Rashid NA, Jothi N. Predicting Generalized Anxiety Disorder among women using random forest approach. *2016 3rd International Conference on Computer and Information Sciences (ICCOINS).* (2016). p. 37–42. doi: 10.1109/ICCOINS.2016.7783185
167. Jothi N, Husain W, Rashid NA, Xin LK. Predicting generalised anxiety disorder among women using decision tree-based classification. *Int J Business Informat Syst.* (2018) 29:75–91. doi: 10.1504/IJBIS.2018.093998
168. Jothi N, Husain W, Rashid NA. Predicting generalized anxiety disorder among women using Shapley value. *J Infect Public Health.* (2020) 14:103–8. doi: 10.1016/j.jiph.2020.02.042
169. Yengil E, Acipayam C, Kokacya MH, Kurhan F, Oktay G, Ozer C. Anxiety, depression and quality of life in patients with beta thalassemia major and their caregivers. *Int J Clin Exp Med.* (2014) 7:2165–72.
170. Goodman AA, Borkin MA, Robitaille TP. New thinking on, and with, data visualization. *ArXiv.* (2018).
171. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. *JMIR Mental Health.* (2016) 3:e7. doi: 10.2196/mental.4984
172. Marshall JM, Dunstan DA, Bartik W. The digital psychiatrist: in search of evidence-based apps for anxiety and depression. *Front Psychiatry.* (2019) 10:831. doi: 10.3389/fpsyt.2019.00831
173. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: The new Medical Research Council guidance. *BMJ.* (2008) 337:a1655. doi: 10.1136/bmj.a1655
174. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials.* (2015) 16:495. doi: 10.1186/s13063-015-1023-4
175. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" *Lancet.* (2005) 365:82–93. doi: 10.1016/S0140-6736(04)17670-8
176. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ.* (2014) 348:g1687. doi: 10.1136/bmj.g1687
177. Muris P, Roodenrijs D, Kelgtermans L, Sliwinski S, Berlage U, Baillieux H, et al. No Medication for My Child! A naturalistic study on the treatment preferences for and effects of cogmed working memory training versus psychostimulant medication in clinically referred youth with ADHD. *Child Psychiatry Human Dev.* (2018) 49:974–92. doi: 10.1007/s10578-018-0812-x

178. Khadjesari Z, Stevenson F, Godfrey C, Murray E. Negotiating the 'grey area between normal social drinking and being a smelly tramp': a qualitative study of people searching for help online to reduce their drinking. *Health Expectat.* (2015) 18:2011–20. doi: 10.1111/hex.12351
179. Marewski JN, Gigerenzer G. Heuristic decision making in medicine. *Dialogues Clin Neurosci.* (2012) 14:77–89. doi: 10.31887/DCNS.2012.14.1/jmarewski
180. Veltman KH. Syntactic and semantic interoperability: new approaches to knowledge and the semantic web. *New Rev Informat Networking.* (2001) 7:159–83. doi: 10.1080/13614570109516975
181. Gordon WJ, Catalini C. Blockchain technology for healthcare: facilitating the transition to patient-driven interoperability. *Comput Struct Biotechnol J.* (2018) 16:224–30. doi: 10.1016/j.csbj.2018.06.003
182. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. (2018). p. 80–9. doi: 10.1109/DSAA.2018.00018
183. Angwin J, Larson J, Mattu S, Kirchner L. *Machine Bias. ProPublica*. (2016). Available online at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
184. Kaadoud IC, Rougier NP, Alexandre F. Knowledge extraction from the learning of sequences in a long short term memory (LSTM) architecture. *ArXiv*. (2019).
185. Kim WC, Mauborgne R. Blue ocean leadership. *Harvard Business Rev.* (2014) 92:60–8, 70, 72 passim.
186. Andersson G. *The Internet and CBT: A Clinical Guide*. CRC Press (2014). doi: 10.1201/b13645
187. Lipton ZC. The mythos of model interpretability. *arXiv*. (2017).
188. Guise, J.-M., Chang C, Viswanathan M, Glick S, Treadwell J, et al. *Systematic Reviews of Complex Multicomponent Health Care Interventions*. Agency for Healthcare Research and Quality (US) (2014). Available online at: http://www.ncbi.nlm.nih.gov/books/NBK194846/
189. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Informat Fusion.* (2020) 58:82–115. doi: 10.1016/j.inffus.2019.12.012
190. Kamishima T, Akaho S, Asoh H, Sakuma J. Fairness-aware classifier with prejudice remover regularizer. In Flach PA. De Bie T, Cristianini N, editors. *Machine Learning and Knowledge Discovery in Databases.* Springer(2012). p. 35–50. doi: 10.1007/978-3-642-33486-3_3
191. Bourne KC. Chapter 19—your organization. In: Bourne KC, editor. *Application Administrators Handbook.* Morgan Kaufmann (2014). p. 329–43. doi: 10.1016/B978-0-12-398545-3.00019-4
192. Rodiya K, Gill P. A review on anonymization techniques for privacy preserving data publishing. *Int J Eng Res Technol.* (2015) 4:228–231. doi: 10.15623/ijret.2015.0411039
193. Kido T, Takadama K. *The Challenges for Interpretable AI for Well-being -Understanding Cognitive Bias and Social Embeddedness. CEUR-WS* 2448 (2019).
194. Guidotti R, Monreale A, Ruggieri S, Turini F, Pedreschi D, Giannotti F. A survey of methods for explaining black box models. *ArXiv.* (2018). doi: 10.1145/3236009
195. Wathelet M, Duhem S, Vaiva G, Baubet T, Habran E, Veerapa E, et al. Factors associated with mental health disorders among university students in France confined during the COVID-19 pandemic. *JAMA Network Open.* (2020) 3:e2025591. doi: 10.1001/jamanetworkopen.2020.25591
196. Goals and Principles for Data Governance. Available online at: http://www.datagovernance.com/adg_data_governance_goals/
197. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell.* (2013) 35:1798–828. doi: 10.1109/TPAMI.2013.50
198. Alasadi SA, Bhaya WS. Review of data preprocessing techniques in data mining. *J Eng Appl Sci.* (2017) 12:4102–7.
199. Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* (2015). p. 427–36. doi: 10.1109/CVPR.2015.7298640
200. Li W, Yang Y, Liu ZH, Zhao YJ, Zhang Q, Zhang L, et al. Progression of mental health services during the COVID-19 outbreak in China. *Int J Biol Sci.* (2020) 16:1732. doi: 10.7150/ijbs.45120
201. Sebastian-Coleman L. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Newnes.* (2012). doi: 10.1016/B978-0-12-397033-6.00020-1
202. Taylor J, Pagliari C. Comprehensive scoping review of health research using social media data. *BMJ Open.* (2018) 8:e022931. doi: 10.1136/bmjopen-2018-022931
203. Selvi U, Pushpa S. A review of big data and anonymization algorithms. *Int J Appl Eng Res.* (2015) 10.
204. Huckvale K, Prieto JT, Tilney M, Benghozi PJ, Car J. Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment. *BMC Med.* (2015) 13:214. doi: 10.1186/s12916-015-0444-y
205. Eysenbach G, Group, C.-E. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. *J Med Internet Res.* (2011) 13:e126. doi: 10.2196/jmir.1923
206. Ayache S, Eyraud R, Goudian N. Explaining black boxes on sequential data using weighted automata. *ArXiv*. (2018).
207. Ma H, Miller C. Trapped in a double bind: Chinese overseas student anxiety during the COVID-19 pandemic. *Health Commun.*(2020) 1–8. doi: 10.1080/10410236.2020.1775439
208. Deshazo JP, Lavallie DL, Wolf FM. Publication trends in the medical informatics literature: 20 years of "Medical Informatics" in MeSH. *BMC Med Inform Decis Mak.* (2009) 9:7. doi: 10.1186/1472-6947-9-7
209. Bucci S, Schwannauer M, Berry N. The digital revolution and its impact on mental health care. *Psychol Psychotherap.* (2019) 92:277–97. doi: 10.1111/papt.12222

Conflict of Interest: AD, AM, and ICK were employed by the company onepoint when the research was conducted.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychiatry

METHODS
published: 17 November 2021
doi: 10.3389/fpsyt.2021.578298

The Cambridge Cognitive and Psychiatric Assessment Kit (CamCOPS): A Secure Open-Source Client–Server System for Mobile Research and Clinical Data Capture

Rudolf N. Cardinal[1,2*†] *and Martin Burchell*[1†]

[1] *Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom,* [2] *Cambridgeshire and Peterborough NHS Foundation Trust, Liaison Psychiatry Service, Cambridge, United Kingdom*

OPEN ACCESS

Edited by:
Martin J. Sliwinski,
The Pennsylvania State University, United States

Reviewed by:
Soyong Eom,
Yonsei University, South Korea
Weidan Pu,
Central South University, China

***Correspondence:**
Rudolf N. Cardinal
rnc1001@cam.ac.uk

†ORCID:
Rudolf N. Cardinal
orcid.org/0000-0002-8751-5167
Martin Burchell
orcid.org/0000-0003-2447-8263

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Psychiatry

Received: *30 June 2020*
Accepted: *19 October 2021*
Published: *17 November 2021*

Citation:
Cardinal RN and Burchell M (2021) The Cambridge Cognitive and Psychiatric Assessment Kit (CamCOPS): A Secure Open-Source Client–Server System for Mobile Research and Clinical Data Capture. Front. Psychiatry 12:578298. doi: 10.3389/fpsyt.2021.578298

CamCOPS is a free, open-source client–server system for secure data capture in the domain of psychiatry, psychology, and the clinical neurosciences. The client is a cross-platform C++ application, suitable for mobile and offline (disconnected) use. It allows touchscreen data entry by subjects/patients, researchers/clinicians, or both together. It implements a large and extensible range of tasks, from simple questionnaires to complex animated tasks. The client uses encrypted data storage and sends data *via* an encrypted network connection to a CamCOPS server. Individual institutional users set up and run their own CamCOPS server, so no data is transferred outside the hosting institution's control. The server, written in Python, provides clinically oriented and research-oriented views of tasks, including the tracking of changes over time. It provides an audit trail, export facilities (such as to an institution's primary electronic health record system), and full structured data access subject to authorization. A single CamCOPS server can support multiple research/clinical groups, each having its own identity policy (e.g., fully identifiable for clinical use; de-identified/pseudonymised for research use). Intellectual property rules regarding third-party tasks vary and CamCOPS has several mechanisms to support compliance, including for tasks that may be permitted to some institutions but not others. CamCOPS supports task scheduling and home testing via a simplified user interface. We describe the software, report local information governance approvals within part of the UK National Health Service, and describe illustrative clinical and research uses.

Keywords: clinical informatics, research data capture, cognitive assessment, psychology, psychiatry, clinical neurosciences, information governance

INTRODUCTION

There are strong potential advantages to the electronic capture of information relevant to cognitive and psychiatric assessment. Measurement-based care improves clinical outcomes (1). Some simple standardized scales are in widespread clinical use, such as for affective disorders or cognitive examination [e.g., (2, 3)], but if the information is captured using pen and paper then its subsequent clinical accessibility and/or availability for research is limited, and

tasks must be scored by hand, taking time and introducing the potential for error. More complex computerized tasks are being translated from research to clinical use [e.g., (4)], but the clinical application of such animated tasks can be limited by practical considerations such as availability. Clinical and research assessments involve the documentation of a considerable quantity of information. Whether in a research or a clinical environment, and whether in an environment using paper-based or electronic health records (EHRs), there are incentives to capture such information electronically and in a standardized and structured fashion (5, 6). These incentives include a potential reduction in the effort of data capture; the ability to reproduce information accurately, legibly, and fast; the ability to appreciate trends over time; and the ability to analyse data for research or administrative purposes later. Information entered directly by patients can be used for screening and other purposes [e.g., (7)]. Rapid electronic systems can also capture information on outcomes that may not otherwise be measured routinely, such as quality of life indicators, used as the basis of many health economic measurements (8).

However, software for this purpose must overcome several potential pitfalls. First, for an application to enter widespread clinical use, it should save clinicians time, or at least place minimal time burden on clinicians. It should be quick to use and available at the bedside, in the clinic, or wherever a clinical or research encounter may take place. Second, users or institutions may be deterred from using software that is proprietary or closed-source (9–12), expensive, or that comes with practical restrictions on the use of raw data. Third, data capture systems are easy to write but harder to secure. There are considerable information security problems that would prohibit many simple applications from being used within a secure environment, as in a clinical context. For example, applications are likely to fall foul of UK National Health Service (NHS) information governance principles if they allow one patient to see another's data; transmit patient-identifiable data (PID) over an insecure e-mail network or *via* an unencrypted network link; use inappropriate cryptographic algorithms; fail to prevent unencrypted PID being backed up automatically from a tablet to commercial "cloud" storage; or use servers hosted on insecure or third-party computers, including those in prohibited jurisdictions (13–18).

We describe a novel client–server software package, the Cambridge Cognitive and Psychiatric Assessment Kit (CamCOPS), which attempts to address these problems. It incorporates a number of common and freely available tasks, and can serve as a basis for the addition of further arbitrary tasks in the future. It is an open-source cross-platform system that uses touchscreen tablet devices or desktop/laptop computers for data capture. Instances of the client application ("app") send their information securely to a central server, owned and controlled by the operating institution. The server provides a "front end" for convenient use by clinicians and researchers, with additional "back-end" facilities to support subsequent research analysis and system interoperability. CamCOPS offers many well-known questionnaires and some more advanced (e.g., animated) tests relevant to cognitive and psychiatric assessment, plus structured and unstructured clinical record-keeping facilities. Data capture can be performed with the app offline, so the system can be used in places with no network reception, such as on domiciliary visits or in unusual radiofrequency environments. The system is compatible with UK NHS information security standards, though compliance with those standards requires other institutional practices as well. As the system is free and open source, we suggest it is suitable for others to use and extend.

DESIGN AND FACILITIES FROM A USER'S PERSPECTIVE

Client–Server Architecture

Data collection and storage is organized around a client–server model (**Figure 1**). Tablet devices or desktop/laptop computers running the CamCOPS app act as one type of client. A clinician/researcher, a patient/subject, or both together can interact with the device to capture information. Upon request, the app then sends these data securely to the server, located within the host institution. The other main type of client is a clinician/researcher using a web browser or other interface to retrieve information from the server. Strict controls, described below, govern the exchange of data between clients and the server.

Subject Identification

The software system is designed to cope with clinical environments that use fully identifiable patient information, and with research environments in which participants are assigned a pseudonym or code and an individual's identity is not obtainable without additional information (the pseudonym-to-identity mapping) stored securely elsewhere. The flexibility to operate in both these environments is achieved by defining the meaning of multiple identification (ID) numbers and specifying the minimum and/or maximum information permitted.

The system defines the following subject identity fields, not all of which need to be used: forename, surname, date of birth, sex/gender (M/F/X) (19), and an arbitrary number of ID number types (e.g., national ID number, hospital ID number, study ID number), plus optional address, e-mail, general practitioner, and "other" details for convenience. The administrator defines the meaning of each of the ID number types. CamCOPS supports data verification for some specific ID number types, such as NHS numbers, which incorporate a checksum.

CamCOPS supports two types of ID policy: an *upload* policy and a *finalizing* policy. The upload policy defines the identity information required for the client app to send data to the server. The finalizing policy defines the identity information required for the app to move data to the server, allowing erasure from the client device (with permanent storage on the server). This two-stage process allows data to be entered for new subjects before that subject is registered on a host institution's systems. Two examples may serve to clarify.

In a research environment using pseudonyms, the administrator might define the meaning of "ID number type 1" to be "Research ID." The upload policy might be "sex AND

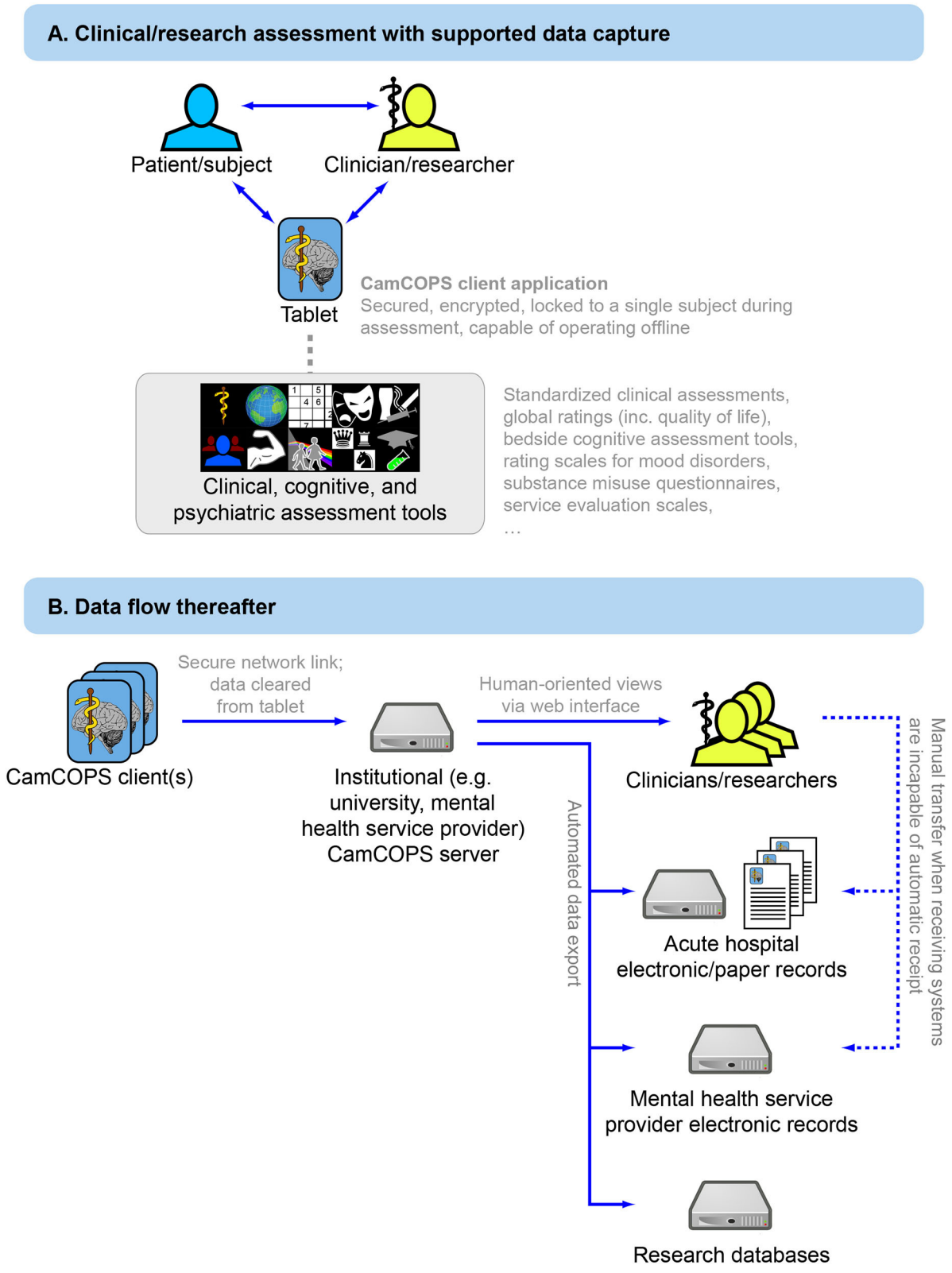


FIGURE 1 | Overview of the CamCOPS data capture system. **(A)** Data capture to the mobile app. **(B)** Subsequent data flow from the mobile device to the institution's CamCOPS server, and thence to individuals viewing or analysing the data, and/or electronic and (if required) paper clinical records.

idnum1". The finalizing policy might be identical. Therefore, the system would accept uploads only if the researcher had entered a subject's sex/gender and research ID number (as defined by the institution or individual research study concerned), but would not require any other information. Indeed, other information might be prohibited, such as "sex AND idnum1 AND NOT (forename OR surname OR dob)".

In a complex clinical environment using fully identifiable records, such as a mental health Trust that operates its own patient numbering scheme but also provides urgent on-call services to several hospitals in its region, the administrator might define "idnum1" to mean "Hospital A number," "idnum2" to mean "Hospital B number," "idnum3" to mean "NHS number," and so on. Suppose Hospital A is the provider institution. The upload policy might be "forename AND surname AND dob AND sex AND (idnum1 OR idnum2 OR idnum3)", and the finalizing policy might be "forename AND surname AND dob AND sex AND idnum1". This would mean a clinician could enter patient details in Hospital B, using Hospital B's number, without yet knowing the number used by their core institution (Hospital A). The system would require a full name, date of birth, sex/gender, and at least one ID number. At that early stage, the clinician could upload the data, and store a properly identified electronic copy in Hospital B's electronic or paper records. On return to their base in Hospital A, the clinician could look up the patient's number in Hospital A's system or register a new patient, and complete the record by filling in the Hospital A number (idnum1). At this point the software would allow the record to be re-uploaded and deleted from the tablet.

Using the CamCOPS Client Application to Capture Data

The starting point of the client app is shown in **Figure 2**. To capture data, the clinician/researcher usually begins by selecting a subject, recording the subject's details according to the identification policies in place. The operator then selects a task and creates a new task instance [current available tasks are listed at (20)]. The task will then run. Typical tasks appear as single-page or multiple-page questionnaires, or animated tasks (**Figure 3**). They range from very simple tasks, such as the Patient Health Questionnaire-9 [PHQ-9; (2)], through tasks with more complex logic, such as the Clinical Interview Schedule—Revised [CISR; (21, 22)], and those with a more complex interface, such as the Addenbrooke's Cognitive Examination—III [ACE-III; (3)], to complex animated tasks such as a three-dimensional intradimensional/extradimensional set-shifting task (23).

Questionnaire-style tasks have a consistent user interface, indicating mandatory/optional data items and permitting progression when mandatory information has been provided. The questionnaire user interface is consistent across platforms (operating systems, OSs). User customization of the interface is limited but includes font sizing for accessibility and language selection (discussed further below).

Some tasks are intrinsically anonymous, in which case they are not associated with any subject information at any stage, visibly or invisibly.

Tasks may collect information from the patient/subject alone, the clinician/researcher alone, or both together. Questionnaire-style tasks provide consistent colour-based visual cues as to the respondent. Tasks that involve the clinician's/researcher's judgement also record the details of the clinician/researcher conducting the assessment. These details may be pre-configured by the operator in advance so they are automatically entered, but may be edited, for example when a tablet-wielding clinician needs to document an assessment conducted by a more senior clinician. Some OSs (e.g., iOS, Android) are not designed for multi-user use, and the CamCOPS client does not offer specific multi-user facilities, but it stores per-user data when running under multi-user OSs.

CamCOPS also supports multimedia facilities in direct and indirect ways. Sound is used in some tasks, and the app can use the device's camera to capture photographs (such as of handwriting or other paper notes). In addition, text fields in CamCOPS can accept input from voice-recognition dictation systems supported by the OS.

The operating mode described above is oriented towards interactive use by a clinician/researcher and patient/subject together or consecutively ("clinician mode"). In addition, CamCOPS also supports a "single user" mode. This is intended for patients/subjects to complete tasks by themselves, for example at home in advance of a clinic appointment or between appointments, or as part of an ongoing research study. To use this mode, the clinical or research team defines one or more task schedules on the server (such as a weekly PHQ-9 for 6 weeks), and registers the subjects. The subjects download the CamCOPS app and enter the server's URL (uniform resource locator) with an identification/security code. The app is presented *via* a highly simplified user interface, and will then offer tasks to the subject automatically according to the defined schedule(s), moving data to the server whenever a task is completed.

Viewing Completed Tasks

Once complete, tasks may be viewed on the client device (e.g., tablet) or the server. Tasks are visible on the tablet until they are moved off it (typically at the point of upload to the server) and are visible to authorized users on the server as soon as they have been uploaded, and indefinitely thereafter.

In the client app, tasks display summary details, such as the total score from a questionnaire, and often also a read-only facsimile of the full task, as seen by the subject or clinician during the task. The facsimile view is provided automatically for all questionnaire-style tasks.

On the server, tasks provide an HTML (hypertext markup language) view, optimized for browsing speed, or a PDF (Portable Document Format) view, optimized for printing (**Figure 3**). Both show the raw captured data, plus summary information calculated automatically. The PDF view adds subject identification information to all pages, making them suitable for printing and direct use in paper-based clinical environments, and provides space for an authenticating physical signature where tasks have been conducted by clinicians (as opposed to tasks that are entirely self-rated by subjects).

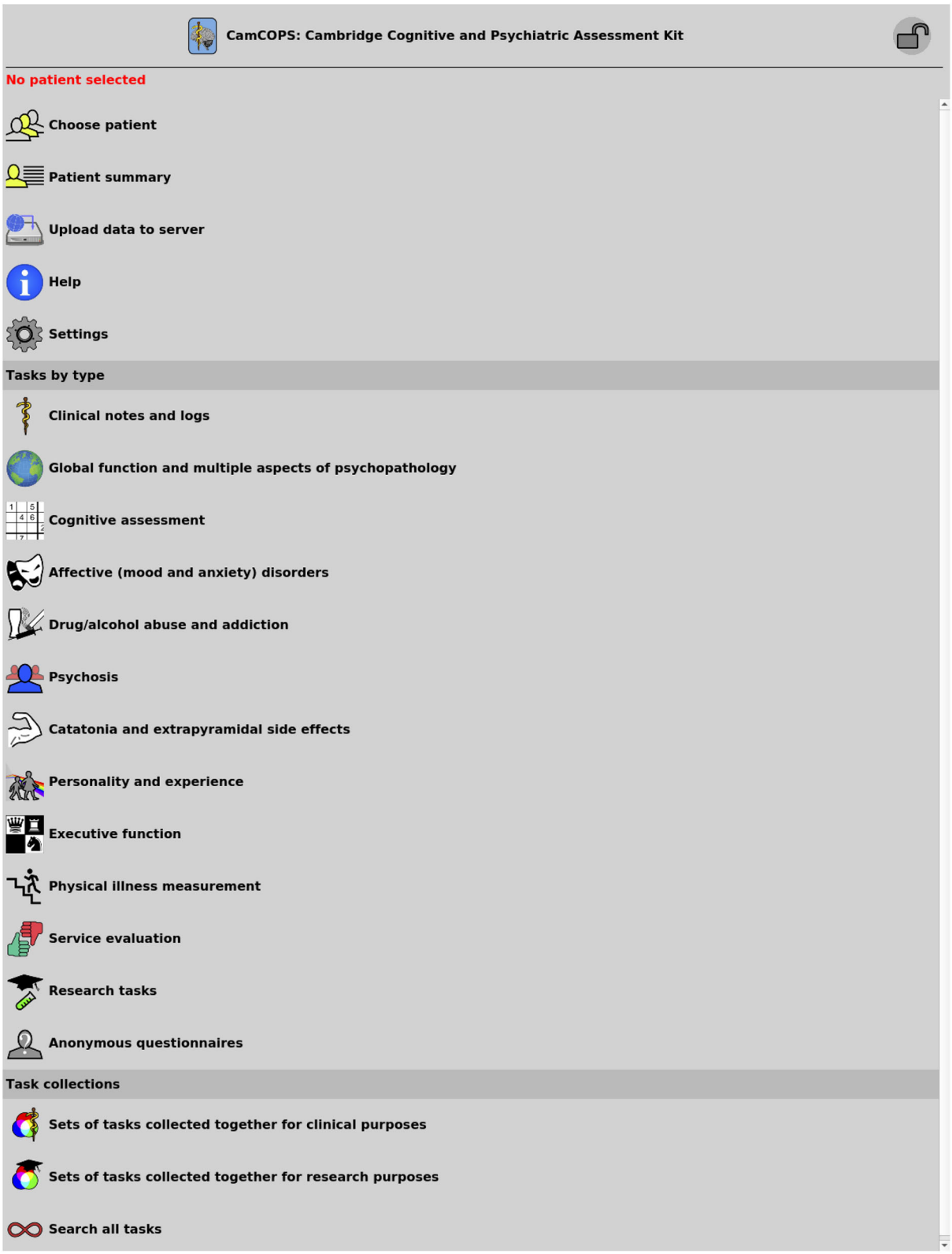


FIGURE 2 | The main menu of the CamCOPS app.

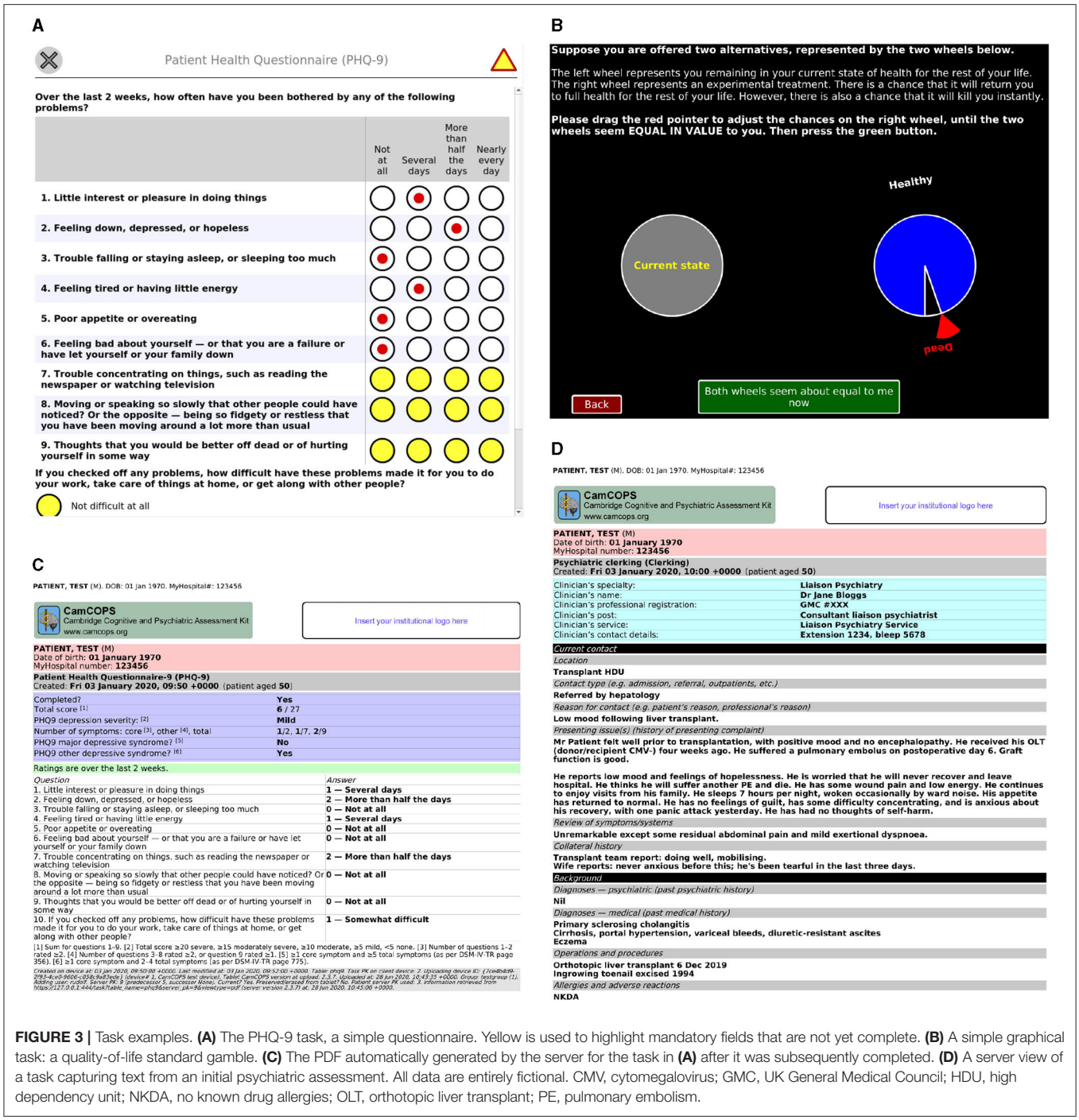


FIGURE 3 | Task examples. **(A)** The PHQ-9 task, a simple questionnaire. Yellow is used to highlight mandatory fields that are not yet complete. **(B)** A simple graphical task: a quality-of-life standard gamble. **(C)** The PDF automatically generated by the server for the task in **(A)** after it was subsequently completed. **(D)** A server view of a task capturing text from an initial psychiatric assessment. All data are entirely fictional. CMV, cytomegalovirus; GMC, UK General Medical Council; HDU, high dependency unit; NKDA, no known drug allergies; OLT, orthotopic liver transplant; PE, pulmonary embolism.

The user may filter tasks by subject, date, task type, and so on, but may also search also by free-text content; thus, for example, all task instances containing the word "overdose" can be searched for, whether those "tasks" relate to initial psychiatric assessment, a progress note, or a comment made by a participant in a research questionnaire.

The server also provides a summary view oriented towards text, and another oriented towards numerical data (**Figure 4**), both available in HTML and PDF format. The clinical text view shows all tasks for a given subject, optionally constrained by date, and shows key text from each task (e.g., summary scores for cognitive assessments or mood questionnaires, or all text for clinical assessments and progress notes), with hyperlinks to the full tasks for further detail. The numerical trackers show trends in numerical information over time in graphical format (such as for mood disorder questionnaire summary scores, or body mass index [BMI]).

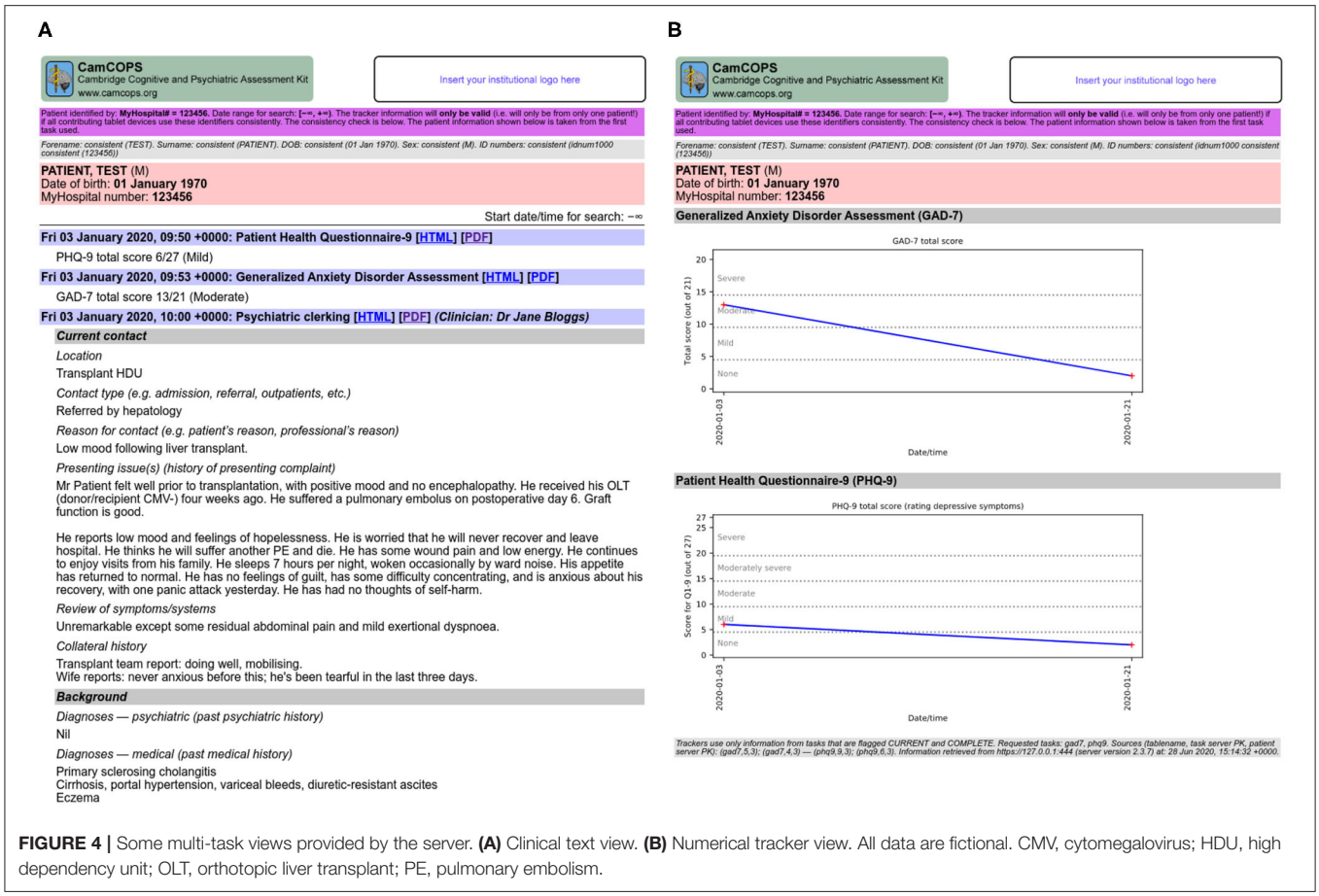


FIGURE 4 | Some multi-task views provided by the server. **(A)** Clinical text view. **(B)** Numerical tracker view. All data are fictional. CMV, cytomegalovirus; HDU, high dependency unit; OLT, orthotopic liver transplant; PE, pulmonary embolism.

Task Information

The online documentation (20) includes help pages for all CamCOPS tasks, hyperlinked to from the app itself. The help pages include details of each task's history and provenance, with links to key studies such as on the task's validity and reliability, where applicable. It remains for end-user clinicians/researchers to establish the applicability of a particular task to a given subject/patient in their context.

Internationalization

Text used by the client and server software is internationalized, supporting arbitrary languages (with current text for English and Danish), and the task framework supports internationalization of individual tasks. Where tasks supported by CamCOPS have been translated and that translation validated, the framework permits the translated versions to be selected automatically. Users choose their preferred language dynamically.

Interfaces for Research

While clinicians typically focus on a single patient at a time, researchers typically analyse data from multiple subjects together. The CamCOPS server allows suitably authorized users to download data in bulk, for exporting to other databases or manipulation in spreadsheets or statistical software. Download formats include Microsoft Excel (XLSX), OpenOffice/LibreOffice (ODS open document spreadsheet format), tab-separated values (TSV), R script (24) (though R can also read a number of other structured formats exported by CamCOPS), textual SQL (structured query language), and binary SQLite format (25). In addition, users can view raw and calculated data in structured XML (extensible markup language) format. Administrators may export data in bulk, including *via* formats suitable for third-party anonymisation tools (26).

Following the DRY ("don't repeat yourself") principle of software engineering (27), CamCOPS stores raw data, not calculated data. For example, PHQ-9 information is stored as a set of answers to each of the 10 questions; the summary scores are not stored but are calculated "live" upon request. BMI information is stored as height and mass, and so on. The method of calculation of summary scores is specific to each task. To simplify research and to reduce errors caused by researchers having to calculate summary scores, CamCOPS calculates these. The system offers a basic research data dump oriented towards convenience, in which most tasks provide a single spreadsheet-style page. This has one row per task instance, including includes raw data, summary scores (calculated at the moment of request), and subject identifiers together.

The CamCOPS server is an interface to a relational database with a well-defined structure. It is conceivable—but in our view highly unlikely—that institutions would wish to give users

direct access to this database (which would circumvent standard security controls). However, for full access to relevant raw data, CamCOPS allows authorized users to download a relational database containing data of interest to and permitted to them, as well as downloading structured data directly to statistical packages such as R (24) (**Figure 5**). Relational database export is more powerful but more complex for users. Summary scores are also provided automatically in this situation, by calculating them as the download is created.

Group System

A given institution may need to capture data in several different contexts. For example, it might provide a number of clinical services. Staff in those services might want to analyse their service's data in isolation but also see data for their patients that has been collected by other clinical services. Simultaneously, the institution might support clinical research using identifiable data. Researchers might typically be allowed to see only the data collected for their subjects as part of their study (while, simultaneously, clinicians looking after those patients might want to see any clinically relevant data, collected as part of the research or otherwise). Finally, the institution might support research using pseudonymised data.

CamCOPS supports these usage scenarios simultaneously via groups. A group might represent, for example, a clinical service or a specific research study. Users belong to one or more groups, and upload data into a specific group at any one time. A group has its own set of ID policies (as above), and may, as a whole, be permitted to see data from specific other groups. Thus, for example, a clinical group might use fully identifiable data according to a certain identification standard, while a research group may use a study-specific pseudonym and prohibit direct identifiers. A researcher might belong to one or more research study groups, and only be permitted to see data collected within them. A different clinical research group might use an ID number type in common with clinical services, and the system can be configured to allow clinicians to see data from all clinical services plus "research" data for the same patients, without researchers being able to see "clinical" data.

While is also possible to run multiple instances of the CamCOPS server, the group system is intended to make this unnecessary for most purposes.

Export Facilities

Individual users may wish to download different subsets of data in various formats (as above), but it may also be desirable to export data systematically from the server. A prototypical example would be the need to copy clinically relevant data to an institution's primary EHR system. CamCOPS supports export in different formats (including PDF, HTML, and XML) and *via* different transmission methods [including *via* HL7 (30), e-mail, and file-based export]. Exports can be scheduled and/or triggered by the arrival of a task on the server. CamCOPS also supports direct export to relational databases, and to REDCap (31) via an open-source interface (32). We are also seeking to improve integration with other EHR systems, via standard information exchange methods such as FHIR (33).

Other Administrative Operations

Subject to permission, users can run reports on the server. These include activity reports and search tools. Group administrators can manage users within their groups, and superusers have full control over the whole system.

To assist compliance with NHS records management procedures (34–37), specific records can be erased of content or deleted entirely by privileged users. All records for a given patient can be deleted entirely, as might be required after a certain number of years have elapsed, or in a research context if a subject withdraws consent. Records can be annotated manually by users with annotation authority (for example, to indicate an error or that the patient disputes its contents) and patient details can be corrected (for example, if a name was misspelled).

IMPLEMENTATION

Software Platforms

The client app is written in C++ (38) using the open-source Qt cross-platform framework (39). CamCOPS has been used on Android devices, iOS devices (e.g., Apple iPad), Windows tablets, and conventional desktop computers (Windows, Linux, macOS). Application data is stored in an encrypted database using SQLCipher (40), based on SQLite (25). Cryptography is provided by OpenSSL (41), developed from SSLeay (42).

The CamCOPS server is cross-platform software written in Python (43). It is supplied with HTTP (hypertext transfer protocol) servers including CherryPy (44) and Gunicorn (45), which may be used directly or *via* a more sophisticated web server such as the Apache HTTP Server (46). It is normally run under Linux (47) (tested with Ubuntu/Debian and CentOS). CamCOPS typically uses the open-source MySQL/MariaDB database (48, 49) but supports others *via* SQLAlchemy (50). A Docker Compose containerized application is provided for consistency and ease of installation (51).

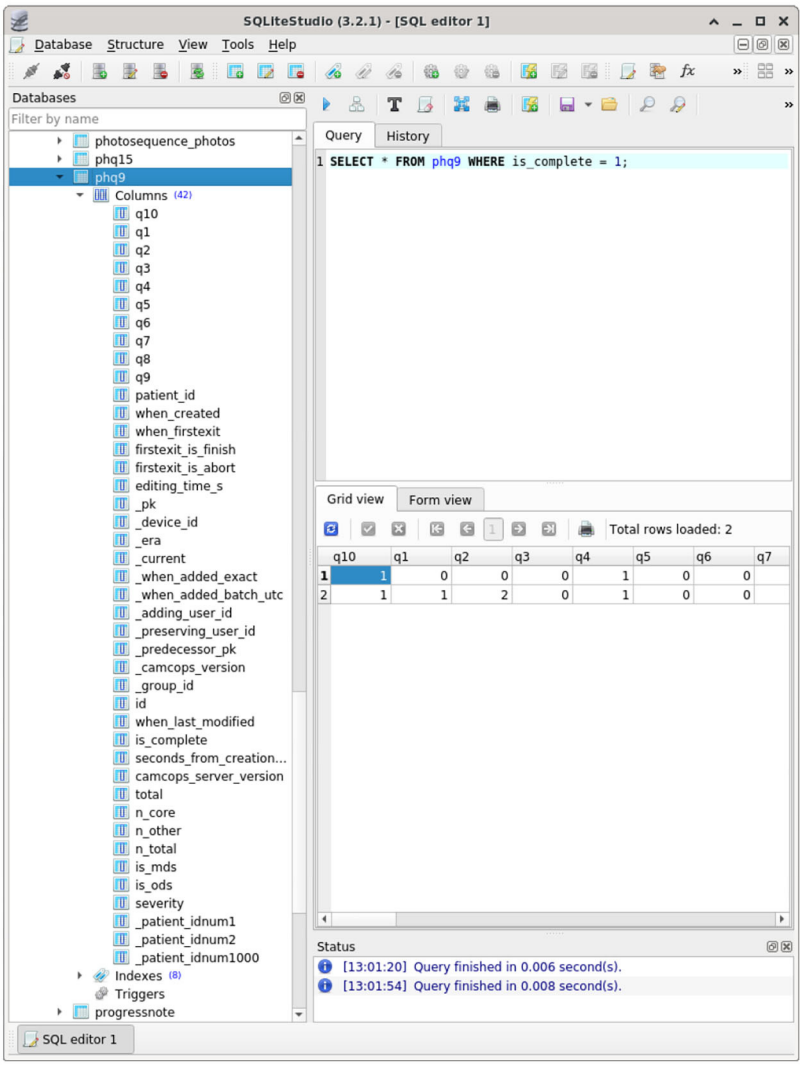
Distribution

Documentation is online (20). The source code and some binaries are available from GitHub (52). The Android client app is available *via* the Android Google Play Store, and the iOS version *via* the Apple App Store. Apple prohibits public distribution, by other routes, of applications that can be installed on arbitrary iOS devices (53).

Data Storage and Synchronization

CamCOPS stores its data using standard relational database mechanisms (54). A simple format is used, with a table to record subject details, a linked table to record ID numbers, and one or more tables for each task, linked to the subject table except in the case of anonymous tasks. The app records the time of last modification for all records. Tasks also record their creation time, the time the task was first exited, and whether the task was completed or aborted at that time. This allows measurement of the time it takes to complete a task. Dates and times captured by tasks are stored in ISO-8601 format, with time zone information and arbitrary temporal precision (by default accurate to 1 ms to allow reaction time recording). Binary large objects (BLOBs)

A



B

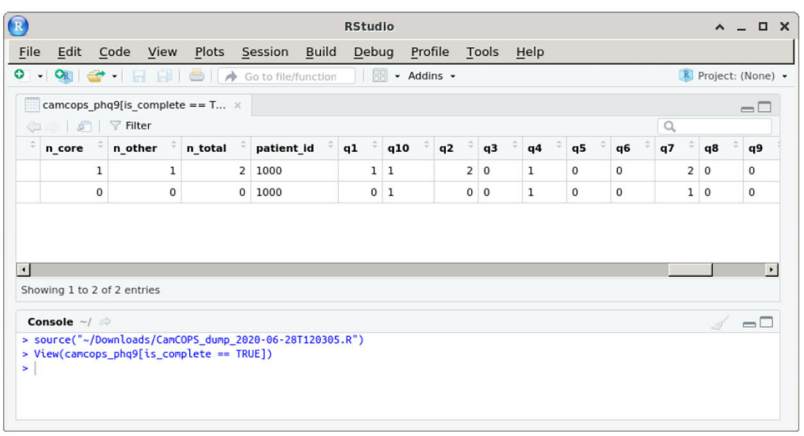


FIGURE 5 | Some research-oriented methods of data access. **(A)** Structured data in a relational database downloaded from CamCOPS. An SQLite database (25) is shown in SQLiteStudio (28). **(B)** Data downloaded and imported directly into R (24), shown inside RStudio (29). All data are fictional.

such as images are stored in the database; this is not definitively better or worse than storage in a filesystem (with the database holding a reference to the file), but storage in the database has the advantage of being easily ACID (atomicity, consistency, isolation, durability) compliant.

Subject identification is one area where CamCOPS departs from the DRY principle (27). The CamCOPS server maintains copies of each device's patient identification records, as of the moment of each upload. It does so because repetition is a key safety feature to ensure correct patient identification in clinical environments, and because the use of additional non-unique identifiers is important for clinical safety. For example, if clinical records only had NHS numbers on them, they would be technically correct but clinically useless, because clinicians think of patients by their name. CamCOPS does not fetch other details; for example, when given an ID number, it does not fetch a patient's forename/surname from a national or institutional database. Therefore, users need to enter this information. Of course, several CamCOPS users can enter data about the same patient on different devices, and it is possible for users to enter incorrect name/number combinations or to misspell names. The appropriate logical mechanism to link multiple records about the same patient is defined by the host institution, but is typically by the use of a single standardized institutional or national ID number. When the CamCOPS server interface combines records, linking them by the desired method (e.g., institutional ID number), it warns the user prominently if any records contain incompatible information (e.g., misspelled names or non-matching dates of birth). Suitably authorized users can correct mistakes (e.g., misspelled names) on the server, once records have been finalized to the server. CamCOPS contains framework code to support validation of subject identity at the point of upload (e.g., against an institutional database), but this has not been used concretely yet.

No history information is stored in the client app's database, but history information is added by the server. Servers distinguish records from different client devices using a unique device identifier. The server also marks uploaded records with a Boolean "current" flag. When a record is re-uploaded, the old record is marked as no longer current, linked to its successor, and its time of removal and removing user recorded, while the new record is marked current, linked to its predecessor, and its time of addition and adding user recorded. This allows a modification history to be followed, and permits linking of contemporaneous information across multiple tables.

The client app can copy data to the server, but may also move data by uploading it, wiping it from local storage, and starting afresh. A "move" may be accomplished for individual anonymous tasks, for all tasks associated with a particular subject or subjects, or for all data on the device. Optionally, basic subject identifiers can be preserved on the device to speed the entry of subsequent data for the same subject. The server manages this "move or copy" capability by adding a further "era" field, which is either the string literal "NOW" (for records still present on the device) or the date/time that the data was uploaded and wiped from the device. Using these mechanisms, which allow the server to store multiple snapshots of a device's state over time, records can be wiped from the device yet remain available on the server, or be modified and "overwritten" on the server, leaving a historical trail of modifications available for inspection.

Uploads are accomplished as atomic transactions; that is, they succeed in their entirety or fail as a whole. This preserves the relational structure of the database in the face of unexpected network disruption.

Hardware Platforms and Costs

CamCOPS has been used on tablets, touchscreen laptops, and conventional laptop/desktop computers. In practice, we have found that hardware keyboards (e.g., Bluetooth keyboards for tablets) are essential for any form of data capture that uses text extensively, such as clinical note-taking, because on-screen keyboards are slow to operate. The choice of tablet may depend on price, on the form of network connectivity desired (e.g., Wi-Fi only vs. Wi-Fi plus 3G/4G cellular data), and on the software distribution model desired. For example, Android tablets can install software from the Google Play Store, but can also install software downloaded from arbitrary web sites. Thus, an institution could download the CamCOPS code, modify it for its own purposes, compile it using the open-source development tools, and distribute it on its own internal or public-facing web site. In contrast, distribution to iPad devices is only permissible *via* the Apple App Store or via internal distribution by organizations or individuals who pay for the Apple iOS Developer Program (53, 55).

CamCOPS is free of charge, but the system as a whole requires some infrastructure. In a university research environment, a simple server installation requires only a single Linux physical server or virtual machine with a network connection, plus a transport layer security (TLS)/secure sockets layer (SSL) X.509 certificate ("SSL certificate") for secure HTTP (HTTPS). In a UK NHS clinical research environment, such a server may need to operate within a secure network, and there may be additional costs for virtual private network (VPN) access to that network from outside. The main additional cost is for client devices, which vary according to user preference (e.g., Android tablet; iPad; Windows tablet; touchscreen laptop). The client devices must be able to communicate with the server (e.g., *via* a wired connection, Wi-Fi, or 3G/4G cellular data).

Performance

The server is optimized for performance using multithreading or multiprocessing and caching systems. The basic overhead of the scripts is very low: a server with an Intel dual-core 3 GHz processor and solid-state disks took 3 ± 1 ms (mean $\pm$ standard deviation) to process an HTTP transaction, retrieve and validate session information from the database, and return the main menu ($n = 100$). Retrieving a PHQ-9 task in HTML format took 9.5 ± 1.9 ms ($n = 100$), including the time taken to audit the request. Registration of a mobile device took 10.6 ± 0.4 ms ($n = 100$) including approximately 6 ms for password cryptography, which is deliberately slow in the *bcrypt* system (56). Performance in practice depends also on the underlying database and hardware; MySQL offers

the option to trade full ACID compliance for performance *via* the *innodb_flush_log_at_trx_commit* option (48), set for speed during the benchmarks given above. The client is similarly optimized for performance, including the use of a multithreaded database handler so that encryption does not slow the user interface.

Writing New Tasks

CamCOPS has >120 tasks and more are regularly added. Many psychiatric assessment scales use a questionnaire style, with multiple-choice fields, yes/no fields, free text, and other common input elements. Other tasks may require significant programming, such as cognitive assessment tasks that present stimuli and measure responses in a time-sensitive or complex way. The CamCOPS platform supports arbitrary tasks by providing a questionnaire-style interface, a *tabula rasa* allowing graphical and arbitrarily complex tasks, or a combination of the two.

Free-form tasks use C++/Qt code to create tasks of arbitrary complexity including visual animations and auditory stimuli. Questionnaire-style tasks use a simpler standardized interface. Questionnaires are built from combinations of elements, including:

- static text, images, lines, and spacing;
- an audio player;
- Boolean fields (NULL/false/true) with associated text or an image;
- a button, capable of executing arbitrary code;
- a canvas for sketching, which can display a background image;
- a countdown, to assist clinicians in timed tasks;
- date, time, and date/time pickers;
- a diagnostic code element, usable with any hierarchical diagnostic code system such as ICD-9-CM (compatible with DSM-IV-TR) or ICD-10 (57, 58);
- multiple-choice (1-from-*n*) questions (MCQs), in a variety of common layouts;
- multiple-response (*k*-from-*n*) questions;
- photographs, taken using the mobile device's camera, also useful for photocopying paper records;
- inline and pop-up pickers (an alternative 1-from-*n* representation);
- discrete and continuous scales represented by sliders;
- a thermometer-style scale;
- fields accepting typed input, with validation for textual or numerical fields;
- containers for laying out other elements.

The software is designed to be extensible. Adding a new questionnaire-style task presently requires (1) a C++ header/source file for the client app, specifying the task's structures and content (see excerpt in **Box 1**); (2) addition of that task to the app's master task list and menu system; (3) addition of strings to a string file in any languages required; and (4) a Python file for the server, specifying the table structure and the HTML content that is automatically used to make the server's HTML and PDF views.

BOX 1 | C++ code snippet illustrating the core of the implementation of a questionnaire-style task, the PHQ-9 (2), within the CamCOPS client app. This task uses some static text, a grid-style set of multiple-choice questions (MCQs) for questions 1–9 that all share a set of answers mapped to the data values 0–3, and a single MCQ for question 10. Calls to the xstring() function yield internationalized (language-/locale-specific) task strings; for example, xstring("q1") in the English locale evaluates to "1. Little interest or pleasure in doing things," while xstring("a3") evaluates to "Nearly every day." See Figure 3A for the resulting task.

```
const NameValueOptions options_q1_9{
    {xstring(''a0''), 0},
    {xstring(''a1''), 1},
    {xstring(''a2''), 2},
    {xstring(''a3''), 3},
};
const NameValueOptions options_q10{
    {xstring(''fa0''), 0},
    {xstring(''fa1''), 1},
    {xstring(''fa2''), 2},
    {xstring(''fa3''), 3},
};
QuPagePtr page((new QuPage{
    (new QuText(xstring(''stem'')))->setBold(true),
    new QuMcqGrid(
        {
            QuestionWithOneField(xstring(''q1''), fieldRef(''
                q1'')),
            QuestionWithOneField(xstring(''q2''), fieldRef(''
                q2'')),
            QuestionWithOneField(xstring(''q3''), fieldRef(''
                q3'')),
            QuestionWithOneField(xstring(''q4''), fieldRef(''
                q4'')),
            QuestionWithOneField(xstring(''q5''), fieldRef(''
                q5'')),
            QuestionWithOneField(xstring(''q6''), fieldRef(''
                q6'')),
            QuestionWithOneField(xstring(''q7''), fieldRef(''
                q7'')),
            QuestionWithOneField(xstring(''q8''), fieldRef(''
                q8'')),
            QuestionWithOneField(xstring(''q9''), fieldRef(''
                q9'')),
        },
        options_q1_9
    ),
    (new QuText(xstring(''finalq'')))->$>$setBold(true),
    new QuMcq(fieldRef(''q10''), options_q10),
})->$>$setTitle(xstring(''title_main'')));
```

INTELLECTUAL PROPERTY MANAGEMENT

The intellectual property in the CamCOPS software must be distinguished from the intellectual property in tasks supported by the CamCOPS platform. The source code for CamCOPS is licensed under the open-source GNU General Public License v3+ (59). The same licence applies to tasks developed *de novo* by us as part of the CamCOPS project. CamCOPS also uses some third-party software libraries (e.g., for cryptography) with open-source licences. We took care to ensure that all other material potentially subject to others' copyright, such as text from tasks developed by others, is not included in the main CamCOPS source code. For example, the code developed by us to present and score a questionnaire is segregated from the text that makes up an individual questionnaire. Furthermore, we have taken care to ensure that all use of tasks within CamCOPS is permitted either by the copyright declarations published with the original versions of the tasks, or by explicit verification for each task. We have not

included content for any task where we are aware of copyright restrictions incompatible with distribution under an open-source licence. CamCOPS supports tasks under the following copyright models, ordered from least to most restrictive.

1. CamCOPS includes a number of freely available tasks. For example, the Patient Health Questionnaire-9 (2) is published with an explicit declaration that it is in the public domain (60), while the copyright to the National Adult Reading Test (61) is held by its author, who kindly gave permission for its free use in perpetuity (62).
2. Some tasks are published with a copyright declaration allowing, for example, free non-commercial use and reproduction with appropriate attribution, but restricting commercial use [e.g., (63)]. CamCOPS includes user-completed fields indicating whether the software is being used for clinical, research, educational, and/or commercial use; each field can take the value "yes," "no," or "unknown." It restricts some tasks on this basis, according to their published permissions. These tasks cannot then be used outside their copyright restrictions without explicit dishonesty by the user, in breach of the CamCOPS terms and conditions of use that all users must acknowledge, and of the tasks' licensing terms. However, it remains the user's responsibility to check that they are legally permitted to use each task, and to comply with any licensing terms.
3. Some tasks allow reproduction for institutions that have paid a license fee or undergone another registration process, but not otherwise. To cope with these, CamCOPS supports a method where the default task is only a data collection tool (as for type 4 below), with copyright-free placeholder strings such as "Question 1." The institution may then choose to install an XML file containing the actual task text on their server instance(s). When the CamCOPS client app registers with the server, it downloads any strings specific to that institution. As these add-on XML files are not distributed with the CamCOPS itself (merely templates), the open-source licensing of CamCOPS does not conflict with the restricted licensing applicable to such tasks. Responsibility for any add-on files rests with the hosting institution, as does compliance with any licensing terms, including any training requirements.
4. In addition, we had a local need to capture information electronically for tasks that are distributed commercially and cannot be distributed under an open-source licence, such as the Beck Depression Inventory (64). For this situation, in an attempt to improve on the research method of typing data by hand into a spreadsheet, we developed "skeleton" questionnaires that refer to the original questions only as "Question 1," "Question 2," and so on. This method allows data to be recorded electronically without including elements subject to copyright, but makes the task implementation useless except to clinicians/researchers who can refer to their own licensed copy of the test.

We note that ascertaining copyright status can be difficult, particularly for older tasks. For example, the Edinburgh Postnatal Depression Scale was published with a notice saying "users may reproduce the scale without further permission providing they respect copyright by quoting the names of the authors, the title and the source of the paper in all reproduced copies" (65), but this instruction has been superseded by a different set of permissions that prohibit unrestricted electronic reproduction (66). In all instances, if we have inadvertently erred in our assessment of a task's copyright status or licensing permissions, we will remove it from CamCOPS with our apologies if we are alerted to the fact.

INFORMATION GOVERNANCE, SECURITY, AND AUDIT

The CamCOPS information governance and security model is multi-layered. It is not sufficient to have a "secure" mobile application; a hosting institution must implement other security measures.

Minimizing Patient-Identifiable Data Held on Mobile Devices

Assuming that identifiable information is used at all, there are two main methods by which the CamCOPS app minimizes the amount of patient-identifiable information held on a mobile device.

First, data exchange with the server is essentially one-way (upload, not download). Therefore, even if all the security measures (see below) were somehow circumvented, possession of a device implies possession of information about at most a few patients, created recently on that device. The app will not retrieve information created on other devices.

Second, its dominant method of uploading is to *move* data to the server, not to *copy* it. Users upload when they choose, and can be prompted whenever a new task is complete. When they upload, they are offered a three-way choice. (1) The "move" option moves details of all patients and their task data to the server, deleting that data irreversibly from the device. If some patients do not meet the server's finalizing criteria, as above, then the user cannot move data until this problem is fixed. (2) The "move, keeping patients" option moves all patients' task data, but it keeps the basic patient details, so the user can add more tasks for these patients later. (3) The "copy" option copies data to the server, though it still "moves" patients or anonymous tasks that the user has explicitly marked as "finished."

Users are encouraged to move data whenever possible. However, the option to copy remains important, as in the multi-hospital example given above: when a patient has been entered using institution B's ID number, information must be uploaded and stored in institution B's records immediately, but institution A's number must later be added before that record can be finalized and moved to the server.

Device Security

Mobile device security is provided without the need for users to encrypt the entire device, since they might inadvertently fail to do so. All CamCOPS data is stored using the 256-bit form of the Advanced Encryption Standard (AES) cipher suite (AES-256) (67).

As a general security feature, not specifically related to or required by CamCOPS for its security, users may also choose to encrypt their devices using a strong password. Android devices allow on-device encryption (for Android version 3 and higher). This encrypts applications' data areas with a passcode (68, 69). Apple iPads and related iOS devices invoke encryption when a passcode is entered (70, 71). Both these platforms have "sandboxes" to prevent one application seeing another's data (71, 72). After device encryption is enabled, the tablet device will require a passcode every time it is turned on or re-activated after its screensaver has activated. Since a misplaced tablet will lock itself, lost or stolen tablets become useless to anyone except their owner. Other OSs provide similar functions.

Application Security

The CamCOPS app has three security modes when running in "clinician mode": Locked, Unlocked, and Privileged. In the Locked mode, the app is locked to a single subject and can only view or add records pertaining to that subject, or anonymous tasks. This mode is designed for a clinician/researcher to hand the device to a subject. It takes a single touch to lock the app, but it takes a password to unlock it. In the Unlocked mode, all data may be viewed and edited. This mode is designed for use by clinicians/researchers. Privileged mode is designed for administrators' use. In Privileged mode, features such as the following are unlocked: configuring the link to a server, registering the device with a server, and (if the device permits) exporting the local database to an insecure storage area such as a removable secure digital (SD) card. (Despite the name, there is nothing intrinsically secure about an SD card.).

CamCOPS requires the app password to start, and to access the encrypted databases. Since data security is prioritized, there is no recovery method if this password is lost: the app would require re-installation, with loss of any data not yet uploaded.

In typical clinical use, an administrator might set up CamCOPS to point to the appropriate institutional server and then give clinicians the "unlock" password but not the privileged-mode password. This would not be impossible for an astute clinician to circumvent, by uninstalling and reinstalling the app, but the clinician is, after all, entrusted with the primary clinical information in any case. In practice, this extra level of security may help to prevent the clinician from misconfiguring the app by accident.

Internally, the app never sends patient-identifiable data to the device's system logging stream, except when authorized *via* a privileged-mode data dump, so a malicious user who plugs the device into a debugging computer, such as via a Universal Serial Bus (USB) cable, will not see patient-identifiable data that way. The CamCOPS app stores its stores its "unlock" and privileged-mode passwords using irreversible *bcrypt* hashes (56)—that is, the passwords themselves are never stored. Moreover, the database in which these hashes are stored is itself encrypted. The administrator may choose, following local institutional policy, whether the CamCOPS app stores the user's server password using reversible encryption or does not store it at all. Storage with encryption is more convenient but less secure, since the password would be potentially vulnerable to a skilled attacker in possession of the CamCOPS app password (and the device's unlock code, if enabled). Not storing the password is more secure, but requires the user to enter the password each time data is uploaded.

Network Link and Server Security

Communication between the client app and the server is secured as follows. The app's network link to the server is constrained to use HTTPS and therefore link encryption. The specific encryption used depends on the web server's configuration; typically, it would be configured to use TLS 1.2 with the AES cipher suite (73). By default, the app will insist on a validated SSL certificate, though this can be turned off by the administrator for low-security environments that use a self-signed ("snake oil") SSL certificate.

Client application instances must register with a server. This serves several purposes. Firstly, the server does not want unauthorized devices uploading to it. Therefore, the server will only accept uploads from registered devices, and requires users to authenticate, with a username previously approved by an administrator for device registration, before accepting registration. Secondly, administrators will not want their clinicians or researchers to upload data to unauthorized servers. Registration is therefore a privileged-mode function. We envisage that in practice, device registration would be managed by an administrator for high-security environments. Thirdly, the server and the app should share a set of ID descriptions and upload/finalizing policies (see "Subject identification" above). The app reads the ID descriptions and policies from the server at registration, and re-checks these before commencing an upload.

The server requires username/password identification before it will accept an upload, and requires that the device be validly registered. Devices are distinguished by a unique device identifier (a long random number). The server accepts incoming data but will not provide unrelated data to the app. Therefore, even a hand-crafted app masquerading as an instance of CamCOPS and in possession of a valid username, password, and device ID cannot download sensitive data *via* the app–server link. The server will not add new fields or tables based on the claims of the uploading agent, and will not upload to reserved tables or fields. The server takes standard precautions against SQL injection (74).

Communication between users and the server *via* the web front end is secured as follows. The web front end is constrained to use HTTPS and therefore link encryption. This requires appropriate configuration of the web server hosting the CamCOPS installation, but is also ensured by CamCOPS through its session security methods. Access is governed by username/password pairs. The server stores all CamCOPS passwords using irreversible hashes (56); passwords themselves are not stored. The only session information stored on the client side is a HTTPS-only session cookie containing a server-generated session ID and token; the token is regenerated by the server at login to prevent session fixation (75). Sessions expire after a defined period of inactivity and cannot be transferred between client Internet Protocol (IP) addresses. Administrators configure a maximum password lifetime. The server will lock user accounts for increasing periods of time in response to multiple

login failures. It will mimic normal login failure behaviour for non-existent usernames, including the time it would normally take for password cryptography, to prevent automated username discovery. Optionally, administrators may require multi-factor authentication, such as *via* e-mail, text message (short message service, SMS), or a third-party authenticator app (e.g., Google Authenticator).

Internally, the server must deal briefly with a clear-text database password, but encapsulates all such code with an error-trapping framework to prevent the password leaking, and promptly discards the password after connecting to its database.

Access to data *via* the server's web front end is governed by user-based and group-based permissions. *Users* themselves may have superuser status (which gives unrestricted access to data and administrative functions via the front end), or be "locked" to a single patient record (when that user belongs to a patient/subject for "single-patient" mode), or be a routine "staff" user. Users may be a member of one or more groups. At any time, one group is selected to receive data uploaded by that user. *Groups* were discussed above. Groups define patient/subject identification criteria (e.g., fully identifiable vs. pseudonymised) and intellectual property restrictions. Groups "own" subsets of data, but groups (and thus their members) can also be granted permission to view data from specific other groups. *User–group associations (group memberships)* are associated with a further set of permissions: to administer the group (e.g., manage users within that group), plus individual permissions to upload data, to register new client devices, to log in *via* the web front end, to view data for multiple subjects when no subject search criteria have been applied, to export data in bulk, to run reports, or to attach notes to uploaded tasks. These permissions provide fine-grained control over what users can see and do, but a security breach of a group administrator account, or even worse a superuser account, would permit large-scale access to CamCOPS data held on the server.

The server must also be secured in other ways that are outside the scope of the CamCOPS system itself but are nevertheless critical. Standard security considerations include limiting physical access to the server; preventing visibility on public networks (e.g., limiting visibility to internal institutional networks or *via* secure VPN access to them); configuring a firewall appropriately; limiting secure shell (SSH) access; ensuring that the web server does not offer CamCOPS data by any route other than via the CamCOPS web front end itself; ensuring that no inappropriate users have access to the back-end database systems stored on the server; ensuring that the server is backed up regularly; ensuring physical security of backups; and ensuring server availability (e.g., in the face of power failure) should this be required.

"Analytics" Security

It is commonplace amongst mobile applications to send information about application usage back to the application's creators. CamCOPS does not do this. No information is sent by the client app except to the chosen institutional server, and no information is exported by the server except as permitted or configured by the local administrator.

Black Hat's Options

It is important to ask of any potentially sensitive system: what would it take to steal its data? Several methods are possible for CamCOPS:

- *Steal a device, the device's OS password, and its CamCOPS app password together.* This would allow existing records, still on that device, to be viewed.
- *Steal a device, the device's OS password, its CamCOPS app password, and its CamCOPS privileged-mode password together.* This would allow records still on that device to be sent to a "dark" server of the attacker's choosing.
- *Steal a user's CamCOPS server password, and a means of accessing the network on which the server is held.* This would allow the attacker to view data on the server (subject to the permissions granted to that user). If the server is on the open Internet, the network security requirement is eliminated, emphasizing the importance of network security for sensitive data, as well as strong passwords. This is the route of attack requiring particular security focus, since a predominant route of data theft is *via* "social engineering" rather than technical methods (76, 77). This risk is mitigated by requiring multi-factor authentication (as above).
- *Break into the server and gain direct access to its database.* This emphasizes the importance of securing the server.

These methods of attack may appear plausible but should not be possible:

- *Steal a device and the device's OS password, "root" the device to bypass factory default access restrictions, and access the tablet's CamCOPS SQLite database directly.* This would yield only CamCOPS app databases encrypted with AES-256.
- *Steal a tablet that has not been properly secured with a device (OS) password, or in other ways bypass the OS security.* As before, without the CamCOPS password, this would yield only an AES-256-encrypted database.
- *Steal a tablet and the tablet's OS password, download the open-source CamCOPS app, modify it, install it over the existing app without deleting the app data (bypassing any OS-specific digital signature checks on software installation), and attempt to use the modified app to export data.* Since the CamCOPS app does not know the password used to encrypt a given user's data, this conveys no benefit to the attacker; the database remains encrypted.

Once a computer is stolen, it can be dismantled. One must therefore consider also the possibility of breaking the encryption. No practical method is known of breaking the AES algorithm used to encrypt tablet data. The US National Security Agency approves AES for US government information classified Secret (for AES-128 or higher) or Top Secret (for AES-192 or higher) (67) and the UK NHS approves it for clinical data (16). CamCOPS uses AES-256. To give a sense of scale, a brute-force attack on an n-bit key takes a mean of $0.5 \times 2^n + 0.5$ cycles; therefore, a 256-bit key would take approximately 1.83×10^{59}

years to discover by this method with a 10 GHz attack frequency. The universe is 1.38×10^{10} years old.

Other Means of Ensuring Security of Patient-Identifiable Data

If a mobile device or other computer can "see" data on a remote server, then those data can be captured, even if by the simple expedient of saving a screenshot or taking a photograph of the device. This applies to any computer program, not just CamCOPS. Therefore, technical constraints are insufficient: users must be prohibited by cultural (institutional and/or legal) constraints from saving or storing patient-identifiable information on mobile devices in non-permitted ways. Similarly, users must be encouraged to look after their computer devices carefully, locking them when not in active use.

Audit Trails

Client-side audit trails are minimal, but the app time-stamps all tasks at their creation, and time-stamps the last modification to any record, as well as collecting information relevant to the time it takes to complete each task. In contrast, there is significant audit logging on the server. The CamCOPS server maintains a number of task-specific tables. To each record, the server adds fields allowing an audit trail. When a record is modified or deleted, the old versions are kept. The server's tables therefore contain a snapshot of each device's current state, and a complete audit trail, whose granularity is the frequency of uploads from a particular device. Access requests to the server *via* the web interface are also audited and logged, as are command-line CamCOPS operations by administrators.

Security Against Data Loss

Crashes in the CamCOPS app should not (and in our experience during development, do not) affect data integrity, because the SQLite back-end, with perhaps 500 million deployments worldwide (78), is designed to cope with this (79, 80). Additionally, only a small quantity of data is ever stored on the device, since data is regularly moved to the server, so the vulnerability to data loss from a device or app fault is in any case small. When the app upload its data, the process is atomic, meaning that the transaction either succeeds as a whole or fails as a whole, and does not leave the databases in a "halfway" state. Data on the server is typically stored using the well-established MySQL/MariaDB database system (48, 49).

Data loss remains possible. Reasons for this may include factors outside the CamCOPS system, such as a server environment that is insufficiently robust to cope with power loss or disaster. An amateurish example would be a server without an uninterruptible power supply (UPS). An example of server failure in an NHS high-availability environment was the Buncefield oil depot explosion on 11 December 2005, which temporarily disabled some laboratory computer systems used by our local acute hospital because a major computing provider was located near that depot.

As with any software system (81), it is also possible that the CamCOPS system might contain undiscovered bugs and therefore lose data. During development, in addition to human testing, several other steps are taken to minimize this possibility. CamCOPS includes an automated unit testing framework. We use a continuous integration (CI) service to run the automated tests every time the server code is changed, thus checking for software regressions, and the CI service also checks against a database of any reported security vulnerabilities in the Python packages used. For the client app, C++ compilation automatically detects some categories of error (82). We have a process of peer review for substantial code changes. In day-to-day operation, the server verifies that task information is complete, and valid (i.e., that all field values are permitted for that task), or warns the user accordingly. It also catches any potential internal errors to ensure that all transactions end in a database commit or a database rollback, meaning that any crashes that might occur within the server do not corrupt data or leave database locks held and block other processes.

However, CamCOPS is not presently accredited to NHS Interoperability Toolkit (ITK) standards or certified as a primary part of a clinical record. Therefore, a core requirement of data security would be to ensure that any information of sufficient importance be copied (e.g., in fully structured or PDF format) promptly from CamCOPS to a certified information storage system, such as an institution's primary EHR. To enable automatic copying of CamCOPS data into a certified information storage system, CamCOPS provides automatic export facilities (as above).

Security and Risk Comparate

One matter that is easily overlooked in discussions of technical security measures is the *relative* security or risk of an electronic approach compared to its alternatives, which are often far from risk-free. In areas with no Internet connectivity, the alternative to storing patient-identifiable data on a mobile device is usually to write it down. Paper-based methods can be less secure than their electronic equivalents (83). In addition, manual scoring of cognitive assessment scales is vulnerable to assessor cognitive error (84–86) and this in itself represents a degree of clinical risk. Paper-based methods can also limit clinical information transfer, if handwriting is unclear or becomes unclear through photocopying or faxing, or if the time required to copy or summarize information means that only a subset of information is transferred.

Legacy Security

Legacy security refers to the possibility that changes in hardware or software render old data inaccessible or unusable, such as when software applications refuse to start after expiry of a licence period. The CamCOPS code is open source, so can be installed, modified, and used freely by anyone, and should only include tasks/questionnaires that are in the public domain or where permission exists to use the task in perpetuity. As a last resort there is a clear procedure should the legal position on a task ever change, allowing removal of disputed content but preservation of all data: namely to remove or replace disallowed text and/or media from the app's and the server's resource files, leaving the code intact. This would result in a stripped-down data capture task and the ability to display and manipulate old data, as

described above. Third-party code and development tools used by CamCOPS are similarly open source.

ADDITIONAL CONSIDERATIONS FOR USE IN CLINICAL PRACTICE AND RESEARCH

Software Regulations and Limitations

While we have tried to ensure that CamCOPS is reliable and accurate, the terms and conditions of use include a disclaimer to the effect that the authors and distributors are not responsible for errors or liable for any consequences of users' reliance upon the content provided with CamCOPS. Content contained in or accessed through CamCOPS should not be relied upon for medical purposes in any way; if medical advice is required, users should seek expert medical assistance. CamCOPS is intended for use under the supervision of medical practitioners or researchers conducting ethically approved academic research.

Regarding the European Union Medical Devices Directive (87): CamCOPS is not intended primarily for the diagnosis and/or monitoring of human disease. It has not yet undergone a conformity assessment under the Medical Devices Directive, and thus cannot be described as or put into service as a medical device. We note that Medical Device approval is typically not required for research software tools, during research where there is no medical purpose for the device (88); such research has its own regulatory controls. Nor is it typically needed for software systems where the software does not interpret data, merely storing and transmitting it without change (for example, Medical Device approval is not needed for word processors, spreadsheets, databases, or e-mail systems that may sometimes contain medical data) (89); many CamCOPS tasks relating to clinical work perform no such interpretation. We are continuing to explore this evolving area of regulation.

Local Clinical and Research Approvals

In addition to these caveats, use within NHS England would require appropriate local NHS Trust approval (17). The CamCOPS system stores small quantities of patient-identifiable data on an encrypted mobile device for a limited period of time. NHS England guidelines allow this possibility subject to (a) strict rules regarding encryption, such as suitable cryptographic algorithms used with strong passwords; (b) all such devices being owned by the Trust, disallowing mobile devices owned by clinicians personally; and (c) Trust Information Governance and Caldicott Guardian approval (13, 16, 90). Device encryption on iPads uses AES-256 (71), while Android uses AES-128 (69, 91); both satisfy NHS encryption guidelines (16). CamCOPS data encryption, as above, is in addition to this. NHS Scotland guidelines classify data using a traffic-light system according to the risk of patient identification and harm or distress caused by loss (92). Patient-identifiable data relating to mental states would be classified as amber or red—likely often red. When applied to the CamCOPS system, which holds information transiently offline on a mobile device, these standards would require NHS-owned devices with whole-disk encryption and a strong password (92).

As noted above, CamCOPS is not a primary EHR system and it is critical that any clinically relevant data be copied to an institution's primary EHR. CamCOPS provides mechanisms for this to occur automatically (see above), subject to the EHR having the capability to receive it (see below for discussion of one possible fallback position with EHRs that do not).

In a research context, information-handling procedures will be directed by an appropriate national or institutional research governance framework [e.g., (93, 94)]. Clinical information governance guidelines are typically at least as stringent as guidelines that govern research with volunteers who have given explicit consent to research, and more stringent than guidelines covering pseudonymised or anonymised records, or non-sensitive information. CamCOPS was therefore designed against clinical information governance standards.

All tasks allowing free-text entry, and many established structured questionnaires in psychiatry, permit the capture of risk-related information, such as about suicidality. If such information is captured without direct supervision by a clinician, it is vital that a clinical service or research study has approved methods for handling such information. Most critically, patients/subjects must be aware that reporting information to an app is not a substitute for talking to their clinical/research team or obtaining emergency health care if required. Users must indicate that they understand this in order to use the app, but appropriate expectations must also be set by the institution operating the software.

Experimental Tasks

CamCOPS is designed to operate as a translational research platform, implementing human-specific and cross-species tasks derived from basic neuroscience research. Some experimental tasks are included in CamCOPS and are clearly labelled as such; more may be added.

EARLY EXPERIENCES

CamCOPS development began in 2012 and the first version of the client, written in the Titanium cross-platform Javascript framework (95), was available in 2013 together with a Python-based server. The system was developed incrementally, except that in 2017 the client was rewritten in C++/Qt for better performance and power, and the server reworked. CamCOPS was first approved for clinical use in October 2014 within Cambridgeshire & Peterborough NHS Foundation Trust (CPFT), and has been used both for clinical and research purposes. It has been deployed for research within CPFT, the University of Cambridge, and at academic institutions in Denmark and Singapore. It has been used on Android tablets including the Asus TF201, Asus TF300T, and Sony Xperia Z2 Tablet, and touchscreen Windows devices including the Microsoft Surface Book 2. Our experience has been that Windows tablets provide familiarity and multi-purpose computing for many users, whilst Android tablets can be cheap. All these operating systems support offline voice-recognition dictation systems, as described above, though we have found physical keyboards considerably more accurate for text entry.

As a clinical example, CamCOPS has been operational in CPFT's perinatal mental health service since 2019, where it is used to record questionnaire data relating to symptoms and service experience. Notably, the version of the EHR system in use did not have the capability to receive an automatic data "feed" from CamCOPS. We therefore used the poor substitute of having CamCOPS automatically e-mail tasks (on receipt) *via* an internal secure e-mail system to an administrative team, who uploaded them to the patient's EHR.

Examples in a research context include the Insight study (96) and MOJO study (Khandaker, NHS research ethics reference 19/EE/0233) examining the relationship between systemic inflammation and mood symptoms, in which CamCOPS has been used to capture a range of data encompassing medical history, affective symptoms including a standardized self-report computerized interview (21, 22), physical symptoms such as fatigue and joint inflammation, and quality-of-life measures.

COMPARISON TO OTHER SYSTEMS; STRENGTHS AND WEAKNESSES

There are a large number of free and commercial applications offering data capture for psychiatry-oriented questionnaires such as the PHQ-9, and similarly many web sites for users to design and offer generic surveys for free or *via* a variety of commercial models. Some systems offer extensively validated complex cognitive assessment tasks *via* a closed-source model with provider-hosted data [e.g., (97, 98)]. Others use a range of data collection techniques (mobile apps, web interfaces, text messaging) to collect information in specific clinical domains, such as for mood monitoring [e.g., (99)] or to detect psychiatric morbidity in general hospital contexts [e.g., (100)]. REDCap (31) is one widely used general-purpose research system, free but not open source (101–103), oriented towards flexible online data capture and using a model where institutions host their own instance (31, 102, 103).

CamCOPS differs from these systems in some ways, and at times complements them. Of course, all major design decisions come with trade-offs. We see the major decisions as follows.

Firstly, CamCOPS is free and open-source software; moreover, it has a "copyleft" licence that ensures derivative works must remain open source. This eliminates direct software costs and allows public scrutiny of the code, but may reduce the incentive for commercialization and commercial support. It also prevents the full incorporation of tasks incompatible with this licensing model. Careful intellectual property review is required with respect to new tasks (see above), though that would be true regardless of the software licence.

Second, we follow the principle of institutional hosting. This offers institutions complete ownership and control of their data, but comes with the burden of having to provide, obtain, or outsource relevant computing infrastructure and some burden of computer administration.

Third, CamCOPS can operate offline. This major design decision reflected our need to operate in offline environments such as on domiciliary visits to mobile phone (cellular data) "black spots" for our network providers, or in acute hospital environments with radiofrequency shielding or lack of Wi-Fi for other reasons. This inevitably excludes the much simpler software model where all testing is performed online *via* a web site, and it brings complexities in development, data security management (discussed above), and deployment (such as upgrading client apps). A benefit is that the client, being written in a high-performance low-level general-purpose programming language, is essentially unrestricted; thus, CamCOPS can and does implement animated tasks, generalized linear modelling, and so forth.

Fourth, we support on-device "registration" of new subjects/patients, and support multiple groups and identification policies. This adds clinical flexibility (e.g., capturing data in relation to an emergency referral prior to administrative patient registration) and supports a variety of clinical and research settings, from fully identified clinical work, to a mix of clinical and research work, to de-identified research. However, it adds complexity and can require more later verification than a model in which all patients are registered in advance on the server according to a unified identity policy. In practice, since the identity policy (or policies) is configured by the local system administrator, this balance is in large part determined by the hosting institution according to its needs.

Fifth, the tight security for data stored transiently on mobile devices, with its principle of data minimization, brings some trade-offs, such as the absence of a view of historical data "on the fly" within the mobile app. If historical CamCOPS information needs to be viewed, that is presently not supported "offline" but only *via* online web access to the server. This may limit utility in some situations.

Sixth, tasks are implemented at present as part of the CamCOPS code base, rather than being user-defined [cf. e.g., (31)]. An advantage is that tasks are developed as "canonical" versions, with their source code open—for example, everyone can check to see if there is a logical error in the implementation of a task. We have also found that the requirement to implement aspects of each task in both C++ and Python serves as an intrinsic cross-check for this kind of error, although it involves some extra work. It also brings the benefit that tasks are unconstrained—that is, they can use any feature of a general-purpose programming language—rather than being constrained by the limitations of a scripting environment, so they can be tailored to achieve a good user interface and experience. The obvious disadvantage is that CamCOPS is not well-suited for the creation of new questionnaires specific to a clinical service or research study on a rapid, *ad hoc* basis (including research workflow tasks such as recording consent); CamCOPS may therefore complement software designed for that purpose in a clinical [e.g., (104)] or research [e.g., (31)] environment. It also requires more programming experience to develop new tasks than simpler systems.

Finally, we note that in the clinical domain there is often tension between different modes of data capture that we see as stemming from a lack of interoperability. Many EHR systems are not designed to be used by patients at all, but are designed for

clinicians to enter data. In the UK, this is changing gradually with the advent of "patient portals" and EHR-connected apps, but while some EHRs can capture basic questionnaire data from patients directly, we know of none that can capture structured data from complex clinician-assisted tasks [e.g., (3)] or animated cognitive assessments [e.g., (23)] directly into the EHR *via* a convenient interface. This creates demand for systems that can, and that situation is likely to persist—primary EHR systems do a lot, but they cannot do everything. Accordingly, we suggest that the future focus in this area should be on using the "best tool for the job"—capturing directly into the EHR as the first preference, but using external tools (such as CamCOPS or others) where required—plus work to improve the integration of external systems and EHRs, so that data flows seamlessly in the most structured way possible as well as the most clinically relevant.

SUMMARY

Regardless of the current and future sophistication of phenotype measurement *via* passive data collection (105), in our view overt data capture will continue to remain central to digital phenotyping in psychiatry. We present CamCOPS, a free and open-source client–server system for direct data capture in the general area of psychiatry, psychology, and the clinical neurosciences. It runs on multiple platforms and emphasizes touchscreen data capture. It has both clinical and research applications and is designed to operate against stringent information governance requirements, with hosting institutions having complete ownership and control of the data they collect. It can operate with fully identifiable or de-identified information. We discuss security concerns that would apply to any system of this kind, and describe the approaches used in CamCOPS. It provides summary views on the data that we believe are useful for clinicians, whilst retaining full structured data for research, and it supports multiple export mechanisms to communicate with other systems. It implements a large and growing family of tasks, ranging from questionnaires to animated cognitive assessments, with techniques to address a range of licensing and intellectual property rules. We discuss its strengths and weaknesses and report on some early practical uses.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://camcops.readthedocs.io/; https://github.com/RudolfCardinal/camcops.

AUTHOR CONTRIBUTIONS

RC designed and wrote CamCOPS (2012−) and drafted the manuscript. MB contributed to the design and development (2019−). Both authors contributed, edited, and approved the final manuscript.

FUNDING

RC was supported by a Wellcome Trust postdoctoral fellowship (091998/Z/10/Z). RC's and MB's research was supported by a UK Medical Research Council (MRC) Mental Health Data Pathfinder grant (MC_PC_17213 to RC). Deployment was supported in part by the UK National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014). The work was conducted within the Behavioural and Clinical Neuroscience Institute, supported by the Wellcome Trust (093875/Z/10/Z) and the MRC (G1000183).

ACKNOWLEDGMENTS

We thank Martin Denton for security advice and technical support; Ed Bullmore for institutional support; Julia Deakin and Hannah Clarke for helpful discussion and piloting; Rob Smithies, Chris Randall, Melanie Coombes, Philip Cave, Mai Wong, Tim Simmance, Gerhard Smith, Cathy Walsh, Mike Bell, and Jane Berezynskyj for assistance in planning and hardware bids; Richard Matt for support with the technical aspects of proposals and for commissioning independent penetration testing; Chess Denman and CPFT's Information Governance Committee for approvals; Jonathon Artingstall for support; Jenny Nelder for project management; Joe Kearney for developing some tasks; Trish Barker-Barrett, Jules Mackenzie and colleagues for the perinatal service deployment; Rosemary Boyle and Ted Krawec for advice; multiple copyright holders (see software documentation) for permission to use specific tasks; and three referees for helpful suggestions.

REFERENCES

1. Guo T, Xiang Y-T, Xiao L, Hu C-Q, Chiu HFK, Ungvari GS, et al. Measurement-based care versus standard care for major depression: a randomized controlled trial with blind raters. *Am J Psychiatry*. (2015) 172:1004–13. doi: 10.1176/appi.ajp.2015.14050652
2. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary care evaluation of mental disorders. Patient health questionnaire. *JAMA J Am Med Assoc*. (1999) 282:1737–44. doi: 10.1001/jama.282.18.1737
3. Hsieh S, Schubert S, Hoon C, Mioshi E, Hodges JR. Validation of the Addenbrooke's Cognitive Examination III in frontotemporal dementia and Alzheimer's disease. *Dement Geriatr Cogn Disord*. (2013) 36:242–50. doi: 10.1159/000351671
4. Blackwell AD, Sahakian BJ, Vesey R, Semple JM, Robbins TW, Hodges JR. Detecting dementia: novel neuropsychological markers of preclinical Alzheimer's disease. *Dement Geriatr Cogn Disord*. (2004) 17:42–8. doi: 10.1159/000074081
5. Professional Record Standards Body. *PRSB Standards for the Structure and Content of Health and Care Records*. (2018). Available online at: https://

www.rcplondon.ac.uk/projects/outputs/standards-clinical-structure-and-content-patient-records (accessed June 27, 2020).
6. Academy of Medical Royal Colleges, Health & Social Care Information Centre. *Standards for the Clinical Structure and Content of Patient Records*. London: Royal College of Physicians (2013).
7. Rayner L, Matcham F, Hutton J, Stringer C, Dobson J, Steer S, et al. Embedding integrated mental health assessment and management in general hospital settings: feasibility, acceptability and the prevalence of common mental disorder. *Gen Hosp Psychiatry*. (2014) 36:318–24. doi: 10.1016/j.genhosppsych.2013.12.004
8. UK National Institute for Health and Care Excellence. *Guide to the Methods of Technology Appraisal 2013*. (2013). Available online at: https://www.nice.org.uk/process/pmg9/chapter/foreword (accessed June 27, 2020).
9. UK National Health Service. *NHS Digital Service Manual: NHS Service Standard: 12. Make New Source Code Open*. (2019). Available online at: https://service-manual.nhs.uk/service-standard/12-make-new-source-code-open (accessed June 29, 2020).
10. Yackel TR. How the open-source development model can improve medical software. *Stud Health Technol Inform*. (2001) 84:68–72. doi: 10.3233/978-1-60750-928-8-68
11. Leong TY, Kaiser K, Miksch S. Free and open source enabling technologies for patient-centric, guideline-based clinical decision support: a survey. *Yearb Med Inform*. (2007) 16, 74–86. doi: 10.1055/s-0038-1638529
12. Shah J, Rajgor D, Pradhan S, McCready M, Zaveri A, Pietrobon R. Electronic data capture for registries and clinical trials in orthopaedic surgery: open source versus commercial systems. *Clin Orthop*. (2010) 468:2664–71. doi: 10.1007/s11999-010-1469-3
13. NHS Connecting for Health. *NHS Information Governance: Guidelines on Use of Encryption to Protect Person Identifiable and sensitive information*. (2008). Available online at: https://www.webarchive.org.uk/wayback/en/archive/20130425190519/https://www.connectingforhealth.nhs.uk/systemsandservices/infogov/security/encryption.pdf (accessed June 27, 2020).
14. NHS Connecting for Health. *IG Toolkit Version 8: Information Security Assurance Requirement 322: Detailed Guidance on Secure Transfers*. (2010). Available online at: https://web.archive.org/web/20211026162854/https://docplayer.net/storage/27/10830978/1635269300/eH88Sxjne7iWx1llUrbfMA/10830978.pdf (accessed June 27, 2020).
15. NHS Connecting for Health. *NHSmail Mobile Configuration Guide: Apple iPhone* London: NHS Connecting for Health (2011).
16. NHS Digital. *Approved Cryptographic Algorithms: Good Practice Guideline*. (2016). Available online at: https://webarchive.nationalarchives.gov.uk/ukgwa/20161021125701/https://systems.digital.nhs.uk/infogov/security/infrasec/gpg/acs.pdf (accessed October 26, 2021).
17. NHS Connecting for Health, British Medical Association. *Joint Guidance on Protecting Electronic Patient Information*. (2008). Available online at: https://datatracker.ietf.org/doc/html/rfc5246 (accessed June 27, 2020).
18. UK. Data Protection Act 2018. (2018). Available online at: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted (accessed October 26, 2021).
19. Australian Passport Office. *Sex and Gender Diverse Passport Applicants*. (2013). Available online at: https://web.archive.org/web/20140912083901/https://www.passports.gov.au/web/sexgenderapplicants.aspx (accessed October 26, 2021).
20. Cardinal RN, Burchell M. *CamCOPS documentation*. (2020). Available online at: https://camcops.readthedocs.io/ (accessed October 26, 2021).
21. Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychol Med*. (1992) 22:465–86. doi: 10.1017/S0033291700030415
22. Lewis G. Assessing psychiatric disorder with a human interviewer or a computer. *J Epidemiol Community Health*. (1994) 48:207–10. doi: 10.1136/jech.48.2.207
23. Rogers RD, Tunbridge EM, Bhagwagar Z, Drevets WC, Sahakian BJ, Carter CS. Tryptophan depletion alters the decision-making of healthy volunteers through altered processing of reward cues. *Neuropsychopharmacol*. (2003) 28:153–62. doi: 10.1038/sj.npp.1300001
24. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing (2019).
25. Hipp DR. *SQLite*. (2000). Available online at: https://www.sqlite.org/ (accessed October 26, 2021).
26. Cardinal RN. Clinical Records Anonymisation and Text Extraction (CRATE): an open-source software system. *BMC Med Inform Decis Mak*. (2017) 17:50. doi: 10.1186/s12911-017-0437-1
27. Hunt A. *The Pragmatic Programmer: From Journeyman to Master*. Reading, MA: Addison-Wesley (2000).
28. Salawa P. *SQLiteStudio*. (2018). Available online at: https://sqlitestudio.pl/ (accessed October 26, 2021).
29. RStudio Team. *RStudio: Integrated Development for R*. Boston, MA: RStudio, PBC (2020). Available online at: https://www.rstudio.com/
30. HL7 International. *Health Level Seven (HL7) version 2*. (2015). Available online at: https://www.hl7.org/ (accessed October 26, 2021).
31. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. (2009) 42:377–81. doi: 10.1016/j.jbi.2008.08.010
32. Burns SS, Browne A, Davis GN, Rimrodt SL, Cutting LE. *PyCap (version 1.0.2)*. Nashville, TN: Vanderbilt University; Childrens Hospital of Philadelphia (2016). Available online at: https://pycap.readthedocs.io/
33. HL7.org. *HL7 FHIR [Fast Healthcare Interoperability Resources] Release 4*. (2019). Available online at: https://hl7.org/fhir/ (accessed October 26, 2021).
34. UK. *Public Records Act 1958*. (1958). Available online at: https://www.legislation.gov.uk/ukpga/Eliz2/6-7/51 (accessed October 26, 2021).
35. UK Department of Health. *Records Management: NHS Code of Practice, Part 1*. (2006). Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/547055/Records_Management_-_NHS_Code_of_Practice_Part_1.pdf (accessed June 27, 2020).
36. UK Department of Health. *Records Management: NHS Code of Practice, Part 2 (2nd edition)*. (2009). Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/547054/Records_Management_-_NHS_Code_of_Practice_Part_2_second_edition.pdf.pdf (accessed June 27, 2020).
37. UK. *Data Protection Act 1998*. (1998). Available online at: https://www.legislation.gov.uk/ukpga/1998/29 (accessed October 26, 2021).
38. Stroustrup B. *The C++ Programming Language*. Reading, MA: Addison-Wesley. (1986).
39. The Qt Company. Qt. (2017). Available online at: https://www.qt.io/ (accessed October 26, 2021).
40. Zetetic, LLC. *SQLCipher*. (2017). Available online at: https://www.zetetic.net/sqlcipher/ (accessed October 26, 2021).
41. The OpenSSL Project. *OpenSSL Toolkit*. (2016). Available online at: https://www.openssl.org/ (accessed October 26, 2021).
42. Young E. *SSLeay*. (1998). Available online at: https://www.cryptsoft.com/ (accessed October 26, 2021).
43. van Rossum G. *Python Reference Manual*. Centrum voor Wiskunde en Informatica. Amsterdam: Netherlands (1995). Available online at: https://www.python.org/ (accessed October 26, 2021).
44. The CherryPy Team. *CherryPy 18.1.0*. (2018). Available online at: https://cherrypy.org/ (accessed October 26, 2021).
45. Chesneau B. *Gunicorn 19.8.1*. (2018). Available online at: https://gunicorn.org/ (accessed October 26, 2021).
46. The Apache Software Foundation. *Apache HTTP Server 2.4.20*. (2016). Available online at: https://httpd.apache.org/ (accessed October 26, 2021).
47. Torvalds L. *Linux*. (1991). Available online at: https://www.linuxfoundation.org/ (accessed October 26, 2021).
48. Oracle Corporation. *MySQL 8.0 Reference Manual*. (2020). Available online at: https://dev.mysql.com/doc/refman/8.0/en/ (accessed June 27, 2020).
49. MariaDB Foundation. *MariaDB Server*. (2020). Available online at: https://mariadb.org/ (accessed October 26, 2021).
50. Bayer M. *SQLAlchemy*. (2016). Available online at: https://www.sqlalchemy.org/ (accessed October 26, 2021).
51. Merkel D. *Docker: Lightweight Linux Containers for Consistent Development and Deployment. Linux J*. (2014). Available online at: https://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment (accessed June 27, 2020).

52. Cardinal RN, Burchell M. *CamCOPS source code*. (2020). Available online at: https://github.com/RudolfCardinal/camcops (accessed October 26, 2021).
53. Apple Inc. *iOS Developer Program Enterprise License Agreement* (2021). Available online at: https://web.archive.org/web/20211020035029/https://developer.apple.com/support/downloads/terms/apple-developer-enterprise-program/Apple-Developer-Enterprise-Program-License-Agreement-20210607-English.pdf (accessed October 26, 2021).
54. Codd EF. A relational model of data for large shared data banks. *Commun ACM*. (1970) 13:377–87. doi: 10.1145/362384.362685
55. Apple Inc. *iOS Developer Program License Agreement* (2021). Available online at: https://web.archive.org/web/20211020044325/https://developer.apple.com/support/downloads/terms/apple-developer-program/Apple-Developer-Program-License-Agreement-20210607-English.pdf (accessed October 26, 2021).
56. Provos N, Mazieres D. A future-adaptable password scheme. In: *Proceedings of 1999 USENIX Annual Technical Conference*. (2020). p. 81-92. Available online at: https://www.usenix.org/legacy/events/usenix99/provos/provos_html/node1.html (accessed June 27, 2020).
57. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines (CDDG)*. (1992). Available online at: https://www.who.int/entity/classifications/icd/en/bluebook.pdf (accessed December 7, 2007).
58. World Health Organization, US National Center for Health Statistics, US Ceters for Medicare and Medicaid Services. *International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)*. (1979). Available online at: https://www.cdc.gov/nchs/icd/icd9cm.htm (accessed June 10, 2014).
59. Free Software Foundation. *GNU General Public License*. (2007). Available online at: https://www.gnu.org/licenses/ (accessed October 26, 2021).
60. Pfizer Inc. *PHQ Screeners*. (2020). Available online at: https://www.phqscreeners.com/ (accessed June 29, 2020).
61. Nelson HE. *National Adult Reading Test (NART): For the Assessment of Premorbid Intelligence in Patients with Dementia: Test Manual*. Windsor: NFER-Nelson (1982).
62. Nelson HE. *Use of the National Adult Reading Test*. Personal communication to Rudolf Cardinal (May 30, 2013).
63. Bell V, Halligan PW, Ellis HD. The Cardiff Anomalous Perceptions Scale (CAPS): a new validated measure of anomalous perceptual experience. *Schizophr Bull*. (2006) 32:366–77. doi: 10.1093/schbul/sbj014
64. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry*. (1961) 4:561–71. doi: 10.1001/archpsyc.1961.01710120031004
65. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. *Br J Psychiatry J Ment Sci*. (1987) 150:782–6. doi: 10.1192/bjp.150.6.782
66. Royal College of Psychiatrists. *The Edinburgh Postnatal Depression Scale*. Personal communication to Rudolf Cardinal (August 6, 2013).
67. National Security Agency CNSS Secretariat. *National Policy on the Use of the Advanced Encryption Standard (AES) to Protect National Security Systems and National Security Information*. (2003). Available online at: https://csrc.nist.gov/groups/ST/toolkit/documents/aes/CNSS15FS.pdf (accessed June 27, 2020).
68. Multiple authors. *Are There Actually Any Advantages to Android Full-Disk Encryption*? (2012). Available online at: https://security.stackexchange.com/questions/10529/are-there-actually-any-advantages-to-android-full-disk-encryption (accessed March 28, 2014).
69. Google Inc., Open Handset Alliance. *Android: Notes on the Implementation of Encryption in Android 3.0*. (2014). Available online at: https://web.archive.org/web/20140530175700/https://source.android.com/devices/tech/encryption/android_crypto_implementation.html (accessed June 27, 2020).
70. Mogull R. *How to Use Your iPad Securely*. (2011). Available online at: https://www.macworld.com/article/1160313/iPad_security.html (accessed June 27, 2020).
71. Apple Inc. *iOS Security*. (2012). Available online at: https://web.archive.org/web/20140405001141/https://www.apple.com/ipad/business/docs/iOS_Security_Oct12.pdf (accessed June 27, 2020).
72. Google Inc., Open Handset Alliance. *Android: System Permissions*. (2020). Available online at: https://developer.android.com/guide/topics/security/permissions.html (accessed June 27, 2020).
73. Network Working Group. *The Transport Layer Security (TLS) Protocol Version 1.2 (RFC 5246)*. (2008). Available online at: https://tools.ietf.org/html/rfc5246 (accessed October 26, 2021).
74. Karwin B. *SQL Antipatterns: Avoiding the Pitfalls of Database Programming*. Raleigh, NC: Pragmatic Bookshelf (2010).
75. The MITRE Corporation. *CWE-384: Session Fixation*. (2008). Available online at: https://cwe.mitre.org/data/definitions/384.html (accessed June 27, 2020).
76. Cullen AJ, Mann I. Hacking the human: countering the socially engineered attack. *J Inf Warf*. (2008) 7:24–35. Available online at: https://www.jstor.org/stable/26486865
77. Mann I. *Hacking the Human: Social Engineering Techniques and Security Countermeasures*. Aldershot: Gower. (2008).
78. Hipp DR. *SQLite: Most Widely Deployed SQL Database*. (2014). Available online at: https://sqlite.org/mostdeployed.html (accessed March 28, 2014).
79. Hipp DR. *SQLite: How SQLite Is Tested*. (2014). Available online at: https://www.sqlite.org/testing.html (accessed June 27, 2020).
80. Hipp DR. *SQLite: Atomic Commit in SQLite*. (2014). Available online at: https://www.sqlite.org/atomiccommit.html (accessed June 27, 2020).
81. Lyu MR. ed. *Handbook of Software Reliability Engineering*. Los Alamitos, CA: IEEE Computer Society Press (1996).
82. Nanz S, Furio CA. A comparative study of programming languages in Rosetta code. *ICSE 15 Proc 37th Int Conf Softw Eng*. (2015) 1:778–88. doi: 10.1109/ICSE.2015.90
83. Barrows RC Jr, Clayton PD. Privacy, confidentiality, and electronic medical records. *J Am Med Inform Assoc*. (1996) 3:139–48. doi: 10.1136/jamia.1996.96236282
84. Belojević G, Ohrström E, Rylander R. Effects of noise on mental performance with regard to subjective noise sensitivity. *Int Arch Occup Environ Health*. (1992) 64:293–301. doi: 10.1007/BF00378288
85. Logie RH, Gilhooly KJ, Wynn V. Counting on working memory in arithmetic problem solving. *Mem Cognit*. (1994) 22:395–410. doi: 10.3758/BF03200866
86. Oberauer K, Demmrich A, Mayr U, Kliegl R. Dissociating retention and access in working memory: an age-comparative study of mental arithmetic. *Mem Cognit*. (2001) 29:18–33. doi: 10.3758/BF03195737
87. Council of the European Union. *Council Directive 93/42/EEC ("Medical Devices Directive"), Amended by Directive 98/79/EC, Directive 2000/70/EC, Directive 2001/104/EC, Regulation (EC) No. 1882/2003, Directive 2007/47/EC*. (2007). Available online at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1993L0042:20071011:en:PDF (accessed June 27, 2020).
88. UK Medicines and Healthcare Products Regulatory Agency. *Clinical Investigations of Medical Devices - Guidance for Manufacturers*. (2020). Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/865135/Guidance_for_mfrs_on_clinical_trials_January_2020.pdf (accessed June 27, 2020).
89. UK Medicines and Healthcare Products Regulatory Agency. *Medical Devices: Software Applications (apps)*. (2020). Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/890025/Software_flow_chart_Ed_1-06_FINAL.pdf (accessed June 27, 2020).
90. NHS Connecting for Health. *NHS Information Governance: Laptop Security Policy*. (2008). Available online at: https://web.archive.org/web/20140820040233/https://www.igt.hscic.gov.uk/WhatsNewDocuments/Exemplar%20Laptop%20Security%20Policy.doc (accessed June 27, 2020).
91. Google Inc., Open Handset Alliance. *Android: Security: Encryption*. (2020). Available online at: https://source.android.com/security/encryption (accessed June 27, 2020).
92. Matheson J. *CEL 25: NHS Scotland Mobile Data Protection Standard*. (2012). Available online at: https://www.sehd.scot.nhs.uk/mels/CEL2012_25.pdf (accessed June 27, 2020).
93. UK Department of Health. *Research Governance Framework for Health and Social Care: Second Edition*. (2005). Available online at: https://www.gov.

uk/government/publications/research-governance-framework-for-health-and-social-care-second-edition (accessed June 27, 2020).
94. University of Cambridge. *University of Cambridge Policy on the Ethics of Research Involving Human Participants and Personal Data*. (2016). Available online at: https://www.research-integrity.admin.cam.ac.uk/files/policy_on_the_ethics_of_research_involving_human_participants_and_personal_data_oct_2016.pdf (accessed June 27, 2020).
95. Appcelerator Inc. *Appcelerator Titanium Mobile Development Environment*. Mountain View, CA (2014). Available online at: https://www.appcelerator.com/Titanium/ (accessed June 27, 2020).
96. Khandaker GM. *IL-6 inhibition In Patients With Depression and Low-Grade Inflammation: The Insight Study*. (2018). Available at: https://www.isrctn.com/ISRCTN16942542 (accessed June 27, 2020).
97. Barnett JH, Blackwell AD, Sahakian BJ, Robbins TW. The Paired Associates Learning (PAL) test: 30 years of CANTAB translational neuroscience from laboratory to bedside in dementia research. *Curr Top Behav Neurosci.* (2016) 28:449–74. doi: 10.1007/7854_2015_5001
98. Cambridge Cognition. *CANTAB*. Cambridge: Cambridge Cognition. (2020). Available online at: https://www.cambridgecognition.com/ (accessed October 26, 2021).
99. Goodday SM, Atkinson L, Goodwin G, Saunders K, South M, Mackay C, et al. The True Colours remote symptom monitoring system: a decade of evolution. *J Med Internet Res.* (2020) 22:15188. doi: 10.2196/15188
100. Lamb RC, Matcham F, Turner MA, Rayner L, Simpson A, Hotopf M, et al. Screening for anxiety and depression in people with psoriasis: a cross-sectional study in a tertiary referral setting. *Br J Dermatol.* (2017) 176:1028–34. doi: 10.1111/bjd.14833
101. Vanderbilt University. *REDCap License Terms*. (2020). Available online at: https://projectredcap.org/partners/termsofuse/ (accessed June 25, 2020).
102. Wikipedia. *REDCap*. (2019). Available online at: https://en.wikipedia.org/wiki/REDCap (accessed June 26, 2020).
103. Vanderbilt University. *REDCap FAQ*. (2020). Available online at: https://projectredcap.org/about/faq/ (accessed June 26, 2020).
104. Microsoft. *Microsoft Forms (Office 365)*. Redmond, WA: Microsoft Corporation (2016).
105. Torous J, Kiang MV, Lorme J, Onnela J-P. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Ment Health.* (2016) 3:e16. doi: 10.2196/mental.5165

Author Disclaimer: The views expressed are those of the author and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

Conflict of Interest: RC consults for Campden Instruments Ltd., in the area of research software and receives royalties from Cambridge University Press, Cambridge Enterprise, and Routledge.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership