frontiers | Research Topics

Digital therapeutics: Using software to treat, manage, and prevent disease

Edited by
Louis N. Awad, Kirsten Smayda, Sabrina R. Taylor, Terry D. Ellis and Tim Campellone

Coordinated by
Brian Harris

Published in
Frontiers in Digital Health
Frontiers in Medicine

ISSN 1664-8714
ISBN 978-2-8325-3657-5
DOI 10.3389/978-2-8325-3657-5

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 therapeutics: Using software to treat, manage, and prevent disease

Topic editors

Louis N. Awad — Boston University, United States
Kirsten Smayda — MedRhythms, United States
Sabrina R. Taylor — MedRhythms, Inc., United States
Terry D. Ellis — Boston University, United States
Tim Campellone — Woebot Labs Inc., United States

Topic coordinator

Brian Harris — MedRhythms, Inc., United States

Citation

Awad, L. N., Smayda, K., Taylor, S. R., Ellis, T. D., Campellone, T., Harris, B., eds. (2023). *Digital therapeutics: Using software to treat, manage, and prevent disease*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3657-5

Table of contents

frontiers | Frontiers in Digital Health

TYPE Editorial
PUBLISHED 25 September 2023
DOI 10.3389/fdgth.2023.1261124

OPEN ACCESS

EDITED BY
Max A. Little,
University of Birmingham, United Kingdom

REVIEWED BY
Stephan H. Schug,
DGG German eHealth Association, Germany
Markus Wolf,
University of Zurich, Switzerland

*CORRESPONDENCE
Kirsten E. Smayda
ksmayda@medrhythms.com

RECEIVED 18 July 2023
ACCEPTED 07 September 2023
PUBLISHED 25 September 2023

CITATION
Smayda KE, Campellone T, Taylor SR, Harris B, Ellis TD and Awad LN (2023) Editorial: Digital therapeutics: using software to treat, manage, and prevent disease.
Front. Digit. Health 5:1261124.
doi: 10.3389/fdgth.2023.1261124

Editorial: Digital therapeutics: using software to treat, manage, and prevent disease

Kirsten E. Smayda[1]*, Tim Campellone[2], Sabrina R. Taylor[1], Brian Harris[1], Terry D. Ellis[3] and Louis N. Awad[3]

[1]MedRhythms, Portland, ME, United States, [2]Woebot Labs Inc., San Francisco, CA, United States, [3]College of Health and Rehabilitation Sciences, Sargent College, Boston University, Boston, MA, United States

KEYWORDS

DTx, digital therapeutics, software, hardware, medical device, intervention, symptom monitoring, clinical tool

Editorial on the Research Topic
Digital therapeutics: using software to treat, manage, and prevent disease

The Research Topic, *Digital Therapeutics: Using Software to Treat, Manage, and Prevent Disease*, was developed to build a collection of evidence that exhibits and incites innovation of novel digital therapeutics (DTx) and their component technologies, and to showcase outcomes that meet the needs of all stakeholders in the process of commercializing digital therapeutics. This Topic and Editorial primarily focuses on the U.S.A. market, although two articles are from teams in Canada, Germany, and Switzerland, and is particularly timely given the state of technological progress and the digital therapeutics industry.

We live in a time when digital technology has advanced such that interventions can be delivered in a clinically meaningful way. Technology can now support high-fidelity data collection (e.g., biometric, physiological, and kinematic), appropriate controls, feedback loops, and detailed visual and auditory resolution—all to support the delivery of an intervention. Macrotrends in healthcare also make it an ideal moment for this Research Topic: an increasing momentum and focus on reimbursement of DTx, and the recent rise of telehealth due to the COVID-19 pandemic. While this Topic focuses on DTx, defined by ISO/TR 11147:2023(en) and the DTx Alliance as "…health software intended to treat or alleviate a disease, disorder, condition, or injury by generating and delivering a medical intervention that has a demonstrable positive therapeutic impact on a patient's health," (1, 2) other products and services within the digital health ecosystem, including patient symptom monitoring and clinical support tools, are included in this topic and represent the landscape within which DTx exist.

This Research Topic highlights work from commercial, academic, and collaborations across both entities and includes a range of clinical and care populations including depression, multiple sclerosis, opioid-based pain management, stroke, language, cognition, dementia, and attention-deficit/hyperactivity disorder. Amongst the articles that describe results of an intervention study, Kulikov et al. addresses the gap between need and access to evidence-based services for adolescent mental health by presenting initial, positive, evidence from a randomized controlled trial on the feasibility and acceptability of a digital therapeutic, Spark, to treat depression in adolescents. Cuyler et al. finds that a 28-day home-based Capnometry Guided Respiratory Intervention could support symptom

reduction and adherence in people with panic disorder and post-traumatic stress disorder. DTx also offer an unprecedented view into engagement patterns during treatment. Heusser et al. describes a machine learning model that measures quality of user interactions and intended use, and provides a helpful contextual framing for DTx in the introduction, as well. Liu et al. characterizes the relationship between engagement/dosage and improvements across 13 skill domains in people who had a stroke that resulted in speech, language, and cognitive deficits, finding that a higher dosage is related to greater improvement in in-home therapy outcomes over 6 months.

Two articles provide original research related to tracking symptoms and outcomes in real-time. Chen et al. sought to "identify divergent factors that influence subjectively and objectively measured cognitive functioning in real time in people with multiple sclerosis" and DiCarlo et al. finds that "SMS texting is a feasible method for gathering outcomes after stroke at scale to evaluate the efficacy of acute stroke treatments." Two articles focus on supporting the care team of patients. Melvin et al. provides a "common nomenclature" to be used by clinicians and developers to support interpretation and application of artificial intelligence models. And Braun et al. describes PHREND®, an algorithm updated with new data and can "predict freedom of relapse and 3-months confirmed disability progression" to support decision-making between patient and clinician.

Two articles provide forward-looking perspectives. In Watson et al., the authors elucidate the gaps within the evolving and dynamic regulatory landscape for how prescription digital therapeutics (PDTs) are currently evaluated for safety and efficacy and regulated by the U.S. Food and Drug Administration (FDA). Russo et al. presents a "theoretical background, rationale, and development plans" for a music-based digital therapeutic to manage agitation and anxiety in people with dementia.

Lastly, one article, Giravi et al., reviews the literature on the "clinical evidence of digital interventions delivered via virtual reality and mobile apps to improve opioid-based analgesia" and concluded that they can improve pain scores compared to "treatment as usual".

The editorial team experienced two learnings, in particular, that might benefit the digital therapeutics industry to consider. First, Institutional Review Boards (IRBs) are important partners in conducting research in both non-commercial and commercial contexts. If a company or research group intends to publish data from commercial users of a product or service, it is important that the researchers still submit their protocol to an IRB for exempt status determination under 45 CFR § 46.104(d)(4), if using de-identified data. This is ideally done prospectively before the data is collected, but can also be done retroactively. Having an exemption determination in-hand can help expedite the review process, and could also be complemented by clear language in a privacy notice or terms of service informing the user that their data may be used for research and publication purposes.

Second, this Research Topic is missing research that includes payers and implementation studies. In the future, we encourage researchers to consider working with payers because of the critical role they play in reimbursing digital therapeutics, and also conducting implementation research to identify and plan for barriers to successful adoption. There were, however, several articles that represented collaborations across types of institutions. For example, the team from Pear Therapeutics co-authored their manuscript with the Devices division at Sanofi, a global pharmaceutical company (Watson et al.); the team at Freespira worked with the Laboratory for the Study of Anxiety Disorders at The University of Texas at Austin (Cuyler et al.); Chen et al. involved a collaboration across Rutgers, University of Illinois, and the Kessler Foundation; and Russo et al., involved a collaboration across multiple universities in Toronto, LUCID Inc., and Right to Music. These collaborations paint an evolving picture of the collaborations needed to bring evidence-based digital therapeutics that are rooted in good science and the reality of bringing DTx to market.

In conclusion, the editorial team is grateful to everyone who submitted their article for consideration in this important Research Topic. The quality and breadth of research in this Topic bolsters the foundation of evidence for not only the products and services represented in this Research Topic, but also the digital therapeutics industry at large. It is the editorial team's earnest hope that the work presented here will add to the momentum for digital therapeutics to be adopted by the many stakeholders in healthcare, including providers, payers, and patients.

Author contributions

KS: Conceptualization, Data curation, Writing – original draft, Writing – review & editing. TC: Writing – review & editing, Writing – original draft. ST: Writing – review & editing, Writing – original draft. BH: Writing – review & editing, Conceptualization. TE: Writing – review & editing. LA: Conceptualization, Writing – review & editing.

Conflict of interest

LA is a paid advisor for MedRhythms. BH is the CEO of and is employed by MedRhythms and holds stock options in the company. ST is employed by MedRhythms and holds stock options in the company. KS is employed by MedRhythms and holds stock options in the company. TC is employed by Woebot Health and holds stock options in Click Therapeutics. TE has received grant funding from MedRhythms.

Publisher's note

References

1. SO/TR 11147:2023(en). Health informatics—Personalized digital health—Digital therapeutics health software systems. (2023). Available at: https://www.iso.org/standard/83767.html (Accessed August 15, 2023).

2. Digital Therapeutics Alliance. What is a DTx?. Available at: https://dtxalliance.org/understanding-dtx/what-is-a-dtx/ (Accessed August 15, 2023).

frontiers in Digital Health

ORIGINAL RESEARCH
published: 11 March 2022
doi: 10.3389/fdgth.2022.856829

PHREND®—A Real-World Data-Driven Tool Supporting Clinical Decisions to Optimize Treatment in Relapsing-Remitting Multiple Sclerosis

Stefan Braune[1]*, *Elisabeth Stuehler*[2], *Yanic Heer*[2], *Philip van Hoevell*[2], *Arnfin Bergmann*[1] and *NeuroTransData Study Group*[1]

[1] NeuroTransData, Neuburg an der Donau, Germany, [2] PwC Data and Analytics, Zurich, Switzerland

OPEN ACCESS

Edited by:
Tim Campellone,
University of California, Berkeley,
United States

Reviewed by:
Leif Simmatis,
University Health Network
(UHN), Canada
Rosalia Dacosta-Aguayo,
University of Barcelona, Spain

***Correspondence:**
Stefan Braune
sbraune@neurotransdata.com

Specialty section:
This article was submitted to
Personalized Medicine,
a section of the journal
Frontiers in Digital Health

Received: 17 January 2022
Accepted: 09 February 2022
Published: 11 March 2022

Citation:
Braune S, Stuehler E, Heer Y, van
Hoevell P, Bergmann A and
NeuroTransData Study Group (2022)
PHREND®—A Real-World
Data-Driven Tool Supporting Clinical
Decisions to Optimize Treatment in
Relapsing-Remitting Multiple
Sclerosis.
Front. Digit. Health 4:856829.
doi: 10.3389/fdgth.2022.856829

Background: With increasing availability of disease-modifying therapies (DMTs), treatment decisions in relapsing-remitting multiple sclerosis (RRMS) have become complex. Data-driven algorithms based on real-world outcomes may help clinicians optimize control of disease activity in routine praxis.

Objectives: We previously introduced the PHREND® (Predictive-Healthcare-with-Real-World-Evidence-for-Neurological-Disorders) algorithm based on data from 2018 and now follow up on its robustness and utility to predict freedom of relapse and 3-months confirmed disability progression (3mCDP) during 1.5 years of clinical practice.

Methods: The impact of quarterly data updates on model robustness was investigated based on the model's C-index and credible intervals for coefficients. Model predictions were compared with results from randomized clinical trials (RCTs). Clinical relevance was evaluated by comparing outcomes of patients for whom model recommendations were followed with those choosing other treatments.

Results: Model robustness improved with the addition of 1.5 years of data. Comparison with RCTs revealed differences <10% of the model-based predictions in almost all trials. Treatment with the highest-ranked (by PHREND®) or the first-or-second-highest ranked DMT led to significantly fewer relapses ($p < 0.001$ and $p < 0.001$, respectively) and 3mCDP events ($p = 0.007$ and $p = 0.035$, respectively) compared to non-recommended DMTs.

Conclusion: These results further support usefulness of PHREND® in a shared treatment-decision process between physicians and patients.

Keywords: multiple sclerosis (MS), personalized medicine, disease modifying agent, real word data, treatment, effectiveness

INTRODUCTION

Shared clinical decision processes in multiple sclerosis (MS) require multidimensional, complex interactions between physicians and patients. There is an asymmetry in knowledge between professionals and laymen regarding available MS therapies, and it is difficult for physicians to clearly convey differences between different treatment options to patients during the limited time of the medical practice visit. This can create diverging treatment expectations between patients and physicians (1–3), impair shared decision processes, and hinder necessary adherence. Personalized data driven clinical prediction tools with informative visualization can facilitate these discussions and improve the joint doctor-patient efforts to implement the individually most effective DMT, yet no such tools were available in the past for routine use.

One barrier to the development of efficient decision-support tools is a lack of relevant data sources. Although the number of choices of different disease-modifying therapies (DMTs) for relapsing-remitting MS (RRMS) with diverse treatment mechanism increases, information from randomized clinical trials (RCTs) in RRMS usually remains limited to a single head-to-head study with one of the available DMTs. The total observation time within such trials is usually 2 years, and no information can be derived regarding next-best treatment options or allowing for patients' preferences. The RCT's two active arms perspective thus provides only limited information for overall longer-term treatment options and more complex treatment requirements in an individual patient.

Real-world data (RWD) and advanced statistical methods are utilized in growing numbers of comparative effectiveness studies aiming to fill this gap (4–8), but also these efforts remain confined to a retrospective cohort view and are not suitable to support personalized decision strategies for clinical routine.

An alternative idea is to base models on "objective" measures and clinical predictors, such as the biomarker neurofilament light chain. This marker has indeed enabled first insights into probable dynamics of relapsing remitting multiple sclerosis (RRMS) also from a cohort perspective [for review see (9)]. It has also shown predictive potential when measured as a *post-hoc* response marker to DMTs (10). In addition, enzyme-linked immunospot assay (ELISPOT) testing of B-cell activity has been shown to successfully predict the likelihood of individual DMT responsiveness to interferons or glatiramer acetate (11). Despite these advances, it appears highly unlikely that single or sets of biomarkers will become available in the foreseeable future to support personalized treatment decisions in all MS patients and for all DMTs.

To meet the multiple demands of improved communication between patients and doctors and data-driven decision making based on real-world experience, NeuroTransData (NTD) and PricewaterhouseCoopers (PwC) embarked on the development of a mathematical algorithm based on real-world data from the NTD MS registry to calculate the probabilities of patients with RRMS in diverse clinical situations to remain free of relapse and free of 3 months confirmed disability progression (3mCDP) for available DMTs.

A previous publication provided comprehensive information on methods, validity and robustness of the predictive models implemented in the web-based tool called "Predictive Health Care with Real-World Evidence in Neurological Disorders" (PHREND®) (12).

In brief, we implemented two hierarchical Bayesian generalized linear models (GLMs) to predict the probabilities of (a) freedom of relapse activity, (b) freedom of 3 months confirmed disability progression (CDP) for every of the currently available DMTs after a switch from a previous DMT. The predictive framework was based on RWD collected in the NTD MS registry consisting of clinical data including patient characteristics and disease history. Predictors used for the predictive models are: age, gender, duration of RRMS, previous therapy and its duration, indicator if one of the two previous therapies was second line, EDSS total score, number of relapses within last 12 months, time since last relapse. Based on these individual information probabilities for both effectiveness parameters can be calculated for a prospective period up to 4 years, scalable at the discretion of the user.

Assessment of the model performance demonstrated that both models provided robust and accurate predictions and that both models generalized to new patients and clinical sites. The predictive relapse model achieved an average out-of-sample C-Index of 0.65 and an average out-of-sample mean squared error (MSE) of 0.76 relapses. The predictive CDP model achieved an average C-Index of 0.58 and an average out-of-sample MSE of 0.12 CDPs. Robustness against different choices of priors was proven by the fact that changing the prior distributions did not influence the predicted therapy ranking.

Accounting for individual clinical patient characteristics, the resulting predictive probabilities are intended to be provided during the discussion of potential change in DMT to support the shared decision process between physicians and patients.

We herein report on the clinical value and further validation of PHREND® with three new sets of results: (1) update of the models' performance over time, after more data of the ongoing NTD MS registry collection has been added regularly to re-train the models since the initial publication in 2020 (12), which was based on a data cut from 2018; (2) external validation by comparing of PHREND® predictions to RCT results based on current models, including new DMTs which have entered the German treatment landscape since the last data cut and were subsequently integrated into the training sets of the model; (3) new assessment of clinical relevance of the recommendations based on whether patients received DMTs recommended by PHREND® or not.

MATERIALS AND METHODS

Database and Parameter

NTD MS Registry

This study employed real-world data recorded in the NTD MS registry. NTD is a Germany-wide network of physicians in the fields of neurology and psychiatry that was founded in 2008. Each practice is certified according to network-specific and ISO 9001

criteria. Compliance with these criteria is audited annually by an external certified audit organization.

Codes uniquely identifying patients are managed by the Institute for Medical Information Processing, Biometry and Epidemiology [Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE)] at the Ludwig Maximilian University in Munich, Germany, acting as an external trust center. Written informed consent is obtained from each patient providing data for the MS registry. The data acquisition protocol described above was approved by the ethical committee of the Bavarian State Medical Association (Bayerische Landesärztekammer; June 14, 2012, ID 12114) and re-approved by the ethical committee of the Medical Association of North-Rhine (Ärztekammer Nordrhein; April 25, 2017, ID 2017071).

Demographic and clinical parameters of MS patients are captured in real time with an average of 3.7 Expanded Disability Status Scale (EDSS) assessments per patient year. Data quality is monitored by the NTD data management team, and data inputs are checked for inconsistencies and errors manually and by using an automated error-analysis program. Additionally, automated and manually executed queries are implemented to check for inconsistencies and request missing information. All data are pseudonymized and pooled to form the NTD MS database.

Data Extraction Period, Numbers of Patients

The current study is based on the same original dataset as described in the previous publication (12) and additional quarterly data cuts extracted from the NTD MS registry between July 1, 2018 and October 2020. After quality control and application of inclusion criteria as described (12), this dataset includes 3,119 patients.

Predictive Models and Selection of Predictors

PHREND® supports treatment decisions for optimization of treatment switches from a current DMT in RRMS, which needs to be discontinued due to lack of efficacy or adverse events, to another therapy. The minimum time between diagnosis of MS and first application of PHREND® must be 6 months. Because the EDSS scale is not linear across its entire range from 0 ("normal") to 10 ("death due to MS"), PHREND® cannot be used if the current EDSS is higher than 6. Based on individual patient characteristics at the time of the intended switch, PHREND® provides predictive probabilities to remain free of relapse and free of EDSS-based 3mCDP under the newly chosen DMT. The probability of staying relapse-free is derived from modeling the number of relapses following a negative binomial distribution and subsequently computing the fraction of predicted count of relapses that equals zero. The probability of staying 3mCDP-free is derived from modeling it as a binary event. Both models account for varying observation time in the training set [i.e., for varying time on DMT (12)]. Information used for the calculations comprises age, gender, EDSS, current therapy and duration of current therapy, number of previous DMTs and information if one of those was already a second-line treatment, time since RRMS diagnosis, number of previous relapses in the last year, and time since the last relapse. The choice of these parameters as predictors was based (1) on availability of sufficient data to train the statistical models on a representative population, (2) routine collection in clinical practice as prerequisite for widespread usability, and (3) proof of impact strength and usefulness of each parameter on the prediction (12).

Internal Validation and Prediction Quality Over Time

Underlying considerations, methods and results of first internal validations of PHREND® were previously communicated (12). Here, the mean square error (MSE) as well as the negative log-likelihood (NLL) are used to assess the goodness-of-fit of the models (i.e., the deviation between observed and predicted outcomes). The C-Index (0 to 1, where "1" indicates perfect predictions) measures the discrimination accuracy of a model and is defined as the proportion of concordant pairs (i.e., predicted outcomes match actual outcomes) divided by the total number of possible evaluation pairs. All three of these measures are computed in-sample and out-of-sample, where either predictions for patients used for training the models or for a set of new and unseen patients are evaluated to understand the model's generalizability. The credible intervals of the resulting model coefficients are indicators for the prediction certainty (i.e., the smaller the more certain the prediction). They are computed empirically, where a large set of models are fitted based on a randomly sampled initialization, and subsequently the range of each coefficient is described by credible intervals using the 90%-intervals from the resulting sets of coefficients. A small interval shows that, despite randomly selected initial values, the observed patient data for training is sufficiently informative to produce similar coefficients.

Because PHREND® is based on data extracted from the routinely used ongoing NTD MS registry (13), there is a steady increase of information (∼1.3% increase of patient numbers per quarter in the year 2020, data not shown here). Therefore, PHREND® is updated on a quarterly basis and prediction quality over time is monitored. For this work, the performance measures as described above were calculated repeatedly for models trained on quarterly updates of the database up to and including October 1, 2020.

External Comparison With RCTs

For external comparisons, PHREND® predictions were compared to results of RCTs to assess consistency with current research results. For this analysis, the published results of the active treatment cohorts of the clinical trials CONFIRM [dimethyl fumarate, glatirameracetate (14)], DEFINE [dimethyl fumarate (15)], REGARD [interferon-ß, glatiramer acetate (16)], TRANSFORMS [interferon-ß, fingolimod (17)], AFFIRM [natalizumab (18)], CLARITY [cladribine (19)], OPERA I and II [ocrelizumab (20)] and TEMSO [teriflunomide (21)] were chosen. For each study population, a comparable cohort in the NTD MS registry was identified by aiming to apply the same inclusion criteria as described in the corresponding study and by

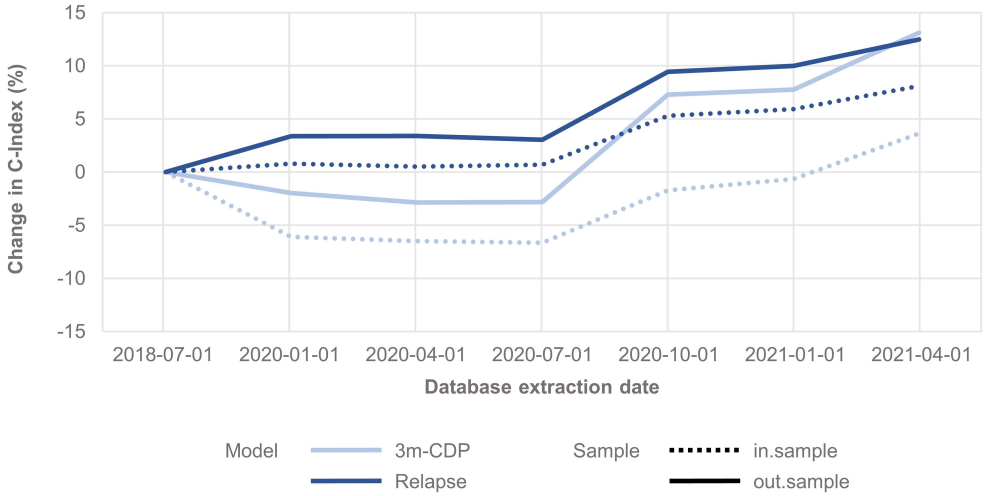


FIGURE 1 | Change in discrimination accuracy over time (C-Index) for the 3mCDP and the Relapse model, with the model performance based on the data extract from July 2018 as reference point. The plots show an increase of discriminative performance over time for the relapse model, whereas the performance of the CDP model initially dropped and only recovered during the last quarters analyzed for this work. Dashed lines show the performance change for the in-sample predictions, i.e., the predictions for patients used for training the model, and solid lines show the performance change when predicting for unseen patients, i.e., they address how well the model generalizes to an unknown population.

subsequently comparing mean values and standard deviations of continuous and distributions of categorial parameter to assure comparability of clinical and demographic baseline characteristics between groups.

Probabilities of staying free of relapses and free of 3mCDP within the clinical study timeframe were predicted using PHREND® models for the corresponding NTD MS cohorts, and their means and 90% credible intervals were compared to published results for each active treatment arm.

Clinical Robustness and Value

PHREND® provides personalized ranking of predictions for all DMTs available in Germany, which are ordered with respect to either highest probability for the patient to be relapse- or 3mCDP-free (**Figure 5**). The clinical usefulness and superiority of outcome of these recommendations made by PHREND® was previously affirmed by comparing therapy effectiveness for selected DMT cohorts where the recommended DMT was prescribed vs. where another DMT was chosen (12).

The current analysis investigates three different scenarios based on all DMTs simultaneously and independently of which substance was ranked highest or lowest: (1) patients taking the highest ranked therapy, (2) patients who took one of the two highest ranked therapies, and (3) patients who took one of the two least ranked therapies, always contrasted with results from patients on any other of the lower or higher-ranked treatments, respectively.

The comparability of patient groups in the analyses is ensured by a preceding propensity-score-based weighted matching (22) based on defined patient characteristics (age, time since diagnosis, previous DMT, duration of previous DMT, number of previous DMTs, indication if one of the previous DMTs was a second line treatment, gender, EDSS, time since last relapse, number of relapses in the last year). Relapse activity and 3mCDP is plotted for both subgroups in boxplots.

Implementation of the PHREND® Algorithm in a Web-Based Application

The web-based PHREND® application was developed using a human-centered-design approach over ten design iterations, in direct collaboration with doctors and patients to provide a clear, intuitively understandable presentation of the calculations for each DMT and the robustness of each probability. The presentation needs to integrate options to reflect upcoming questions in the shared decision process regarding their impact on choices between DMTs. PHREND® can be used as a standalone solution with clinical data being entered manually per patient or as part of the patient management platform DESTINY® (13) with automated data transfers.

RESULTS

Internal Validation and Prediction Quality Over Time

The analysis shows that the span of credible intervals of the models' coefficients decreased consistently over time (**Supplementary Figures 1, 2**). The discrimination accuracy (C-Index) of the models over time [with the published performance (12) as a reference point] increased for the relapse model for both in-sample and out-of-sample predictions (**Figure 1**). After an initial decrease, the discrimination accuracy for the 3mCDP model also increased with the availability of new data, outperforming the initially published model in the last quarter (out-of-sample). The apparent increase of the performance in Oct

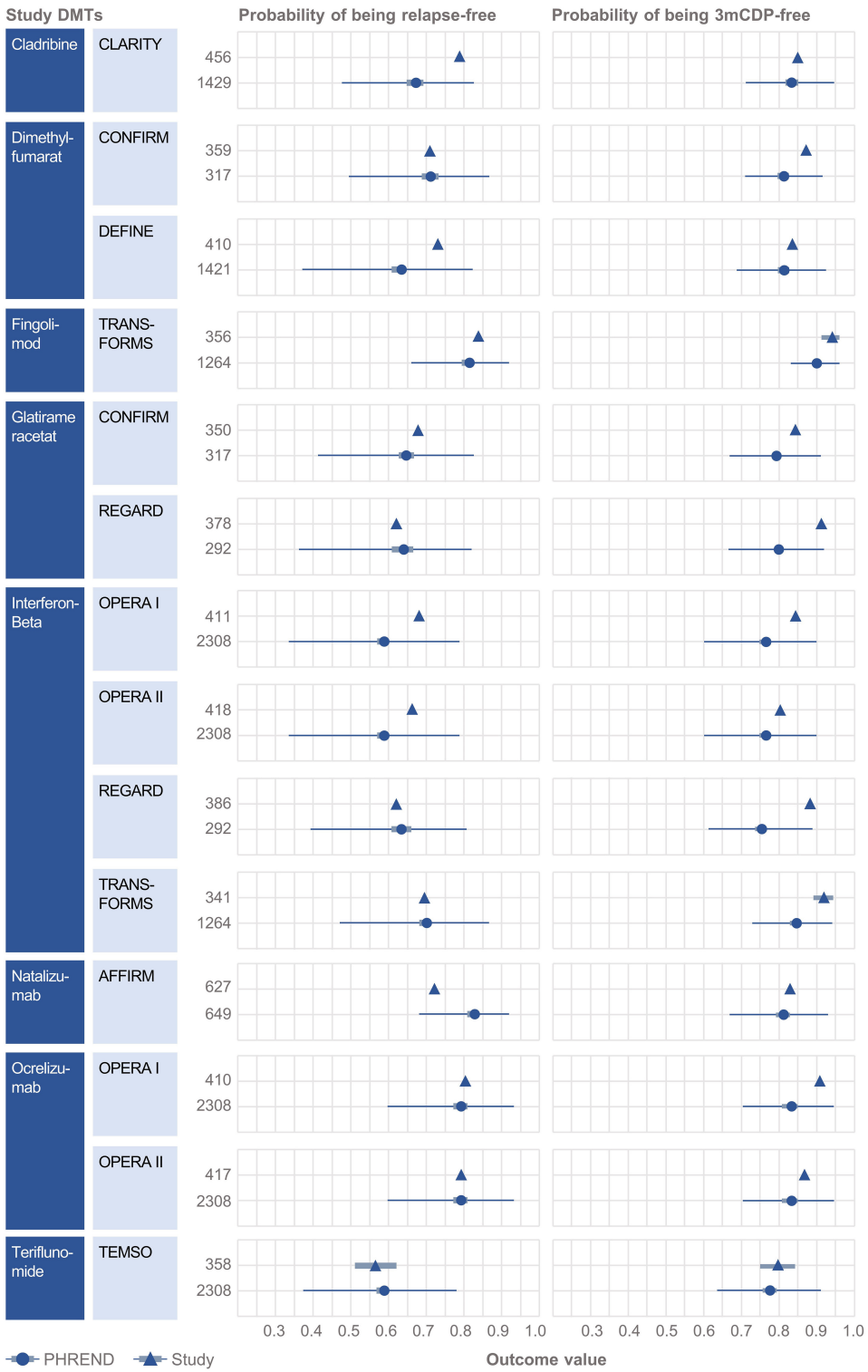


FIGURE 2 | External validation of the PHREND models using outcomes from eight clinical studies. A black triangle denotes the mean outcome reported by the clinical study, and the bar its 90%-confidence interval (if available). In analogy the PHREND model results are shown in red dots, with additional information on range of predictions. The wider bars show the 90%-confidence interval of the prediction. The thin horizontal bars show the range between the 5 and 95% quantile of all predictions. The numbers represent the patients included into the respective analysis. CONFIRM (14), DEFINE (15), REGARD (16), TRANSFORMS (17), AFFIRM (18), CLARITY (19), OPERA I and II (20) and TEMSO (21).

2020 was driven by the inclusion of new treatments ocrelizumab and cladribine, due to short observation time and informative priors used for training the models. This effect was observed to even out in the subsequent quarters.

Comparison With External RCT Data

Clinical baseline characteristics (MS duration, age, EDSS, relapses within previous 12 months, time since last relapse) were highly consistent between the NTD MS registry and RCT study populations (**Supplementary Figures 3, 4**). Probabilities predicted by PHREND® for being relapse free and 3mCDP free mostly approximated the results reported by the corresponding clinical study (**Figure 2**). Almost all differences were smaller than 10% between the predicted and the real study results, with the exemption for relapse activity with cladribine [CLARITY (19)], and 3mCDP with glatiramer acetate [REGARD (16)] and interferon-ß [REGARD (16)].

Clinical Consistency and Value

PHREND® provides a real-world data-driven ranking of therapies for both endpoints (see **Figure 5**). A the group level, the probability of staying relapse-free after propensity-score based weighting was statistically significantly higher ($p < 0.001$) for

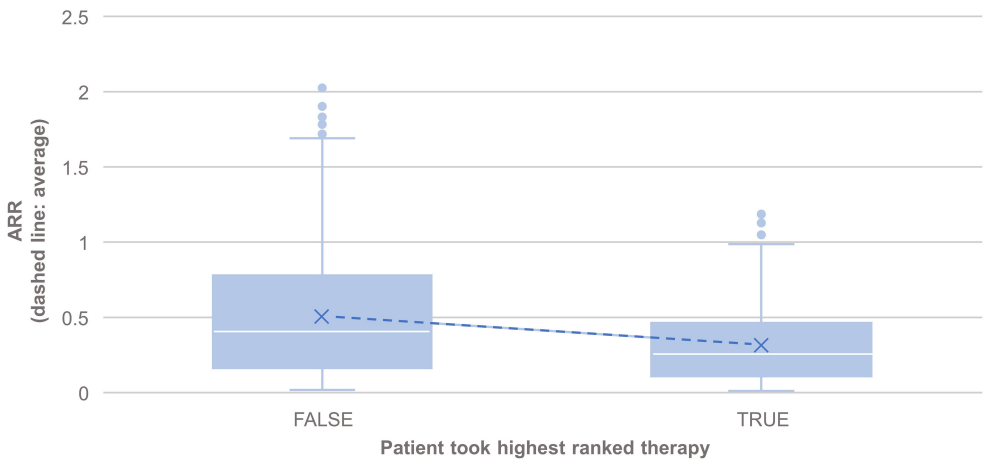


FIGURE 3 | Comparison of the annualized relapse rate after propensity score matching of groups for the relapse model, based on the disease courses as observed in the registry. $N = 495$ of 3,119 patients took the highest ranked DMT recommended by PHREND (case "TRUE"). ARR was significantly lower in this group compared to patients, who did not follow the reconditions and chose another DMT ("FALSE, $N = 2{,}624$", $p < 0.001$). The blue points show mean ARR for each subgroup, with a line to visualize the resulting slope.

FIGURE 4 | Comparison of the annualized relapse rate after propensity score matching of groups for the relapse model, based on the disease courses as observed in the registry. $N = 1{,}165$ of 3,119 patients took one of the first two DMTs recommended by PHREND (left), and $N = 1{,}076$ of 3,119 patients took one of the two least ranked DMTs (right). The blue points show the mean ARR for each subgroup, with a line to visualize the resulting slope in comparison to patients with other DMT decisions. ARR was statistically significantly lower when following the recommendation of the two highest ranked therapies (left, $p < 0.001$), and significantly higher for patients on one of the two least ranked DMTs (right, $p < 0.001$).

TABLE 1 | Comparison of therapy effectiveness for relapse and 3mCDP models in propensity score matched patient groups receiving DMTs recommended by PHREND vs. other than recommended DMTs.

DMT*	Model	Slope coefficient[a]	Sample size treated with recommended DMT*	Sample size treated with other than recommend DMT*	*p*-value
Highest ranked DMT	Relapse	−0.5193	495	2,624	<0.001
Highest ranked DMT	3mCDP	−0.4544	570	2,549	0.007
First or second highest ranked DMTs	Relapse	−0.4130	1,165	1,954	<0.001
First or second highest ranked DMTs	3mCDP	−0.2377	1,191	1,928	0.035
One of the two least ranked DMTs	Relapse	0.3695	1,076	2,043	<0.001
One of the two least ranked DMTs	3mCDP	0.3018	952	2,167	0.009

[a]*Derived from a survey-weighted negative binomial generalized linear model (Relapse) or a survey-weighted binomial generalized linear model (3mCDP), where negative sign indicates lower disease activity. DMT*, one or more disease modifying therapies that were ranked according to description.*

patients taking the highest-ranked therapy as opposed to any other of the lower ranked treatments (**Figure 3**).

The results also showed statistically significantly lower annual relapse rate (ARR) in patients who received one of the two highest ranked DMTs as recommended by PHREND® (n = 1,954, $p < 0.001$), and statistically significantly higher ARR in patients who had received one of the two least-recommended DMTs (n = 2,043, $p < 0.001$) (**Figure 4**). Statistically significant superiority was also found when the effects of personalized DMT selection were evaluated for the risk of CDP (**Table 1**).

Implementation of the PHREND® Algorithm in a Web-Based Application

Probabilities of outcomes under each available DMT are graphically displayed as natural frequencies in a ranked manner according to the results of the predictive calculations (**Figure 5**). This presentation of the probabilities corresponds to the current state of research in medical communication and was tested to be well understood by physicians and patients. The length of the prediction period can be chosen between 2 and 6 years. 90% credible intervals are displayed for each prediction to provide information on the homogeneity of the single results and to demonstrate possible overlap between outcomes. The smaller the interval, the more reliable is the prediction for the individual patient. To support the workflow of the shared decision process between physicians and patients, personal preferences such as family planning, route of administration and others can be incorporated. In these cases, not-suitable DMT options are shaded to allow the demonstration of the impact of certain preferences on the available DMT spectrum to choose from and the possible consequences regarding effectiveness of treatment options.

On-demand, deeper insights on the factors contributing to a single prediction are provided on an extra page (**Figure 6**). Results can be stored as PDF file as hand-out for the patient and for documentation purposes in medical record systems.

DISCUSSION

The previously published, initial assessment of the model performance demonstrated mathematically robust and accurate predictions based on the C-index and MSE (12). Models predicting freedom of relapse and of 3mCDP were shown to generalize to new patients and clinical sites and were robust against different choices of the priors and against sample size. In the current work, the same measures were analyzed based on quarterly updated database extractions, demonstrating the robustness of the models' predicted effectiveness probabilities with about 1.3% per quarter new patient data over time and also after the addition of the new DMTs cladribine and ocrelizumab. In parallel with this evidence of increasing accuracy of the models, the reliable performance of the model is demonstrated herein though consistently decreasing spans of credible intervals of the models' coefficients over time. These observations underscore the essential necessity for ongoing monitoring of the database and provide example metrics to ensure the models' performance and consistency. With continued application of the routines described herein, the beneficial effect of increasing data on the quality of the predictive probabilities can be expected to continue.

In an additional step toward external validation, this study showed that differences were smaller than 10% between predicted probabilities of PHREND® for freedom of relapse activity and 3mCDP, respectively, in NTD MS registry real-world patient cohorts compared with results from prospective RCT cohorts in a total of 14 active arms derived from 8 RCTs. The evident similarity of predictions of PHREND® based on real-world data and the published results from prospectively captured, blinded data from RCTs provide a meaningful external confirmation of the robustness of the algorithms employed. Observed differences of the comparisons in relapse activity in the CLARITY study (19) and 3mCDP in OPERA (20) and REGARD (16) trials likely reflect differences in patient characteristics not captured in the available cohort information, because pairwise single-patient based matching was not possible. For example, CLARITY (19) included 75% treatment-naïve patients, while all patients from the NTD MS registry were switching DMTs. Additional

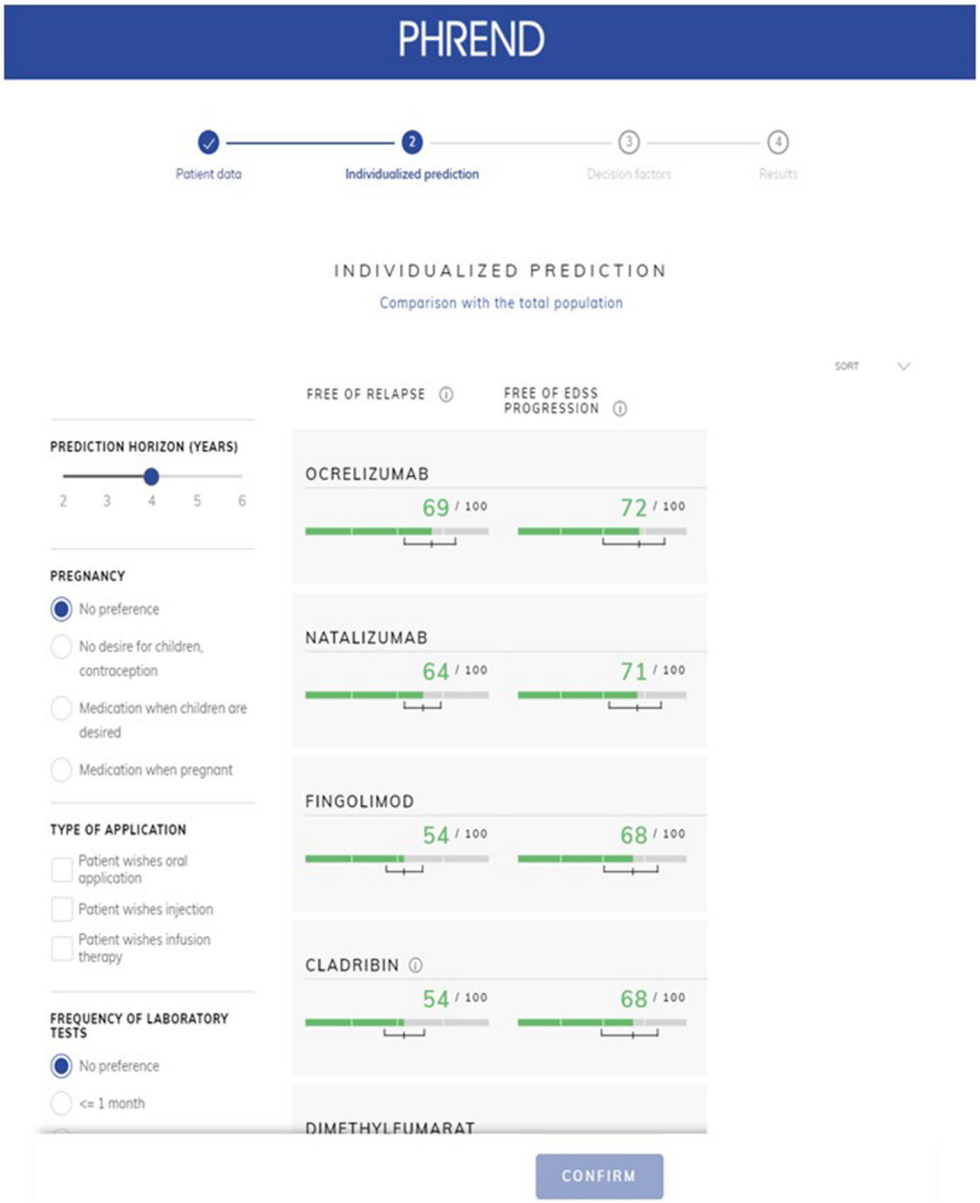


FIGURE 5 | PHREND display of resulting personalized predictions. The green horizontal bars show the individual probability of staying free of relapse (left column) or free of 3mCDP (right column) in the chosen time period (upper left corner, here 4 years). Additionally, a 90% credible interval is displayed for each prediction, i.e., the outcome is within this range with 90% probability. Disease modifying therapie's generic names are displayed in the application.

external validations are planned to further explore the validity of PHREND® in an ongoing endeavor to understand and improve predictive accuracy.

When comparing patients who, based on a combination of clinical consideration and individual preferences, chose the highest-ranked, or one of the two top-ranked DMTs, with patients who did not, the clinical course in both approaches was statistically significantly better regarding frequency of relapse activity and 3mCDP than in the comparator group with lower-ranked DMTs. Conversely, the opposite effect was shown with statistically significantly negative effects on both effectiveness parameter, if one of the two lowest-ranked DMTs was chosen compared to the top-ranked DMTs by the algorithm. This demonstrates the real-world accuracy of the mathematical algorithm developed in identifying the optimal DMTs in individual patients based on patient-specific parameters and real-world practice situations. Further validation is necessary with external non-NTD personalized patient data

FIGURE 6 | PHREND display of parameter distribution and impact of biographic and medical factors on prediction. On the left side, the patient's personal characteristics are shown within the distribution derived from the whole patient population. On the right side, the impact of each characteristic of the patient for the personalized prediction of being relapse-free and disability progression-free is shown.

to evaluate robustness and generalizability of the algorithm to other datasets.

It is important to note that PHREND® does not intend to automatize the medical decision process but to provide additional information beyond cohort-based study results and intuition. Integrated into the complex shared-decision process, it empowers physicians and patients to select optimal DMTs individually by providing data-driven, quantified outcome probabilities. Initial feedback obtained from NTD clinicians and patients indicate that the integration of PHREND® into the shared-decision process results in a more structured, rational communication process, which can reduce fears and avoidance patterns in patients and provide a base for a time-efficient shared-decision process. It remains to be evaluated, how this experience of a personalized transparent therapy decision can contribute to patients' motivation and adherence. Mandatory CE certification for PHREND® as medical tool is currently being obtained. Registration for PREND® is restricted to physicians, because it is an integral part of a medical process (https://www.neurotransdata.com/en/destiny#phrend).

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 Bavarian Medical Board (Bayerische

Landesärztekammer; June 14, 2012, ID 12114), Medical Board North-Rhine (Ärztekammer Nordrhein; April 25, 2017, ID 2017071). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SB: concept, analysis, and interpretation and writing. ES: concept, data preparation, analysis, and interpretation and writing. YH: analysis and interpretation. PH: concept, interpretation, supervision, and project administration. AB: concept, project administration, and funding. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by NeuroTransData GmbH, Neuburg/Donau, Germany and PricewaterhouseCoopers, Zurich, Switzerland. The commercial entities of NeuroTransData and PricewaterhouseCoopers had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.856829/full#supplementary-material

REFERENCES

1. Mohr DC, Goodkin DE, Likosky W, Gatto N, Neilley LK, Griffin C, et al. Therapeutic expectations of patients with multiple sclerosis upon initiating interferon beta-1b: Relationship to adherence to treatment. *Mult Scler.* (1996) 2:222–6. doi: 10.1177/135245859600200502
2. Rothwell PM, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. *BMJ.* (1997) 314:1580–3. doi: 10.1136/bmj.314.7094.1580
3. Neuhaus M, Calabrese P, Annoni JM. Decision-making in multiple sclerosis patients: a systematic review. *Mult Scler Int.* (2018) 2018:7835952. doi: 10.1155/2018/7835952
4. Braune S, Grimm S, van Hoevell P, Freudensprung U, Pellegrini F, Hyde R, et al. Comparative effectiveness of delayed-release dimethyl fumarate versus interferon, glatiramer acetate, teriflunomide, or fingolimod: results from the German NeuroTransData registry. *J Neurol.* (2018) 265:2980–92. doi: 10.1007/s00415-018-9083-5
5. Laplaud DA, Casey R, Barbin L, Debouvrie M, De Séze J, Brassat D, et al. Comparative effectiveness of teriflunomide vs dimethyl fumarate in multiple sclerosis. *Neurology.* (2019) 93:e635–46. doi: 10.1212/WNL.0000000000007938
6. Hillert J, Tsai JA, Nouhi M, Glaser A, Spelman T. A comparative study of teriflunomide and dimethyl fumarate within the Swedish MS registry. *Mult Scler.* (2021) 28:237–46. doi: 10.1177/13524585211019649
7. Reder AT, Arndt N, Roman C, Geremakis C, Mendoza JP, Su R, et al. Real-world propensity score comparison of treatment effectiveness of peginterferon beta-1a vs. subcutaneous interferon beta-1a, glatiramer acetate, and teriflunomide in patients with relapsing-remitting multiple sclerosis. *Mult Scler Relat Disord.* (2021) 51:10293. doi: 10.1016/j.msard.2021.102935
8. Lorscheider J, Benkert P, Lienert C, Hänni P, Derfuss T, Kuhle J, et al. Comparative analysis of dimethyl fumarate and fingolimod in relapsing-remitting multiple sclerosis. *J Neurol.* (2021) 268:941–9. doi: 10.1007/s00415-020-10226-6
9. Ferreira-Atuesta C, Reyes S, Giovanonni G, Gnanapavan S. The evolution of neurofilament light chain in multiple sclerosis. *Front Neurosci.* (2021) 15:642384. doi: 10.3389/fnins.2021.642384
10. Delcoigne B, Manouchehrinia A, Barro C, Benkert P, Michalak Z, Kappos L, et al. Blood neurofilament light levels segregate treatment effects in multiple sclerosis. *Neurology.* (2020) 94:e1201–12. doi: 10.1212/WNL.0000000000009097
11. Tacke S, Braune S, Rovituso DM, Ziemssen T, Lehmann PV, Dikow H, et al. B-Cell activity predicts response to glatiramer acetate and interferon in relapsing-remitting multiple sclerosis. *Neurol Neuroimmunol Neuroinflamm.* (2021) 8:e980. doi: 10.1212/NXI.0000000000000980
12. Stühler E, Braune S, Lionetto F, Heer Y, Jules E, Westermann C, et al. Framework for personalized prediction of treatment response in relapsing remitting multiple sclerosis. *BMC Med Res Methodol.* (2020) 20:24. doi: 10.1186/s12874-020-0906-6
13. Bergmann A, Stangel M, Weih M, van Hövell P, Braune S, Köchling M, et al. Development of registry data to create interactive doctor-patient platforms for personalized patient care, taking the example of the DESTINY system. *Front Digit Health.* (2021) 3:633427. doi: 10.3389/fdgth.2021.633427
14. Fox RJ, Miller DH, Phillips JT, Hutchinson M, Havrdova E, Kita M, et al. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *N Engl J Med.* (2012) 367:1087–97. doi: 10.1056/NEJMoa1206328
15. Gold R, Kappos L, Arnold DL, Bar-Or A, Giovannoni G, Selmaj K, et al. Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. *N Engl J Med.* (2012) 367:1098–107. doi: 10.1056/NEJMoa1114287
16. Mikol DD, Barkhof F, Chang P, Coyle PK, Jeffrey DR, Schwid SR, et al. Comparison of subcutaneous interferon beta-1a with glatiramer acetate in patients with relapsing multiple sclerosis (the REbif vs glatiramer acetate in relapsing MS disease [REGARD] study): a multicentre, randomised, parallel open-label trial. *Lancet Neurol.* (2008) 7:903–14. doi: 10.1016/S1474-4422(08)70200-X
17. Cohne JA, Barkhof F, Comi G, Hartung HP, Khatro BO, Montalban X, et al. Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. *N Engl J Med.* (2010) 362:402–15. doi: 10.1056/NEJMoa0907839
18. Polman CH, O'Connor PW, Havrdova E, Hutchinson M, Kappos L, Miller DH, et al. A randomized, placebo-controlled trial of natalizumab. *N Engl J Med.* (2006) 354:899–910. doi: 10.1056/NEJMoa044397
19. Giovanonni G, Comi G, Cook S, Rammohan K, Rieckmann P, Sorensen PS, et al. A placebo-controlled trial of oral cladribine for relapsing multiple sclerosis. *N Engl J Med.* (2010) 362:416–26. doi: 10.1056/NEJMoa0902533
20. Hauser SL, Bar-Or A, Comi G, Giovannoni G, Hartung PH, Hemmer B, et al. Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. *N Engl J Med.* (2017) 376:221–34. doi: 10.1056/NEJMoa1601277
21. O'Connor P, Wolinsky JS, Confavreux C, Comi G, Kappos L, Olsson TP, et al. Randomized Trial of Oral Teriflunomide. *N Engl J Med.* (2011) 365:1293–303. doi: 10.1056/NEJMoa1014656
22. Ridgeway G, McCaffrey D, Morral A, Burgette L, Griffin BA. *Toolkit for Weighting and Analysis of Nonequivalent Groups: A Tutorial for the Twang Package.* Santa Monica, CA: RAND Corporation of nonequivalent groups: a tutorial for the twang package (2017). Available online at: https://cran.r-project.org/web/packages/twang/index.html.

Conflict of Interest: SB, AB, and all members of the NeuroTransData Study Group are working self-employed as medical doctors in outpatient practices. All of them are members of the NeuroTransData doctors' network (https://www.neurotransdata.com/en/about-us). SB has received honoraria from Kassen4rztliche Vereinigung Bayern and HMOs for patient care; honoraria for consulting, project management, clinical studies, and lectures from Biogen,

CSL Behring, NeuroTransData, Novartis, Roche and Thieme Verlag; honoraria and expense compensation as board member of NeuroTransData. AB has received consulting fees from advisory board, speaker, and other activities for NeuroTransData; honoraria and expense compensation for project management and clinical studies from Novartis and Servier. ES, YH, and PH are employees of PricewaterhouseCoopers, Z, Switzerland.

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 | Frontiers in Digital Health

MINI REVIEW
published: 25 April 2022
doi: 10.3389/fdgth.2022.872675

Artificial Intelligence in Perioperative Medicine: A Proposed Common Language With Applications to FDA-Approved Devices

Ryan L. Melvin[1]*, *Matthew G. Broyles*[1], *Elizabeth W. Duggan*[1], *Sonia John*[1], *Andrew D. Smith*[2] *and Dan E. Berkowitz*[1]

[1] *Department of Anesthesiology and Perioperative Medicine, University of Alabama at Birmingham, Birmingham, AL, United States,* [2] *Department of Radiology, University of Alabama at Birmingham, Birmingham, AL, United States*

As implementation of artificial intelligence grows more prevalent in perioperative medicine, a clinician's ability to distinguish differentiating aspects of these algorithms is critical. There are currently numerous marketing and technical terms to describe these algorithms with little standardization. Additionally, the need to communicate with algorithm developers is paramount to actualize effective and practical implementation. Of particular interest in these discussions is the extent to which the output or predictions of algorithms and tools are understandable by medical practitioners. This work proposes a simple nomenclature that is intelligible to both clinicians and developers for quickly describing the interpretability of model results. There are three high-level categories: *transparent*, *translucent*, and *opaque.* To demonstrate the applicability and utility of this terminology, these terms were applied to the artificial intelligence and machine-learning-based products that have gained Food and Drug Administration approval. During this review and categorization process, 22 algorithms were found with perioperative utility (in a database of 70 total algorithms), and 12 of these had publicly available citations. The primary aim of this work is to establish a common nomenclature that will expedite and simplify descriptions of algorithm requirements from clinicians to developers and explanations of appropriate model use and limitations from developers to clinicians.

Keywords: artificial intelligence, AI, machine learning, algorithm, FDA approval

OPEN ACCESS

Edited by:
Kirsten Smayda, MedRhythms, United States

Reviewed by:
Bertalan Mesko, The Medical Futurist Institute, Hungary

***Correspondence:**
Ryan L. Melvin
rmelvin@uabmc.edu

Specialty section:
This article was submitted to Personalized Medicine, a section of the journal Frontiers in Digital Health

Received: *09 February 2022*
Accepted: *29 March 2022*
Published: *25 April 2022*

Citation:
Melvin RL, Broyles MG, Duggan EW, John S, Smith AD and Berkowitz DE (2022) Artificial Intelligence in Perioperative Medicine: A Proposed Common Language With Applications to FDA-Approved Devices. Front. Digit. Health 4:872675. doi: 10.3389/fdgth.2022.872675

INTRODUCTION

The list of medical uses for Artificial Intelligence (AI) and Machine Learning (ML) is expanding rapidly (1). Recently, this trend has been particularly true for anesthesiology and perioperative medicine (2, 3). Deriving utility from these algorithms requires medical practitioners and their support staff to sift through a deluge of technical and marketing terms (3). This paper provides an aid for separating the signal of utility from the noise of jargon.

This work proposes a straightforward nomenclature for describing the interpretability and appropriate use of AI/ML products that will be intuitive to developers and clinicians alike. The applicability and utility of this terminology of these terms is then applied to Food and Drug Administration (FDA) approved AI/ML algorithms (1) with perioperative utility. Such a standardized language may speed discussion and understanding among technical developers and

clinical users. To this end, there are three standardized terms for conveying interpretability and indicating the appropriate use of systems and products based on AI/ML. The terms are *transparent*, *translucent*, and *opaque*.

Opaque describes a system that (a) estimates non-linearly applied parameters that require advanced analysis (external to the product itself) to be understood, (b) estimates such a large set of parameters that a human cannot interpret them unless aided by a tool external to the product, or (c) provides a prediction with no indication as to the reason for the prediction. *Translucent* describes a product that incorporates non-specific methods for assisting the end user in understanding possible reasons for a prediction that would otherwise be categorized as "opaque." *Transparent* describes a product that estimates a linearly applied parameter or relatively small set of parameters and is implemented to indicate to the user how much each feature influences the output or indicates the rationale for a prediction (e.g., providing exact weights for features considered). Alternatively, a transparent system's prediction is easily verifiable at the time the prediction is made using information the system provides.

For example, consider a clinician wanting to predict future glucose values for a patient (4). A careful developer for such a model would be concerned with several factors including whether the clinician desires simply a prediction (opaque) or a prediction with an explanation (translucent or transparent). If an explanation is needed, the developer may wonder how specific to each patient the explanation needs to be. Would a list of factors considered be sufficient (translucent), or does the clinician need to know exactly how much each factor contributed to the prediction for each patient (transparent)?

This imagined developer seeks to understand the level of interpretability required for this algorithm. However, there is also a tradeoff that the developer is considering. Requiring more interpretability limits the types of algorithms that can be used. The deep learning algorithms that currently automate driving, image recognition, and recently folded protein structure estimation (among many other things) tend to lack interpretability (opaque). On the other end of the spectrum, generalized linear models (such as logistic regression) tend toward high interpretability (transparent) but often have less accuracy than deep learning algorithms.

The situation also works in reverse. Consider a developer attempting to explain a new model to a randomly selected clinician. The clinician may wonder how an algorithm works or why it makes certain predictions. For some products, such questions are easy to answer (transparent). For others, it is exceedingly difficult (opaque). And the distinction between such systems at times seems arbitrary. Rather than developers and clinicians continuously engaging in this discussion *de novo* for every project and collaboration, presented here are three standardized terms for conveying the interpretability and indicating the appropriate use of AI/ML models—*transparent*, *translucent*, and *opaque*.

These ideas behind the terms are not new (5–7). Phrases such as "glass box," (8) "white box," (9) "gray box," (10) "interpretable," (11) "explainable," (6, 8, 11–15) and "black box" (8, 10–14) are often used to describe the complexity of algorithms from a somewhat technical perspective. Proposed here are less technically intended terms meant to describe the perspective of the end user rather than the developer—a distinction discussed later with several examples. In addition to being first and foremost clinician-friendly, these terms are intended to convey sufficient technical information to developers for understanding the types of algorithms appropriate for the desired use case. Note that these terms do not describe the underlying mathematics of an algorithm or even the technical details of a particular implementation; rather, they describe the experience of the end user.

For readers familiar with technical usages of "black box," (8, 10–14) our usage of *opaque* is similar with the exception it focuses on the experience of the end user and what they reasonably know or are presented with by a specific algorithm implementation. For readers familiar with the concept of "Explainable AI," the "explainable" piece often refers to a secondary technology applied to a trained AI/ML model that extracts information about why the model makes certain predictions (6, 8, 11–15). An algorithm that makes use of such a technology and shows the output to an end user would be classified under our nomenclature as *translucent*.

METHODS AND DEFINITIONS

The initial source of algorithms for consideration in this review is a constantly updated online database of FDA-approved algorithms. At the time of this writing, the database contained 70 such algorithms (1). As reported by the primary citation for the database (1), the majority of algorithms in this database were approved with 510(k) clearance. Other approval methods seen in the database are *de novo* pathway clearance and premarket approval clearance. The database makes broad categorizations of applicable fields for these algorithms. The fields most represented are Radiology, Cardiology, and Internal Medicine/General Practice.

Perioperative medicine is not explicitly mentioned in the database. Therefore, the categorization of "perioperative utility" in this review was made under the best judgment of the authors. The primary purpose of this work, though, is to establish a nomenclature, using the algorithms labeled as having perioperative utility in examples of applying this terminology. Of the 70 algorithms in the database, 22 were determined to have perioperative utility.

Each record in the database includes the name of the algorithm and the parent company. Using this information, along with the details in the corresponding FDA announcements themselves, journal reviewed articles describing the function of these algorithms were sought. This search included—but was not limited to—searching the parent company's website for mentions of journal articles. From this search, citations for 12 of the 22 algorithms were found.

The categories applied to these 12 algorithms were *opaque*, *translucent*, and *transparent*. **Table 1** summarizes these categories. *Opaque* describes a system that (a) estimates

TABLE 1 | Summary of category definitions.

Category and example usage	Defining features
Opaque **Concrete example:** A clinician desires a prediction of future glucose values for a patient with no need to understand how the prediction was made	• Estimates non-linearly applied parameters that require advanced analysis to be understood • OR estimates such a large set of parameters that a human cannot interpret them unaided • OR provides a prediction with no indication as to the reason for the prediction
Translucent **Concrete example:** A clinician desires a prediction of future glucose values for a patient with an explanation of what clinical factors were involved in making the prediction	• Includes techniques for explaining the predictions from an otherwise opaque algorithm • Examples: • Plotting non-linear functions of features • Variable importance methods • Providing a list of factors considered
Transparent **Concrete example:** A clinician desires a prediction of future glucose values for a patient with an explanation of exactly how much each involved factor contributed to the prediction for each patient	• Estimates a linearly applied parameter or relatively small set of parameters that indicate how much each feature influences the output • OR indicates the specific rationale for a prediction • Example**:** providing the exact features and weights responsible for a given prediction • OR provides a prediction that is easily verifiable at the time of the prediction

non-linearly applied parameters that require advanced analysis (external to the product itself) to be understood, (b) estimates such a large set of parameters that a human cannot interpret them unless aided by a tool external to the product, or (c) provides a prediction with no indication as to the reason for the prediction. That is, an opaque system meets at least one of the criteria (a), (b), or (c). The general theme of this definition is whether the end consumer of the product's prediction also receives some measure of explanation as to why the specific prediction was made. Succinctly, if the user does not receive such an explanation, then the system is categorized as "opaque."

Translucent describes a product that incorporates non-specific methods for assisting the end user in understanding possible reasons for a prediction that would otherwise be categorized as "opaque." Examples include plotting non-linear functions of features, variable importance methods, and providing a list of factors considered. The general theme for the "translucent" category is that non-specific information about factors influencing the prediction is provided.

A system predicting diabetes diagnosis that considers weight, age and diet in its algorithm is translucent; if modified to provide the relative weights of each of these factors in making the diagnostic prediction, the algorithm would be considered transparent. The first case provides non-specific prediction factors (translucent) while the second includes specific information for the end-user (transparent). Similarly, for image recognition, placing the corresponding image (or wave form) next to a predicted label or highlighting a segment of an image corresponding to a prediction for image recognition would be *translucent*. Again, the distinction between *translucent* and *transparent* is non-specific vs. specific rationale for the prediction from the perspective of the end user. For comparison, the term *explainable* is used to describe tools used by a developer to make the output of a product more easily interpretable (16–18). While *translucent* conveys a similar idea, we emphasize that its definition is from the perspective of the end user and the kind of information a particular system based on an AI/ML algorithm provides to them. For an exploration of various meanings and uses of "Explainability" in the context of artificial intelligence and machine learning, see the work by Bhatt et al. (17).

Transparent describes a product that estimates a linearly applied parameter or relatively small set of parameters and is implemented to indicate to the user how much each feature influences the output or indicates the rationale for a prediction (e.g., providing exact weights for features considered). Alternatively, a transparent system's prediction is easily verifiable at the time the prediction is made using information the system provides. Consider the previous example of system that predicts a diabetes diagnosis and indicates BMI, age, and diet. This system is *translucent*, but an algorithm that provides the weights used for these features would be *transparent.*

EXAMPLES OF FDA-APPROVED ALGORITHMS

Through the process described above, three of the 12 products with a located citations and clear perioperative utility were categorized as opaque (**Table 2**). These are RhythmAnalytics from Biofourmis Singapore Pte. Ltd., and the Guardian Connect System from Medtronic. RhythmAnalytics along with its underlying Biovitals Analytics Engine monitors, which categorizes cardiac arrhythmias via a convolutional neural network—a deep learning technique—that consumes wavelet transforms and short-time Fourier transforms of electrocardiogram (ECG) signals (19). Such a deep learning technique is inherently opaque, as the number of weights and their non-linear combination makes unaided understanding of the reasons for a prediction not feasible. The Guardian Connect System alerts users to interstitial glucose levels outside of a specified range. The system offers two alert types, "threshold" and "predictive." The "threshold" alerts are transparent in that they indicate if the sensor glucose reading is above or below a threshold. The "predictive" alerts indicate whether glucose is predicted to be outside the specified range within the next 10–60 min (4). While the manufacturer's user guide (20) explains that these predictions are formula-based, prediction-based alerts are provided using the name of the prediction, not the reasons for the prediction. The categorizations presented here are based on the information provided to a user when a prediction is provided. Therefore, this system is opaque.

In the translucent category are products that provide a prediction alongside an upfront list of features considered—but no specific weights—or a non-specific visualization of signal

TABLE 2 | System classifications.

Name of device or algorithm	Online database description (1)	Description based on primary citation	Classification
Biovitals analytics engine	"Cardiac monitor"	Detects prolonged QT interval (19)	Opaque
Rhythm analytics	"Monitoring cardiac arrhythmias"	Deep learning to classify rhythm (19)	Opaque
Guardian connect system	"Predicting blood glucose changes"	Predicts glucose levels outside of the normal range and gives predictive alerts (4, 20)	Opaque
eMurmer ID	"Heart murmur detection"	Determines if murmurs are innocent or pathologic (22)	Translucent
physIQ heart rhythm and respiratory module	"Detection of atrial fibrillation"	Uses patients' own baseline to detect changes (21)	Translucent
DreaMed	"Managing type 1 diabetes"	Recommends insulin doses (23, 24)	Translucent
ECG app	"Detection of atrial fibrillation"	Watch-based atrial fibrillation detection (25)	Transparent
FibriCheck	"Cardiac monitor"	Smartphone atrial fibrillation detection (26, 27)	Transparent
Irregular rhythm notification feature	"Detection of atrial fibrillation"	Smartphone irregular rhythm notification (28)	Transparent
WAVE clinical platform	"Monitoring vital signs"	Remote vital sign monitoring and alerts (29, 30)	Transparent
EchoMD automated ejection fraction software	"Echocardiogram analysis"	Helps place echo device. Calculates ejection fraction (31)	Transparent
Caption guidance	"Software to assist medical professionals in the acquisition of cardiac ultrasound images"	Works with EchoMD above (31)	Transparent

used in making the prediction. Of the 12 systems, three fit the criteria for this category (**Table 2**). For example, the physIQ Heart Rhythm and Respiratory Module from physIQ Inc. collapses improving or declining factors related heart failure into a single index. The signals related to this index are viewable by patients and providers along with the corresponding calculated index (21), which led to categorizing this algorithm as translucent. Another cardiac monitoring product, eMurmer ID from CSD Labs GmbH, predicts whether murmurs are pathologic. Along with the prediction, the algorithm shows the systolic and diastolic phases considered; however, the specifics of the predictions "are protected under proprietary regulations," (22) leading to inclusion of this algorithm in the translucent category. Likewise, the DreaMed algorithm from DreaMed Diabetes, Ltd., produces reports with important features related to its insulin dose recommendations (23, 24), placing it in the translucent category.

The remaining six (of 12) products fall in the transparent category (**Table 2**). Of these, three are implementations of atrial fibrillation detection that provide either annotated signals along with predictions in the case of ECG App from Apple Inc. (25) and FibriCheck from Qompium NV (26, 27) or use explicit "if… then…" rules in the case of the Irregular Rhythm Notification Feature from Apple Inc. (28). Similarly, the Wave Clinical Platform from Excel Medical Electronics LLC provides vital sign alerts based on a set of rules and provides the reason for the alert when triggered (29, 30). Also, in the transparent category is the EchoMD Automated Ejection Fraction Software which integrates with the Caption Guidance system from Caption Health Inc. to indicate when an a echocardiogram transducer is correctly placed for the automated calculation of ejection fraction (31). Since the reasons for the provided feedback are intuitively obvious (the transducer is or is not physically placed correctly), these systems are transparent.

Additional Examples and Future Work

Here we have applied our proposed common language to FDA-approved algorithms, as initial examples of how these terms might be used in clinical contexts. We recognize there are many tools and devices in the world of perioperative medicine used for patient care, education, research, quality improvement, and operations. Covering all of these would be at least a book-length task, well beyond the scope of this project. However, we consider the next step in developing this nomenclature to be a review article addressing the (additional) most common algorithms for patient care. As a small sample, such a review article might include products such as BIS (Bispectral Index) (32, 33), Sedline (34), Datex-Ohmeda Entropy (35), and Edwards Hemosphere (36).

Beyond these additional examples in perioperative medicine, these terms are immediately extensible to other medical fields. While all examples provided herein dealt with algorithms surrounding surgery, note that the definitions themselves (**Table 1**) are agnostic to any medical subfield.

Additionally, some commentary seems warranted with respect to proprietary algorithms. Indeed, in this review, some products were given opaque or translucent classifications, which may change if the algorithmic details and/or source code for such products were ever released by the intellectual property owners. This dimension of the nomenclature whereby a products categorization could be changed by a public release of information further emphasizes that these terms are from the perspective of the end user rather than technical details.

CONCLUSIONS

This work presents a nomenclature for describing algorithm implementations and applies it to several examples in the

literature. This terminology is composed of three high-level categories: *transparent*, *translucent*, and *opaque*. These terms are applied the point of view of the clinician. To indicate how these terms can be used to categorize AI systems, AI/ML systems with FDA are presented as examples. A database of these examples that will be updated as new systems gain FDA approval is available at https://sites.uab.edu/periop-datascience/algo-database.

This nomenclature aids in understanding the appropriate use of models. In high-risk situation, the requirement for accuracy may be paramount. Alternatively, in high-profile situations, predictions may need to be explainable to stakeholders. For example, the FDIC in the United States requires financial institutions to develop "conceptually sound" (37) models. An assessment of conceptual soundness would be easiest for transparent models and most difficult for opaque models.

The primary values of common nomenclature are expediting and simplifying descriptions of model requirements and appropriate use between clinicians and developers.

AUTHOR CONTRIBUTIONS

RM performed the initial algorithm search and filtering and prepared the first manuscript draft. MB, ED, and SJ helped assess the perioperative utility of each algorithm and revised the manuscript. AS suggested the framework for the project, helped assess the appropriate category of each algorithm, and revised the manuscript. DB helped organize the project team, helped assess the perioperative utility of each algorithm, and revised the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

1. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med.* (2020) 3:1–8. doi: 10.1038/s41746-020-00324-0
2. Mathis MR, Kheterpal S, Najarian K. Artificial intelligence for anesthesia: what the practicing clinician needs to know. *Anesthesiology.* (2018) 129:619–22. doi: 10.1097/ALN.0000000000002384
3. Maheshwari K, Ruetzler K, Saugel B. Perioperative intelligence: applications of artificial intelligence in perioperative medicine. *J Clin Monit Comput.* (2020) 34:625–8. doi: 10.1007/s10877-019-00379-9
4. Abraham SB, Arunachalam S, Zhong A, Agrawal P, Cohen O, McMahon CM. Improved real-world glycemic control with continuous glucose monitoring system predictive alerts. *J Diabetes Sci Technol.* (2019) 15:91–7. doi: 10.1177/1932296819859334
5. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *HHS Public Access.* (2019) 2:749–60. doi: 10.1038/s41551-018-0304-0
6. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI-Explainable artificial intelligence. *Sci Robot.* (2019) 4:eaay7120.
7. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol.* (2019) 19:1–18. doi: 10.1186/s12874-019-0681-4
8. Rai A. Explainable AI: from black box to glass box. *J Acad Mark Sci.* (2020) 48:137–41. doi: 10.1007/s11747-019-00710-5
9. Loyola-Gonzalez O. Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view. *IEEE Access.* (2019) 7:154096–113. doi: 10.1109/ACCESS.2019.2949286
10. Feldman JM, Kuck K, Hemmerling T. Black box, gray box, clear box? How well must we understand monitoring devices? *Anesth Analg.* (2021) 132:1777–80. doi: 10.1213/ANE.0000000000005500
11. Gaur M, Faldu K, Sheth A. Semantics of the black-box: can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Comput.* (2021) 25:51–9. doi: 10.1109/MIC.2020.3031769
12. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access.* (2018) 6:52138–60. doi: 10.1109/ACCESS.2018.2870052
13. Zednik C. Solving the black box problem: a normative framework for explainable artificial intelligence. *Philos Technol.* (2021) 34:265–88. doi: 10.1007/s13347-019-00382-7
14. Rudin C, Radin J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Sci Rev.* (2019) 1:1–9. doi: 10.1162/99608f92.5a8a3a3d
15. Kundu S, AI. in medicine must be explainable. *Nat Med.* (2021) 27:1328. doi: 10.1038/s41591-021-01461-z
16. Holzinger A. From machine learning to explainable AI. In: *DISA 2018 - IEEE World Symp Digit Intell Syst Mach Proc.* Košice, Slovakia (2018) 55–66. doi: 10.1109/DISA.2018.8490530
17. Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, et al. Explainable machine learning in deployment. In: *FAT* 2020 - Proc 2020 Conf Fairness, Accountability, Transpar.* Barcelona, Spain (2020) 648–57. doi: 10.1145/3351095.3375624
18. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. *Advances in Neural Information Processing Systems 30*. Long Beach, CA: Curran Associates, Inc. (2017). p. 4765–74. Available online at: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
19. Rajput KS, Wibowo S, Hao C, Majmudar M. *On Arrhythmia Detection by Deep Learning and Multidimensional Representation.* (2019). Available online at: http://arxiv.org/abs/1904.00138
20. Medtronic. *Guardian Connect.* Medtronic (2017). Available online at: https://guardianconnect.medtronic-diabetes.co.uk
21. Stehlik J, Schmalfuss C, Bozkurt B, Nativi-Nicolau J, Wohlfahrt P, Wegerich S, et al. Continuous wearable monitoring analytics predict heart failure hospitalization: the link-hf multicenter study. *Circ Hear Fail.* (2020) 13:e006513. doi: 10.1161/CIRCHEARTFAILURE.119.006513
22. Lai LSW, Redington AN, Reinisch AJ, Unterberger MJ, Schriefl AJ. Computerized automatic diagnosis of innocent and pathologic murmurs in pediatrics: a pilot study. *Congenit Heart Dis.* (2016) 11:386–95. doi: 10.1111/chd.12328
23. Nimri R, Dassau E, Segall T, Muller I, Bratina N, Kordonouri O, et al. Adjusting insulin doses in patients with type 1 diabetes who use insulin pump and continuous glucose monitoring: variations among countries and physicians. *Diabetes Obes Metab.* (2018) 20:2458–66. doi: 10.1111/dom.13408
24. Nimri R, Oron T, Muller I, Kraljevic I, Alonso MM, Keskinen P, et al. Adjustment of insulin pump settings in type 1 diabetes management: advisor pro device compared to physicians' recommendations. *J Diabetes Sci Technol.* (2022) 16:364–72. doi: 10.1177/1932296820965561
25. Turakhia MP, Desai M, Hedlin H, Rajmane A, Talati N, Ferris T, et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: the Apple Heart Study. *Am Heart J.* (2019) 207:66–75. doi: 10.1016/j.ahj.2018.09.002
26. Selder JL, Proesmans T, Breukel L, Dur O, Gielen W, van Rossum AC, et al. Assessment of a standalone photoplethysmography (PPG) algorithm for detection of atrial fibrillation on wristband-derived data. *Comput Methods Programs Biomed.* (2020) 197:105753. doi: 10.1016/j.cmpb.2020.105753
27. Verbrugge FH, Proesmans T, Vijgen J, Mullens W, Rivero-Ayerza M, Van Herendael H, et al. Atrial fibrillation screening with photoplethysmography through a smartphone camera. *Europace.* (2019) 21:1167–75. doi: 10.1093/europace/euz119

28. Ip JE. Evaluation of cardiac rhythm abnormalities from wearable devices. *JAMA*. (2019) 321:1098. doi: 10.1001/jama.2019.1681
29. Hravnak M, Devita MA, Clontz A, Edwards L, Valenta C, Pinsky MR. Cardiorespiratory instability before and after implementing an integrated monitoring system. *Crit Care Med.* (2011) 39:65–72. doi: 10.1097/CCM.0b013e3181fb7b1c
30. Tarassenko L, Hann A, Young D. Integrated monitoring and analysis for early warning of patient deterioration. *Br J Anaesth.* (2006) 97:64–8. doi: 10.1093/bja/ael113
31. Schneider M, Bartko P, Geller W, Dannenberg V, König A, Binder C, et al. A machine learning algorithm supports ultrasound-naïve novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF. *Int J Cardiovasc Imaging.* (2021) 37:577–86. doi: 10.1007/s10554-020-02046-6
32. Morimoto Y, Hagihira S, Koizumi Y, Ishida K, Matsumoto M, Sakabe T. The relationship between bispectral index and electroencephalographic parameters during isoflurane anaesthesia. *Anesth Analg.* (2004) 98:1336–40. doi: 10.1213/01.ANE.0000105867.17108.B6
33. Connor CW. A Forensic disassembly of the BIS monitor. *Anesth Analg.* (2020) 131:1923–33. doi: 10.1213/ANE.0000000000005220
34. Drover D, Ortega HRR. Patient state index. *Best Pract Res Clin Anaesthesiol.* (2006) 20:121–8. doi: 10.1016/j.bpa.2005.07.008
35. Viertiö-Oja H, Maja V, Särkelä M, Talja P, Tenkanen N, Tolvanen-Laakso H, et al. Description of the entropy algorithm as applied in the datex-ohmeda S/5 entropy module. *Acta Anaesthesiol Scand.* (2004) 48:154–61. doi: 10.1111/j.0001-5172.2004.00322.x
36. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology.* (2018) 129:663–74. doi: 10.1097/ALN.0000000000002300
37. FDIC. *Supervisory Guidance on Model Risk Management.* (2017). Available online at: https://www.fdic.gov/news/financial-institution-letters/2017/fil17022a.pdf

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.

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 | Frontiers in Digital Health

MINI REVIEW
published: 13 June 2022
doi: 10.3389/fdgth.2022.884047

Adjunct Digital Interventions Improve Opioid-Based Pain Management: Impact of Virtual Reality and Mobile Applications on Patient-Centered Pharmacy Care

Hayam Y. Giravi[1]*, *Zack Biskupiak*[2], *Linda S. Tyler*[3] and *Grzegorz Bulaj*[2]*

[1] University of Utah College of Pharmacy, Salt Lake City, UT, United States, [2] Department of Medicinal Chemistry, L.S. Skaggs College of Pharmacy, University of Utah, Salt Lake City, UT, United States, [3] Department of Pharmacotherapy, L.S. Skaggs College of Pharmacy, University of Utah, Salt Lake City, UT, United States

Digital therapeutics (DTx, mobile medical apps, software as a medical device) are rapidly emerging as clinically effective treatments for diverse chronic diseases. For example, the Food and Drug Administration (FDA) has recently authorized a prescription virtual reality (VR) app for treatment of moderate to severe low back pain. The FDA has also approved an adjunct digital therapy in conjunction with buprenorphine for opioid use disorder, further illustrating opportunities to integrate digital therapeutics with pharmacotherapies. There are ongoing needs to disseminate knowledge about advances in digital interventions among health care professionals, policymakers, and the public at large. This mini-review summarizes accumulating clinical evidence of digital interventions delivered *via* virtual reality and mobile apps to improve opioid-based analgesia. We identified relevant randomized controlled trials (RCTs) using Embase and PubMed databases which reported pain scores with a validated pain scale (e.g., visual analog scales, graphic rating scale, numeric rating scale) and use of a digital intervention in conjunction with opiates. Among identified RCTs, the majority of studies reported improved pain scores in the digital intervention group, as compared to "treatment as usual" group. Our work suggests that VR and mobile apps can be used as adjunct digital therapies for pain management. We discuss these findings in the context of how digital health technologies can transform patient-centered pharmacy care.

Keywords: pharmacotherapy, analgesics, mHealth, smartphone apps, therapeutic video games, serious video games, opioid epidemic, health care

OPEN ACCESS

Edited by:
Tim Campellone, University of California, Berkeley, United States

Reviewed by:
Shabbir Syed Abdul, Taipei Medical University, Taiwan

***Correspondence:**
Hayam Y. Giravi
hayam.giravi@pharm.utah.edu
Grzegorz Bulaj
bulaj@pharm.utah.edu

Specialty section:
This article was submitted to Personalized Medicine, a section of the journal Frontiers in Digital Health

Received: *25 February 2022*
Accepted: *16 May 2022*
Published: *13 June 2022*

Citation:
Giravi HY, Biskupiak Z, Tyler LS and Bulaj G (2022) Adjunct Digital Interventions Improve Opioid-Based Pain Management: Impact of Virtual Reality and Mobile Applications on Patient-Centered Pharmacy Care. Front. Digit. Health 4:884047. doi: 10.3389/fdgth.2022.884047

INTRODUCTION

Pain management is a complex, multifaceted challenge that has become a major public health crisis, with an estimated 126.1 million US adults suffering from pain (1). In 2016, over 60 million patients filled or refilled one or more prescriptions for opioid analgesics (1). Although opioid-based analgesia is frequently used to treat both acute and chronic pain, health care professionals (physicians, physician assistants, pharmacists, and nurses) have limited knowledge on opioid analgesic therapies (2). In addition to inadequate pain relief, the use of opioids for

pain management is challenged by significant adverse effects including physical dependence, tolerance, sedation, dizziness, constipation, nausea, vomiting, and respiratory depression (3). Trends in opioid prescription and the associated mortality continue to be problematic not only in the US, but also in other countries (4, 5). Multimodal approaches for pain management such as combination therapy with both nonpharmacological means in addition to traditional pharmacological therapeutics can be effective in achieving optimal control of pain (6, 7). Many aspects of pain management and the opioid epidemic may effectively be addressed by shifting clinical practice to using more non-pharmacological and non-invasive treatments (8), including "digital analgesics" interventions (9–13) and mobile apps to support opioid tapering (14, 15).

Digital health technologies encompass diverse software-based tools which can improve health and therapy outcomes for many chronic diseases. Digital therapeutics (DTx), also known as mobile medical applications, are software-based interventions intended to treat specific medical conditions (16–18). To provide evidence-based therapies, DTx receive marketing authorization (software as a medical device, or SaMD) from regulatory agencies. In the US, the FDA has approved and cleared several digital therapeutics for the treatment of diabetes (type 1 and 2), ADHD, asthma, COPD, chronic low back pain, chronic insomnia, substance use disorder and opioid use disorder. It is also noteworthy that two non-profit organizations, namely The Digital Medicine Society (www.dimesociety.org/) and The Digital Therapeutics Alliance (dtxalliance.org/) are dedicated to advance and promote this rapidly evolving branch of digital health.

Pioneering work on the SnowWorld virtual reality (VR) video game for burn patients illustrates early efforts to bring digital interventions for pain to clinical practice (19–22). There has been an increasing number of clinical studies on VR and mobile apps to improve pain management and relief (6, 9–11, 23–39). For example, a 12-week RCT of a multidisciplinary back pain mobile app (Kaia) showed significant reduction of pain intensity in patients with non-specific low back pain (40). In 2020, the FDA granted a breakthrough medical device designation to a VR app, RelieVRx (previously named EaseVRx), for treatment of intractable low back pain and treatment-resistant fibromyalgia. In 2021, RelieVRx received the FDA authorization for marketing a prescription virtual reality app for treatment of moderate to severe low back pain (41). These advances in digital interventions highlight opportunities for their use as adjunct therapies in combination with diverse analgesic drugs.

To the best of our knowledge, there are currently no review studies focused on effects of digital interventions on opioid-based pain management. Given the increasing number of clinical studies on VR and mobile apps for pain, there is a need for systematic reviews and meta-analyses (SR/MA) of the impact of DTx on different types of pain in combination with analgesic medications. The objectives of this mini-review are: (1) to summarize findings from currently published RCTs focused on adjunct digital interventions (VR and mobile apps) for opioid-based pain management, and (2) to encourage future SR/MA studies on adjunct digital interventions for pain in combination with specific analgesic drugs, including opioids, NSAIDs and others. We further discuss our findings in the context of how digital therapeutics can impact patient-centered pharmacy care.

ADJUNCT DIGITAL INTERVENTIONS FOR OPIOID-BASED ANALGESIA

In order to identify adjunct digital interventions for pain management in conjunction with opioid-based analgesia, EMBASE and PubMed databases were searched for relevant RCTs, systematic reviews and meta-analyses. Database search with keywords "pain," "acute pain," "chronic pain," "cancer pain," "burn pain," "postoperative pain," "pain management," "virtual reality," "VR," "web-based," "phone app," "mobile app," "opioid" and "digital therapeutics" identified nine RCTs which met the following inclusion criteria: (1) reported digital interventions were compared to pharmacological interventions alone, (2) reported pain scores with a validated pain scale (e.g., visual analog scales, graphic rating scale, numeric rating scale), and (3) reported use of concomitant opioids. Studies that did not explicitly report use of opioids or use of a validated pain scale were excluded. In addition to searching the databases, we also examined RCTs evaluated in recent systematic reviews and meta-analyses on digital interventions for pain for those clinical studies that matched inclusion/exclusion criteria mentioned above (30, 32–34).

As summarized in **Table 1**, our search yielded nine RCTs which met inclusion and exclusion criteria. A majority of RCTs examined effects of digital interventions in burn pain patients (a total of $n = 227$), whereas two studies were focused on cancer pain. Regarding types of digital interventions, a vast majority of studies used VR apps. Eight studies demonstrated significant reduction in one or more pain outcomes (19, 42–45, 47–49), whereas one RCT reported no significant changes in pain intensity, as compared to the control groups. Based on the RCTs listed in **Table 1**, these findings suggest that adjunct digital interventions can improve pain scores or reduce medication use in opioid-based analgesia.

Two additional RCTs investigated digital interventions in pain patients taking opioid analgesics, but they did not meet all three inclusion criteria (the comparator was not pharmacological treatment alone) (13, 50). In one RCT examining digital intervention in breast cancer surgery patients, the treatment group showed significant reduction of time (by **5** days) toward cessation of opioid medications, as compared to the control group (digital health education) (13). In another RCTs, VR app intervention (as compared to standard iPad use) did not change postoperative pain scores nor opioid consumption in pediatric patients (50). As discussed below, with more ongoing RCTs studying digital interventions and opioid-based analgesia in pain patients, our results justify near-future SR/MA study to evaluate clinical efficacy of adjunct VR and mobile apps in conjunction with analgesics to improve pain management.

DISCUSSION

There are ongoing needs to mitigate the opioid crisis in the United States (51, 52). To increase awareness about potential

TABLE 1 | Summary of randomized controlled trials of digital interventions in patients with acute or chronic pain.

Author	Study design; Duration or number of sessions	Pain type; Pain scale	Population (*n*), intervention, and comparator description	Concomitant medication(s)	Results
Bani Mohammed et al. (42)	Prospective RCT 40 sessions	Cancer Pain VAS[a]	**Population ($n = 80$):** Women (ages 30–70 years) with breast cancer **Intervention ($n = 40$):** VR (Ocean Rift interactive game or Happy Place non-interactive video) plus morphine **Comparator ($n = 40$):** Oral or intravenous morphine alone	IV or oral morphine	One session of adjunct VR resulted in a significant reduction in pain scores when compared to morphine alone (mean post-VAS score: iVR 0.33 vs. control 4.84; $p < 0.001$)
Carrougher et al. (19)	Within-subject RCT 78 sessions	Burn Pain GRS[b]	**Population ($n = 39$):** Adult burn patients (ages 21–57 years) who required PT **Intervention:** VR (SnowWorld) with pharmacological analgesia during PT **Comparator:** Pharmacological analgesia alone	Oral methadone or OxyContin and a preprocedural short-acting opioid (e.g., oxycodone)	Adjunctive VR significantly reduced worst pain scores by 27% (VR 40 ± 30 vs. control 55 ± 25; $p = 0.004$)
Hoffman et al. (43)	Within-subject RCT 22 sessions	Burn Pain GRS[a]	**Population ($n = 11$):** Pediatric and adult patients (ages 9–40 years) with burns requiring hospitalization **Intervention:** VR (SnowWorld) with pharmacological analgesia during dressing changes **Comparator:** Pharmacological analgesia alone	Standard opioid analgesics and benzodiazepines	Mean pain ratings were lower with adjunct iVR when compared to the control group for all 3 pain measures (worst pain, time spent thinking about pain, and pain unpleasantness); differences were all statistically significant ($p < 0.05$)
Hoffman et al. (44)	Within-subject RCT 24 sessions	Burn Pain VAS[b]	**Population ($n = 12$):** Adult burn patients (ages 19–47 years) **Intervention:** VR (SpiderWorld) with pharmacological analgesia during PT **Comparator:** Pharmacological analgesia alone	Long-acting opioids (typically OxyContin)	All 12 participants reported statistically significant less pain with adjunct VR distraction (worst pain: VR 19.92 vs. control 42, $p = 0.002$; average pain: VR 14.67 vs. control 36.33, $p = 0.002$)
Maani et al. (45)	Within-subject RCT 24 sessions	Burn Pain GRS[a]	**Population ($n = 12$):** US soldiers (ages 20–27 years) with burn wounds **Intervention:** VR (SnowWorld) with pharmacological analgesia during wound debridement **Comparator:** Pharmacological analgesia alone	Fast acting opioids and/or ketamine	Significant difference in mean worst pain scores >7 (iVR 5.67 vs. control 8.33; $p = 0.043$); no significant difference between groups for mild to moderate pain (VR 4.17 vs. control 3.33)
Morris et al. (46)	Within-subject RCT 22 sessions	Burn Pain NRS[a]	**Population ($n = 11$):** Adult burn patients (ages 23–54 years) undergoing physiotherapy **Intervention:** VR (eMagin Z800 3DVisor; game: Chicken Little) with pharmacological analgesia during PT **Comparator:** Pharmacological analgesia alone	Morphine and acetaminophen/ codeine (Dolorol Forte) to all eligible subjects; ibuprofen was given to two subjects	No significant difference in pain reduction between both groups (mean difference = 2.09; 95% CI-0.67 to 4.85, $p = 0.13$)
Schmitt et al. (47)	Within-subject RCT 1 to 5 days	Burn Pain GRS[b]	**Population ($n = 54$):** Hospitalized pediatric (ages 6–19 years) burn patients undergoing physical therapy **Intervention:** VR (SnowWorld) with pharmacological analgesia during PT **Comparator:** Pharmacological analgesia/sedation alone	Oral opioid (e.g., hydromorphone, fentanyl lozenge) +/− oral benzodiazepine (e.g., midazolam)	Significant reduction in cognitive (decreased by 44%), affective (decreased by 32%), and sensory pain (decreased by 27%) with adjunct immersive VR ($p < 0.05$)
Sharar et al. (48)	Within-subject RCT 146 sessions	Burn Pain GRS[b]	**Population ($n = 88$):** Pediatric and adult patients (ages 6–65 years) who required postburn PT **Intervention:** VR (SnowWorld) with pharmacological analgesia during PT **Comparator:** Pharmacological analgesia alone	Systemic opioid and/or benzodiazepine	Significant decrease in worst pain intensity scores in the VR group (VR 43.5 ± 3.5 vs. control 54.2 ± 3.1; $p = 0.003$)

(Continued)

TABLE 1 | Continued

Author	Study design; Duration or number of sessions	Pain type; Pain scale	Population (*n*), intervention, and comparator description	Concomitant medication(s)	Results
Yang et al. (49)	Prospective RCT 4 weeks	Cancer Pain NRS[a]	**Population (*n* = 58):** Adults (ages 18–75 years) with cancer-related pain **Intervention (*n* = 31):** Pain Guard mobile app **Comparator (*n* = 27):** Standard pharmaceutical care	Oxycodone, morphine, methadone, and/or tramadol	Pain Guard significantly decreased the frequency of breakthrough cancer pain (Pain Guard: median 3, IQR 2–7 vs. control: median 13, IQR 9.5–14, $p < 0.001$) and lead to a higher rate of pain remission ($p < 0.001$) with fewer adverse events reported

Only RCTs which compared digital interventions with pharmacological interventions alone are included in this table.
GRS, graphic rating scale; IQR, interquartile range; iVR, immersive virtual reality; n, number of participants; NRS, numerical rating scale; PT, physical therapy; RCT, randomized controlled trial; VAS, visual analog scale; VR, virtual reality; vs., versus.
[a]*Measured on a 0- to 10-cm scale.*
[b]*Measured on a 0- to 100-mm scale.*

TABLE 2 | Examples of mobile and VR applications for pain management.

Developing company	Available applications	Description of application	Mechanism of action	Clinical data
AppliedVR	*RelieVRx*; *EaseVRx*	Marketed for the treatment of moderate to severe low back pain. Manage pain *via* immersive experience, guide patients to desirable clinical outcomes. Opioid sparing clinical treatment. Participants in their VR intervention for 2 weeks endorsed reduced pain catastrophizing scores as well as reduced overall pain.	Theories stemming from Cognitive Behavioral Therapy (CBT) employed in tandem with VR.	Garcia et al. (54) Garcia et al. (9) Spiegel et al. (55)
BreatheVR	*BreatheVR*	BreatheVR is a companion application for the Gear VR and Oculus GO VR setups. 8 of 10 participants in the initial pilot study all reported significant reductions in pain after only short periods of time using BreatheVR.	Deep breathing techniques in combination with a specifically designed relaxation VR landscape.	Mevlevioglu et al. (56)
Flowly	*Flowly*	The Flowly mobile application manages pain using theories from biofeedback in combination with VR to encourage pain management for patients and teach lasting techniques. Participants in their initial trials reported lower pain scores, lower pain catastrophizing scores, and reported needing lower dosages of their opioid medication to manage pain following the intervention.	Use of VR in combination with Flowly's mobile application to teach techniques of biofeedback, promoting pain management.	Flowly (57)
Kaia Health	*Kaia Health*	Musculoskeletal pain care with the use of custom physical therapy or rehabilitation exercise programs. Users report reduction in pain symptoms, reduction in stress symptoms, and further benefits. Accessible, clinical grade PT from the comfort of home.	Use of AI algorithms to guide physical therapy and rehabilitation sessions. Established PT methods such as progressive muscle relaxation. Used in combination with VR for best results.	Biebl et al. (58) Priebe et al. (25)

CBT, cognitive behavioral therapy; VR, virtual reality; PT, physical therapy.

benefits of digital interventions for pain management, this mini-review project focused on whether virtual reality and mobile apps can improve opioid-based analgesia. Our findings suggest that VR applications can offer clinical-grade interventions for opioid-sparing pain management, and are in accord with conclusions from a recent systematic-review and meta-analysis that "*Virtual reality is an effective pain reduction measurement added to analgesics for burn patients undergoing dressing change or physical therapy*." (32). The FDA authorization to market RelieVRx as a prescription virtual reality pain treatment further emphasizes opportunities to combine digital interventions with analgesics (41). It is noteworthy that clinical evidence for digital interventions in pain management is still limited and needs additional multi-center RCTs to validate their clinical efficacy and effectiveness in patients with various pain conditions (30–34, 53).

As shown in **Table 2**, there are several VR and mobile applications currently available for patients and health care providers as tools for improving pain management. RelieVRx has received the FDA authorization (through de novo regulatory pathway) to be marketed as a prescription virtual reality pain treatment for adult patients with chronic low back pain (41). Kaia Health is a mobile app intended for adults with acute or chronic, non-specific musculoskeletal pain, which received class II medical device status in Europe, while is marketed in the US under the FDA enforcement discretion. Flowly VR and biofeedback app is presented as "opioid-sparing pain management device" (www.flowly.world/), but to the best of our knowledge, Flowly has not received the FDA authorization as a medical device, as of writing this mini-review. While digital health technologies are rapidly evolving and expanding, we believe that this article will encourage health care professionals to explore opportunities to integrate digital interventions with pharmacotherapies for improved pain management.

Bringing digital interventions for pain to clinical practice is challenged by complexity of workflow in pain management (59). Mobile apps have been recognized as opportunities to improve pharmacy practice (60–63). Pharmacists often work in interdisciplinary care teams and make recommendations to both providers and patients about pharmacologic and non-pharmacologic interventions, including pain management (64–66). Given an important role of pharmacists in opioid stewardship and prevention of future opioid crisis (67, 68), we hypothesize here that pharmacists recommendations to integrate digital therapeutics with opioid-based analgesia will improve outcomes of opioid tapering programs (69–73). Recently, the Academy of Managed Care Pharmacy convened a forum that brought digital therapeutic innovators, payers, pharmacy benefit managers, and other key stakeholders to discuss the role of digital interventions as therapeutic options (74). While implementation of digital health technologies within health care systems is both inevitable and challenging (75–77), it will be important for payers to consider their health care coverage, especially as more evidence emerges with the potential opportunity of lowering overall health care costs and increasing clinical outcomes. An initial cost-effectiveness analysis of the reimbursement rate for digital therapeutics for low back pain suggests economic benefits for health care in Germany (78).

Integration of digital interventions with drug-based therapies is illustrated by the FDA approval of a prescription adjunct digital therapeutic, namely reSET-O® PDT, in conjunction with buprenorphine for opioid use disorder (OUD). This adjunct digital intervention was shown to improve therapy and health care outcomes, including cost-effectiveness (79–84). From the perspective of long-term therapy outcomes for chronic diseases, patients could benefit from research and development of both adjunct digital therapeutics and drug+digital combination therapies (using drug-device combination product regulatory pathway, where drug is combined with a mobile app approved as SaMD) (18, 85–89). Although drug+digital combination therapies offer a full integration of pharmacotherapy and non-pharmacological intervention, to the best of our knowledge there are no currently known such drug-device combination products. Other future prospects for improved patient-centered pain management may include integration of drug-based analgesia with patient education delivered *via* digital health technologies (29, 90, 91), and integration of digital health technologies with self-care and therapeutic home environment (92).

A limitation of this mini-review is a lack of systematic review methodology and meta-analysis, thus precluding to draw evidence-based conclusions on effectiveness of digital interventions for opioid-based analgesia. Given that clinical studies on digital interventions for reduction of opioid use in pain management is a very active area of research (e.g., from ClinicalTrials.gov: NCT04139564, NCT04010266, NCT03851042, NCT04273919, NCT04416555 and others), it is prudent to wait for more published results from all relevant RCTs. Another limitation of this project is a focus on opioid-based treatments, rather than on opioids and non-steroidal anti-inflammatory drugs (NSAIDs). This is due to a limited number of clinical studies which report use of specific pain medications when evaluating VR or mobile apps in pain management. We hope that despite these limitations, this mini-review will raise awareness on how digital interventions can improve patient-centered pharmacy care for pain and for other medical conditions.

Given complex and unmet needs to address the opioid crisis (52, 93), this review supports several actionable recommendations to be considered. Educating health care professionals, patients and policymakers about the FDA-approved VR and mobile apps for pain should be led by both patient advocacy groups (e.g., The American Chronic Pain Association and the US Pain Foundation) and professional organizations (e.g., The American College of Physicians and The American Academy of Neurologists). Integrated healthcare systems and hospitals can create VR simulation centers for patient education about their diagnosis and treatment options including digital interventions (94, 95). Educating pharmacists, nurses and physician assistants about digital health technologies will accelerate clinical workflow redesign to incorporate their "internal champions" roles in decision making for pain management (64–66, 77, 96). For opioid prescription and tapering for chronic pain, revisions and updates to the CDC guidelines and payer pharmacy coverage should include the use of digital therapeutics for pain relief and management (97). Lastly, increasing social media campaigns (98, 99), and direct-to-consumer advertising of VR and mobile apps for pain will expand public awareness about digital therapeutics, and will also impact prescribing practices in the future (100, 101).

CONCLUSION

Our mini-review suggests that both VR and mobile apps can be used as adjunct digital therapies in conjunction with opioid-based analgesics for pain management. Such interventions, which are applicable to hospital, hospital at home and stay-at-home care, can improve patient-centered pharmacy care and opioid tapering outcomes. Rapidly evolving digital health technologies create opportunities to integrate pharmacotherapies with non-pharmacological treatments for pain, while regulatory approval of commercially available digital interventions as DTx

for pain management is critical for reimbursement and health care implementation.

AUTHOR CONTRIBUTIONS

HG and GB: conceptualization, literature search and review, and manuscript writing. ZB: literature search and review and manuscript writing. LT: literature review and manuscript writing. All authors contributed to the article and approved the submitted version.

FUNDING

GB acknowledges a research support by the ALSAM Foundation Grant.

REFERENCES

1. Hagemeier NE. Introduction to the opioid epidemic: the economic burden on the healthcare system and impact on quality of life. *Am J Manag Care*. (2018) 24:S200–S6.
2. Williamson C, Martin BJ, Argoff C, Gharibo C, McCarberg B, Atkinson T, et al. Pain management and opioid therapy: persistent knowledge gaps among primary care providers. *J Pain Res.* (2021) 14:3223–34. doi: 10.2147/JPR.S316637
3. Benyamin R, Trescot AM, Datta S, et al. Opioid complications and side effects. *Pain Physician*. (2008) 11(2 Suppl):S105–20. doi: 10.36076/ppj.2008/11/S105
4. Kurdi A. Opioids and gabapentinoids utilisation and their related-mortality trends in the United Kingdom primary care setting, 2010–2019: a cross-national, population-based comparison study. *Front. Pharmacol.* (2021) 12:732345. doi: 10.3389/fphar.2021.732345
5. Bedson J, Chen Y, Hayward RA, Ashworth J, Walters K, Dunn KM, et al. Trends in long-term opioid prescribing in primary care patients with musculoskeletal conditions: an observational database study. *Pain.* (2016) 157:1525–31. doi: 10.1097/j.pain.0000000000000557
6. Shebib R, Bailey JF, Smittenaar P, Perez DA, Mecklenburg G, Hunter S. Randomized controlled trial of a 12-week digital care program in improving low back pain. *NPJ Digit Med.* (2019) 2:1. doi: 10.1038/s41746-018-0076-7
7. Amorim AB, Pappas E, Simic M, Ferreira ML, Jennings M, Tiedemann A, et al. Integrating Mobile-health, health coaching, and physical activity to reduce the burden of chronic low back pain trial (IMPACT): a pilot randomised controlled trial. *BMC Musculoskelet Disord.* (2019) 20:71. doi: 10.1186/s12891-019-2454-y
8. Qaseem A, Wilt TJ, McLean RM, Forciea MA, Clinical Guidelines Committee of the American College of P. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the american college of physicians. *Ann Intern Med*. (2017) 166:514–530. doi: 10.7326/M16-2367
9. Garcia LM, Birckhead BJ, Krishnamurthy P, Sackman J, Mackey IG, Louis RG, et al. An 8-week self-administered at-home behavioral skills-based virtual reality program for chronic low back pain: double-blind, randomized, placebo-controlled trial conducted during COVID-19. *J Med Internet Res.* (2021) 23:e26292. doi: 10.2196/26292
10. Won AS, Bailey J, Bailenson J, Tataru C, Yoon IA, Golianu B. Immersive virtual reality for pediatric pain. *Children*. (2017) 4:52. doi: 10.3390/children4070052
11. Irvine AB, Russell H, Manocchia M, Mino DE, Cox Glassen T, Morgan R, et al. Mobile-Web app to self-manage low back pain: randomized controlled trial. *J Med Internet Res.* (2015) 17:e1. doi: 10.2196/jmir.3130
12. Pronk Y, Peters M, Sheombar A, Brinkman JM. Effectiveness of a mobile eHealth app in guiding patients in pain control and opiate use after total knee replacement: randomized controlled trial. *JMIR Mhealth Uhealth.* (2020) 8:e16415. doi: 10.2196/16415
13. Darnall BD, Ziadni MS, Krishnamurthy P, Flood P, Heathcote LC, Mackey IG, et al. "My surgical success": effect of a digital behavioral pain medicine intervention on time to opioid cessation after breast cancer surgery-a pilot randomized controlled clinical trial. *Pain Med.* (2019) 20:2228–37. doi: 10.1093/pm/pnz094
14. Magee M, Gholamrezaei A, McNeilage AG, Dwyer L, Sim A, Ferreira M, et al. Evaluating acceptability and feasibility of a mobile health intervention to improve self-efficacy in prescription opioid tapering in patients with chronic pain: protocol for a pilot randomised, single-blind, controlled trial. *BMJ Open.* (2022) 12:e057174. doi: 10.1136/bmjopen-2021-057174
15. Magee MR, McNeilage AG, Avery N, Glare P, Ashton-James CE. mHealth interventions to support prescription opioid tapering in patients with chronic pain: qualitative study of patients' perspectives. *JMIR Form Res.* (2021) 5:e25969. doi: 10.2196/25969
16. Patel NA, Butte AJ. Characteristics and challenges of the clinical pipeline of digital therapeutics. *NPJ Digit Med.* (2020) 3:159. doi: 10.1038/s41746-020-00370-8
17. Shuren J, Patel B, Gottlieb S, FDA. Regulation of mobile medical apps. *JAMA*. (2018) 320:337–8. doi: 10.1001/jama.2018.8832
18. Sverdlov O, van Dam J, Hannesdottir K, Thornton-Wells T. Digital therapeutics: an integral component of digital innovation in drug development. *Clin Pharmacol Ther.* (2018) 104:72–80. doi: 10.1002/cpt.1036
19. Carrougher GJ, Hoffman HG, Nakamura D, Lezotte D, Soltani M, Leahy L, et al. The effect of virtual reality on pain and range of motion in adults with burn injuries. *J Burn Care Res.* (2009) 30:785–91. doi: 10.1097/BCR.0b013e3181b485d3
20. Hoffman HG. Virtual-reality therapy. *Sci Am.* (2004) 291:58–65. doi: 10.1038/scientificamerican0804-58
21. Hoffman HG, Seibel EJ, Richards TL, Furness TA, Patterson DR, Sharar SR. Virtual reality helmet display quality influences the magnitude of virtual reality analgesia. *J Pain.* (2006) 7:843–50. doi: 10.1016/j.jpain.2006.04.006
22. Hoffman HG, Chambers GT, Meyer WJ 3rd, Arceneaux LL, Russell WJ, Seibel EJ, et al. Virtual reality as an adjunctive non-pharmacologic analgesic for acute burn pain during medical procedures. *Ann Behav Med.* (2011) 41:183–91. doi: 10.1007/s12160-010-9248-7
23. Bailey JF, Agarwal V, Zheng P, Smuck M, Fredericson M, Kennedy DJ, et al. Digital care for chronic musculoskeletal pain: 10,000 participant longitudinal cohort study. *J Med Internet Res.* (2020) 22:e18250. doi: 10.2196/18250
24. Thurnheer SE, Gravestock I, Pichierri G, Steurer J, Burgstaller JM. Benefits of mobile apps in pain management: systematic review. *JMIR Mhealth Uhealth.* (2018) 6:e11231. doi: 10.2196/11231
25. Priebe JA, Haas KK, Moreno Sanchez LF, Schoefmann K, Utpadel-Fischler DA, Stockert P, et al. Digital treatment of back pain versus standard of care: the cluster-randomized controlled trial, rise-uP. *J Pain Res.* (2020) 13:1823–38. doi: 10.2147/JPR.S260761
26. Chi B, Chau B, Yeo E, Ta P. Virtual reality for spinal cord injury-associated neuropathic pain: Systematic review. *Ann Phys Rehabil Med.* (2018) 62:49–57. doi: 10.1016/j.rehab.2018.09.006
27. Tashjian VC, Mosadeghi S, Howard AR, Lopez M, Dupuy T, Reid M, et al. Virtual reality for management of pain in hospitalized patients: results of a controlled trial. *JMIR Ment Health.* (2017) 4:e9. doi: 10.2196/mental.7387
28. Darnall BD, Krishnamurthy P, Tsuei J, Minor JD. Self-administered skills-based virtual reality intervention for chronic pain: randomized controlled pilot study. *JMIR Form Res.* (2020) 4:e17293. doi: 10.2196/17293
29. Garcia LM, Birckhead BJ, Krishnamurthy P, Mackey I, Sackman J, Salmasi V, et al. Three-month follow-up results of a double-blind, randomized placebo-controlled trial of 8-week self-administered at-home behavioral skills-based virtual reality (VR) for chronic low back pain. *J Pain.* (2021) 23:822–840. doi: 10.1016/j.jpain.2021.12.002
30. Huang Q, Lin J, Han R, Peng C, Huang A. Using virtual reality exposure therapy in pain management: a systematic review and meta-analysis of randomized controlled trials. *Value Health.* (2022) 25:288–301. doi: 10.1016/j.jval.2021.04.1285

31. Lewkowicz D, Slosarek T, Wernicke S, Winne A, Wohlbrandt AM, Bottinger E. Digital therapeutic care and decision support interventions for people with low back pain: systematic review. *JMIR Rehabil Assist Technol.* (2021) 8:e26612. doi: 10.2196/26612
32. Luo H, Cao C, Zhong J, Chen J, Cen Y. Adjunctive virtual reality for procedural pain management of burn patients during dressing change or physical therapy: a systematic review and meta-analysis of randomized controlled trials. *Wound Repair Regen.* (2019) 27:90–101. doi: 10.1111/wrr.1
33. Chuan A, Zhou JJ, Hou RM, Stevens CJ, Bogdanovych A. Virtual reality for acute and chronic pain management in adult patients: a narrative review. *Anaesthesia.* (2021) 76:695–704. doi: 10.1111/anae.15202
34. Zheng C, Chen X, Weng L, Guo L, Xu H, Lin M, et al. Benefits of Mobile Apps for Cancer Pain Management: Systematic Review. *JMIR Mhealth Uhealth.* (2020) 8:e17055. doi: 10.2196/17055
35. O'Connor S, Mayne A, Hood B. Virtual reality-based mindfulness for chronic pain management: a scoping review. *Pain Manag Nurs.* (2022). doi: 10.1016/j.pmn.2022.03.013
36. Grassini S. Virtual reality assisted non-pharmacological treatments in chronic pain management: a systematic review and quantitative meta-analysis. *Int J Environ Res Public Health.* (2022) 19. doi: 10.3390/ijerph19074071
37. Găină MA, Szalontay AS, Ştefănescu G, Bălan GG, Ghiciuc CM, Bolo? A, et al. State-of-the-art review on immersive virtual reality interventions for colonoscopy-induced anxiety and pain. *J Clin Med.* (2022) 11:1670. doi: 10.3390/jcm11061670
38. He ZH, Yang HM, Dela Rosa RD, De Ala MB. The effects of virtual reality technology on reducing pain in wound care: a meta-analysis and systematic review. *Int Wound J.* (2022). doi: 10.1111/iwj.13785[Epub ahead of print].
39. Nagpal AS, Raghunandan A, Tata F, Kibler D, McGeary D. Virtual reality in the management of chronic low back pain: a scoping review. *Front Pain Res.* (2022) 3:856935. doi: 10.3389/fpain.2022.856935
40. Toelle TR, Utpadel-Fischler DA, Haas KK, Priebe JA. App-based multidisciplinary back pain treatment versus combined physiotherapy plus online education: a randomized controlled trial. *Npj Digit Med.* (2019) 2:34. doi: 10.1038/s41746-019-0109-x
41. Rubin R. Virtual reality device is authorized to relieve back pain. *JAMA.* (2021) 326:2354. doi: 10.1001/jama.2021.22223
42. Bani Mohammad E, Ahmad M. Virtual reality as a distraction technique for pain and anxiety among patients with breast cancer: a randomized control trial. *Palliative and Supportive Care.* (2019) 17:29–34. doi: 10.1017/S1478951518000639
43. Hoffman HG, Patterson DR, Seibel E, Soltani M, Leahy L, Sharar SR. Virtual reality pain control during burn wound debridement in the hydrotank. *Clin J Pain.* (2008) 24:299–304. doi: 10.1097/AJP.0b013e318164d2cc
44. Hoffman HG, Patterson DR, Carrougher GJ. Use of virtual reality for adjunctive treatment of adult burn pain during physical therapy: a controlled study. *Clin J Pain.* (2000) 16:244–50. doi: 10.1097/00002508-200009000-00010
45. Maani CV, Hoffman HG, Morrow M, Maiers A, Gaylord K, McGhee LL, et al. Virtual reality pain control during burn wound debridement of combat-related burn injuries using robot-like arm mounted VR goggles. *Journal of Trauma: Injury, Infection and Critical Care.* (2011) 71:S125–S30. doi: 10.1097/TA.0b013e31822192e2
46. Morris LD, Louw QA, Crous LC. Feasibility and potential effect of a low-cost virtual reality system on reducing pain and anxiety in adult burn injury patients during physiotherapy in a developing country. *Burns.* (2010) 36:659–64. doi: 10.1016/j.burns.2009.09.005
47. Schmitt YS, Hoffman HG, Blough DK, Patterson DR, Jensen MP, Soltani M, et al. A randomized, controlled trial of immersive virtual reality analgesia, during physical therapy for pediatric burns. *Burns.* (2011) 37:61–8. doi: 10.1016/j.burns.2010.07.007
48. Sharar SR, Carrougher GJ, Nakamura D, Hoffman HG, Blough DK, Patterson DR. Factors influencing the efficacy of virtual reality distraction analgesia during postburn physical therapy: preliminary results from 3 ongoing studies. *Arch Phys Med Rehabil.* (2007) 88:S43–S9. doi: 10.1016/j.apmr.2007.09.004
49. Yang J, Weng L, Chen Z, Cai H, Lin X, Hu Z, et al. Development and testing of a mobile app for pain management among cancer patients discharged from hospital treatment: randomized controlled trial. *JMIR mHealth and uHealth.* (2019) 7:e12542. doi: 10.2196/12542
50. Specht BJ, Buse CR, Phelps JR, Phillips MR, Chiavacci SD, Harrell LE, et al. Virtual reality after surgery—a method to decrease pain after surgery in pediatric patients. *Am Surg.*
51. Stoicea N, Costa A, Periel L, Uribe A, Weaver T, Bergese SD. Current perspectives on the opioid crisis in the US healthcare system: a comprehensive literature review. *Medicine.* (2019) 98:e15425. doi: 10.1097/MD.0000000000015425
52. Humphreys K, Shover CL, Andrews CM, Bohnert ASB, Brandeau ML, Caulkins JP, et al. Responding to the opioid crisis in North America and beyond: recommendations of the Stanford-Lancet Commission. *Lancet.* (2022) 399:555–604. doi: 10.1016/S0140-6736(21)02252-2
53. Pfeifer AC, Uddin R, Schröder-Pfeifer P, Holl F, Swoboda W, Schiltenwolf M. Mobile application-based interventions for chronic pain patients: a systematic review and meta-analysis of effectiveness. *J Clin Med.* (2020) 9:3557. doi: 10.3390/jcm9113557
54. Garcia LM, Darnall BD, Krishnamurthy P, Mackey IG, Sackman J, Louis RG, et al. Self-administered behavioral skills-based at-home virtual reality therapy for chronic low back pain: protocol for a randomized controlled trial. *JMIR Res Protoc.* (2021) 10:e25291. doi: 10.2196/25291
55. Spiegel B, Fuller G, Lopez M, Dupuy T, Noah B, Howard A, et al. Virtual reality for management of pain in hospitalized patients: a randomized comparative effectiveness trial. *PLoS ONE.* (2019) 14:e0219115. doi: 10.1371/journal.pone.0219115
56. Mevlevioglu D, Murphy D, Tabirca S. Visual respiratory feedback in virtual reality exposure therapy: a pilot study. *ACM International Conference on Interactive Media Experiences.* Virtual Event, USA: Association for Computing Machinery (2021). p. 1–6. doi: 10.1145/3452918.3458799
57. Flowly. *Benefits of Virtual Reality Biofeedback for Pain Management* (2021).
58. Biebl JT, Rykala M, Strobel M, Kaur Bollinger P, Ulm B, Kraft E, et al. App-based feedback for rehabilitation exercise correction in patients with knee or hip osteoarthritis: prospective cohort study. *J Med Internet Res.* (2021) 23:e26658. doi: 10.2196/26658
59. Sarkar U, Lee JE, Nguyen KH, Lisker S, Lyles CR. Barriers and facilitators to the implementation of virtual reality as a pain management modality in academic, community, and safety-net settings: qualitative analysis. *J Med Internet Res.* (2021) 23:e26623. doi: 10.2196/26623
60. Aungst TD. Integrating mHealth and mobile technology education into the pharmacy curriculum. *Am J Pharm Educ.* (2014) 78:19. doi: 10.5688/ajpe78119
61. Aungst TD. Medical applications for pharmacists using mobile devices. *Ann Pharmacother.* (2013) 47:1088–95. doi: 10.1345/aph.1S035
62. Aungst TD, Miranda AC, Serag-Bolos ES. How mobile devices are changing pharmacy practice. *Am J Health Syst Pharm.* (2015) 72:494–500. doi: 10.2146/ajhp140139
63. AMCP partnership forum: digital therapeutics-what are they and where do they fit in pharmacy and medical benefits? *J Manag Care Spec Pharm.* (2020) 26:674–81. doi: 10.18553/jmcp.2020.19418
64. Boren LL, Locke AM, Friedman AS, Blackmore CC, Woolf R. Team-based medicine: incorporating a clinical pharmacist into pain and opioid practice management. *PM&R.* (2019) 11:1170–7. doi: 10.1002/pmrj.12127
65. Kang I, Urick B, Vohra R, Ives TJ. Physician-pharmacist collaboration on chronic non-cancer pain management during the opioid crisis: a qualitative interview study. *Res Social Adm Pharm.* (2019) 15:1027–31. doi: 10.1016/j.sapharm.2019.04.052
66. Giannitrapani KF, Glassman PA, Vang D, McKelvey JC, Thomas Day R, Dobscha SK, et al. Expanding the role of clinical pharmacists on interdisciplinary primary care teams for chronic pain and opioid management. *BMC Fam Pract.* (2018) 19:107. doi: 10.1186/s12875-018-0783-9
67. Salwan A, Hagemeier NE, Tudiver F, Dowling-McClay K, Foster KN, Arnold J, et al. Community pharmacist engagement in opioid use disorder prevention and treatment behaviors: a descriptive analysis. *J Am Pharm Assoc.* (2003) 60:e173–e8. doi: 10.1016/j.japh.2020.06.008
68. Chisholm-Burns MA, Spivey CA, Sherwin E, Wheeler J, Hohmeier K. The opioid crisis: Origins, trends, policies, and the roles of pharmacists. *Am J Health Syst Pharm.* (2019) 76:424–35. doi: 10.1093/ajhp/zxy089

69. Firemark AJ, Schneider JL, Kuntz JL, Papajorgji-Taylor D, Dickerson JF, Thorsness LA, et al. "We need to taper." Interviews with clinicians and pharmacists about use of a pharmacy-led opioid tapering program. *Pain Med.* (2021) 22:1213–22. doi: 10.1093/pm/pnaa442
70. Kuntz JL, Schneider JL, Firemark AJ, Dickerson JF, Papajorgji-Taylor D, Reese KR, et al. A pharmacist-led program to taper opioid use at kaiser permanente northwest: rationale, design, and evaluation. *Perm J.* (2020) 24:19.216. doi: 10.7812/TPP/19.216
71. Page J, Traver R, Patel S, Saliba C. Implementation of a proactive pilot health plan-driven opioid tapering program to decrease chronic opioid use for conditions of the back and spine in a medicaid population. *J Manag Care Spec Pharm.* (2018) 24:191–6. doi: 10.18553/jmcp.2018.24.3.191
72. Hundley L, Spradley S, Donelenko S. Assessment of outcomes following high-dose opioid tapering in a Veterans Healthcare System. *J Opioid Manag.* (2018) 14:89–101. doi: 10.5055/jom.2018.0436
73. Darnall BD, Fields HL. Clinical and neuroscience evidence supports the critical importance of patient expectations and agency in opioid tapering. *Pain.* (2022) 163:824–6. doi: 10.1097/j.pain.0000000000002443
74. AMCP Partnership Forum Develops Steps to Strengthen Evaluation of Digital Therapeutics: Academy of Managed Care Pharmacy (2021).
75. Kubo A, Kurtovich E, McGinnis M, Aghaee S, Altschuler A, Quesenberry C Jr, et al. A randomized controlled trial of mhealth mindfulness intervention for cancer patients and informal cancer caregivers: a feasibility study within an integrated health care delivery system. *Integr Cancer Ther.* (2019) 18:1534735419850634. doi: 10.1177/1534735419850634
76. Avalos LA, Aghaee S, Kurtovich E, Quesenberry C Jr, Nkemere L, McGinnis MK, et al. A mobile health mindfulness intervention for women with moderate to moderately severe postpartum depressive symptoms: feasibility study. *JMIR Ment Health.* (2020) 7:e17405. doi: 10.2196/17405
77. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. *Npj Digit Med.* (2022) 5:13. doi: 10.1038/s41746-022-00557-1
78. Lewkowicz D, Wohlbrandt AM, Bottinger E. Digital therapeutic care apps with decision-support interventions for people with low back pain in germany: cost-effectiveness analysis. *JMIR Mhealth Uhealth.* (2022) 10:e35042. doi: 10.2196/35042
79. Velez FF, Colman S, Kauffman L, Ruetsch C, Anastassopoulos K. Real-world reduction in healthcare resource utilization following treatment of opioid use disorder with reSET-O, a novel prescription digital therapeutic. *Expert Rev Pharmacoecon Outcomes Res.* (2021) 21:69–76. doi: 10.1080/14737167.2021.1840357
80. Maricich YA, Xiong X, Gerwien R, Kuo A, Velez F, Imbert B, et al. Real-world evidence for a prescription digital therapeutic to treat opioid use disorder. *Curr Med Res Opin.* (2021) 37:175–83. doi: 10.1080/03007995.2020.1846023
81. Maricich YA, Gerwien R, Kuo A, Malone DC, Velez FF. Real-world use and clinical outcomes after 24 weeks of treatment with a prescription digital therapeutic for opioid use disorder. *Hosp Pract.* (1995) 2021:1–8. doi: 10.1080/21548331.2021.1974243
82. Velez FF, Huang D, Mody L, Malone DC. Five-year budget impact of a prescription digital therapeutic for patients with opioid use disorder. *Expert Rev Pharmacoecon Outcomes Res.* (2022) 1–9. doi: 10.1080/14737167.2022.2016396
83. Velez FF, Malone DC. Cost-Effectiveness analysis of a prescription digital therapeutic for the treatment of opioid use disorder. *J Mark Access Health Policy.* (2021) 9:1966187. doi: 10.1080/20016689.2021.1966187
84. Velez FF, Luderer HF, Gerwien R, Parcher B, Mezzio D, Malone DC. Evaluation of the cost-utility of a prescription digital therapeutic for the treatment of opioid use disorder. *Postgrad Med.* (2021) 133:421–7. doi: 10.1080/00325481.2021.1884471
85. Bulaj G. Combining non-pharmacological treatments with pharmacotherapies for neurological disorders: a unique interface of the brain, drug-device, and intellectual property. *Front Neurol.* (2014) 5:126. doi: 10.3389/fneur.2014.00126
86. Metcalf CS, Huntsman M, Garcia G, Kochanski AK, Chikinda M, Watanabe E, et al. Music-enhanced analgesia and antiseizure activities in animal models of pain and epilepsy: toward preclinical studies supporting development of digital therapeutics and their combinations with pharmaceutical drugs. *Front Neurol.* (2019) 10:277. doi: 10.3389/fneur.2019.00277
87. Raijada D, Wac K, Greisen E, Rantanen J, Genina N. Integration of personalized drug delivery systems into digital health. *Adv Drug Deliv Rev.* (2021) 176:113857. doi: 10.1016/j.addr.2021.113857
88. Bulaj G, Clark J, Ebrahimi M, Bald E. From precision metapharmacology to patient empowerment: delivery of self-care practices for epilepsy, pain, depression and cancer using digital health technologies. *Front Pharmacol.* (2021) 12:612602. doi: 10.3389/fphar.2021.612602
89. Afra P, Bruggers CS, Sweney M, Fagatele L, Alavi F, Greenwald M, et al. Mobile software as a medical device (SaMD) for the treatment of epilepsy: development of digital therapeutics comprising behavioral and music-based interventions for neurological disorders. *Front Hum Neurosci.* (2018) 12:171. doi: 10.3389/fnhum.2018.00171
90. Darnall BD, Roy A, Chen AL, Ziadni MS, Keane RT, You DS, et al. Comparison of a single-session pain management skills intervention with a single-session health education intervention and 8 sessions of cognitive behavioral therapy in adults with chronic low back pain: a randomized clinical trial. *JAMA Netw Open.* (2021) 4:e2113401. doi: 10.1001/jamanetworkopen.2021.13401
91. Ziadni MS, Gonzalez-Castro L, Anderson S, Krishnamurthy P, Darnall BD. Efficacy of a single-session "empowered relief" zoom-delivered group intervention for chronic pain: randomized controlled trial conducted during the COVID-19 pandemic. *J Med Internet Res.* (2021) 23:e29672. doi: 10.2196/29672
92. Huntsman DD, Bulaj G. Healthy dwelling: design of biophilic interior environments fostering self-care practices for people living with migraines, chronic pain, and depression. *Int J Environ Res Public Health*. (2022) 19:2248. doi: 10.3390/ijerph19042248
93. Volkow ND, Blanco C. The changing opioid crisis: development, challenges and opportunities. *Mol Psychiatry.* (2021) 26:218–33. doi: 10.1038/s41380-020-0661-4
94. Bekelis K, Calnan D, Simmons N, MacKenzie TA, Kakoulides G. Effect of an immersive preoperative virtual reality experience on patient reported outcomes: a randomized controlled trial. *Ann Surg.* (2017) 265:1068–73. doi: 10.1097/SLA.0000000000002094
95. Chen G, Zhao Y, Xie F, Shi W, Yang Y, Yang A, et al. Educating outpatients for bowel preparation before colonoscopy using conventional methods vs virtual reality videos plus conventional methods: a randomized clinical trial. *JAMA Netw Open.* (2021) 4:e2135576. doi: 10.1001/jamanetworkopen.2021.35576
96. Lagisetty P, Smith A, Antoku D, Winter S, Smith M, Jannausch M, et al. A physician-pharmacist collaborative care model to prevent opioid misuse. *Am J Health Syst Pharm.* (2020) 77:771–80. doi: 10.1093/ajhp/zxaa060
97. Togun AT, Karaca-Mandic P, Wurtz R, Jeffery MM, Beebe T. Association of 3 CDC opioid prescription guidelines for chronic pain and 2 payer pharmacy coverage changes on opioid initiation practices. *J Manag Care Spec Pharm.* (2021) 27:1352–64. doi: 10.18553/jmcp.2021.27.10.1352
98. Allen HG, Stanton TR, Di Pietro F, Moseley GL. Social media release increases dissemination of original articles in the clinical pain sciences. *PLoS ONE.* (2013) 8:e68914. doi: 10.1371/journal.pone.0068914
99. Suman A, Armijo-Olivo S, Deshpande S, Marietta-Vasquez J, Dennett L, Miciak M, et al. A systematic review of the effectiveness of mass media campaigns for the management of low back pain. *Disabil Rehabil.* (2021) 43:3523–51. doi: 10.1080/09638288.2020.1743777
100. Beilfuss S, Linde S. Pharmaceutical opioid marketing and physician prescribing behavior. *Health Econ.* (2021) 30:3159–85. doi: 10.1002/hec.4424
101. Mackey TK, Cuomo RE, Liang BA. The rise of digital direct-to-consumer advertising? comparison of direct-to-consumer advertising expenditure trends from publicly available data sources and global policy implications. *BMC Health Serv Res.* (2015) 15:236. doi: 10.1186/s12913-015-0885-1

Conflict of Interest: GB is a founder and owner of OMNI Self-care, LLC, a health promotion company creating digital content for disease self-management and is a

co-inventor on two issued US patents 9,569,562 and 9,747,423 "Disease Therapy Game Technology" and patent-pending application "Multimodal Platform for Treating Epilepsy". These patents are related to digital health technologies, and are owned by the University of Utah.

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.

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 | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 17 November 2022
DOI 10.3389/fdgth.2022.976001

OPEN ACCESS

EDITED BY
Kirsten Smayda,
MedRhythms, United States

REVIEWED BY
Michael Musker,
University of South Australia, Australia
Salman Razzaki,
Zinc VC, United Kingdom

*CORRESPONDENCE
Robert N. Cuyler
cuyler@freespira.com

SPECIALTY SECTION
This article was submitted to Personalized Medicine, a section of the journal Frontiers in Digital Health

RECEIVED 22 June 2022
ACCEPTED 21 October 2022
PUBLISHED 17 November 2022

CITATION
Cuyler RN, Katdare R, Thomas S and Telch MJ (2022) Real-world outcomes of an innovative digital therapeutic for treatment of panic disorder and PTSD: A 1,500 patient effectiveness study.
Front. Digit. Health 4:976001.
doi: 10.3389/fdgth.2022.976001

Real-world outcomes of an innovative digital therapeutic for treatment of panic disorder and PTSD: A 1,500 patient effectiveness study

Robert N. Cuyler[1]*, Rahul Katdare[2], Simon Thomas[2] and Michael J. Telch[3]

[1]Freespira, Inc., Houston, TX, United States, [2]Freespira, Inc., Kirkland, WA, United States, [3]Laboratory for the Study of Anxiety Disorders, University of Texas, Austin, TX, United States

Objective: Prior clinical trials have shown consistent clinical benefit for Capnometry Guided Respiratory Intervention (CGRI), a prescription digital therapeutic for the treatment of panic disorder (PD) and post-traumatic stress disorder (PTSD). The purpose of this study is to report real-world outcomes in a series of patients treated with the intervention in clinical practice.
Design: This paper reports pre- and post-treatment self-reported symptom reduction, measures of respiratory rate and end-tidal carbon dioxide levels, drop-out and adherence rates drawn from an automatic data repository in a large real-world series of patients receiving CGRI for panic disorder and PTSD.
Setting: Patients used the intervention in their homes, supported by telehealth coaching.
Participants: Patients meeting symptom criteria for panic disorder ($n = 1{,}395$) or posttraumatic stress disorder ($n = 174$) were treated following assessment by a healthcare professional.
Intervention: Capnometry Guided Respiratory Intervention is a 28-day home-based treatment that provides breath-to-breath feedback of respiratory rate and exhaled carbon dioxide levels, aimed at normalizing respiratory style and increasing patients' mastery for coping with symptoms of stress, anxiety, and panic. Health coaches provide initial training with weekly follow up during the treatment episode. Remote data upload and monitoring facilitates individualized coaching and aggregate outcomes analysis.
Main outcome measures: Self-reported Panic Disorder Severity Scale (PDSS) and the Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5) scores were obtained at pre-treatment and post-treatment.
Results: Panic disorder (PD) patients showed a mean pre-to-post-treatment reduction in total PDSS scores of 50.2% ($P < 0.001$, $d = 1.31$). Treatment response rates for PD (defined as a 40% or greater reduction in PDSS total scores) were observed in 65.3% of the PD patients. PTSD patients showed a pre-to-post-treatment reduction in total PCL-5 scores of 41.1% ($P < 0.001$, $d = 1.16$). The treatment response rate for PTSD (defined as a ≥10-point reduction in PCL-5 scores) was 72.4%. In an additional analysis of response at the individual level, 55.7% of panic disorder patients and 53.5% of PTSD patients were classified as treatment responders using the Reliable Change Index. Patients with both normal and below-normal baseline exhaled CO_2

levels experienced comparable benefit. Across the 28-day treatment period, mean adherence rates of 74.8% (PD) and 74.9% (PTSD) were recorded during the 28-day treatment. Dropout rates were 10% (PD) and 11% (PTSD) respectively.

Conclusions: The results from this cohort of 1,569 patients treated with the CGRI intervention demonstrate significant rates of symptom reduction and adherence consistent with prior published clinical trials. The brief duration of treatment, high adherence rates, and clinical benefit suggests that CGRI provides an important addition to treatment options for panic disorder and PTSD.

KEYWORDS

CGRI, panic disorder, PTSD, digital therapeutic, telehealth, biofeedback, carbon dioxide hypersensitivity

Introduction

Panic disorder (PD) and Post-traumatic Stress Disorder (PTSD) are common and often become chronic behavioral health conditions. Lifetime prevalence for isolated panic attacks is reported at 22.7% and lifetime prevalence for full criteria panic disorder is estimated at 4.8% (1). Estimates of lifetime PTSD prevalence range from 3.4% to 8.0% in the general population and 7.7% to 13.4% in veterans (2). The most widely utilized and recommended current treatments are psychotropic medications and/or psychotherapy. In the case of panic disorder and recurrent panic symptoms, review of pharmacologic treatments shows substantial rates of inadequate response, with many patients experiencing chronic relapsing conditions (3). When the nature of actual care delivered in routine clinical practice rather than published trials is examined, most patients with panic symptoms are seen in the general medical sector and the majority of patients do not receive evidence-based pharmacologic or psychotherapeutic treatments (1). Similarly, PTSD often takes a chronic course with up to a third of individuals remaining symptomatic a decade after trauma exposure (4). A meta-analysis of placebo-controlled studies of cognitive behavioral therapies show small to medium effect sizes for PD and PTSD as compared to more robust results for OCD, social anxiety disorder, and acute stress disorder (5).

A common limitation of psychotherapeutic and pharmacologic approaches to PD and PTSD is that neither address the role of respiratory physiology and breathing style. A very useful review by Boulding and colleagues (6) examines relevant literature and proposes a useful classification of respiratory styles (termed "dysfunctional breathing") implicated in a spectrum of health conditions including panic. More specifically, a substantial body of work posits situational as well as chronic dysfunctional breathing as risk factors in panic attacks and the subsequent development of panic disorder (7). Evidence supporting the respiratory dysregulation hypothesis comes from a substantial body of work linking CO_2 hypersensitivity to panic attacks and panic disorder, initiated in large part by Klein's conceptualization of a faulty suffocation alarm (8, 9). An important "marker" of this hypersensitivity comes from studies of carbon dioxide challenge testing. In experimental lab settings, researchers have established that compared to healthy controls, most panic sufferers (and many close relatives) react with pronounced panic symptoms, including fear and physiological distress, when exposed to single or repeated breaths of CO_2- enriched air (10–15).

A smaller body of work has identified similar reactivity for individuals with PTSD. A double-blind, randomized control study of reactivity to CO_2 challenge showed that diagnosed PTSD patients were highly reactive to inhaled 35% CO_2 but not to a placebo gas mixture, while healthy controls were largely unaffected (16). In addition, soldiers who demonstrated high distress during CO_2 challenge were found to be at higher risk than non-reactors for developing PTSD symptoms during deployment to Iraq (17).

This evidence of CO_2 hypersensitivity as a common risk factor for both panic and PTSD, as well as evidence of a bidirectional relationship between the two conditions (18), provided a compelling rationale that led our treatment development team to develop Freespira®, a digital therapeutic specifically targeting normalization of dysfunctional breathing patterns *via* Capnometry-Guided Respiratory Intervention (CGRI); the intervention received FDA-clearance for treatment of panic disorder in 2013 and in 2018 for PTSD.

Origins and previous efficacy trials

Research conducted by Meuret and colleagues (19) established a treatment protocol (Capnometry Assisted Respiratory Therapy, or CART) using feedback of respiratory rate and exhaled CO_2. This trial showed significant and sustained symptomatic improvement in panic disorder severity with reported 93% treatment response (≥40% reduction in Panic Disorder Severity Scale (PDSS) scores) one-year post-treatment and large effect size (Cohen's $d = 2.6$). This trial used commercially available capnometers and cassette-tape-recorded pacing tones for delivery of the CART protocol.

The CGRI intervention described in this paper represents an adaptation of the core Meuret protocol, using different instrumentation as well as imbedded data capture and remote review capabilities. A multi-center benchmarking study (20) offering CGRI in four independent anxiety treatment centers showed one-year response rates (≥40% reduction in PDSS scores) of 82% in treatment completers and large effect size (Cohen's $d = 2.3$). Additionally, the authors identified subsets of participants classified as hypocapnic or normocapnic based on baseline averages of $etCO_2$. Hypocapnic subjects experienced greater increases in exhaled carbon dioxide levels at post-treatment but the authors determined that the intervention produced equivalent clinical benefit post-treatment for normocapnic as well as hypocapnic subjects.

A health economic outcome study (21) undertaken by Highmark Health and Allegheny Health Network reported 91% response rates and 68% remission rates (PDSS scores ≤5) one-year post-CGRI treatment. Highmark, the insurer of the participant patients, compiled cost data by comparing paid healthcare claims (all sources) for the one year prior to and one year following the 28-day treatment. The study reported a 35% reduction in overall paid claims, a 65% reduction in emergency department costs, and a 68% reduction in pharmaceutical costs for the study participants.

A real-world study (22) was conducted in an employer-sponsored health clinic. CGRI was offered to patients seeking treatment for panic-related symptoms, following identification by primary care or behavioral health staff clinicians. In this case series, 18 participants with panic showed mean PDSS decreases of 7.2 scale points, with 67% showing significant reductions in PDSS scores (≥40%). Participants additionally showed decreases in behavioral health visits post-intervention.

An open label clinical trial (23) offering CGRI for treatment of PTSD was conducted at the Palo Alto VA, with enrollment open to both veterans and civilians. Mean Clinician-Administered PTSD Scale for DSM-5 (CAPS-5) scores declined significantly from pre- to post-treatment (49.5 to 27.1; Cohen's $d = 1.3$). Moreover, at six-month follow-up, 50% were rated as "in remission"(based on post-treatment CAPS-5 showing significant reduction from baseline, an absolute score ≤ 25, plus no longer meeting DSM-5 criteria for PTSD as rated by a study clinician). Treatment completers averaged 77% adherence (i.e., completion of 43 of 56 recommended sessions). Similar to the prior PD trial (20), hypocapnic subjects significantly increased $etCO_2$ levels from pre- to post-treatment. Both hypocapnic and normocapnic cohorts experienced significant reductions in CAPS-5 scores at six-month follow up, with a larger effect size (2.3) for the hypocapnic group compared to the normocapnic group (0.8).

Proposed mechanism(s) of action

Several potential mechanisms of action have been proposed for the CART/CGRI protocols. As noted above (21, 23), symptomatic improvement in both panic and PTSD is associated with the ability of hypocapnic users to normalize $etCO_2$ levels. Meuret et al. (24), in a randomized trial comparing CART with a cognitive therapy, found evidence of equivalent effectiveness but specific benefit in normalizing CO_2 in the CART group while identifying perceived control (but not increase in $etCO_2$) as a putative mechanism in the cognitive condition. Meuret and colleagues (25) additionally suggested that repeated exposure to respiratory distress may have led to an attenuation of respiratory distress *via* induction of dyspnea during the treatment.

Feinstein (26) and colleagues in a recent paper provide elegant synthesis and conceptualization of CO_2 hypersensitivity, neuroanatomy, acid-base balance, and the role of chemoreceptors in anxiety conditions. The authors introduce the concept of "apnea induced anxiety", which they interpret as a "an evolutionarily determined manifestation of the broader freezing response that the amygdala is well-known to coordinate". One implication of the Feinstein paper is the possibility that, in addition to the role of desensitization and development of enhanced sense of control and self-efficacy described above, the CART/CGRI intervention may also function to inhibit abrupt, de-stabilizing spikes in CO_2 and pH (27) provoked by dysfunctional breathing that induce anxiety and avoidance behaviors.

Study rationale

This paper reports treatment effectiveness data on patients treated with CGRI in clinical practice. The large pool of completed treatments available here represents an opportunity to evaluate real world effectiveness of CGRI as a follow-up to the prior clinical trials.

Materials and methods

Treatment device and protocol

CGRI teaches a specific breathing style *via* a system providing real-time feedback of respiratory rate (RR) and exhaled carbon dioxide ($etCO_2$) levels facilitated by health coaching and data capture. The CGRI system described in this paper combines: (a) a proprietary sensor for measurement of respiratory data, (b) an app-based respiratory feedback protocol pre-loaded on a tablet running Android 4.0 or higher (c) secure automatic data capture of adherence, physiological,

and symptom severity metrics, and (d) telehealth training/coaching to educate and support patient use of the system.

The respiratory sensor (see **Figure 1**) measures breath-to-breath RR and $etCO_2$ sampled *via* a small diameter nasal cannula. During a treatment session, real-time physiologic parameters are calculated by the sensor and transmitted to the Bluetooth®-connected tablet running dedicated, proprietary software. Respiratory data are graphically and numerically displayed, and audio/text instructions are provided by the tablet app. Per FDA clearance, prospective patients are authorized for the treatment by a licensed healthcare provider who affirms presence of the relevant conditions (panic disorder, panic attacks, or PTSD) and absence of contra-indications such as pregnancy, severe COPD or unstable psychiatric condition.

Twice-daily sessions for 28 days are recommended. Each 17-min session comprises three stages: (a) two minutes of ***baseline*** respiratory measurement (patients are instructed to breathe as usual with eyes closed), (b) 10 min of respiratory ***pacing*** measurement (patients are instructed to breathe in sync with a rising and falling audio tone and to adjust respiratory volume guided by display of $etCO_2$ levels relative to the normal range, and (c) five minutes of ***transition*** measurement (maintain paced breathing and $etCO_2$ level without cueing by audio tones). The rationale of this final phase is to "stamp in" the targeted respiratory style with reduced feedback, thus engendering self-management skills that promote awareness of the onset of dysregulated breathing and the capacity to substitute the learned breathing style. An actual patient example of RR and $etCO_2$ graphs at baseline, 7 days, and 28 days is seen in **Figure 2**. Note the progressive normalization of $etCO_2$ values and slowing/stabilization of respiratory rate during the 28-day course of treatment.

Patients are classified at the initiation of treatment as panic or PTSD based on their initial clinical assessment, with all patients receiving an identical treatment protocol. Patients complete a baseline Panic Disorder Severity Scale (PDSS) (28) or PTSD Checklist for DSM-5 (PCL-5) (29) followed by self-report PDSS measurements (panic patients) or PCL-5 at post-treatment *via* embedded scale questions on the tablet computer.

An initial 45-min secure video teleconference is conducted with an assigned health coach who provides: (a) the treatment

FIGURE 1
CGRI system.

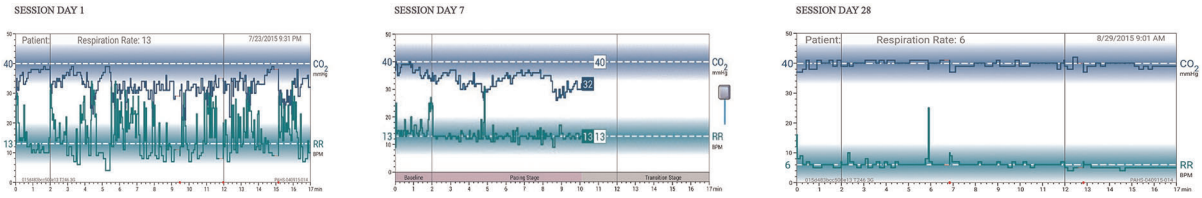


FIGURE 2
Respiratory feedback display.

rationale, (b) determination of patient goals/expectations, (c) education regarding diaphragmatic breathing and respiratory targets, (d) instructions for using the sensor/tablet, and (e) observation/feedback while the patient undertakes an initial session. Weekly 10–15-min follow-up sessions with the health coach review the prior week's sessions (available to the patient on the tablet and the coach on a secure portal) and provide coaching for continued progress. Although interactive video is the preferred method for these follow-up sessions, communication *via* phone or text can be substituted based on patient preference and progress. Weekly coaching notes are reviewed by clinical management and an end-of-treatment summary report consisting of initial and final session graphs, adherence information, coach observations, and symptom changes during treatment is sent to collaborating clinicians.

Remote monitoring and data capture

Following each session, data are uploaded from the tablet to a secure server. Data include breath-by-breath physiological data (mean respiratory rates and mean $etCO_2$ values for each of the session stages), self-reported symptoms as indexed by the PDSS or PCL-5 surveys, and images of the session graph (identical to what was seen by the patient on the tablet). Uploaded session data are maintained in a database with query and viewing tools facilitating longitudinal review of sessions by the assigned health coach. This review provides valuable information for identifying and addressing issues related to adherence problems, difficulties attaining respiratory targets, and symptom changes. Aggregation of patient data on a monthly, quarterly, and annual basis are conducted to track overall adherence, patient attrition, and clinical response rates.

Methods

Participants and source

This sample is comprised of 1,569 patients treated with CGRI between September 01, 2017 and September 16, 2021, drawn from a larger pool of 3,050 total patients treated with the intervention since first availability. Participant demographics are seen in **Table 1**. As consumers of routine clinical care, patients included insured and self-pay participants and were not paid for participation in this study. Self-report clinical rating surveys were not embedded into the tablet software until September 2017. Therefore, the 1,481 patients treated prior to availability of the in-app survey are excluded from this analysis. These 1,569 patients with pre/post surveys represent 89% of treatment completers during the study period, meaning that completed end-of-treatment surveys were missing for 11% of patients. During the study

TABLE 1 Demographic breakdown by diagnostic group.

	PD ($N = 1,395$)		PTSD ($N = 174$)	
	M	SD	*M*	SD
Age	39.2	13.9	40.9	14.9
Gender	*N*	%	*N*	%
Women	1,060	76	127	73
Men	335	24	46	26
Unknown	0	0	1	1

time frame, 10% (279) of PD patients and 11% (36) of PTSD patients were classified as dropouts, having completed fewer than 6 total sessions. Reasons for drop out were not systematically recorded.

As per FDA requirements, patients were authorized for treatment by a licensed healthcare provider. Authorizing clinicians included both independent practicing professionals as well as contracted, state-licensed professionals who obtained a health history, confirmed the absence of contraindications, and obtained a pre-treatment PDSS or PCL-5 to determine eligibility and baseline symptom severity. Individuals who were under the age of 13, pregnant, had COPD or other advanced respiratory illness, inadequately controlled seizures or asthma, active suicidal ideation, schizophrenia, or active psychosis were screened out. Authorizing clinicians required individuals with medical complexity or Covid-19 history and residual respiratory symptoms to obtain additional medical clearance from a personal physician.

A diagram of patient flow is detailed in **Figure 3**. For the purposes of this analysis, data were de-identified and compiled in aggregate from a secure database. At initiation of treatment, patients gave consent for inclusion in research conducted *via* analysis of de-identified data such as reported here in their acceptance of pre-treatment terms and conditions. Retrospective IRB-exempt status was granted for this analysis of de-identified data by the Institutional Review Board, University of Texas at Austin (IRB ID-STUDY00003542).

Measures

Patient physiological metrics were uploaded *via* Wi-Fi or cellular LTE to a secure server at the completion of each session and maintained in a database. Respiratory metrics (average respiratory rate and etCO2) were captured for each of the three stages of the 17-min sessions.

Self-reported panic symptom severity was measured using the 7-item Panic Disorder Severity Scale. Self-reported PTSD symptom severity was measured using the 20-item PCL-5. Baseline measures for both PD and PTSD scales were

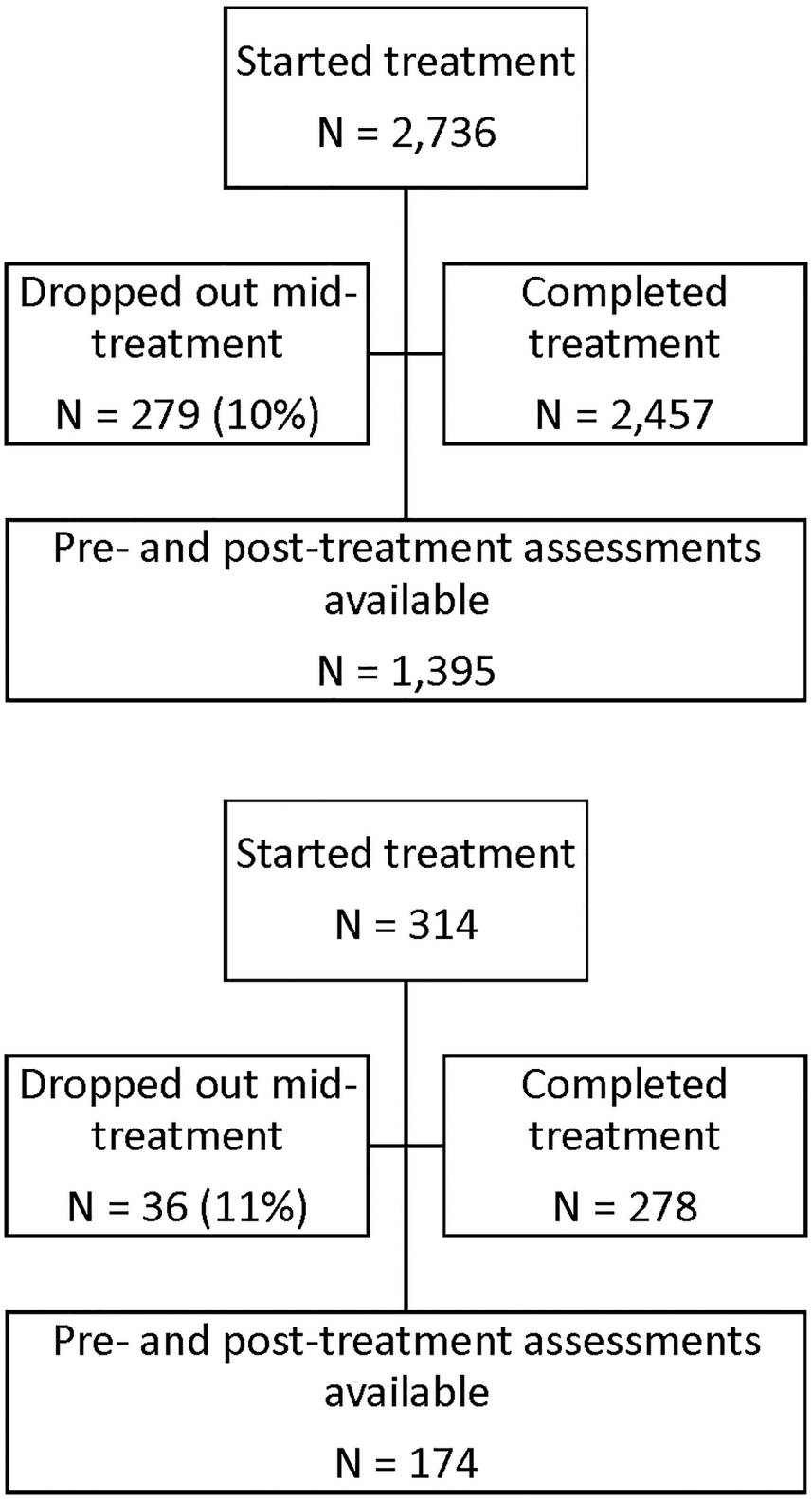


FIGURE 3
Patient flow.

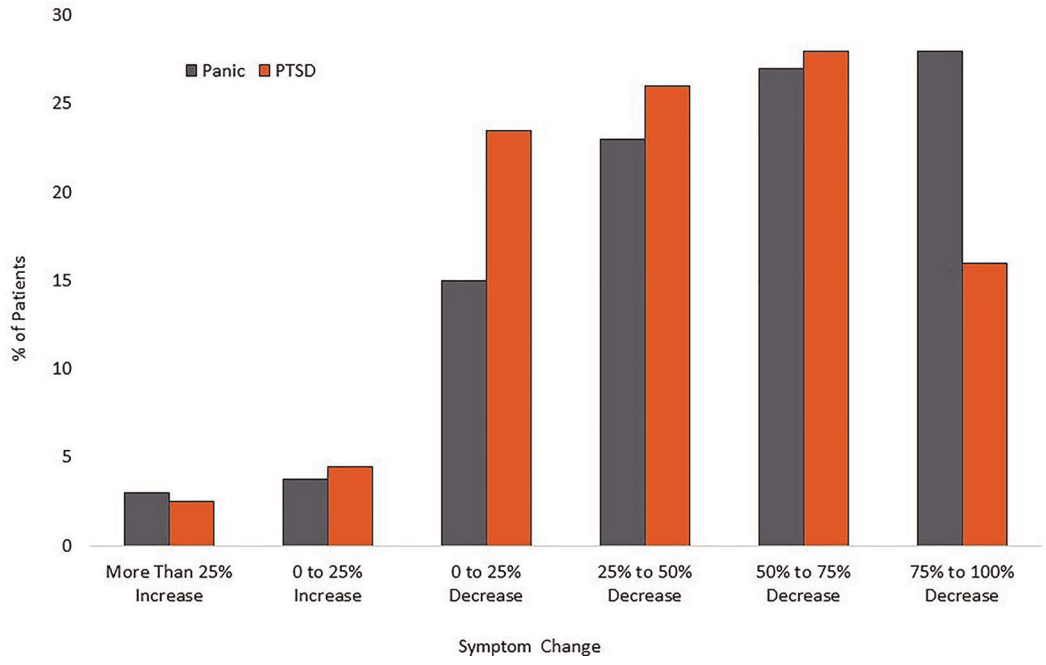


FIGURE 4
Distribution of symptom reduction.

obtained by the referring clinician or recorded during an assessment/authorization interview by a licensed healthcare professional. The post-treatment assessment of both the PDSS and the PCL-5 were administered on-screen *via* the tablet computer. Patients were classified as normocapnic ($etCO_2 \geq 37$) or hypocapnic ($etCO_2 < 37$) from the average baseline-stage of the first at-home treatment sessions for purposes of examining the role of changes in respiratory characteristics over the course of treatment.

Statistical methods

For non-dropout panic patients with both pre-and-post treatment survey scores, clinical response was defined as a 40% or greater reduction in scores on the PDSS; remission was defined as a score of five or less on the PDSS (30). For non-dropout PTSD patients, treatment response was defined as a reduction of PCL-5 score ≥ 10 points (31). Proportions of participants with the desired outcome and associated 95% lower bounds were estimated. Changes in mean scores were compared using a *t*-test and effect sizes calculated. A modified intent to treat analysis was also performed which included all patients who were trained on the treatment during the time frame in which pre- and post-treatment scales were available.

In order to more fully triangulate the construct of *clinically meaningful change,* we utilized the Jacobsen et al. (32) two-pronged statistical approach for determining clinically significant improvement for each participant. Briefly, this index first requires the assessment of whether each participant's magnitude of pre-to-post change is statistically reliable. This step is accomplished by calculating the Reliable Change Index (RCI) for each participant. For those participants showing statistically reliable improvement (RCI = 1.96 or greater), a determination is then made as to whether the participant's posttreatment score is closer to the distribution of scores for patients without the targeted disorder (PD or PTSD) or whether patient's post-treatment score continues to fall within the distribution of scores for the PD and PTSD disordered groups.

Treatment adherence was calculated by determining the proportion of the 56 recommended CGRI sessions completed over the course of the study, based on objective data automatically captured to the cloud-based server. Because some patients completed more than the required number of respiratory sessions, we coded all patients who completed 56 or more sessions as 100% compliant; for all others, we calculated adherence as the number of completed sessions divided by 56. Treatment dropouts were defined as patients completing ≤ 6 sessions; these patients are included in the intent to treat analysis.

10.3389/fdgth.2022.976001

Results

Symptom severity

For the PD cohort, the mean PDSS score declined from 14.7 (sd = 5.8) at baseline to 7.2 (sd = 5.7) at post-treatment. This 7.5-point decline represents a 50% decrease, with a large effect size (Cohen's $d = 1.3$). PDSS reduction of at least 40% was attained by 911 patients [65.3% (95% CI-62.7%–67.8%)]. Scores reflecting likely remission on the PDSS were recorded for 577 patients [41.4% (95%CI-38.7%–44.0%)]. Calculation of the Reliable Change Index classified 55.7% [95% CI-53.0%–58.3%] of participants as treatment responders. A modified intent to treat analysis of PD patients ($n = 1{,}610$) identified 979 patients as treatment responders [60.8% (95% CI-58.4%–63.2%)] and 609 patients as achieving remission [37.8% (95% CI-35.5%–40.3%)]. Calculation of the Reliable Change Index classified 51.9% [95% CI-49.5%–56.5%] of participants as treatment responders.

For the PTSD cohort, the mean PCL-5 score dropped from 47.9 (sd = 15.4) at baseline to 28.2 (sd = 18.4) at posttreatment. This 19.7-point change represents a 41.1% decrease and a large effect size (Cohen's $d = 1.16$). Within the PTSD cohort, 126 patients [72.4% (95%CI-65.0%–78.9%)] had a PCL-5 reduction of at least 10 scale points. Calculation of the Reliable Change Index classified 53.5% [95% CI- 46.0%–60.9%] of participants as treatment responders. A modified intent to treat analysis of PTSD patients ($n = 246$) recorded response rates of 56.9% [95% CI-50.5%–63.2%]. Calculation of the Reliable Change Index classified 40.6% [95%CI-34.5%–47.1%] of participants as treatment responders. Distribution of changes in symptom scores can be seen in **Figure 4**. Results are tabulated in **Tables 2** (Completer Sample) and **Table 3** (Intent to Treat Sample).

TABLE 2 Pre to post changes on primary outcomes for each diagnostic group (completer sample).

Outcome	PD ($N = 1395$)		PTSD ($N = 174$)	
	PDSS		PCL-5	
	M	SD	*M*	SD
Baseline	14.7	5.8	47.9	15.4
Posttreatment	7.2	5.7	28.2	18.4
P-Value	<0.001		<0.001	
Effect Size – Cohen's *D*	1.31		1.16	
Reliable Change (%)	55.7		53.5	
Average Adherence (%)	74.8		74.9	

TABLE 3 Pre to post changes on primary outcomes for each diagnostic group (intent-to-treat sample).

Outcome	PD ($N = 1610$)		PTSD ($N = 246$)	
	PDSS		PCL-5	
	M	SD	*M*	SD
Baseline	14.7	5.7	48.8	15.0
Posttreatment	7.8	5.9	34.7	20.6
P-Value	<0.001		<0.001	
Effect Size – Cohen's *D*	1.21		0.78	
Reliable Change (%)	51.9		40.6	
Average Adherence (%)	71.9		68.6	

Adherence

The PD group averaged 42 completed sessions of the recommended 56, a 75% adherence rate, while the PTSD cohort averaged 42 completed sessions of the recommended 56, representing a 75% adherence rate. Distribution of overall adherence can be seen in **Figure 5**. The relationship between adherence and symptom reduction (as measured by percent of participants reaching clinically significant symptom reduction) is illustrated in **Figure 6**.

Respiratory parameters

Of the 1,395 panic completers, 900 (65%) were classified at baseline as normocapnic and 495 (35%) were classified as hypocapnic. Pre- to post-treatment $etCO_2$ mean changed from 32.8 (sd = 2.73) to 36.8 (sd = 3.96) at post-treatment for the hypocapnic group and 39.6 (sd = 2.50) to 39.9 (sd = 3.57) for the normocapnic group. The 1.15 effect size for the hypocapnic group exceeded the 0.12 value for the normocapnic group. When effect sizes for PDSS symptom reductions were calculated, the normocapnic and hypocapnic groups each showed large effect sizes (1.29 and 1.32, respectively).

Of the 174 PTSD completers, 115 (66%) were classified at baseline as normocapnic and 59 (34%) were classified as hypocapnic. Pre- to post-treatment mean $etCO_2$ changed from 33.7 (sd = 2.05) to 37.2 (sd = 3.69) at post-treatment for the hypocapnic group and 39.3 (sd = 2.50) to 39.5 (sd = 3.66) for the normocapnic group. The 1.19 effect size for $etCO_2$ increase in the hypocapnic group exceeded the 0.08 value for the normocapnic group. When effect sizes for PCL-5 symptom reductions were calculated, the normocapnic and hypocapnic groups each showed large effect sizes (1.12 and 1.25, respectively). **Table 4** details the relationship of baseline $etCO_2$ levels to outcomes.

Discussion

Our primary aim in this report is to present real-world effectiveness data for CGRI in clinical practice. Statistical

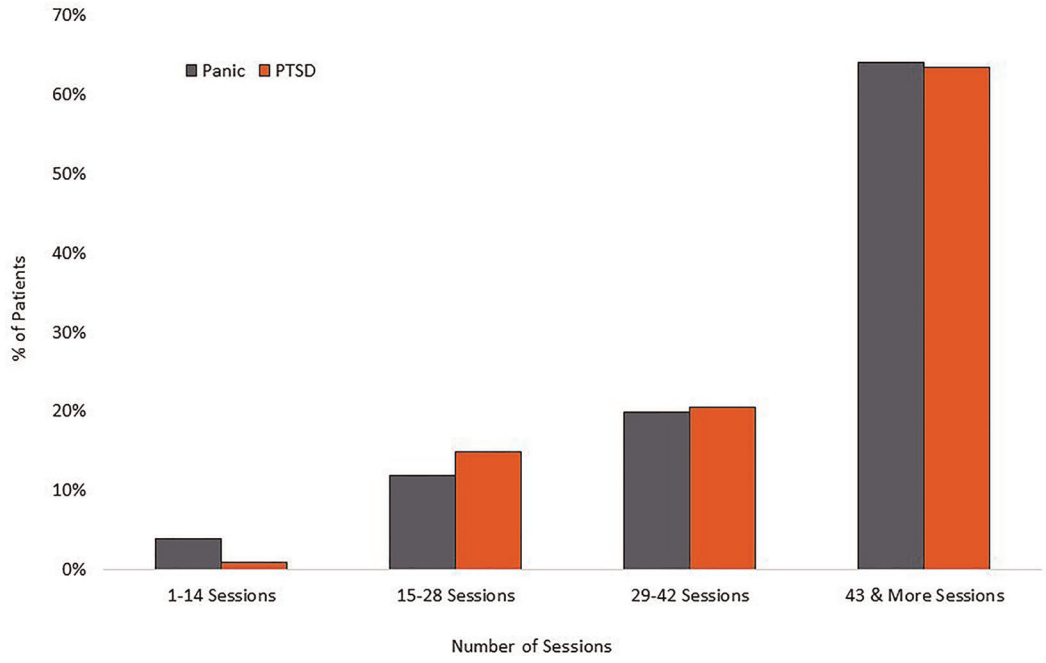


FIGURE 5
Treatment adherence.

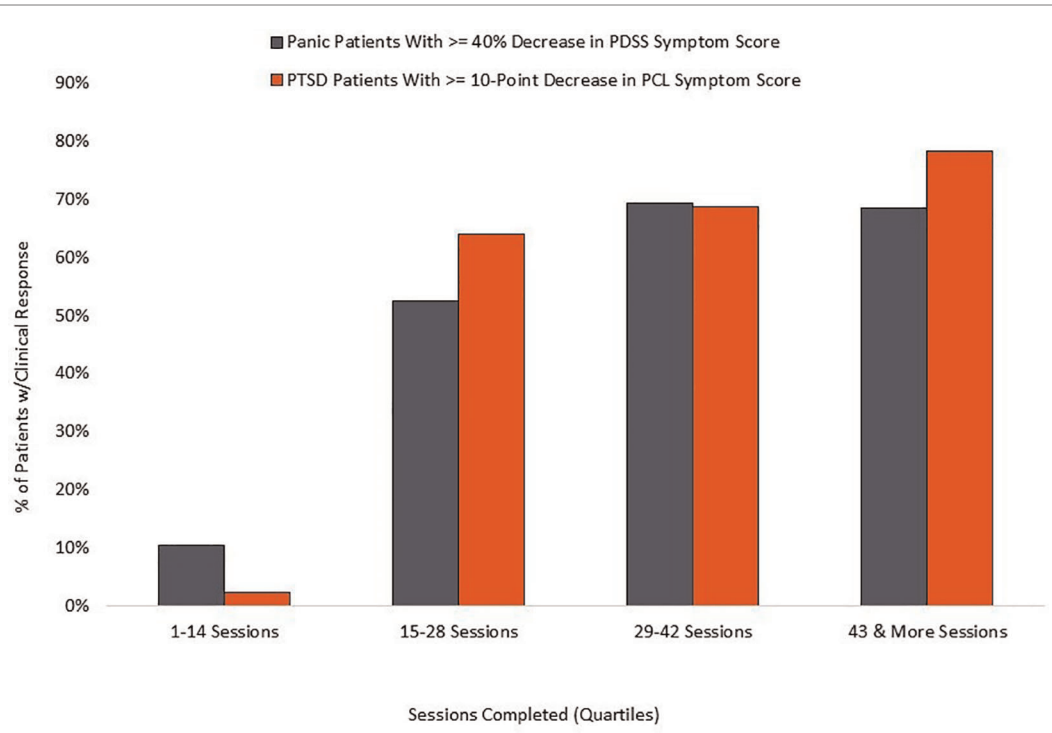


FIGURE 6
Relationship of symptom reduction to adherence.

TABLE 4 Normocapnic vs. hypocapnic subjects.

Condition	Panic Disorder		PTSD	
$etCO_2$ Status	Normocapnic	Hypocapnic	Normocapnic	Hypocapnic
N=	900	495	115	59
Mean (sd) Baseline etCO2	39.58 (2.50)	32.84 (2.73)	39.23 (2.62)	33.66 (2.05)
Mean (sd) Final etCO2	39.94 (3.57)	36.76 (3.96)	39.49 (3.66)	37.22 (3.69)
P value change	0.013	<0.001	0.53	<0.001
D Value for $etCO_2$ Change	0.112	1.15	0.083	1.193
Mean (sd) Symptom Reduction	51.1% (5.92)	48.57% (6.32)	40.3% (18.13)	42.58% (13.48)
P value change	<0.001	<0.001	<0.001	<0.001
D Value for Symptom Change	1.29	1.32	1.12	1.25

analyses of outcomes in this cohort of over 1,500 patients reflect significant and clinically meaningful symptom decreases in both panic and PTSD groups, with large effect sizes when comparing mean pre-treatment to post-treatment scores. As points of reference, the clinical outcomes documented in this data set are comparable to those seen in the prior published trials. The mean 7.5-point PDSS decline compares favorably with the 9.4 recorded by Tolin et al. (20), with large effect size in each report (1.31 vs. 2.3 respectively). The 19.7-point pre-post PCL-5 decline exceeds levels accepted for clinically significant change, and the final 28.2 mean score falls at the lower range of the 28–37 cutoff scores commonly used as threshold identifying likely presence of PTSD (31). Although the PCL-5 and CAPS-5 are different instruments, network analysis methods (33) suggests that the two scales provide comparable measurement of PTSD symptoms. The large effect size for the PTSD cohort in the current analysis (1.16) approximates the 1.3 obtained in Ostacher et al. (23).

As seen in the prior CGRI trials (20, 23), patients who began treatment with hypocapnic $etCO_2$ levels showed significantly greater increases in mean exhaled carbon dioxide levels following treatment than normocapnic subjects. These results conform to expectation. Normocapnic subjects would not be expected to increase these levels beyond 40 mmHg as patients are coached by the intervention to target $etCO_2$ around this range value. With identical directions, the final $etCO_2$ values for the hypocapnic subgroups on the other hand increased by approximately 4 mmHg, representing goal attainment and significant normalization of this respiratory measure. Prior studies have shown that both normocapnic and hypocapnic subjects achieve significant clinical benefit after use of this intervention. Similar results are seen in the present study, with large effect sizes (range 1.11 to 1.32) across the normocapnic and hypocapnic subgroups for each condition.

An RCT comparing Meuret's CART protocol with a cognitive therapy arm (24) demonstrated significant and comparable benefit from both treatments, but concluded that $etCO_2$ increases were a significant mediator of change in the CART condition but not in the cognitive therapy condition. The author concluded that changes in $etCO_2$ are directly responsible for some of the symptom reduction in the respiratory therapy. It is unlikely that a single mechanism of action is responsible for the clinical benefit seen consistently with the CGRI and CART protocols, with individuals with normal and depressed $etCO_2$ baseline levels appear to benefit comparably.

It is possible that lower-than-normal $etCO_2$ levels during a single, brief baseline measurement is an inadequate surrogate for dysfunctional breathing in symptomatic individuals. Rapid and variable respiratory rate as well as significant decreases in $etCO_2$ during the "pacing" phase of the CGRI treatment are commonly observed but not analyzed in the present study. These metrics are the subject of planned subsequent analyses. In addition, significant cardio-respiratory instability has been observed during the onset of panic attacks (34), raising the possibility that individuals learning respiratory control *via* the CART/CGRI interventions may develop skills that inhibit symptom escalation at the point of interoceptive awareness of respiratory distress or in response to triggering external events. Feinstein and colleagues' work discussed earlier (26) also suggests that respiratory stability may function within neural networks to suppress the apnea-induced anxiety that the authors implicate in symptom surge and learned avoidance. In summary, the continuing evidence of benefit to hypocapnic as well as normocapnic individuals suggests that a single measure of low baseline $etCO_2$ level does not function as a meaningful biomarker for treatment response. Planned future research will look at reactivity to CO_2 challenge and other measures of dysfunctional breathing as well as non-respiratory potential predictive biomarkers.

The relationship between completed sessions and symptom reduction (see **Figure 6**) suggests a distinct "dose/response" relationship. Negligible symptom reduction is detected in patients completing fewer than 15 session, with clinical response rising robustly in the next quartile of participants and plateauing or increasing by the final quartile, marked by completion of 43 or more sessions. Standard coaching protocol is to advocate for maximum adherence, and the 75%

adherence rate documented here suggests that most patients respond positively to those recommendations. We observe that some patients show rapid respiratory control and symptom reduction within the first two weeks of treatment, and it is possible that adherence in this subset may decline for patients who are experiencing rapid symptom relief and are able to stabilize breathing without twice-daily formal practice. However, the current data does not examine this potential relationship, nor do we know whether maintenance of symptom reduction beyond immediate post-treatment may be diminished in patients who fail to complete a threshold mark, suggested by the current data to be in the range of 60%–70% adherence.

The real-world outcomes reported here are appropriately compared to established treatments. For panic conditions, antidepressants and benzodiazepines are considered first line medications. Risk of abuse, side effect burden, and risk of relapse following discontinuation are commonly reported challenges with psychopharmacology (35, 36). The National Center for PTSD (37) has moved medications to a second line option for PTSD, with strong first -line recommendations for manualized trauma-focused psychotherapies. While cognitive behavioral therapies are widely considered to have the strongest evidence base for psychotherapeutic treatment of the anxiety disorders, limitations are noted in key reviews of CBT regarding response rates and tolerability (38–43). In contrast, benefits of CGRI in the context of this study include brief duration of treatment, at-home administration, clinically significant symptom reduction, and favorable adherence/dropout rates.

Limitations

Several limitations of this effectiveness study deserve comment. First, as is the case with all non-randomized treatment effectiveness studies, threats to internal validity (e.g., selection, regression to the mean, etc.) cannot be ruled out. However, there are several features of this study that increase confidence that the symptom reduction observed in both cohorts was likely due to the CGRI intervention as opposed to extraneous factors. For instance, the low drop-out rates observed for both cohorts (<11%) were well below that observed in most psychotherapy RCT's for panic disorder or PTSD, thus reducing the likelihood that patient attrition was biasing the treatment response rate. Moreover, both PD and PTSD tend to show a chronic clinical course without treatment (44, 45), thus helping to rule out regression to the mean or spontaneous remission as likely candidates for explaining the observed symptom reduction.

As with most effectiveness studies, inclusion criteria were relaxed and geared towards clinicians' judgement of patient suitability for treatment rather than symptom cutoff scores or other trial enrollment criteria. This resulted in some patients scoring in the marginal range of symptom severity at pre-treatment. However, the average mean PDSS and PCL-5 scores at entry were at clinically significant levels and comparable to those reported in prior CGRI trials. These data suggest that the screening and authorization processes largely enroll patients with the intended conditions.

A third limitation of this open-label trial is the absence of an active control condition or a stringent respiratory control intervention such as false respiratory feedback. Future studies are needed to disentangle whether CGRI-induced symptom changes are mediated by changes in respiratory parameters (i.e., respiration rate and $etCO_2$ levels) in addition to or as opposed to alternative putative mechanisms such as expectancy effects, desensitization to dyspnea, or change in self-efficacy for controlling symptoms. A fourth limitation of this study is the absence of extended follow-up outcome assessments, thus precluding conclusions regarding the durability of the CGRI-induced symptom changes. However, prior trials have reported sustained treatment benefit at six to twelve-month follow ups (20, 21, 23). The size of the PTSD cohort is substantially less than that for PD, which reflects the more recent FDA clearance for PTSD for this treatment. Although encouraging, additional review of outcomes for CGRI in PTSD is warranted to determine if the response rates seen in this analysis remain consistent as treatment volumes increase.

Finally, many potential prognostic variables (e.g., comorbid psychiatric and medical conditions, prior history of treatment, duration of disorder, etc.) were not included as part of data capture. Data concerning concurrent treatment with psychotherapy or medication were not obtained, thus presenting an important confound in the study, i.e., whether the benefits obtained were independent of or synergistic with other therapies.

Challenges and treatment enhancements

While positive clinical benefit and adherence levels are described in the paper, experience and patient feedback point to certain areas needing improvement. The CGRI protocol provides the same instructions and performance targets regardless of baseline patient characteristics. As an example, a patient with significantly below-normal $etCO_2$ and/or rapid, unstable respiratory rate is given the same set of instructions and targets as a patient with more normal baseline respiratory style. Perhaps as a consequence, some patients experience distressing air hunger in the early stages of treatment, as reduced respiratory volume is the behavior necessary to raise $etCO_2$ in hypocapnic users. With coaching support and education, many individuals tolerate this side effect and persist, while others may discontinue treatment. A related complaint comes from individuals who are self-described "perfectionists" who are frustrated with their inability to hit the 40 mmHg target initially and express dissatisfaction or distress related to

lack of perceived success. No systematic method for identifying or ameliorating these or other reasons for non-compliance is currently in place. Evaluation of CGRI in randomized control studies with sham or active control arms will be important in validating the accumulating evidence from open label and real-world trials discussed in this paper.

Future product development intends to broaden the scope of demographic and health data obtained upon registration, which may allow for greater precision in predicting which patients are likely to respond to this intervention. Additionally, gamification and individualization of the treatment protocol represents an important opportunity to improve engagement and perhaps enhance outcomes. Such efforts may optimize application to sub-populations such as adolescents and individuals with attentional difficulties.

Conclusions

Clinically meaningful symptom reductions in both PD and PTSD patients were achieved using the CGRI treatment. The symptom reductions reported here are consistent with prior published research that tracked outcomes to six months or one year and provide encouraging evidence of clinical effectiveness when the treatment is delivered outside of a formal research setting. The embedded data analytic capacities provide automatic compilation of key outcome metrics such as those reported in this paper. Dissemination of real-world data such as these are vital for evaluating the viability (clinical benefit as well as engagement) of emerging treatments such as CGRI. Adoption of prescription digital therapeutics such as CGRI hold promise for expanding access and patient choice in the treatment of panic disorder and PTSD.

Data availability statement

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

Ethics statement

Retrospective IRB-exempt status was granted for this review of de-identified data by the Institutional Review Board, University of Texas at Austin (IRB ID- STUDY00003542). The patients/participants provided written informed consent to participate in research as part of signed terms and conditions when treatment was initiated.

Author contributions

RNC planned and implemented the study design and supervised delivery of clinical care. Drafting of manuscript was done by RNC and MJT. Data acquisition and statistical analysis was performed by RK and ST. All authors contributed to the article and approved the submitted version.

Funding

This study was internally funded by Freespira, Inc. in the form of salaries, equipment, and supplies.

Acknowledgments

The authors are grateful for the exceptional customer service and patient care delivered by the Freespira team, as well as the engineering, system development and data management teams that created and refined this intervention.

Conflict of interest

RNC, RK, and ST are employees of Freespira, Inc. and receive compensation by way of salary and equity. MJT is a Scientific Advisor of Freespira, Inc. and receives compensation by way of stock options.

Publisher's note

References

1. Kessler RC, Chiu WT, Jin R, Ruscio AM, Shear K, Walters EE. The epidemiology of panic attacks, panic disorder, and agoraphobia in the national comorbidity survey replication. *Arch Gen Psychiatry*. (2006) 63(4):415–24. doi: 10.1001/archpsyc.63.4.415

2. Schein J, Christy Houle C, Urganus A, Cloutier M, Patterson-Lomba O, Wang Y, et al. Prevalence of post-traumatic stress disorder in the United States: a systematic literature review. *Curr Med Res Opin*. (2021) 37(12):2151–61. doi: 10.1080/03007995.2021.1978417

3. Perna G. Panic disorder: from psychopathology to treatment. *Clin Psychopharmacol.* (2012) 18:82–91.

4. Kessler RC, Sonnega A, Bromet E, Hughes M, Nelson CB. Posttraumatic stress disorder in the national comorbidity survey. *Arch Gen Psychiatry.* (1995) 52:1048. doi: 10.1001/archpsyc.1995.03950240066012

5. Carpenter JK, Andrews LA, Witcraft SM, Powers MB, Smits JAJ, Hofmann SG. Cognitive behavioral therapy for anxiety and related disorders: a meta-analysis of randomized placebo-controlled trials. *Depress Anxiety.* (2018) 35 (6):502–14. doi: 10.1002/da.22728

6. Boulding R, Stacey R, Niven R, Fowler SJ. Dysfunctional breathing: a review of the literature and proposal for classification. *Eur Respir Rev.* (2016) 25:287–94. doi: 10.1183/16000617.0088-2015

7. Perna G, Iannone G, Torti T, Caldirola D. Panic disorder, is it really a mental disorder? From body functions to the homeostatic brain. In: AE Nardi, CR Freire, editors. *Panic disorder: Neurobiological and treatment aspects*. Switzerland: Springer International Publishing (2016). p. 93–112.

8. Klein DF. False suffocation alarms, spontaneous panics, and related conditions. An integrative hypothesis. *Arch Gen Psychiatry.* (1993) 50:306–17. doi: http://dx.doi.org/10.1001/archpsyc.1993.01820160076009

9. Klein DF. Panic disorder and agoraphobia: hypothesis hothouse. *J Clin Psychiatry.* (1996) 57(Suppl 6):21–7. PMID: 8647794

10. Do Amaral JMX, Spadaro PTM, Pereira VM, de Oliveia e Silva AC, Nardi AE. The carbon dioxide challenge test in panic disorder: a systematic review of preclinical and clinical research. *Rev Bras Psiquiatr.* (2013) 35:318–31. doi: 10.1590/1516-4446-2012-1045

11. Coryell W, Pine D, Fyer A, Klein D. Anxiety responses to CO_2 inhalation in subjects at high-risk for panic disorder. *J Affect Disord.* (2006) 92(1):63–70. doi: 10.1016/j.jad.2005.12.045

12. Schmidt NB, Zvolensky MJ. Anxiety sensitivity and CO_2 challenge reactivity as unique and interactive prospective predictors of anxiety pathology. *Depress Anxiety.* (2007) 24(8):527–36. doi: 10.1002/da.20267

13. Bystritsky A, Craske M, Maidenberg E, Vapnik T, Shapiro D. Autonomic reactivity of panic patients during a CO_2 inhalation procedure. *Depress Anxiety.* (2000) 11(1):15–26. doi: 10.1002/(sici)1520-6394(2000)11:1<15::aid-da3>3.0.co;2-w

14. Beck JG, Shipherd JC, Zebb BJ. Fearful responding to repeated CO_2 inhalation: a preliminary investigation. *Behav Res Ther.* (1996) 34(8):609–20. doi: 10.1016/0005-7967(96)00039-3

15. Horwath E, Adams P, Wickramaratne P, Pine D, Weissman MM. Panic disorder with smothering symptoms: evidence for increased risk in first-degree relatives. *Depress Anxiety.* (1997) 6(4):147–53. doi: 10.1002/(sici)1520-6394(1997)6:4<147::aid-da3>3.0.co;2-9

16. Kellner M, Muhtz C, Nowack S, Leichsenring I, Wiedemann K, Yassouridis A. Effects of 35% carbon dioxide (CO_2) inhalation in patients with post-traumatic stress disorder (PTSD): a double-blind, randomized, placebo-controlled, cross-over trial. *J Psychiatr Res.* (2018) 96:260–4. doi: 10.1016/J.JPSYCHIRES.2017.10.019

17. Telch MJ, Rosenfield D, Lee HJ, Pai A. Emotional reactivity to a single inhalation of 35% carbon dioxide and its association with later symptoms of posttraumatic stress disorder and anxiety in soldiers deployed to Iraq. *Arch Gen Psychiatry.* (2012) 69(11):1161–8. doi: 10.1001/archgenpsychiatry.2012.8

18. Berenz EC, York TP, Bing-Canar H, Amstadler AB, Mezuk B, Gardner CO, et al. Time course of panic disorder and posttraumatic stress disorder onsets. *Soc Psychiatry Psychiatr Epidemiol.* (2019) 54(5):639–47. doi: 10.1007/s00127-018-1559-1

19. Meuret AE, Wilhelm FH, Ritz T, Roth W. Feedback of end-tidal pCO_2 as a therapeutic approach for panic disorder. *J Psychiatr Res.* (2008) 42(7):560–8. doi: 10.1016/j.jpsychires.2007.06.005

20. Tolin DF, McGrath PB, Hale LR, Weiner DN, Gueorguieva R. A multisite benchmarking trial of capnometry guided respiratory intervention for panic disorder in naturalistic treatment settings. *Appl Psychophysiol Biof.* (2017) 42 (1):51–8. doi: 10.1007/s10484-017-9354-4

21. Kaplan A, Mannarino A, Nickell PV. Evaluating the impact of freespira on panic disorder patients' health outcomes and healthcare costs within the allegheny health network. *Appl Psychophysiol Biof.* (2020) 45(3):175–81. doi: 10.1007/s10484-020-09465-0

22. Madhusudhan DK, Glied KN, Nguyen E, Rose J, Bravata DM. Real-world evaluation of a novel technology-enabled capnometry-assisted breathing therapy for panic disorder. *J Ment Health Clin Psychol.* (2020) 4(4):39–46. doi: 10.29245/2578-2959/2020/4.1220

23. Ostacher MA, Fischer E, Bowen ER, Lyu J, Robbins DJ, Suppes T. Investigation of a capnometry guided respiratory intervention in the treatment of posttraumatic stress disorder. *Appl Psychophysiol Biof.* (2021) 46(4):367–76. doi: 10.1007/s10484-021-09521-3

24. Meuret AE, Rosenfield D, Seidel A, Bhaskara L, Hofmann SG. Respiratory and cognitive mediators of treatment change in panic disorder: evidence for intervention specificity. *J Consult Clin Psychol.* (2010) 78(5):691–704. doi: 10.1037/a0019552

25. Meuret AE, Ritz T, Wilhelm FH, Roth WT, Rosenfield D. Hypoventilation therapy alleviates panic by repeated induction of dyspnea. *Biol Psychiatry Cogn Neurosci Neuroimaging.* (2018) 3(6):539–45. doi: 10.1016/j.bpsc.2018.01.010

26. Feinstein JS, Gould D, Khalsa SS. Amygdala-driven apnea and the chemoreceptive origin of anxiety. *Biol Psychiatry.* (2022) 170:1–13. doi: 10.1016/j.biopsycho.2022.108305

27. Ziemann AE, Allen JE, Dahdaleh N, Drebot II, Coryell MW, Wunsch AM, et al. The amygdala is a chemosensor that detects carbon dioxide and acidosis to elicit fear behavior. *Cell.* (2009) 139(5):1012–21. doi: 10.1016/j.cell.2009.10.029

28. Shear MK, Brown TA, Barlow DH, Money R, Sholomskas DE, Woods SW, et al. Multicenter collaborative panic disorder severity scale. *Am J Psychiatry.* (1997) 154(11):1571–157. doi: 10.1176/ajp.154.11.1571

29. Blevins CA, Weathers FW, Davis MT, Witte TK, Domino JL. The posttraumatic stress disorder checklist for *DSM-5* (PCL-5): development and initial psychometric evaluation. *J Trauma Stress.* (2015) 28(6):489–98. doi: 10.1002/jts.22059

30. Furukawa TA, Shear MK, Barlow DH, Gorman JM, Woods SW, Money R, et al. Evidence-based guidelines for interpretation of the panic disorder severity scale. *Depress Anxiety.* (2009) 26(10):922–9. doi: 10.1002/da.20532

31. PTSD Checklist for DSM-5 (PCL-5) https://istss.org/clinical-resources/assessing-trauma/ptsd-checklist-dsm-5 (Accessed June 2, 2022).

32. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol.* (1991) 59 (1):12–9. doi: 10.1037/0022-006X.59.1.12

33. Moshier SJ, Bovin MJ, Gay NG, Wisco BE, Mitchell KS, Lee DJ, et al. Examination of posttraumatic stress disorder symptom networks using clinician-rated and patient-rated data. *J Abnorm Psychol.* (2018) 127(6):541–7. doi: 10.1037/abn0000368

34. Meuret AE, Rosenfield D, Wilhelm FH, Zhou E, Conrad A, Ritz T, et al. Do unexpected panic attacks occur spontaneously? *Biol Psychiatry.* (2011) 70 (10):985–91. doi: 10.1016/j.biopsych.2011.05.027

35. Marchesi C. Pharmacological management of panic disorder. *Neuropsychiatr Dis Treat.* (2008) 4(1):93–106. doi: 10.2147/ndt.s1557

36. Freire RC, Amrein R, Mochcovitch MD, Dias GP, Machado S, Versiani M, et al. A 6-year posttreatment follow-up of panic disorder patients: treatment with clonazepam predicts lower recurrence than treatment with paroxetine. *J Clin Psychopharmacol.* (2017) 37(4):429–34. doi: 10.1097/JCP.0000000000000740

37. VA/DoD Clinical Practice Guideline for the Management of Posttraumatic Stress Disorder and Acute Stress Disorder. (2017). Available at: www.healthquality.va.gov/guidelines/MH/ptsd/VADoDPTSDCPGFinal012418.pdf (Accessed May 11, 2022).

38. Sloan DM, Marx BP, Lee DJ, Resick PA. A brief exposure-based treatment vs cognitive processing therapy for posttraumatic stress disorder: a randomized noninferiority clinical trial. *JAMA Psychiatry.* (2018) 75(3):233–9. doi: 10.1001/jamapsychiatry.2017.4249

39. Haagen JF, Smid GE, Knipscheer JW, Kleber RJ. The efficacy of recommended treatments for veterans with PTSD: a metaregression analysis. *Clin Psychol Rev.* (2015) 40:184–94. doi: 10.1016/j.cpr.2015.06.008

40. Steenkamp MM, Litz BT, Marmar CR. First-line psychotherapies for military-related PTSD. *JAMA.* (2020) 323(7):656–7. doi: 10.1001/jama.2019.20825

41. Watkins LE, Sprang KR, Rothbaum BO. Treating PTSD: a review of evidence-based psychotherapy interventions. *Front Behav Neurosci.* (2018) 12:258. doi: 10.3389/fnbeh.2018.00258

42. Fernandez E, Salem D, Swift JK, Ramtahal N. Meta-analysis of dropout from cognitive behavioral therapy: magnitude, timing, and moderators. *J Consult Clin Psychol.* (2015) 83(6):1108–22. doi: 10.1037/ccp0000044

43. Bentley KH, Cohen ZD, Kim T, Bullis JR, Nauphal M, Cassiello-Robbins C, et al. The nature timing, and symptom trajectories of dropout from transdiagnostic and single-diagnosis cognitive-behavioral therapy for anxiety disorders. *Behav Ther.* (2021) 52(6):1364–76. doi: 10.1016/j.beth.2021.03.007

44. Pollack MH, Marzol PC. Panic: course, complications and treatment of panic disorder. *J Psychopharmacol.* (2000) 14(2 Suppl 1):S25–30. doi: 10.1177/02698811000142S104

45. Bremner JD, Southwick SM, Darnell A, Charney DS. Chronic PTSD in Vietnam combat veterans: course of illness and substance abuse. *Am J Psychiatry.* (1996) 153(3):369–75. doi: 10.1176/ajp.153.3.369

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 12 January 2023
DOI 10.3389/fmed.2022.1049686

OPEN ACCESS

EDITED BY
Kirsten Smayda,
MedRhythms, United States

REVIEWED BY
Davide Maria Cammisuli,
Catholic University of the Sacred Heart, Italy
David Jing-Piao Lin,
Harvard Medical School, United States
Michael Young,
Kansas State University, United States

*CORRESPONDENCE
Michelle H. Chen
✉ michelle.chen2@rutgers.edu

SPECIALTY SECTION
This article was submitted to Precision Medicine, a section of the journal Frontiers in Medicine

RECEIVED 20 September 2022
ACCEPTED 28 December 2022
PUBLISHED 12 January 2023

CITATION
Chen MH, Cherian C, Elenjickal K, Rafizadeh CM, Ross MK, Leow A and DeLuca J (2023) Real-time associations among MS symptoms and cognitive dysfunction using ecological momentary assessment. *Front. Med.* 9:1049686. doi: 10.3389/fmed.2022.1049686

Real-time associations among MS symptoms and cognitive dysfunction using ecological momentary assessment

Michelle H. Chen[1,2]*, Christine Cherian[1], Karen Elenjickal[1], Caroline M. Rafizadeh[3,4], Mindy K. Ross[5], Alex Leow[5] and John DeLuca[3,4]

[1]Institute for Health, Health Care Policy and Aging Research, Rutgers University, New Brunswick, NJ, United States, [2]Department of Neurology, Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ, United States, [3]Kessler Foundation, East Hanover, NJ, United States, [4]Department of Physical Medicine and Rehabilitation, New Jersey Medical School, Rutgers University, Newark, NJ, United States, [5]Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, United States

Introduction: Multiple sclerosis (MS) is characterized by a wide range of disabling symptoms, including cognitive dysfunction, fatigue, depression, anxiety, pain, and sleep difficulties. The current study aimed to examine real-time associations between non-cognitive and cognitive symptoms (latter measured both objectively and subjectively in real-time) using smartphone-administered ecological momentary assessment (EMA).

Methods: Forty-five persons with MS completed EMA four times per day for 3 weeks. For each EMA, participants completed mobile versions of the Trail-Making Test part B (mTMT-B) and a finger tapping task, as well as surveys about symptom severity. Multilevel models were conducted to account for within-person and within-day clustering.

Results: A total of 3,174 EMA sessions were collected; compliance rate was 84%. There was significant intra-day variability in mTMT-B performance ($p < 0.001$) and levels of self-reported fatigue ($p < 0.001$). When participants reported depressive symptoms that were worse than their usual levels, they also performed worse on the mTMT-B ($p < 0.001$), independent of upper extremity motor functioning. Other self-reported non-cognitive symptoms were not associated with real-time performance on the mTMT-B [$p > 0.009$ (Bonferroni-corrected)]. In contrast, when self-reported fatigue ($p < 0.001$), depression ($p < 0.001$), anxiety ($p < 0.001$), and pain ($p < 0.001$) were worse than the individual's typical levels, they also reported more severe cognitive dysfunction at the same time. Further, there was a statistical trend that self-reported cognitive dysfunction (not mTMT-B performance) predicted one's self-reported sense of accomplishment in real-time.

Discussion: The current study was the first to identify divergent factors that influence subjectively and objectively measured cognitive functioning *in real time* among persons with MS. Notably, it is when symptom severity was worse than the individual's usual levels (and not absolute levels) that led to cognitive fluctuations, which supports the use of EMA in MS symptom monitoring.

KEYWORDS

multiple sclerosis (MS), experience sampling, cognitive impairment, depression, anxiety, fatigue, pain, sleep

1. Introduction

Multiple Sclerosis (MS) is a demyelinating, neurodegenerative disorder of autoimmune causes that disrupts the central nervous system (CNS). It is among the most common neurological diseases, and its age of onset typically occurs between 20 and 50 years (1). MS is accompanied by a range of symptoms, including cognitive dysfunction, fatigue, pain, mood changes, sleep problems, weakness, motor problems, and visual impairment (2).

Cognitive dysfunction, perhaps the most disabling manifestation of MS, is present in approximately 45–60% of MS cases (3, 4). Deficits in learning and memory as well as information processing speed are the most prevalent cognitive deficits in MS (5). Difficulties are also evidenced in complex attention, executive functioning, working memory, and visuospatial functions (5). Such impairments can affect everyday tasks of individuals with MS, disrupting their quality of life, overall wellbeing, and physical and social functioning (6).

Multiple sclerosis symptom severity can fluctuate throughout the day and week (7, 8), which is not captured by traditional clinical tools that ask patients to rate their average symptoms over a period of time (e.g., over the past week or month). The retrospective nature of these inventories can introduce recall bias (9), which is especially problematic for a population with memory difficulties such as MS. There is a need for real-time assessment of MS symptoms, which will improve our understanding of day-to-day symptom variability and inter-symptom associations, as well as advance the development of individualized MS treatment recommendations.

Ecological momentary assessment (EMA) is an approach that repeatedly samples an individual's experiences in real time (e.g., asking them to report their symptom severity weekly, daily, or even every few hours) (10). By assessing real-time MS symptom severity several times per day, EMA allows for direct examination of within-person dynamics and diurnal symptom patterns. EMA has been widely used in studying behavioral health and psychological symptoms such as mood, addiction, and wellbeing (10). However, few MS studies have used this paradigm. Available, albeit limited, MS studies using EMA have shown good feasibility with relatively high compliance rates among their participants, ranging from 83 to 91% (7, 11). The current study will add to this emerging literature.

As with other MS symptoms, cognitive functioning is variable and can fluctuate on a daily basis due internal (e.g., stress) (12) or external triggers (e.g., temperature) (13). With the advent of mobile technology, cognitive assessment can now be easily administered through an individual's smartphone. When combined with EMA, mobile cognitive testing permits the study of real-time associations among cognition, everyday tasks and environment, and other related symptoms (14). For example, an EMA study conducted in middle-aged and older adults with HIV found that engagement in cognitively stimulating activities was associated with better executive functioning and verbal learning, while engagement in more passive activities resulted in worse executive functioning and verbal learning performance (14).

Among the limited literature using EMA in MS, most investigations focused on fatigue. These studies have shown substantial within-person variability in fatigue intensity (7, 15), which justifies the use of EMA in this population. Only one research group has examined a broad range of symptoms, including cognitive dysfunction, depressed mood, fatigue, and pain, as well as inter-symptom associations (6, 7, 16). A study conducted by this group found that poorer cognitive functioning was preceded by worsening within-day pain and fatigue (16). However, cognitive functioning in this study was based on self-report. Given that studies have shown that self-reported cognitive dysfunction do not always correlate with objectively measured cognition (17, 18), more research is needed to clarify the associations between non-cognitive MS symptoms and both subjective and objective cognitive outcomes. Notably, the study found that it was only within-person changes (or "state") in symptom ratings that were associated with other symptoms, and there were no cross-symptom associations in mean symptom levels across time points (or "trait") (16). The state aspect of a symptom refers to transient fluctuations at a point in time that can be affected by situational contexts (e.g., being more anxious than usual because of a doctor's appointment). The trait aspect of a symptom represents the typical pattern for an individual (e.g., usual levels of anxiety) (19). Given the high sampling frequency, EMA enables such separations.

The current study aimed to use smartphone-administered EMA to investigate and characterize real-time relationships between non-cognitive and cognitive symptoms among persons with MS. We expect that deviations in non-cognitive MS symptoms from individuals' typical levels will be associated with real-time cognitive changes. The current study will address the limitations of prior studies by measuring cognitive functioning both objectively and subjectively.

2. Materials and methods

2.1. Participants

Participants were recruited through online advertisements on the National MS Society and Kessler Foundation websites and social media. Interested prospective participants would contact the research team and undergo a brief phone screening to ascertain eligibility. Inclusion criteria included: (1) ownership of an iPhone, (2) access to a desktop or laptop computer that is at least 13 inches in screen size, (3) English is primary language, and (4) self-reported diagnosis of MS by a medical professional. Exclusion criteria consisted of: (1) self-reported diagnosis of neurological conditions other than MS, (2) self-reported diagnosis of serious mental illness (e.g., schizophrenia, bipolar disorder), (3) self-reported diagnosis of attention-deficit/hyperactive disorder or specific learning disorder, (4) self-reported problems with substance misuse, (5) presence of sensory or motor difficulties that would interfere with validity of study assessments (self-reported and through examiner observation), and (6) self-reported MS relapse/exacerbation symptoms within the month prior to enrollment. The study was approved by the Kessler Foundation Institutional Review Board. All participants provided electronic written informed consent through Research Electronic Data Capture (REDCap) tools (20, 21), hosted by New Jersey Medical School, Rutgers University.

2.2. Procedures

Data collection took place between April 2021 and February 2022. Due to the coronavirus disease 2019 (COVID-19) pandemic (22), all study procedures were conducted virtually. Participants

TABLE 1 Ecological momentary assessment protocol.

Construct	Question/task	Assessment modality	Administration frequency
Fatigue	What is your level of fatigue right now on a scale of 0 (no fatigue)—10 (extremely severe fatigue)?	Self-report	3×/day
Depression	What is your level of depression right now on a scale of 0 (not at all depressed)—10 (extremely depressed)?	Self-report	3×/day
Anxiety	What is your level of anxiety right now on a scale of 0 (not at all anxious)—10 (extremely anxious)?	Self-report	3×/day
Upper extremity weakness	What is your level of upper extremity weakness on a scale of 0 (no weakness)—10 (extremely severe weakness)?	Self-report	3×/day
Pain	What is your level of pain right now on a scale of 0 (no pain)—10 (worst pain imaginable)?	Self-report	3×/day
Overall cognitive function	What is your level of cognitive function right now on a scale of 0 (good: my thinking is sharp and quick)—10 (bad: my thinking is very difficult or slow)?	Self-report	3×/day
Executive function	Mobile version of the Trail-Making Test part B (25); participants connected consecutive numbers and letters in order; completion time was used as primary outcome in multilevel models	Performance-based	3×/day
Upper extremity motor speed	Mobile version of a Finger Tapping task (30); participants tapped two fingers of the same hand alternatively for 10 s as quickly as possible; this was done for both right and left hands; average number of taps across two hands was used as covariates in multilevel models	Performance-based	3×/day
Sleep	How many hours of sleep did you get last night? Did you have difficulty falling asleep (yes or no)?	Self-report	1×/day
Accomplishment	To what extent were you able to accomplish everything you wanted to do today on a scale of 0 (I was unable to accomplish anything I wanted to do today)—10 (I was able to accomplish everything I wanted to do today)?	Self-report	1×/day

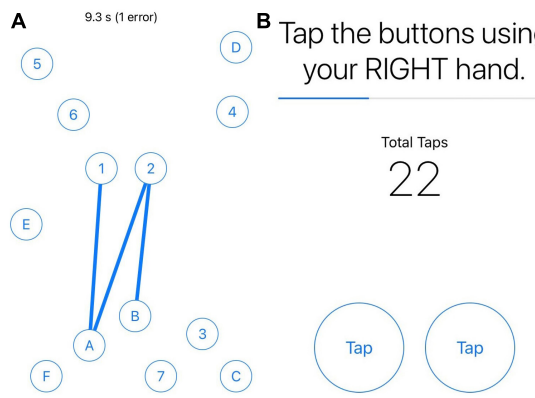


FIGURE 1
Screenshots of mTMT-B and finger tapping tasks. Panel **(A)** is screenshot of the mobile Trail-Making Test part B (mTMT-B) task. Panel **(B)** is a screenshot of the mobile finger tapping task. Users performed one trial with their right hand, followed by a second with their left hand.

completed a virtually administered neuropsychological battery and self-report inventories at baseline. After the baseline assessment, they were instructed on downloading and using the study app (23–25). Then participants were asked to complete EMAs four times per day for 3 weeks. EMAs consisted of brief self-report ratings and performance-based tasks delivered through the participant's smartphone.

2.3. Baseline assessment

Participants completed a virtually administered baseline assessment (*via* videoconferencing), which consisted of a brief battery of neuropsychological tests and phone-based Expanded Disability Status Scale (EDSS) (26). The neuropsychological battery included the oral version of the Symbol Digit Modalities Test (SDMT) (27), which is considered a gold standard clinical trial endpoint for MS-related cognitive dysfunction (28) and was used in the current study to characterize cognitive status (other neuropsychological measures were not used and therefore omitted in this paper). On the SDMT, participants were provided with a key of nine symbol-digit pairs. They were instructed to call out numbers associated with symbols presented in the test stimulus set one at a time as quickly as they could within 90 s. SDMT measures processing speed, with higher scores indicating faster processing speed. Raw scores for SDMT were converted to *z*-scores using normative data from Strober et al. (29). For the phone version of EDSS, assessment of ambulation and functional systems were obtained *via* self-report based on procedures outlined in Lechner-Scott et al. (26). EDSS is the standard method for assessing neurological disability among persons with MS and ranges between 0 and 10 with 0.5 increments (e.g., 0 = no disability, 2.5 = mild disability, 6.0 = requiring a walking aid, 9.0 = confined to bed).

2.4. Ecological momentary assessment protocol

EMAs were administered using the BiAffect app (23–25), which was available for download for iOS devices through the Apple app store. There were four EMAs per day during the 3-week monitoring period. The first three EMAs each day focused on self-reported symptom severity in real-time and performance on smartphone-based cognitive and motor tasks (see **Table 1** for details on EMA measures used in this study) including part B of the Trail-Making Test (mTMT-B) (25) and a finger tapping task (30). For the mTMT-B task, participants were asked to connect and alternate between numbers and letters consecutively and quickly on the screen (see **Figure 1A**). For the mobile finger tapping task, participants were asked to tap two fingers of the same hand alternatively as quickly as possible for 10 s; they performed one trial with their right hand and another trial with their left hand (see **Figure 1B**). Symptom severity was based on the Visual Analogue Scale (VAS) (31), which is

TABLE 2 Demographic and clinical characteristics of the sample.

Age: mean years (SD); range	41.69 (13.39); 20–70
Sex	
Female	41 (91.11)
Male	4 (8.89)
Education: number (proportion)	
High school graduate or fewer years of education	4 (8.89)
Some college with no degree or associate's degree	12 (26.67)
Bachelor's degree	17 (37.78)
Master's degree	10 (22.22)
Doctoral degree	1 (2.22)
Prefer not to answer	1 (2.22)
Race/ethnicity: number (proportion)	
Non-Hispanic white	33 (73.33)
Non-Hispanic black	5 (11.11)
Hispanic/Latino(a)	3 (6.67)
Asian	3 (6.67)
Prefer not to answer	1 (2.22)
MS disease course: number (proportion)	
Relapsing-remitting	39 (86.67)
Primary progressive	3 (6.67)
Secondary progressive	2 (4.44)
Not sure	1 (2.22)
MS disease duration: mean years (SD); range	11.06 (9.30); 4.38 months—29.95 years
EDSS	
0–2: number (proportion)	5 (11.11)
2.5: number (proportion)	12 (26.67)
3.0: number (proportion)	14 (31.11)
3.5–4.5: number (proportion)	11 (24.45)
> 4.5: number (proportion)	3 (6.66)
SDMT: *z*-score (SD)	-1.46 (1.34)

MS, multiple sclerosis; SD, standard deviation; EDSS, Expanded Disability Status Scale; SDMT, Symbol Digit Modalities Test.

commonly used in EMA research, including EMA studies conducted in MS (6, 7, 11). The last EMA of the day asked for reports that only required one response per day (e.g., sleep, sense of accomplishment). Throughout the monitoring period, participants were prompted to complete EMAs through text messages (with reminders to complete them on the study app). They were told to complete each EMA within 1 h of the prompt if not exactly at the prompted time. The first three EMAs were approximately equally spaced in time throughout the day (first in the morning, second in mid-day/early afternoon, and third in late afternoon/early evening) based on the participant's individual sleep-wake cycle. If a participant had a different schedule for the weekend, their EMA schedule was adjusted accordingly. The last EMA of the day was administered about 1–2 h before the participant's bedtime.

TABLE 3 Intraclass correlations (ICCs) for each symptom rating/performance.

	Between-person ICC	Between-day ICC within persons
mTMT-B competition time	0.62	0.11
Self-reported cognitive dysfunction	0.59	0.17
Self-reported fatigue	0.62	0.07
Self-reported depressive symptoms	0.66	0.17
Self-reported anxiety	0.64	0.15
Self-reported pain	0.76	0.08
Self-reported number of hours slept	0.54	N/A
Self-reported difficulties falling asleep	0.64	N/A

ICCs were calculated based on null (unconditional) models with only subject and day random intercepts. Since sleep questions were only administered once per day, there were no between-day ICCs.

2.5. Statistical analysis

All analyses were conducted in R version 4.2.1. Descriptive statistics were used to determine demographic and clinical characteristics of the sample. Multilevel models were used to examine intraday variability and associations among EMA measures, in order to account for within-subject and within-day clustering. All multilevel models included random intercepts for the subject (to account for within-person clustering) as well as random intercepts for concatenation of subject and day variables (e.g., day 1 for subject 0001 is 00011, day 2 for subject 0001 is 00012, etc.; to account for within-day clustering) (32), except for variables collected for only once per day (i.e., sleep, sense of accomplishment) which only included the subject's intercept. All models were fit using the restricted maximum likelihood approach, which is the recommended default method by the R packages lme4 (33) and lmerTest (34).

2.5.1. Compliance to EMA and intraclass correlations

For the first three EMAs of each day (which were time-sensitive), we included the responses in the final dataset if the EMA was completed within 2 h before or after the scheduled time. If participants completed multiple EMA measures within each scheduled period, the first complete response was used. For the last EMA of the day (not time-sensitive), we used the first complete response submitted after the third time-sensitive EMA. Compliance was defined as the ratio of completed EMAs within the specified time periods out of the total number of required EMAs. ICCs for each symptom rating and performance was calculated based on the null (unconditional) multilevel models (with only the random intercepts without fixed effects). ICCs signified the proportions of between-person (in this case, random variance for the subject ID variable) and between-day (in this case, random variance for the concatenated subject and day variable) variances out of the total random variance for each outcome.

2.5.2. Separating state and trait aspects of symptom rating/performance

We separated state (how each symptom varied from the individual's typical level) and trait (each individual's typical level

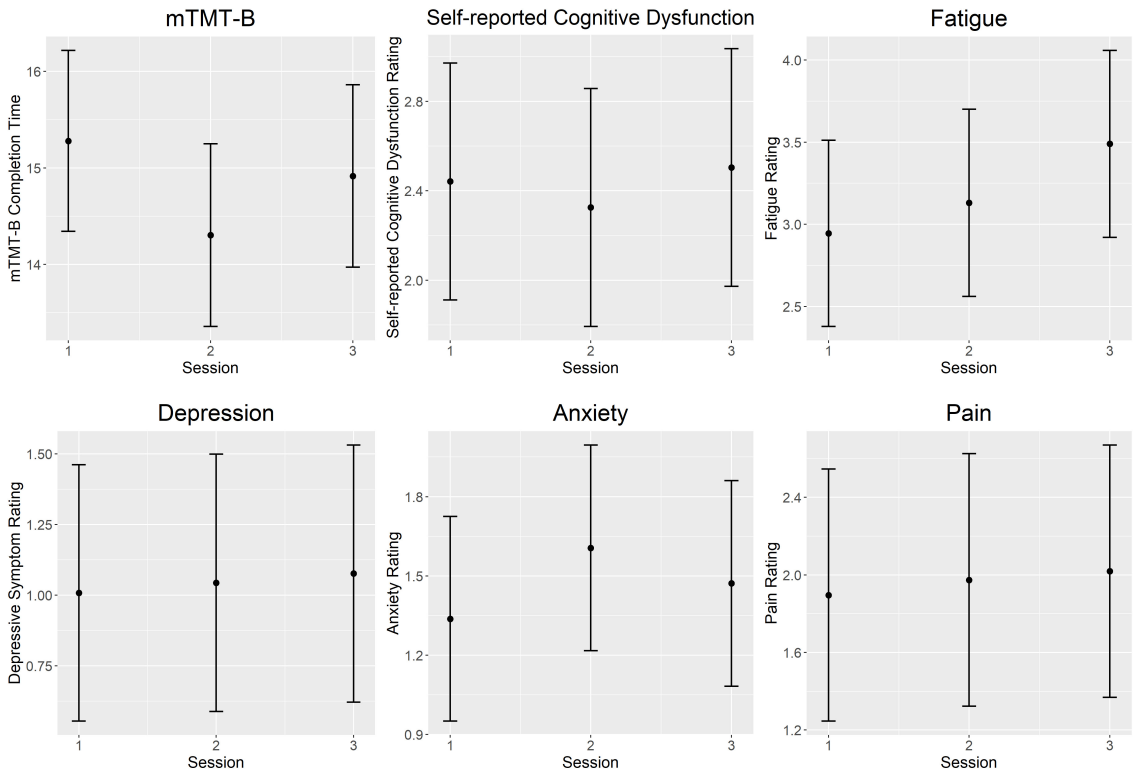


FIGURE 2
Intra-day fluctuations of symptom severity. Both objectively and subjectively measured cognition was worse in the morning and end of day compared to the middle of the day. Anxiety ratings showed the opposite trend and peaked at mid-day. Fatigue ratings increased steadily throughout the day. There were no significant intra-day variations in ratings of pain and depressive symptoms. Each plot represents predicted values from multilevel models for the session fixed effect. Error bars represent 95% confidence intervals. mTMT-B, mobile Trail-Making Test part B.

of symptom severity) aspects of each symptom rating/performance score using participant-mean centering (35). First, for each symptom rating/performance, the scores for different EMA sessions were averaged within each participant, creating the participant means which were also the trait aspect of that symptom rating/performance (e.g., each individual's typical level of depressive symptoms). Then, we centered each EMA score around the participant mean; this was the state aspect of each symptom rating/performance for each EMA session (e.g., when depressive symptoms were more or less severe than the individual's typical level of depression). For self-reported difficulties falling asleep, since it is a binary variable (not continuous), we did not separate their state and trait aspects because there were no "participant means."

2.5.3. Intra-day fluctuations in symptom severity

Multilevel models were conducted to evaluate symptom fluctuations over time. In these models, each MS symptom rating as well as mTMT-B completion time were outcomes, and session number was fixed effect predictor. For mTMT-B, the model was adjusted for age, mean bilateral finger tapping performance (number of taps), state and trait upper extremity weakness rating, and measurement number (to account for practice effects). For self-reported symptom ratings other than depressive symptoms (self-reported cognitive dysfunction, fatigue, anxiety, pain, and sleep), models were adjusted for state and trait depressive symptom ratings (to account for response bias due to depression).

2.5.4. Real-time associations between symptom ratings and cognitive functioning in real time

Multilevel models were used to determine real-time associations between non-cognitive symptom ratings (fatigue, depression, anxiety, pain, and sleep; state and trait aspects of each symptom as well as their interactions as fixed effect predictors in each model) and measures of cognition (mTMT-B completion time and self-reported cognitive dysfunction; each as outcome in separate model). As in previous models, models with mTMT-B completion time as outcome were adjusted for age, mean bilateral finger tapping performance, state and trait upper extremity weakness rating, and measurement number. Models with self-reported cognitive dysfunction rating as outcome were adjusted for state and trait depressive symptom ratings.

2.5.5. Real-time associations between cognitive functioning and self-reported sense of accomplishment in real time

Multilevel models were used to examine real-time associations between cognitive functioning (mTMT-B and self-reported cognitive dysfunction) and perceived sense of accomplishment, with state and trait aspects of the former as fixed effect predictors and latter as outcome. Models were adjusted for state and trait depressive symptom ratings.

2.5.6. Multiple comparison corrections

Since each set of analyses answered an independent question, we adjusted for multiple comparisons using Bonferroni correction for each outcome separately (instead of adjusting for all models

TABLE 4 Model estimates for intra-day symptom fluctuations.

Contrast	Standardized coefficient	95% confidence intervals	*P*-value
mTMT-B completion time			
Session 2 vs. session 1	−0.17	−0.23 to −0.11	< 0.001*
Session 3 vs. session 1	−0.06	−0.12 to 0.00	0.041
Session 3 vs. session 2	0.10	0.04 to 0.17	< 0.001*
Self-reported cognitive dysfunction			
Session 2 vs. session 1	−0.05	−0.10 to 0.01	0.076
Session 3 vs. session 1	0.03	−0.03 to 0.08	0.334
Session 3 vs. session 2	0.07	0.02 to 0.13	0.008*
Self-reported fatigue			
Session 2 vs. session 1	0.07	0.01 to 0.13	0.025
Session 3 vs. session 1	0.21	0.15 to 0.27	< 0.001*
Session 3 vs. session 2	0.14	0.07 to 0.20	< 0.001*
Self-reported depressive symptoms			
Session 2 vs. session 1	0.02	−0.03 to 0.08	0.412
Session 3 vs. session 1	0.04	−0.01 to 0.10	0.111
Session 3 vs. session 2	0.02	−0.03 to 0.08	0.461
Self-reported anxiety			
Session 2 vs. session 1	0.12	0.07 to 0.18	< 0.001*
Session 3 vs. session 1	0.06	0.01 to 0.11	0.020
Session 3 vs. session 2	−0.06	−0.12 to -0.01	0.025
Self-reported pain			
Session 2 vs. session 1	0.03	−0.01 to 0.07	0.162
Session 3 vs. session 1	0.05	0.01 to 0.09	0.026
Session 3 vs. session 2	0.02	−0.03 to 0.06	0.436

All models included random intercepts for subject and an aggregated subject and day variable. Models with mTMT-B as outcomes included age, mean bilateral finger tapping performance, state and trait upper extremity weakness rating, and measurement number as fixed effects. Models with self-reported symptom ratings other than depressive symptoms included state and trait depressive symptom ratings as fixed effects. mTMT-B, mobile Trail-Making Test part B. *Denotes significant comparisons at Bonferroni-corrected $p = 0.008$ level.

conducted in the study). For intra-day variation in symptoms (section "2.5.3. Intra-day fluctuations in symptom severity"), associations between non-cognitive symptoms and mTMT-B performance (section "2.5.4. Real-time associations between symptom ratings and cognitive functioning in real time"), and associations between non-cognitive symptoms and self-reported cognitive dysfunction (section "2.5.4. Real-time associations between symptom ratings and cognitive functioning in real time"), six models were conducted for each question, so the Bonferroni-corrected *p*-value threshold is 0.05/6 = 0.009. For predictors of sense of accomplishment (section "2.5.5. Real-time associations between cognitive functioning and self-reported sense of accomplishment in real time"), two models were conducted, so the Bonferroni-corrected *p*-value threshold s 0.05/2 = 0.025.

3. Results

The study sample consisted of 45 participants with MS, who completed 3,174 EMA sessions across the 3-week monitoring period. Compliance to EMA was 84%. Table 2 summarizes demographic and clinical characteristics of the sample. The sample was, on average, middle-aged and consisted of primarily females and non-Hispanic whites. Majority of the sample completed at least some college. Relapsing-remitting disease course was the dominant phenotype. Disease duration was heterogenous, ranging between several months to almost 30 years. Based on self-report, most participants had EDSS scores between 2.5 and 4.5, which signified the ability to ambulate without aid with some degrees of limitation. Compared to a normative sample, participants in this study had mild to moderate processing speed impairment (*z*-score approaching 1.5 standard deviations below the mean). Between-person and between-day ICCs based on unconditional multilevel models for each symptom rating/performance are summarized in Table 3. Across symptoms, approximately two-thirds of the random variance was attributed to between-person variability relative to within-person variability. Between-day variability was small within each person.

3.1. Intra-day fluctuations in symptom severity

Figure 2 illustrates intra-day fluctuations in various MS symptom severity, and Table 4 summarizes the associated model estimates. mTMT-B completion time and fatigue ratings showed the most variation across sessions each day. Anxiety ratings showed significant variation in the earlier part of the day (sessions 1 vs. 2), while self-reported cognitive dysfunction showed significant variation in the latter part of the day (sessions 2 vs. 3). Depression and pain ratings did not significantly vary across the day.

3.2. Real-time associations between symptom ratings and cognitive functioning

Figure 3 illustrates real-time associations between non-cognitive symptom ratings and mTMT-B performance, and Table 5 summarizes partial model estimates for the state and trait symptom variables (see Supplementary Table 1 for full model estimates). Figure 4 illustrates real-time associations between non-cognitive symptom ratings and self-reported cognitive dysfunction rating, and Table 6 summarizes partial model estimates for the state and trait symptom variables (see Supplementary Table 2 for full model estimates). Among all non-cognitive symptom ratings (both state and trait), only more severe state depressive symptoms were associated with slower mTMT-B completion time. On the other hand, for self-reported cognitive dysfunction, state fatigue, depressive symptoms, anxiety, and pain were all significant predictors, with higher severity in non-cognitive symptoms correlating with more severe self-reported cognitive dysfunction. None of the trait symptom levels, except for depressive symptoms, were significantly associated with self-reported cognitive dysfunction.

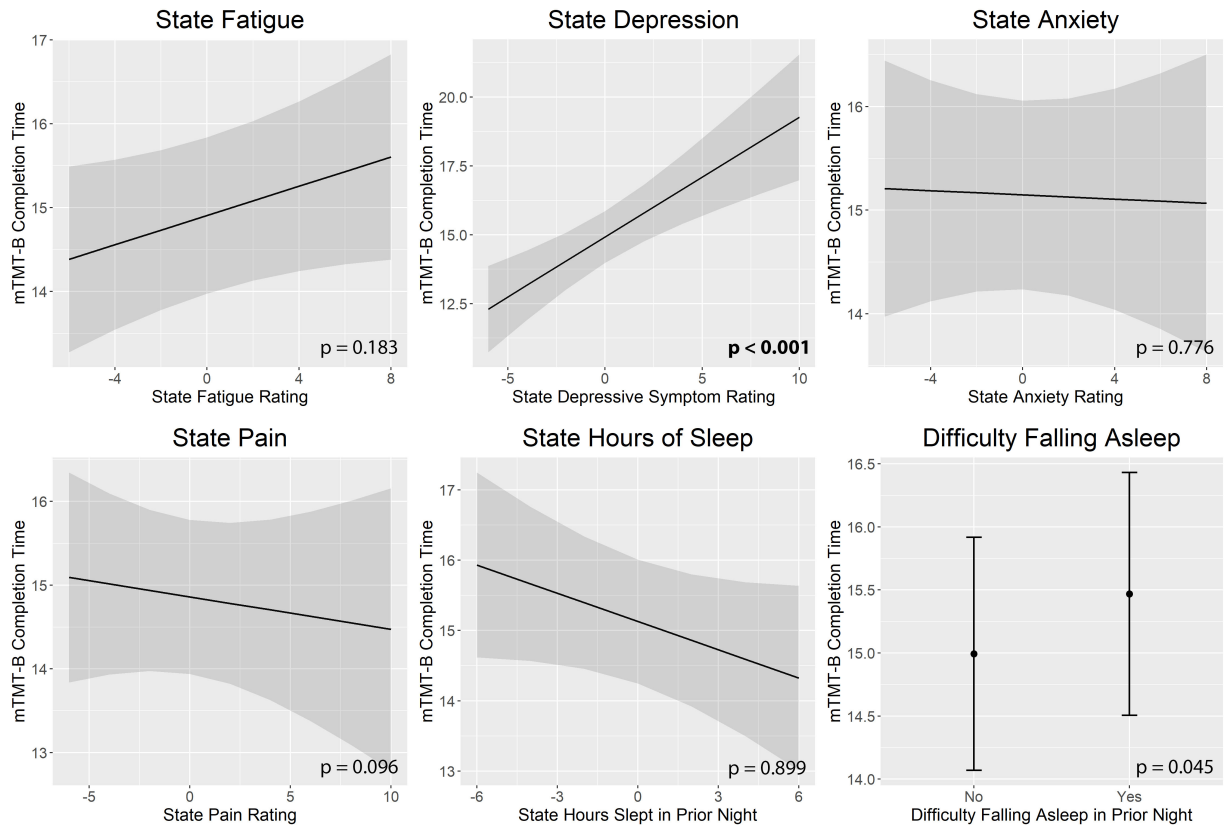


FIGURE 3
Real-time associations between non-cognitive symptom ratings and mTMT-B performance. Among all non-cognitive symptom ratings, only more severe state depressive symptoms was associated with slower mTMT-B completion time. State and trait aspects of each symptom was tested together in the same model; plots show the marginal effects of the state variables. All models included subject and concatenated subject and day variable as random intercepts; and age, mean bilateral finger tapping performance, state and trait upper extremity weakness rating, and measurement number as fixed effects. mTMT-B, mobile Trail-Making Test part B. Error bands represent 95% confidence intervals.

3.3. Real-time associations between cognitive functioning and perceived sense of accomplishment in real time

There was a statistical trend of lower level of state self-reported cognitive dysfunction (but not trait) correlating with higher perceived sense of accomplishment. Neither state nor trait mTMT-B completion time was associated with perceived sense of accomplishment. See Table 7 for model estimates.

4. Discussion

The current study examined real-time cognitive functioning among persons with MS using EMA. It is the first MS EMA study to include objectively measured cognitive functioning—in this case, executive functioning—in real-time. We found that when fatigue, depression, anxiety, and pain were more severe than the individual's usual levels ("state" as opposed to "trait"), the individual reported more cognitive dysfunction. In contrast, objectively measured executive functioning seemed specifically sensitive to state depressive symptoms. Further, there was a trend that self-reported cognitive dysfunction predicted lower perceived sense of accomplishment more than objectively measured executive dysfunction. These results demonstrated divergent factors that influence subjectively and objectively measured cognitive functioning in real time and is the first of such investigation in the MS population. Our results confirmed cross-sectional studies linking cognition with fatigue, depression, anxiety, and pain among persons with MS (36–40), and further extended these studies by establishing real-time associations (more temporally precise) within the real-life context (more ecologically valid).

Results of the current study illustrated the importance of assessing state, and not just trait, aspect of each symptom when considering inter-symptom relationships. We found many significant associations with state variables and almost none with trait variables. This may explain why there were inconsistent findings among cross-sectional studies (focusing on trait), where sometimes certain MS symptoms were associated with other symptoms and sometimes such associations were absent. Only an EMA framework enables investigations into state variations in symptoms. This study confirmed the feasibility of utilizing EMA to assess a range of MS symptoms within the real-world context and showed comparable compliance rates (> 80%) as previous, albeit limited number of, studies (7, 11). Thus, it may be feasible to integrate this form of assessment into routine clinical practice. Current standard of MS care involves once-per-year evaluations, which do not take into account of symptom variability between visits. Even when providers ask about these variations, the responses are likely influenced by recall

TABLE 5 Model estimates for real-time associations between non-cognitive symptom ratings and mTMT-B performance.

Variable	Standardized coefficient	95% confidence intervals	*P*-value
State fatigue	0.03	−0.004 to 0.05	0.183
Trait fatigue	−0.05	−0.21 to 0.12	0.581
State depressive symptoms	0.08	0.04 to 0.12	< 0.001*
Trait depressive symptoms	−0.04	−0.19 to 0.11	0.605
State anxiety	-2.44×10^{-03}	−0.04 to 0.03	0.776
Trait anxiety	−0.13	−0.30 to 0.05	0.160
State pain	-8.14×10^{-03}	−0.04 to 0.02	0.096
Trait pain	−0.09	−0.29 to 0.10	0.349
State number of hours slept in prior night	−0.02	−0.05 to 0.00	0.899
Trait number of hours slept in prior night	−0.12	−0.26 to 0.02	0.104
Difficulties falling asleep in prior night (yes vs. no) in prior night	0.08	0.002 to 0.16	0.045

State and trait aspects of each symptom was tested together in the same model (along with their interaction). All models included random intercepts for subject and an aggregated subject and day variable; and age, mean bilateral finger tapping performance, state and trait upper extremity weakness rating, and measurement number as fixed effects. mTMT-B, mobile Trail-Making Test part B. *Denotes significant comparisons at Bonferroni-corrected $p = 0.008$ level.

bias, especially for a population with known memory impairment such as MS. Other disciplines such as sleep medicine have already demonstrated the clinical feasibility and utility of EMA in the form of sleep diaries that patients have to complete daily for 1–2 weeks. Therefore, it is feasible for such practice to be integrated into MS care, particularly with aid from mobile technologies.

Remote monitoring of neurologic and cognitive symptoms using EMA and smartphone-based cognitive assessment may be extended to other populations as well. Such investigations have already begun in populations such as individuals with HIV (41) and Parkinson's disease (42). Besides subjective EMA surveys and smartphone-based cognitive assessments, objective data on motor fluctuations (43) and psychological symptoms (44) can also be gathered using smartphone sensors in the ambulatory setting. These methods are not dissimilar to established remote monitoring practices used in cardiac (e.g., Holtzer monitor) and diabetes (e.g., continuous glucose monitoring) care. In the age of personalized medicine, remote monitoring will provide patients and clinicians with real-world, temporally rich data needed for individualized treatments and recommendations.

Inclusion of both subjectively and objectively measured cognitive functioning is a strength of the current study. Previous MS studies have found that subjective and objective cognitive functioning do not always correlate (17, 18), and subjective appraisal of one's own cognition relates more strongly to affective symptomology (especially depression) than objective performance (18, 45). Our results help delineate differential factors that influence subjective and objective cognitive outcomes in real time. Further, given the known association between depression and subjective symptom reports (18, 45), we controlled for depression (both state and trait) in analyses with subjective symptom reports as outcomes. Thus, we can conclude that in addition to the clear associations between state depressive symptoms and cognitive dysfunction, state fatigue, anxiety, and pain symptoms were also related to self-reported cognitive deterioration in real time, independent of depression status.

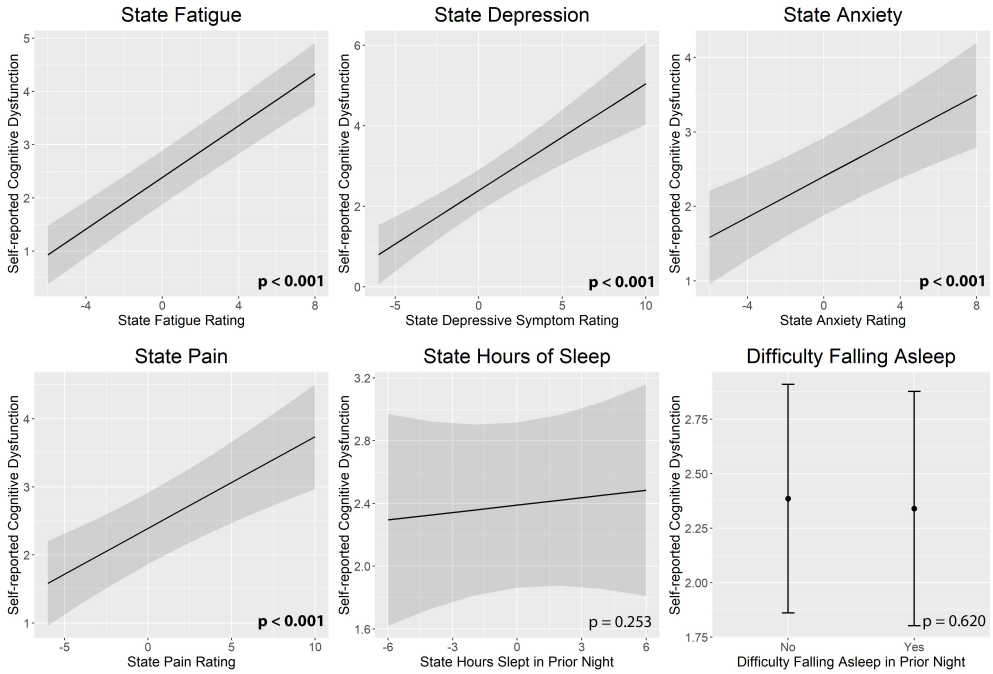


FIGURE 4
Real-time associations between non-cognitive symptom ratings and self-reported cognitive dysfunction. State fatigue, depressive symptoms, anxiety, and pain were all significant predictors of self-reported cognitive dysfunction. State and trait aspects of each symptom was tested together in the same model; plots show the marginal effects of the state variables. All models included subject and day number as random intercepts and state and trait depressive symptoms as fixed effects. Error bands represent 95% confidence intervals.

In contrast, objectively measured executive dysfunction was only related to state depressive symptoms, independent of upper extremity motor functioning. It is possible that the self-report question captured a broader sense of cognitive dysfunction than the objective measure (mTMT) which specifically tapped into processing speed and executive functioning. Our results were consistent with prior MS literature that found a particularly unique association between depression and executive functioning (46–48).

Only one research group has examined a range of MS symptoms using EMA with an adequate sample size (6, 7, 16), while others focused on fatigue (11, 15, 49) or had a very limited sample size (8). We found that fatigue most consistently increased in severity throughout the day, which was consistent with a prior study (7). Further, we found significant inter-symptom associations especially among self-report measures, which was also concordant with the prior literature (16). Compared to works by this research group (6, 7, 16), the present research further added an objective cognitive measure and self-report of anxiety symptoms, which had not been previously investigated using EMA in MS.

Of note, the current study focused on deviations in symptom severity from the individual's typical level (state vs. trait) and not absolute levels. This is an important context for the objective cognitive outcome (mTMT-B) used in this study because we were unable to determine whether individual instances of decline were clinically significant. In fact, there are currently no well-validated mobile cognitive measures with robust norms that take into account of repeated measurements and allow users to determine level of clinical impairment or decline. This is an area requiring future investigations. One promising effort is the National Institute on Aging (NIA)-funded Mobile Toolbox (50), which consists of a suite of mobile tasks validated against gold standard measures with population norms generated. The project is currently in its beta testing phase and will be eventually made available to external researchers. That being said, the current study's version of mTMT-B has been validated with the traditional paper-and-pencil version of TMT in a small sample (25), and practice effects were accounted for in our analyses. For subjective measures, we used the VAS (31) as frequently used in other EMA investigations. But unlike cross-sectional self-report measures with established clinical ranges, severity levels as determined by the VAS are individualized and their relations to other disease characteristics are unknown. Thus, we focused on changes in symptom severity from individual's typical levels as determined by the VAS.

The current study is limited by the relatively low levels of symptom severity reported by our sample. On average, participants were reporting symptom levels below 4 on a 10-point scale. This may be due to the fact that many of our participants had relatively chronic and stable disease course. Future studies should aim to recruit participants with more active disease in order to fully capture intra-day clinical fluctuations. That being said, even with relatively low levels of symptom severity, we still found significant intra-day fluctuations in objective cognitive performance and fatigue ratings.

Anther limitation is the predominance of female sex and relapsing-remitting disease course within our sample, which is consistent with prevalence rates in the general MS population but may restrict our ability to generalize our findings to minority populations such as males with MS and those with progressive disease courses. Future studies may consider oversampling these minority groups to confirm our findings.

TABLE 6 Model estimates for real-time associations between non-cognitive symptom ratings and self-reported cognitive dysfunction.

Variable	Standardized coefficient	95% confidence intervals	*P*-value
State fatigue	0.17	0.15 to 0.20	< 0.001*
Trait fatigue	0.20	−0.03 to 0.42	0.066
State depressive symptoms	0.12	0.08 to 0.16	< 0.001*
Trait depressive symptoms	0.27	0.10 to 0.43	0.005*
State anxiety	0.08	0.05 to 0.11	< 0.001*
Trait anxiety	−0.08	−0.37 to 0.20	0.493
State pain	0.07	0.04 to 0.10	< 0.001*
Trait pain	−0.03	−0.27 to 0.21	0.806
State number of hours slept in prior night	6.61×10^{-03}	−0.02 to 0.04	0.253
Trait number of hours slept in prior night	−0.07	−0.27 to 0.13	0.361
Difficulties falling asleep in prior night (yes vs. no) in prior night	−0.02	−0.10 to 0.06	0.620

State and trait aspects of each symptom was tested together in the same model (along with their interaction). All models included random intercepts for subject and an aggregated subject and day variable (except for sleep variables which only included the subject intercept) and state and trait depressive symptoms as fixed effects. *Denotes significant comparisons at Bonferroni-corrected $p = 0.008$ level.

TABLE 7 Model estimates for real-time associations between cognitive functioning and self-reported sense of accomplishment.

Variable	Standardized coefficient	95% confidence intervals	*P*-value
State mTMT-B completion time	-3.17×10^{-03}	−0.04 to 0.04	0.608
Trait mTMT-B completion time	−0.15	−0.34 to 0.04	0.120
State self-reported cognitive dysfunction	0.04	0.00 to 0.07	0.027*
Trait self-reported cognitive dysfunction	−0.14	−0.35 to 0.07	0.205

State and trait cognitive functioning was tested together in the same model. All models included subject as random intercepts and state and trait depressive symptoms as fixed effects. *Denotes statistical trend (at Bonferroni-corrected $p = 0.025$ level).

Further, there may be a selection bias in our sample since only iPhone users were eligible for our study. While smartphone use is fairly ubiquitous in the U.S. [85% of Americans own smartphones (51)], there may be socioeconomic differences among individuals who use iPhones compared to Android devices. Finally, while EMA is advantageous over retrospective self-report because it minimizes recall bias, it is important to note that besides the mTMT-B and finger tapping tasks, all other symptoms were evaluated subjectively. Future studies may explore real-time objective measures for mood and fatigue through smartphone (e.g., GPS, call/text logs) and other wearable sensors (e.g., heart rate, skin conductance, sleep patterns).

In conclusion, the current study was the first to identify divergent factors that influence subjectively and objectively measured cognitive

functioning in real time. While self-reported cognitive dysfunction was associated with a range of non-cognitive symptoms and self-reported sense of accomplishment, objectively measured executive functioning was only associated with depressive symptoms. Notably, we found that only state aspects of non-cognitive MS symptoms (and not trait) were associated with cognitive fluctuations, which supports the use of EMA in MS symptom monitoring.

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 the Kessler Foundation. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MC designed the study, obtained the funding, analyzed the data, and wrote the full draft of the manuscript (except for section "1. Introduction"). CC wrote the section "1. Introduction" of the manuscript. KE and CR performed the data cleaning and calculations. MR managed and processed the study app data. AL headed the team that created the study app and managed app data, obtained the funding, and provided the feedback on data analysis and manuscript draft. JD obtained the funding and provided the feedback for the analysis and manuscript draft. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by the New Jersey Health Foundation (PC 7–21), the Robert Wood Johnson Foundation (a New Venture Fund/MoodChallenge for ResearchKit), and the National Multiple Sclerosis Society (MB-1606-08779).

Conflict of interest

AL was a cofounder of KeyWise AI, currently a consultant for Otsuka US, and on the medical board of Buoy 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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.1049686/full#supplementary-material

References

1. Mirmosayyeb O, Brand S, Barzegar M, Afshari-Safavi A, Nehzat N, Shaygannejad V, et al. Clinical characteristics and disability progression of early-and late-onset multiple sclerosis compared to adult-onset multiple sclerosis. *J Clin Med.* (2020) 9:1326. doi: 10.3390/jcm9051326

2. Crayton H, Rossman H. Managing the symptoms of multiple sclerosis: a multimodal approach. *Clin Ther.* (2006) 28:445–60. doi: 10.1016/j.clinthera.2006.04.005

3. Chiaravalloti N, DeLuca J. Cognitive impairment in multiple sclerosis. *Lancet Neurol.* (2008) 7:1139–51. doi: 10.1016/S1474-4422(08)70259-X

4. Guimarães J, Sá M. Cognitive dysfunction in multiple sclerosis. *Front Neurol.* (2012) 3:74. doi: 10.3389/fneur.2012.00074

5. Grzegorski T, Losy J. Cognitive impairment in multiple sclerosis–a review of current knowledge and recent research. *Rev Neurosci.* (2017) 28:845–60. doi: 10.1515/revneuro-2017-0011

6. Kratz A, Braley T, Foxen-Craft E, Scott E, Murphy IJ, Murphy S. How do pain, fatigue, depressive, and cognitive symptoms relate to well-being and social and physical functioning in the daily lives of individuals with multiple sclerosis? *Arch Phys Med Rehabil.* (2017) 98:2160–6. doi: 10.1016/j.apmr.2017.07.004

7. Kratz A, Murphy S, Braley T. Ecological momentary assessment of pain, fatigue, depressive, and cognitive symptoms reveals significant daily variability in multiple sclerosis. *Arch Phys Med Rehabil.* (2017) 98:2142–50. doi: 10.1016/j.apmr.2017.07.002

8. Kasser S, Goldstein A, Wood P, Sibold J. Symptom variability, affect and physical activity in ambulatory persons with multiple sclerosis: understanding patterns and time-bound relationships. *Disabil Health J.* (2017) 10:207–13. doi: 10.1016/j.dhjo.2016.10.006

9. Raphael K. Recall bias: a proposal for assessment and control. *Int J Epidemiol.* (1987) 16:167–70.doi: 10.1093/ije/16.2.167

10. Wrzus C, Neubauer A. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. *Assessment.* (2022): [Epub ahead of print]. doi: 10.1177/10731911211067538

11. Powell D, Liossi C, Schlotz W, Moss-Morris R. Tracking daily fatigue fluctuations in multiple sclerosis: ecological momentary assessment provides unique insights. *J Behav Med.* (2017) 40:772–83. doi: 10.1007/s10865-017-9840-4

12. McEwen B, Sapolsky R. Stress and cognitive function. *Curr Opin Neurobiol.* (1995) 5:205–16. doi: 10.1016/0959-4388(95)80028-X

13. Leavitt V, Sumowski J, Chiaravalloti N, DeLuca J. Warmer outdoor temperature is associated with worse cognitive status in multiple sclerosis. *Neurology.* (2012) 78:964–8. doi: 10.1212/WNL.0b013e31824d5834

14. Campbell L, Paolillo E, Heaton A, Tang B, Depp C, Granholm E, et al. Daily activities related to mobile cognitive performance in middle-aged and older adults: an ecological momentary cognitive assessment study. *JMIR mHealth uHealth.* (2020) 8:e19579.doi: 10.2196/19579

15. Heine M, van den Akker L, Blikman L, Hoekstra T, Van Munster E, Verschuren O, et al. Real-time assessment of fatigue in patients with multiple sclerosis: how does it relate to commonly used self-report fatigue questionnaires? *Arch Phys Med Rehabil.* (2016) 97:1887.e–94.e. doi: 10.1016/j.apmr.2016.04.019

16. Kratz A, Murphy S, Braley T. Pain, fatigue, and cognitive symptoms are temporally associated within but not across days in multiple sclerosis. *Arch Phys Med Rehabil.* (2017) 98:2151–9. doi: 10.1016/j.apmr.2017.07.003

17. Goverover Y, Kalmar J, Gaudino-Goering E, Shawaryn M, Moore N, Halper J, et al. The relation between subjective and objective measures of everyday life activities in persons with multiple sclerosis. *Arch Phys Med Rehabil.* (2005) 86:2303–8. doi: 10.1016/j.apmr.2005.05.016

18. Julian L, Merluzzi N, Mohr D. The relationship among depression, subjective cognitive impairment, and neuropsychological performance in multiple sclerosis. *Mult Scler J.* (2007) 13:81–6. doi: 10.1177/1352458506070255

19. Schmitt M, Blum G. State/Trait Interactions. In: Zeigler-Hill V, Shackelford TK editors. *Encyclopedia of personality and individual differences*. Cham: Springer (2020). p. 5206–9. doi: 10.1007/978-3-319-24612-3_1922

20. Harris P, Taylor R, Thielke R, Payne J, Gonzalez N, Conde J. 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

21. Harris P, Taylor R, Minor B, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Infor.* (2019) 95:103208. doi: 10.1016/j.jbi.2019.103208

22. Hensen B, Mackworth-Young C, Simwinga M, Abdelmagid N, Banda J, Mavodza C, et al. Remote data collection for public health research in a COVID-19 era: ethical implications, challenges and opportunities. *Health Policy Plan.* (2021) 36:360–8. doi: 10.1093/heapol/czaa158

23. Vesel C, Rashidisabet H, Zulueta J, Stange J, Duffecy J, Hussain F, et al. Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: a BiAffect iOS study. *J Am Med Infor Assoc.* (2020) 27:1007–18. doi: 10.1093/jamia/ocaa057

24. Zulueta J, Piscitello A, Rasic M, Easter R, Babu P, Langenecker S, et al. Predicting mood disturbance severity with mobile phone keystroke metadata: a biaffect digital phenotyping study. *J Med Internet Res.* (2018) 20:e241. doi: 10.2196/jmir.9775

25. Ross M, Demos A, Zulueta J, Piscitello A, Langenecker S, McInnis M, et al. Naturalistic smartphone keyboard typing reflects processing speed and executive function. *Brain Behav.* (2021) 11:e2363. doi: 10.1002/brb3.2363

26. Lechner-Scott J, Kappos L, Hofman M, Polman C, Ronner H, Montalban X, et al. Can the expanded disability status scale be assessed by telephone? *Mult Scler J.* (2003) 9:154–9. doi: 10.1191/1352458503ms884oa

27. Smith A. *Symbol digit modalities test (SDMT)*. Los Angeles, CA: Western Psychological Services (1982).

28. Strober L, DeLuca J, Benedict R, Jacobs A, Cohen J, Chiaravalloti N, et al. Symbol digit modalities test: a valid clinical trial endpoint for measuring cognition in multiple sclerosis. *Mult Scler J.* (2019) 25:1781–90. doi: 10.1177/1352458518808204

29. Strober L, Bruce J, Arnett P, Alschuler K, Lebkuecher A, Di Benedetto M, et al. A new look at an old test: normative data of the symbol digit modalities test–oral version. *Mult Scler Relat Disord.* (2020) 43:102154. doi: 10.1016/j.msard.2020.102154

30. Bot B, Suver C, Neto E, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci Data.* (2016) 3:160011. doi: 10.1038/sdata.2016.11

31. Carlsson A. Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. *Pain.* (1983) 16:87–101. doi: 10.1016/0304-3959(83)90088-X

32. Kleiman E. Understanding and analyzing multilevel data from real-time monitoring studies: an easily-accessible tutorial using R. *PsyArXiv.* [Preprint]. (2017). doi: 10.31234/osf.io/xf2pw

33. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *arXiv.* [preprint] arXiv:14065823. (2014). doi: 10.18637/jss.v\break067.i01 doi: 10.18637/jss.v067.i01

34. Kuznetsova A, Brockhoff P, Christensen R. lmerTest package: tests in linear mixed effects models. *J Stat Softw.* (2017) 82:1–26. doi: 10.18637/jss.v082.i13

35. Enders C, Tofighi D. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol Methods.* (2007) 12:121. doi: 10.1037/1082-989X.12.2.121

36. Feinstein A, Magalhaes S, Richard J, Audet B, Moore C. The link between multiple sclerosis and depression. *Nat Rev Neurol.* (2014) 10:507. doi: 10.1038/nrneurol.2014.139

37. Krupp L, Elkins L. Fatigue and declines in cognitive functioning in multiple sclerosis. *Neurology.* (2000) 55:934–9. doi: 10.1212/WNL.55.7.934

38. Diamond B, Johnson S, Kaufman M, Graves L. Relationships between information processing, depression, fatigue and cognition in multiple sclerosis. *Arch Clin Neuropsychol.* (2008) 23:189–99. doi: 10.1016/j.acn.2007.10.002

39. Leavitt V, Brandstadter R, Fabian M, Katz Sand I, Klineova S, Krieger S, et al. Dissociable cognitive patterns related to depression and anxiety in multiple sclerosis. *Mult Scler J.* (2020) 26:1247–55. doi: 10.1177/1352458519860319

40. Benson C, Kerr B. Pain and cognition in multiple sclerosis. In: Taylor BK, Finn DP editors. *Behavioral neurobiology of chronic pain*. Berlin: Springer (2014). p. 201–15. doi: 10.1007/7854_2014_309

41. Moore R, Campbell L, Delgadillo J, Paolillo E, Sundermann E, Holden J, et al. Smartphone-based measurement of executive function in older adults with and without HIV. *Arch Clin Neuropsychol.* (2020) 35:347–57. doi: 10.1093/arclin/acz084

42. Weizenbaum E, Fulford D, Torous J, Pinsky E, Kolachalama V, Cronin-Golomb A. Smartphone-based neuropsychological assessment in Parkinson's disease: feasibility, validity, and contextually driven variability in cognition. *J Int Neuropsychol Soc.* (2022) 28:401–13. doi: 10.1017/S1355617721000503

43. Zhan A, Mohan S, Tarolli C, Schneider R, Adams J, Sharma S, et al. Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score. *JAMA Neurol.* (2018) 75:876–80. doi: 10.1001/jamaneurol.2018.0809

44. Torous J, Staples P, Barnett I, Sandoval L, Keshavan M, Onnela J. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *NPJ Digit Med.* (2018) 1:15. doi: 10.1038/s41746-018-0022-8

45. Goverover Y, Chiaravalloti N, DeLuca J. The relationship between self-awareness of neurobehavioral symptoms, cognitive functioning, and emotional symptoms in multiple sclerosis. *Mult Scler J.* (2005) 11:203–12. doi: 10.1191/1352458505ms1153oa

46. Feinstein A. Mood disorders in multiple sclerosis and the effects on cognition. *J Neurol Sci.* (2006) 245:63–6. doi: 10.1016/j.jns.2005.08.020

47. Grech L, Kiropoulos L, Kirby K, Butler E, Paine M, Hester R. The effect of executive function on stress, depression, anxiety, and quality of life in multiple sclerosis. *J Clin Exp Neuropsychol.* (2015) 37:549–62. doi: 10.1080/13803395.2015.1037723

48. Arnett P, Higginson C, Randolph J. Depression in multiple sclerosis: relationship to planning ability. *J Int Neuropsychol Soc.* (2001) 7:665–74. doi: 10.1017/S1355617701766027

49. Kim E, Lovera J, Schaben L, Melara J, Bourdette D, Whitham R. Novel method for measurement of fatigue in multiple sclerosis: real-time digital fatigue score. *J Rehabil Res Dev.* (2010) 47:477–84. doi: 10.1682/JRRD.2009.09.0151

50. Mobile Toolbox. *Mobile toolbox.* (2022). Available online at: https://mobiletoolbox.org/ (accessed August 30, 2022).

51. Pew Resarch Center. *Mobile fact sheet*. (2021). Available online at: https://www.pewresearch.org/internet/fact-sheet/mobile/ (accessed August 30, 2022).

frontiers | Frontiers in Digital Health

TYPE Perspective
PUBLISHED 20 January 2023
DOI 10.3389/fdgth.2023.1064115

OPEN ACCESS

EDITED BY
Terry D. Ellis,
Boston University, United States

REVIEWED BY
Joe Wherton,
University of Oxford, United Kingdom

*CORRESPONDENCE
Frank A. Russo
russo@torontomu.ca

SPECIALTY SECTION
This article was submitted to Connected Health, a section of the journal Frontiers in Digital Health

RECEIVED 10 October 2022
ACCEPTED 02 January 2023
PUBLISHED 20 January 2023

CITATION
Russo FA, Mallik A, Thomson Z, de Raadt St. James A, Dupuis K and Cohen D (2023) Developing a music-based digital therapeutic to help manage the neuropsychiatric symptoms of dementia.
Front. Digit. Health 5:1064115.
doi: 10.3389/fdgth.2023.1064115

Developing a music-based digital therapeutic to help manage the neuropsychiatric symptoms of dementia

Frank A. Russo[1,2,3]*, Adiel Mallik[3], Zoe Thomson[3], Alexander de Raadt St. James[3], Kate Dupuis[4] and Dan Cohen[5]

[1]Department of Psychology, Toronto Metropolitan University, Toronto, ON, Canada, [2]KITE, Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada, [3]LUCID Inc., Toronto, ON, Canada, [4]Center for Elder Research, Sheridan College, Oakville, ON, Canada, [5]Right to Music, New York, NY, United States

The greying of the world is leading to a rapid acceleration in both the healthcare costs and caregiver burden that are associated with dementia. There is an urgent need to develop new, easily scalable modalities of support. This perspective paper presents the theoretical background, rationale, and development plans for a music-based digital therapeutic to manage the neuropsychiatric symptoms of dementia, particularly agitation and anxiety. We begin by presenting the findings of a survey we conducted with key opinion leaders. The findings highlight the value of a music-based digital therapeutic for treating neuropsychiatric symptoms, particularly agitation and anxiety. We then consider the neural substrates of these neuropsychiatric symptoms before going on to evaluate randomized control trials on the efficacy of music-based interventions in their treatment. Finally, we present our development plans for the adaptation of an existing music-based digital therapeutic that was previously shown to be efficacious in the treatment of adult anxiety symptoms.

KEYWORDS

digital therapeutics, dementia, neuropsychiatric symptoms, anxiety, agitation, music, artificial intelligence

Introduction

According to the World Health Organization (2022), there are 55 million people living with dementia worldwide with 10 million new cases annually. The same report estimates the global cost of dementia at 1.3 trillion USD (1). These costs are expected to surpass 2.8 trillion USD by 2030 as the number of people living with dementia rises. Approximately half of the global cost of dementia is attributable to the informal care provided by family members and friends who commonly shoulder tremendous physical, emotional and financial pressures (2). A sizeable minority of people living with dementia in industrialized societies will eventually be placed in long-term care (nursing or assisted care) homes. The proportion of total costs incurred in these homes that can be attributed to dementia has been estimated at 64% (3). Even before the COVID-19 pandemic, the professionals in these homes were chronically overloaded leading them to experience high levels of caregiver burden, and in some cases, moral injury, which has been defined as the perpetration, failure to prevent, or observation of morally-transgressive acts (4).

From the perspective of people living with dementia, their caregivers and the broader healthcare system, there is an urgent need to develop new, easily scalable modalities of support. Digital therapeutics (DTx) represent one such modality being considered. The focus of development in DTx for dementia has been cognitive stimulation (5), typically in the form of reminiscence therapy (6) or brainwave entrainment (7). The objective of such therapeutics is to directly slow the rate of

cognitive decline (8). While we believe that interventions that directly target cognitive outcomes are clinically important, we also believe there is an urgent need to target non-cognitive outcomes. These outcomes have been less well studied in the context of DTx but they have the potential to contribute to patient and caregiver wellbeing, while lowering the costs of care (9).

Our team is currently undertaking the development of a music-based DTx that builds on core AI that we originally developed to mitigate anxiety (https://www.lucidtherapeutics.com). While the anecdotal evidence for the power of music in dementia abounds, the evidence base is still in its early days and tends to be focused on cognitive outcomes. To better understand the potential impact of a music-based DTx on cognitive and non-cognitive outcomes we started our development path by surveying key opinion leaders.

Survey of key opinion leaders

In early 2022, our team undertook a qualitative study with key opinion leaders to gauge the potential value of developing a music-based DTx for dementia. In addition to defining the value proposition, we were interested in specific outcomes that were judged to be feasible, inclusive of cognitive and non-cognitive outcomes. Participants included 7 payers and 12 health-care practitioners specializing in geriatric and dementia care. Payers included medical directors ($n = 5$), pharmacy directors ($n = 1$), and an innovation officer ($n = 1$) associated with health plans that are based in the United States. All payers had experience with the evaluation of DTx for coverage and reimbursement. Health-care practitioners (HCPs) included neurologists ($n = 5$), geriatricians ($n = 4$), and psychiatrists ($n = 3$), all of whom had significant experience in treating Alzheimer's disease (AD) and other forms of dementia (50 patients or more in the last 3 months). Most of the HCPs surveyed had experience with use of DTx in treatment (75%), and about half had some experience in recommending music therapy for patients (58.3%). Both payers and HCPs expressed the view that there was a strong clinical case for a therapy/intervention that would target non-cognitive aspects of dementia. In particular, they identified the *neuropsychiatric symptoms* (10) as being a non-cognitive target outcome that might be well addressed by a music-based DTx. Neuropsychiatric symptoms are extremely common in dementia, affecting as much as 97% of patients (11) and have been associated with reduced quality of life (12), as well as the progression of cognitive decline (13, 14).

Neuropsychiatric symptoms: prevalence, caregiver challenge, and neural substrates

In descending order of frequency, neuropsychiatric symptoms of dementia include apathy, depression, agitation, psychosis, and sleep disturbances (15). Agitation is especially frequent (80%) in residents of long-term care homes and in those who are in moderate to severe stages of the disease (16) but can also affect many individuals (60%) with mild dementia or mild cognitive impairment (MCI) (17). According to the key opinion leaders we surveyed, agitation is the most challenging symptom with respect to patient management. This perspective is consistent with prior surveys conducted with caregivers. One study of American caregivers found agitation to be more distressing than apathy or depression (18). The same conclusion was reached in a study of caregivers conducted in Japan (19). The Japanese study also found that agitation was more likely to contribute to caregiver burnout than other neuropsychiatric symptoms.

In addition to the challenge that agitation presents for caregivers it has also been associated with the progression of cognitive decline in patients (13, 14). From a biopsychosocial model of cognitive aging (20), this association may be attributable to neurotoxic factors that manifest due to stress arising from frequent bouts of agitation (see (21). In the case of patients living in long-term care homes, the association may also be due to side effects of the antipsychotic medications that are commonly prescribed to treat agitation. A meta-regression involving data from ten studies found a strong linear correlation between antipsychotic treatment duration and change in cognition, with greater declines under antipsychotic treatment compared to placebo (22).

Risperidone, a commonly prescribed second-generation antipsychotic with the strongest evidence base for treating agitation and anxiety appears to have no adverse effects on cognition when prescribed as indicated for short-term use (23–25). However, side effects of risperidone include an elevated risk of ischemic stroke and transient ischemic attacks (26), which elevate mortality risk. Studies of other commonly prescribed second-generation antipsychotics, such as olanzapine, have shown some level of risk for cognition, especially in the case of participants with lower cognitive functioning at baseline (27). Haloperidol, a first-generation antipsychotic that continues to be prescribed for agitation is less efficacious than risperidone (28), has potent sedative effects, and was determined to be the riskiest of all pharmacological interventions with respect to mortality (29). In summary, while chronic agitation may hasten the progression of cognitive decline, the existing pharmacological approaches have limited efficacy and can carry significant risks to physical health (30) and cognitive health (22). The healthcare practitioners we surveyed were particularly interested in the development of DTx that would serve as complementary or low-risk alternatives to pharmacological treatment in the management of agitation and anxiety.

Some researchers have characterized agitation as the external manifestation of anxiety (31, 32). Up to 80% of people living with mild-to-moderate dementia experience anxiety (15). Anxiety may even be a risk factor for developing dementia; the risk of conversion to dementia nearly doubles when anxiety symptoms are present in people living with mild cognitive impairment (33). While agitation is more of an external behavior that is readily observable, anxiety is an internal state that can be hidden from plain view. It has been conceptualized as consisting of cognitive (i.e., worry about future threats) and somatic (i.e., bodily tension) components (34). Anxiety tends to be more common in the early stages of the "dementia journey", while agitation is more common in later stages (35). Although the anxiety does not appear to be causally related to later agitation as has often been proposed (32), there is a clear association between the two constructs across stages of disease (36).

The 'Uncertainty and Anticipation Model of Anxiety' (UAMA) posits that anxiety is a set of expected emotional, cognitive, and behavioural responses to the uncertainty of potential future threats,

often coupled with fear (37, 38). The UAMA model proposes that activity in the frontal cortex (dorsomedial prefrontal and orbitofrontal) is responsible for generating probabilistic estimates of future events and expected costs (37). The model also proposes that the amygdala plays a central role in the transmission and interpretation of anxiety and fear. In addition to afferents from the frontal cortex, the amygdala is known to receive afferents from the thalamus, periaqueductal gray, and entorhinal cortex (38–40). People living with a diagnosis of dementia tend to experience a great deal of uncertainty because of the unknown of how their illness will progress and not knowing what threats may await them (41). Neural degradation in frontal areas supporting working memory may further predispose individuals living with dementia to experience anxiety.

In the case of AD, the most common form of dementia, anxiety is associated with damage to subcortical regions which includes atrophy in the amygdala (38, 42) and the entorhinal cortex (43). Cases of more severe anxiety are associated with hyperfusion of the anterior cingulate cortex, decreased grey matter volume in the right precuneus, inferior parietal, left parahippocampal, posterior cingulate gyrus, left insula, and bilateral putamen lobes (37, 38) and hypometabolism in the bilateral entorhinal, anterior hippocampus, left superior temporal and insula regions (38, 44). Positron emission tomography (PET) studies indicate that individuals living with AD and comorbid anxiety possess higher amyloid deposits than those without in the precuneus-posterior cingulate, frontal, parietal, and anterior cingulate cortex (45).

As shown in **Figure 1**, individuals living with AD that present with agitation show severe dysfunction in many of the same brain regions that are implicated in anxiety, including the amygdala, hippocampus, anterior cingulate, posterior cingulate, and insula (46). This pattern of dysfunction agrees with the pattern of disease progression wherein the propensity for anxiety is higher in earlier stages while the propensity for agitation is higher in later stages once more severe brain dysfunction, particularly dysfunction in frontal areas, has occurred (35). Although the type and onset of neural degradation that occurs in the frontal lobes will vary by type of dementia (47), at later stages of the disease, these degradations may uniformly result in the failure to downregulate autonomic arousal in response to uncertainty (37). Taken together, the available evidence suggests that anxiety and agitation are independent but related constructs whose propensity will be influenced as a function of neuropsychiatric disease progression. It stands to reason that the two types of neuropsychiatric symptoms may benefit from similar types of intervention. In recent years, music has emerged as a particularly important intervention for neuropsychiatric symptoms, especially with respect to anxiety and agitation. When used in this context, music may be regarded as a fundamental technology that can be personalized and systematically leveraged to downregulate autonomic arousal arising from uncertainty.

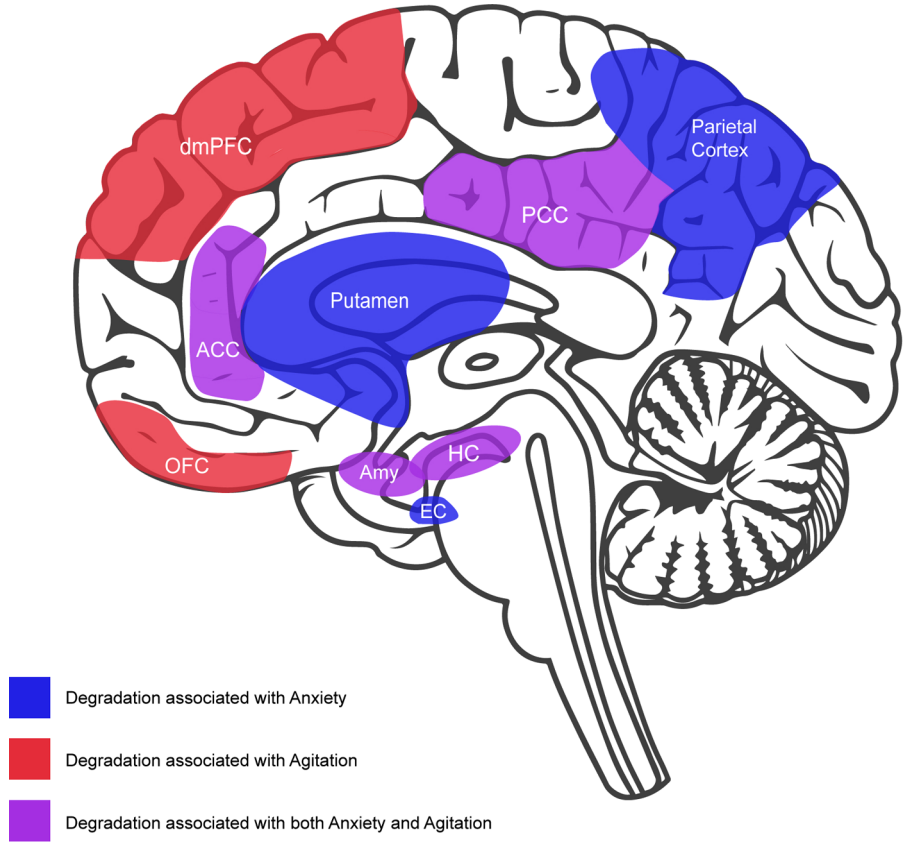


FIGURE 1
Midsaggital view of brain featuring neural degradation that has been associated with anxiety and agitation in dementia: anxiety (blue) = putamen, parietal cortex, and entorhinal cortex (EC); agitation (red) = orbitorfrontal cortex (OFC); dorsolateral prefrontal cortex (dmPFC); anxiety and agitation (purple) = anterior cingulate cortex (ACC), posterior cingulate cortex (PCC), amygdala (Amy), hippocampus (HC), and insula (not depicted in this view).

Are music-based interventions indicated for the treatment of anxiety and agitation symptoms related to dementia?

Qualitative research that has examined the impact of music-based interventions on people living with dementia has revealed significant improvements in quality of life in patients and caregivers. Notable benefits in patients include increased social engagement and reductions in anxiety and agitation (48–54). To further understand these benefits and for whom they may accrue, we conducted a literature search focusing on randomized controlled trials (RCTs) that investigated the use of music-based treatments on anxiety or agitation symptoms in participants with MCI and/or dementia. Of the 15 RCTs found, three RCTs were excluded due to low fidelity (55) and no personalization of music/music therapy (56, 57). Of the remaining 13 RCTs, 12 of them reported a significant reduction in anxiety and/or agitation in the music/music therapy treatment arm (**Table 1**). The most parsimonious interpretation regarding the mechanism of action is a downregulation of autonomic arousal, owing to a shift in balance from sympathetic to parasympathetic activity over the course of music listening (70). Overall, the effect sizes (Cohen's D) trended larger in studies that recruited participants with mild to moderate dementia. It is also notable that the one study that failed to find a significant reduction in anxiety and agitation was limited to participants with severe dementia (64). Thus, based on the available evidence we may conclude that agitation and anxiety symptoms are well-indicated for music-based treatment, especially in participants with mild to moderate dementia.

Current standard of care with respect to therapeutic uses of music

Accumulating evidence demonstrates that people living with dementia enjoy music and may benefit from making music, interacting with music through movement, and passive listening to music (71, 72). It has been shown that people living with AD can retain memory for melodies and lyrics (73), and that when activated, these memories facilitate the retrieval of autobiographical memories (74). The mechanisms underlying these music and memory phenomena are not completely understood but appear to depend on the encoding of musical memories in structures and networks that are resilient to neural degeneration (75). It seems likely that these music and memory phenomena depend in part on dopaminergic activity in the ventral striatum triggered by auditory-reward network connectivity, which is well preserved in MCI but less so in AD (76).

Music and memory phenomena have sparked widespread interest in popular culture, particularly following the release of the 2014 documentary film *Alive Inside*. The film tells the story of MUSIC & MEMORY, a non-profit that facilitates the use of music players for people living with dementia in long-term care homes. In an often-cited highlight of the film, a long-term care home resident named Henry is jolted out of his catatonic state into a charming, articulate, and engaged lover of music. The film also chronicles the struggle that founder Dan Cohen had in convincing healthcare professionals and administrators about the value of music in care for people living with dementia. MUSIC & MEMORY program has since expanded and has been imitated with variations all over the world and validated in a number of large-scale trials (77, 78).

In community and long-term care homes, therapeutic music can be observed in various modalities from music listening to support mood and reminiscence (79), singing to promote health and social wellbeing (80), and music-making coordinated by a licensed music therapist (81). Regardless of the modality, a scoping review of the literature suggests that personalized music is systematically more effective than non-personalized music (82, 83). Personalization is likely important because familiar and preferred music leads to greater activation of dopaminergic and opioid pathways in the ventral striatum than nonfamiliar music (84, 85), and in the case of AD this reward activation is also associated with greater functional connectivity in corticocortical and corticocerebellar networks (86). Personalization is further supported by the fact that people with AD have dysfunctional dopaminergic (87) and opioid transmission (88) and stimulating production of endogenous opioids through music (85) or other means may have a beneficial impact on anxiety and agitation over the near-term and potentially slowing the progression of disease over the long-term (88).

Because personalization is so important to the effectiveness of music it stands to reason that a limiting factor in scalability of any effective music program will be the time and effort required to personalize music for a given individual. This may be especially challenging when the caregiver has limited experience with the person living with dementia and/or the individual has limited communication abilities. A licensed music therapist would be able to cultivate some level of personalization through careful interaction and observation with an individual. However, there are barriers to accessing music therapists, which limits the benefits that may be obtained from music engagement. A music-based DTx will help bridge the gap by offering the level of personalization required for optimal outcomes in the absence of a licensed music therapist.

The music based DTx is not intended to diminish or replace the benefit that trained music therapists may have in both group and individual music therapy sessions. It cannot augment or alter the benefit patients receive from the therapeutic "relationship" that results from co-engagement of both passive and active music interventions, whether those occur in the presence of music therapists or caregivers without this training. It may, however, increase the opportunities for caregivers of all types to advance the therapeutic relationship by coming to learn about their patients in new ways. Lastly, digital interventions that capture continuous physiological data reflecting patient experience with music (e.g., heart-rate variability, pupillometry) may empower caregivers with a level of insight about response to music that would not otherwise be possible, which may inform all manner of music-based interventions including traditional music therapy as well as multi-modal interventions involving music such as augmented reality applications.

To this end, the user interface of the DTx will be designed in a manner that is conducive to operation by either a caregiver in the community (e.g., family member) or a professional in a long-term care home (e.g., music or recreation therapist). There will be no

TABLE 1 Overview of dementia and music therapy randomized control trials.

Study	Conditions	Music Intervention	Dementia Stage	Outcome Measures	Main Findings	Effect Size (Cohen's D)
(58)	Music vs. Passive Control (standard care)	Songs familiar to participants were used. 30 min twice a week for 6 weeks.	Diagnosed with dementia but able to follow simple instructions (assume mild to moderate)	Rate of Anxiety in Dementia Scale (RAID), Cohen-Mansfield Agitation Inventory (CMAI)	Decreased anxiety; no difference in agitation	0.90
(59)	Music Therapy (MT) vs. Active Control (reading)	Individualized MT method used. Music chosen according to patients' personal tastes.	Mild to moderate AD type dementia	Hamilton Anxiety Scale, Geriatric Depression Scale (GDS)	Decreased anxiety; decreased depression.	1.23–2.48 for anxiety.
(60)	Music vs. Active Control (multi-sensory stimulation)	Music therapist selected music for each patient considering patients' musical tastes	Severe dementia	CMAI, RAID, Cornell Scale for Depression in Dementia (CSDD)	Both groups had a decrease in anxiety; no effect for either group on agitation.	N/A
(61)	Music therapy (MT) vs. Active Control (recreation activities)	Music therapist selected music to incite pleasant memories and reduce agitation based on musical parameters (slow tempo etc.) (not personalized)	DSM-IV diagnosis of dementia with high level of behavioural problems indicated by Cohen-Mansfield Agitation Inventory >44; broad range but predominantly moderate to severe	CMAI	Both interventions showed a decrease in agitation; but no statistical difference between the treatment and control groups	N/A
(62)	Music Intervention vs. Passive Control (standard care)	Popular music was selected from the time of patients' youth. Not individually personalized.	Diagnosed with dementia DSM-IV and older than 65; broad range but predominantly moderate	C-CMAI	Reduction in agitated behaviour, physically aggressive behaviour	N/A
(63)	Music listening vs. Controls (singing vs. standard care)	Used familiar songs for music listening and singing groups not individually personalized to participants	Those unable to complete study measures were excluded (assume mild to moderate)	Mini-Mental State Exam (MMSE), Cornell-Brown Scale for Quality of Life (CBS)	Music listening and singing groups had reduction in behavioural symptoms including agitation	N/A
(64)	Music Therapy (MT) vs. Controls (individualized music listening vs. standard care)	MT was based on interactions (verbal/instrumental) between the patient and therapist. IML contained music selections made by music therapist based on interviews with patient's caregivers and the patient.	Moderate to severe	MMSE, Neuropsychiatric Inventory (NPI), CBS, CSDD	Overall behavioural assessment did not show significant differences between treatment and controls. Significant improvement over time in anxiety subscale for all groups. No effect on agitation.	N/A
(65)	Music Therapy (MT) vs. Passive Control (standard care)	MT was based on interactions (verbal/instrumental) between the patient and therapist. Well known songs were used in addition to improvisation.	Display at least two neuropsychiatric symptoms but no significant health problems (assume mild to moderate)	NPI-NH, dementia care mapping (DCM)	Caregivers reported an improvement in managing anxiety and other dementia symptoms. Increased observed well-being in MT group, reduced neuropsychiatric symptoms in MT group.	2.32–2.69
(66)	Music Therapy (MT) vs. Active Control (recreational activities)	Unknown level of personalization. Residents actively engage in music making/singing or listening to music that the music therapist plays or sings.	Nursing home residents with diagnosis of dementia (not enough information to infer stage)	NPI-Q	Agitation and other dementia behaviours were significantly reduced in MT group according to NPI-Q.	N/A
(67)	Music Therapy (MT) vs. Active Control (recreational activities)	MT was based on interactions (verbal/instrumental) between the patient and therapist.	Moderate to severe	NPI, MMSE	Agitation, anxiety, and other dementia behaviours/symptoms were significantly improved according to NPI total score.	0.53–1.8
(68)	Music Therapy (MT) vs. Active Control (recreational activities)	MT based on non-verbal model and on sound-music improvisation using musical instruments.	Severe	NPI, MMSE, Barthel Index (BI)	Agitations, delusions and apathy significantly improved in Music therapy group but not in control group.	0.63
(69)	Music Therapy (MT) vs. Passive Control (standard care)	Individual music therapy: vocal or instrumental improvising, singing, dancing, or listening to familiar or unknown songs	Moderate to severe	CMAI, ADRQL	Agitation decreased significantly during music therapy compared to control.	0.50

10.3389/fdgth.2023.1064115

expectations imposed regarding caregivers' level of experience with music, nor will there be any requirement of familiarity with the music preferences of the person they are caring for.

Adaptation of a music-based DTx for managing anxiety

Our team has recently published an RCT study on the efficacy of a music-based DTx developed by LUCID (https://www.lucidtherapeutics.com) for the treatment of anxiety (89). The system incorporates an AI called Affective Music Recommendation System (AMRS) (90); based on the iso-principle from music therapy (90, 91), which is a form of personalization that is independent of experience or preference. This approach to mood regulation suggests that the mood regulating properties of music may be enhanced if the mood of the music approximates an individual's initial emotional state before it is changed to the target state (92). The iso principle has been indicated in prior research to be more effective than other musical sequences at reducing tension (93). To develop AMRS, it was necessary to begin by curating training data that labeled music with respect to the arousal and valence dimensions of the Russell's Circumplex model of emotion (94). In this model, arousal refers to the activation aspect of felt emotion, ranging from calm to excited, and valence refers to the hedonic aspect of felt emotion, ranging from pleasant to unpleasant.

In LUCID's existing DTx (VIBE), the participant is asked to input their current mood using a 2-dimensional grid representing arousal and valence dimensions. Based on this input and the user's target emotional state (e.g., calm), the machine learning algorithm within the application predicts the optimal sequence of tracks to produce mood induction in the listener from their current emotional state to the target state. This machine learning algorithm uses reinforcement learning techniques and is trained on real-world data correlating the quantitative features of musical excerpts and sequences alongside the emotional responses induced by them in listeners.

Developing a music-based DTx for people living with dementia

Our existing DTx (VIBE) was designed to reduce anxiety. This objective was realized through the interaction of two AI systems: BioMIR (Biological Music Information Retrieval) and AMRS (as described above). The BioMIR system extracts insights about the emotional states that are likely to be evoked by pieces of music in people living with dementia. The AMRS considers the information from the BioMIR system to generate playlists that are personalized to each user. People living with dementia often have a diminished ability to attribute mental states to music, inclusive of emotion (95). For example, a musical excerpt that may calm a healthy person down may have a different effect on a person living with dementia. Therefore, to aid in the development of a new AI for affective music recommendation in this population (AMRS-D), we had to start by obtaining training data from older participants experiencing cognitive decline. To that end, we recently completed a training database study in collaboration with the Centre for Elder Research at Sheridan College with 32 participants living with MCI or early-stage dementia.

Participants were asked to listen to and make self-report judgments of valence, arousal, and absorption. Absorption was defined as the extent to which attentional resources were allocated to the music while listening (96, 97). It is our expectation that higher levels of musical absorption will lead to increased potency of a music-based intervention for mood regulation. While this hypothesis has not yet been validated with respect to state absorption in people living with dementia, it has been validated for trait absorption in a young adult population (see (98, 99)). These subjective judgments were collected alongside a variety of biometric measures which will allow us to use industry-standard machine learning methods to develop a fully closed-loop music recommendation system that can be driven by physiological data alone independent of user input (100). Going forward, the effectiveness of the DTx may be further enhanced by embedding beat stimulation in the theta range (90) with the expectation of increasing the extent of reduction in anxiety and agitation (89).

After the development of AMRS-D and the absorption module, we will begin an exploratory trial to assess the useability of the new system, including early indications of safety and efficacy. The exploratory trial will recruit participants living with dementia in the community by way of their caregivers. The DTx will be used by caregivers on a scheduled daily basis rather than in response to anxiety or agitation in the patient. The exploratory trial will provide an opportunity to solicit qualitative feedback that will lead to product refinement and pave the way for a future clinical proof-of-concept study. In this future proof-of-concept study, adherence and clinically relevant measures of efficacy will be tracked including the Neuropsychiatric Inventory Questionnaire (NPI-Q (101);, Behavioral Pathology in Alzheimer's Disease Rating Scale (BEHAVE-AD) (102), State-Trait Inventory of Cognitive and Somatic Anxiety (STICSA) (103) and the Cohen-Mansfield Agitation Inventory (104).

Discussion

In this paper we have outlined the rationale for a new music-based DTx that LUCID is developing to support the neuropsychiatric symptoms of dementia. This development represents an expansion of our prior work in adult anxiety (89) to help support a related indication in the context of persons living with dementia (i.e., agitation). We have argued that a music based DTx will be effective in mitigating both anxiety and agitation in this population. We envision our music-based intervention as having efficacy at all stages of disease but with the focus of benefits on anxiety in the early stages, and in agitation in the later stages. The DTx will be proactive rather than reactive. Scheduled daily use of the system is projected to lead to improvements in patient and caregiver outcomes and reduced costs of care. The DTx will respect individual preference for music through a personalization module that will eventually be implementable in the absence of caregiver input. The available evidence suggests that patients will find intrinsic benefits in listening to familiar music on its own, independent of

the anticipated mood-regulating properties emphasized through our survey of key opinion leaders. While the primary outcome of our research on the efficacy of the DTx in our future proof of concept study will be mitigation of anxiety and agitation, the personalization module is expected to lead to dopaminergic and opioid activity in the reward system *via* auditory-reward network connectivity (76, 85). We envision that this music based DTx will be used and implemented by healthcare professionals or family caregivers. Special attention will be devoted to overcoming tensions that may arise due to onboarding or protocol adherence. This approach to development is expected to yield direct benefits for patients, while reducing caregiver burden and the escalating costs associated with the greying of the world.

Data availability statement

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

Author contributions

FAR and AM conceptualized the manuscript and wrote the original draft. ZT and AD were responsible for the administration and formal analysis of the survey with key opinion leaders. KD, ZT, AD, and DC all made contributions to the conceptualization of the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

We thank KH for her contributions towards creating the **Figure 1** of the manuscript. We also would like to thank Rhiannon Ueberholz for her help in proofreading the manuscript.

Conflict of interest

FR has served as an advisor for LUCID since 2018 and as Chief Science Officer since 2021. He has been granted stock options, which may qualify him to financially benefit from commercial applications of the technology considered here. AM, ZT and AR are full-time employees of LUCID and have also been granted stock options, which may qualify them to financially benefit from commercial applications of the technology considered here. 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.

Publisher's note

References

1. Organization WH. Dementia (2022) [updated 2022–09–20; cited 2022 2022–09–24]. Available at: https://www.who.int/news-room/fact-sheets/detail/dementia

2. Reinhard SC, Given B, Petlick NH, Bemis A. Supporting family caregivers in providing care. *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*. (2008.

3. Pedroza P, Miller-Petrie MK, Chen C, Chakrabarti S, Chapin A, Hay S, et al. Global and regional spending on dementia care from 2000 to 2019 and expected future health spending scenarios from 2020 to 2050: an economic modelling exercise. *eClinMed*. (2022) 45:101337. doi: 10.1016/j.eclinm.2022.101337

4. Litz BT, Kerig PK. Introduction to the special issue on moral injury: conceptual challenges, methodological issues, and clinical applications. *J Trauma Stress*. (2019) 32 (3):341–9. doi: 10.1002/jts.22405

5. Woods B, Aguirre E, Spector AE, Orrell M. Cognitive stimulation to improve cognitive functioning in people with dementia. *Cochrane Database Syst Rev*. (2012) 2:1–53. doi: 10.1002/14651858.CD005562.pub2

6. Lazar A, Thompson H, Demiris G. A systematic review of the use of technology for reminiscence therapy. *Health Educ Behav*. (2014) 41(1_suppl):51S–61S. doi: 10.1177/1090198114537067

7. Tichko P, Kim JC, Large E, Loui P. Integrating music-based interventions with gamma-frequency stimulation: implications for healthy ageing. *Eur J Neurosci*. (2022) 55(11–12):3303–23. doi: 10.1111/ejn.15059

8. Sedghizadeh MJ, Hasani H, Lahijanian M, Aghajan H, Vahabi Z. Entrainment of gamma oscillations by auditory chirp stimulation in Alzheimer's Disease patients. *Alzheimer's & Dementia*. (2020) 16(S5):e043198. doi: 10.1002/alz.043198

9. Abbadessa G, Brigo F, Clerico M, De Mercanti S, Trojsi F, Tedeschi G, et al. Digital therapeutics in neurology. *J Neurol*. (2022) 269(3):1209–24. doi: 10.1007/s00415-021-10608-4

10. Koumakis L, Chatzaki C, Kazantzaki E, Maniadi E, Tsiknakis M. Dementia care frameworks and assistive technologies for their implementation: a review. *IEEE Rev Biomed Eng*. (2019) 12:4–18. doi: 10.1109/RBME.2019.2892614

11. Lanctôt KL, Amatniek J, Ancoli-Israel S, Arnold SE, Ballard C, Cohen-Mansfield J, et al. Neuropsychiatric signs and symptoms of Alzheimer's Disease: new treatment paradigms. *Alzheimer's & Dementia: Transl Res & Clin Interv*. (2017) 3(3):440–9. doi: 10.1016/j.trci.2017.07.001

12. González-Salvador T, Lyketsos CG, Baker A, Hovanec L, Roques C, Brandt J, et al. Quality of life in dementia patients in long-term care. *Int J Geriatr Psychiatry*. (2000) 15(2):181–9. doi: 10.1002/(SICI)1099-1166(200002)15:2<181::AID-GPS96>3.0.CO;2-I

13. Peters ME, Schwartz S, Han D, Rabins PV, Steinberg M, Tschanz JT, et al. Neuropsychiatric symptoms as predictors of progression to severe Alzheimer's Dementia and death: the cache county dementia progression study. *Am J Psychiatry*. (2015) 172(5):460–5. doi: 10.1176/appi.ajp.2014.14040480

14. Rosenberg PB, Mielke MM, Appleby BS, Oh ES, Geda YE, Lyketsos CG. The association of neuropsychiatric symptoms in mci with incident dementia and Alzheimer disease. *Am J Geriatr Psychiatry*. (2013) 21(7):685–95. doi: 10.1016/j.jagp.2013.01.006

15. Lyketsos CG, Lopez O, Jones B, Fitzpatrick AL, Breitner J, DeKosky S. Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairmentresults from the cardiovascular health study. *JAMA*. (2002) 288(12):1475–83. doi: 10.1001/jama.288.12.1475

16. Carrarini C, Russo M, Dono F, Barbone F, Rispoli MG, Ferri L, et al. Agitation and dementia: prevention and treatment strategies in acute and chronic conditions. *Front Neurol*. (2021) 12:1–18. doi: 10.3389/fneur.2021.644317

17. Van der Mussele S, Le Bastard N, Saerens J, Somers N, Mariën P, Goeman J, et al. Agitation-Associated behavioral symptoms in mild cognitive impairment and Alzheimer's Dementia. *Aging Ment Health*. (2015) 19(3):247–57. doi: 10.1080/13607863.2014.924900

18. Fauth EB, Gibbons A. Which behavioral and psychological symptoms of dementia are the most problematic? Variability by prevalence, intensity, distress ratings, and associations with caregiver depressive symptoms. *Int J Geriatr Psychiatry*. (2014) 29 (3):263–71. doi: 10.1002/gps.4002

19. Hiyoshi-Taniguchi K, Becker CB, Kinoshita A. What behavioral and psychological symptoms of dementia affect caregiver burnout? *Clin Gerontol*. (2018) 41(3):249–54. doi: 10.1080/07317115.2017.1398797

20. Spector A, Orrell M. Using a biopsychosocial model of dementia as a tool to guide clinical practice. *Int Psychogeriatr*. (2010) 22(6):957–65. doi: 10.1017/S1041610210000840

21. Lupien SJ, Juster R-P, Raymond C, Marin M-F. The effects of chronic stress on the human brain: from neurotoxicity, to vulnerability, to opportunity. *Front Neuroendocrinol*. (2018) 49:91–105. doi: 10.1016/j.yfrne.2018.02.001

22. Wolf A, Leucht S, Pajonk F-G. Do antipsychotics lead to cognitive impairment in dementia? A meta-analysis of randomised placebo-controlled trials. *Eur Arch Psychiatry Clin Neurosci*. (2017) 267(3):187–98. doi: 10.1007/s00406-016-0723-4

23. Corbett A, Burns A, Ballard C. Don't use antipsychotics routinely to treat agitation and aggression in people with dementia. *BMJ: British Medical Journal*. (2014) 349:g6420. doi: 10.1136/bmj.g6420

24. Rainer MK, Masching AJ, Ertl MG, Kraxberger E, Haushofer M. Effect of risperidone on behavioral and psychological symptoms and cognitive function in dementia. *J Clin Psychiatry*. (2001) 62(11):894–900. doi: 10.4088/JCP.v62n1110

25. Yunusa I, El Helou ML. The use of risperidone in behavioral and psychological symptoms of dementia: a review of pharmacology, clinical evidence, regulatory approvals, and off-label use. *Front Pharmacol*. (2020) 11:1–7. doi: 10.3389/fphar.2020.00596

26. Shin J-Y, Choi N-K, Jung S-Y, Lee J, Kwon JS, Park B-J. Risk of ischemic stroke with the use of risperidone, quetiapine and olanzapine in elderly patients: a population-based, case-crossover study. *J Psychopharmacol*. (2013) 27(7):638–44. doi: 10.1177/0269881113482530

27. Deberdt WG, Siegal A, Ahl J, Meyers AL, Landbloom R. Effect of olanzapine on cognition during treatment of behavioral and psychiatric symptoms in patients with dementia: a post-hoc analysis. *Int J Geriatr Psychiatry*. (2008) 23(4):364–9. doi: 10.1002/gps.1885

28. Suh G-H, Greenspan AJ, Choi S-K. Comparative efficacy of risperidone versus haloperidol on behavioural and psychological symptoms of dementia. *Int J Geriatr Psychiatry*. (2006) 21(7):654–60. doi: 10.1002/gps.1542

29. Maust DT, Kim HM, Seyfried LS, Chiang C, Kavanagh J, Schneider LS, et al. Antipsychotics, other psychotropics, and the risk of death in patients with dementia: number needed to harm. *JAMA Psychiatry*. (2015) 72(5):438–45. doi: 10.1001/jamapsychiatry.2014.3018

30. Ringman JM, Schneider L. Treatment options for agitation in dementia. *Curr Treat Options Neurol*. (2019) 21(7):30. doi: 10.1007/s11940-019-0572-3

31. Mintzer JE, Brawman-Mintzer O. Agitation as a possible expression of generalized anxiety disorder in demented elderly patients: toward a treatment approach. *J Clin Psychiatry*. (1996) 57(Suppl 7):55–63.; discussion 73-5.

32. Twelftree H, Qazi A. Relationship between anxiety and agitation in dementia. *Aging Ment Health*. (2006) 10(4):362–7. doi: 10.1080/13607860600638511

33. Palmer K, Berger AK, Monastero R, Winblad B, Bäckman L, Fratiglioni L. Predictors of progression from mild cognitive impairment to Alzheimer disease. *Neurology*. (2007) 68(19):1596–602. doi: 10.1212/01.wnl.0000260968.92345.3f

34. 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 (3):313–32. doi: 10.1017/S1352465808004232

35. Bierman EJM, Comijs HC, Jonker C, Beekman ATF. Symptoms of anxiety and depression in the course of cognitive decline. *Dement Geriatr Cogn Disord*. (2007) 24 (3):213–9. doi: 10.1159/000107083

36. Liu KY, Costello H, Reeves S, Howard R. The relationship between anxiety and incident agitation in Alzheimer's Disease. *J Alzheimer's Dis*. (2020) 78:1119–27. doi: 10.3233/JAD-200516

37. Grupe DW, Nitschke JB. Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat Rev Neurosci*. (2013) 14 (7):488–501. doi: 10.1038/nrn3524

38. Chen Y, Dang M, Zhang Z. Brain mechanisms underlying neuropsychiatric symptoms in Alzheimer's Disease: a systematic review of symptom-general and –specific lesion patterns. *Mol Neurodegener*. (2021) 16(1):38. doi: 10.1186/s13024-021-00456-1

39. Tagai K, Nagata T, Shinagawa S, Nemoto K, Inamura K, Tsuno N, et al. Correlation between both morphologic and functional changes and anxiety in Alzheimer's Disease. *Dement Geriatr Cogn Disord*. (2014) 38(3-4):153–60. doi: 10.1159/000358822

40. Nour M, Jiao Y, Teng G-J. Neuroanatomical associations of depression, anxiety and apathy neuropsychiatric symptoms in patients with Alzheimer's Disease. *Acta Neurol Belg*. (2021) 121(6):1469–80. doi: 10.1007/s13760-020-01349-8

41. Samsi K, Manthorpe J. Care pathways for dementia: current perspectives. *Clin Interv Aging*. (2014) 9:2055–63. doi: 10.2147/cia.S70628

42. Poulin SP, Dautoff R, Morris JC, Barrett LF, Dickerson BC. Amygdala atrophy is prominent in early Alzheimer's Disease and relates to symptom severity. *Psychiatry Res: Neuroimaging*. (2011) 194(1):7–13. doi: 10.1016/j.pscychresns.2011.06.014

43. Mah L, Binns MA, Steffens DC. Anxiety symptoms in amnestic mild cognitive impairment are associated with medial temporal atrophy and predict conversion to Alzheimer disease. *Am J Geriatr Psychiatry*. (2015) 23(5):466–76. doi: 10.1016/j.jagp.2014.10.005

44. Hashimoto H, Monseratt L, Nguyen P, Feil D, Harwood D, Mandelkern MA, et al. Anxiety and regional cortical glucose metabolism in patients with Alzheimer's Disease. *J Neuropsychiatry Clin Neurosci*. (2006) 18(4):521–8. doi: 10.1176/jnp.2006.18.4.521

45. Bensamoun D, Guignard R, Furst AJ, Derreumaux A, Manera V, Darcourt J, et al. Associations between neuropsychiatric symptoms and cerebral amyloid deposition in cognitively impaired elderly people. *J Alzheimer's Dis*. (2016) 49:387–98. doi: 10.3233/JAD-150181

46. Rosenberg PB, Nowrangi MA, Lyketsos CG. Neuropsychiatric symptoms in Alzheimer's Disease: what might be associated brain circuits? *Mol Asp Med*. (2015) 43-44:25–37. doi: 10.1016/j.mam.2015.05.005

47. Murtha S, Cismaru R, Waechter R, Chertkow H. Increased variability accompanies frontal lobe damage in dementia. *J Int Neuropsychol Soc*. (2002) 8(3):360–72. doi: 10.1017/S1355617702813170

48. Lipe AW. Using music therapy to enhance the quality of life in a client with Alzheimer's Dementia: a case study. *Music Ther Perspect*. (1991) 9(1):102–5. doi: 10.1093/mtp/9.1.102

49. Olderog Millard KA, Smith JM. The influence of group singing therapy on the behavior of Alzheimer's Disease patients. *J Music Ther*. (1989) 26(2):58–70. doi: 10.1093/jmt/26.2.58

50. Shively C, Henkin L. Music and movement therapy with Alzheimer's Victims. *Music Ther Perspect*. (1986) 3(1):56–8. doi: 10.1093/mtp/3.1.56

51. Murphy K, Liu WW, Goltz D, Fixsen E, Kirchner S, Hu J, et al. Implementation of personalized music listening for assisted living residents with dementia. *Geriatr Nurs (Minneap)*. (2018) 39(5):560–5. doi: 10.1016/j.gerinurse.2018.04.001

52. Sorrell JM. Music as a healing art in dementia care. *J Psychosoc Nurs Ment Health Serv*. (2018) 56(7):15–8. doi: 10.3928/02793695-20180619-04

53. Yao C-T, Lee B-O, Hong H, Su Y-C. Evaluation of the music therapy program interventions on agitated behavior for people with dementia in Taiwan institutional care. *Educ Gerontol*. (2022):1–12. doi: 10.1080/03601277.2022.2099076

54. Dahms R, Eicher C, Haesner M, Mueller-Werdan U. Influence of music therapy and music-based interventions on dementia: a pilot study. *J Music Ther*. (2021) 58(3): e12–36. doi: 10.1093/jmt/thab005

55. Kwak J, Anderson K, O'Connell Valuch K. Findings from a prospective randomized controlled trial of an individualized music listening program for persons with dementia. *J Appl Gerontol*. (2020) 39(6):567–75. doi: 10.1177/0733464818778991

56. Ceccato E, Vigato G, Bonetto C, Bevilacqua A, Pizziolo P, Crociani S, et al. Stam protocol in dementia:a multicenter, single-blind, randomized, and controlled trial. *Am J Alzheimer's Dis & Other Demen®*. (2012) 27(5):301–10. doi: 10.1177/1533317512452038

57. Cooke ML, Moyle W, Shum DHK, Harrison SD, Murfield JE. A randomized controlled trial exploring the effect of music on agitated behaviours and anxiety in older people with dementia. *Aging Ment Health*. (2010) 14(8):905–16. doi: 10.1080/13607861003713190

58. Sung H-C, Lee W-L, Li T-L, Watson R. A group music intervention using percussion instruments with familiar music to reduce anxiety and agitation of institutionalized older adults with dementia. *Int J Geriatr Psychiatry*. (2012) 27 (6):621–7. doi: 10.1002/gps.2761

59. Guétin S, Portet F, Picot MC, Pommié C, Messaoudi M, Djabelkir L, et al. Effect of music therapy on anxiety and depression in patients with Alzheimer's Type dementia: randomised, controlled study. *Dement Geriatr Cogn Disord*. (2009) 28(1):36–46. doi: 10.1159/000229024

60. Sánchez A, Maseda A, Marante-Moar MP, de Labra C, Lorenzo-López L, Millán-Calenti JC. Comparing the effects of multisensory stimulation and individualized music sessions on elderly people with severe dementia: a randomized controlled trial. *J Alzheimer's Dis*. (2016) 52:303–15. doi: 10.3233/JAD-151150

61. Vink AC, Zuidersma M, Boersma F, de Jonge P, Zuidema SU, Slaets JPJ. The effect of music therapy compared with general recreational activities in reducing agitation in people with dementia: a randomised controlled trial. *Int J Geriatr Psychiatry*. (2013) 28(10):1031–8. doi: 10.1002/gps.3924

62. Lin Y, Chu H, Yang C-Y, Chen C-H, Chen S-G, Chang H-J, et al. Effectiveness of group music intervention against agitated behavior in elderly persons with dementia. *Int J Geriatr Psychiatry*. (2011) 26(7):670–8. doi: 10.1002/gps.2580

63. Särkämö T, Laitinen S, Numminen A, Kurki M, Johnson JK, Rantanen P. Pattern of emotional benefits induced by regular singing and music listening in dementia. *J Am Geriatr Soc*. (2016) 64(2):439–40. doi: 10.1111/jgs.13963

64. Raglio A, Bellandi D, Baiardi P, Gianotti M, Ubezio MC, Zanacchi E, et al. Effect of active music therapy and individualized listening to music on dementia: a multicenter randomized controlled trial. *J Am Geriatr Soc*. (2015) 63(8):1534–9. doi: 10.1111/jgs.13558

65. Hsu MH, Flowerdew R, Parker M, Fachner J, Odell-Miller H. Individual music therapy for managing neuropsychiatric symptoms for people with dementia and their carers: a cluster randomised controlled feasibility study. *BMC Geriatr*. (2015) 15(1):84. doi: 10.1186/s12877-015-0082-4

66. Vink AC, Zuidersma M, Boersma F, de Jonge P, Zuidema SU, Slaets JP. Effect of music therapy versus recreational activities on neuropsychiatric symptoms in elderly adults with dementia: an exploratory randomized controlled trial. *J Am Geriatr Soc*. (2014) 62(2):392–3. doi: 10.1111/jgs.12682

67. Raglio A, Bellelli G, Traficante D, Gianotti M, Ubezio MC, Villani D, et al. Efficacy of music therapy in the treatment of behavioral and psychiatric symptoms of dementia. *Alzheimer Dis & Associated Disorders*. (2008) 22(2):158–62. doi: 10.1097/WAD.0b013e3181630b6f

68. Raglio A, Bellelli G, Traficante D, Gianotti M, Ubezio MC, Gentile S, et al. Efficacy of music therapy treatment based on cycles of sessions: a randomised controlled trial. *Aging Ment Health*. (2010) 14(8):900–4. doi: 10.1080/13607861003713158

69. Ridder HMO, Stige B, Qvale LG, Gold C. Individual music therapy for agitation in dementia: an exploratory randomized controlled trial. *Aging Ment Health*. (2013) 17(6):667–78. doi: 10.1080/13607863.2013.790926

70. Labbé E, Schmidt N, Babin J, Pharr M. Coping with stress: the effectiveness of different types of music. *Appl Psychophysiol Biofeedback*. (2007) 32(3):163–8. doi: 10.1007/s10484-007-9043-9

71. Baird A, Samson S. Chapter 11 - music and dementia. In: E Altenmüller, S Finger, F Boller, editors. *Progress in brain research*. 217. Amsterdam, Netherlands: Elsevier (2015). p. 207–35.

72. Sihvonen AJ, Särkämö T, Leo V, Tervaniemi M, Altenmüller E, Soinila S. Music-Based interventions in neurological rehabilitation. *The Lancet Neurol*. (2017) 16(8):648–60. doi: 10.1016/S1474-4422(17)30168-0

73. Cuddy LL, Duffin JM, Gill SS, Brown CL, Sikka R, Vanstone AD. Memory for melodies and lyrics in Alzheimer's Disease. *Music Percept*. (2012) 29(5):479–91. doi: 10.1525/mp.2012.29.5.479

74. Foster NA, Valentine ER. The effect of auditory stimulation on autobiographical recall in dementia. *Exp Aging Res*. (2001) 27(3):215–28. doi: 10.1080/03610730130020866 4

75. Peck KJ, Girard TA, Russo FA, Fiocco AJ. Music and memory in Alzheimer's Disease and the potential underlying mechanisms. *J Alzheimers Dis*. (2016) 51(4):949–59. doi: 10.3233/jad-150998

76. Wang D, Belden A, Hanser SB, Geddes MR, Loui P. Resting-State connectivity of auditory and reward systems in Alzheimer's Disease and mild cognitive impairment. *Front Hum Neurosci*. (2020) 14:1–10. doi: 10.3389/fnhum.2020.00280

77. Bakerjian D, Bettega K, Cachu AM, Azzis L, Taylor S. The impact of music and memory on resident level outcomes in California nursing homes. *J Am Med Dir Assoc*. (2020) 21(8):1045–50.e2. doi: 10.1016/j.jamda.2020.01.103

78. McCreedy EM, Yang X, Baier RR, Rudolph JL, Thomas KS, Mor V. Measuring effects of nondrug interventions on behaviors: music & memory pilot study. *J Am Geriatr Soc*. (2019) 67(10):2134–8. doi: 10.1111/jgs.16069

79. Cunningham S, Brill M, Whalley JH, Read R, Anderson G, Edwards S, et al. Assessing wellbeing in people living with dementia using reminiscence music with a Mobile app (memory tracks): a mixed methods cohort study. *J Healthc Eng*. (2019) 2019:8924273. doi: 10.1155/2019/8924273

80. Good A, Kreutz G, Choma B, Fiocco A, Russo F, Organization WH. The singwell project protocol: the road to understanding the benefits of group singing in older adults. *Public Health Panorama*. (2020) 6(1):141–6.

81. Rio R. A community-based music therapy support group for people with Alzheimer's Disease and their caregivers: a sustainable partnership model. *Front Med (Lausanne)*. (2018) 5:1–7. doi: 10.3389/fmed.2018.00293

82. Leggieri M, Thaut MH, Fornazzari L, Schweizer TA, Barfett J, Munoz DG, et al. Music intervention approaches for Alzheimer's Disease: a review of the literature. *Front Neurosci*. (2019) 13:1–8. doi: 10.3389/fnins.2019.00132

83. Tomaino CM. *Music on their minds: A qualitative study of the effects of using familiar music to stimulate preserved memory function in persons with dementia*. New York, New York, USA: New York University (1998).

84. Salimpoor VN, Benovoy M, Larcher K, Dagher A, Zatorre RJ. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nat Neurosci*. (2011) 14(2):257–62. doi: 10.1038/nn.2726

85. Mallik A, Chanda ML, Levitin DJ. Anhedonia to music and mu-opioids: evidence from the administration of naltrexone. *Sci Rep*. (2017) 7:41952. doi: 10.1038/srep41952

86. King JB, Jones KG, Goldberg E, Rollins M, MacNamee K, Moffit C, et al. Increased functional connectivity after listening to favored music in adults with Alzheimer dementia. *J Prev Alzheimer's Dis*. (2019) 6(1):56–62. doi: 10.14283/jpad.2018.19

87. Engelborghs S, Vloeberghs E, Le Bastard N, Van Buggenhout M, Mariën P, Somers N, et al. The dopaminergic neurotransmitter system is associated with aggression and agitation in frontotemporal dementia. *Neurochem Int*. (2008) 52(6):1052–60. doi: 10.1016/j.neuint.2007.10.018

88. Cai Z, Ratka A. Opioid system and Alzheimer's Disease. *NeuroMol Med*. (2012) 14(2):91–111. doi: 10.1007/s12017-012-8180-3

89. Mallik A, Russo FA. The effects of music & auditory beat stimulation on anxiety: a randomized clinical trial. *PLOS ONE*. (2022) 17(3):e0259312. doi: 10.1371/journal.pone.0259312

90. Labbé A, McMahon Z, Thomson Z. Music as Medicine: Lucid Science+Technology White Paper. [White Paper]. In press (2021)).

91. Altshuler IM. A Psychiatrist's Experience with music as a therapeutic agent. In: Schoen Schullian DM, M., editor. *Music and medicine*. New York: Schuman, Inc. (1948). p. 69–76.

92. Heiderscheit A, Madson A. Use of the iso principle as a central method in mood management: a music psychotherapy clinical case study. *Music Ther Perspect*. (2015) 33(1):45–52. doi: 10.1093/mtp/miu042%J

93. Rider MS. Entrainment mechanisms are involved in pain reduction, muscle relaxation, and music-mediated imagery. *J Music Ther*. (1985) 22(4):183–92. doi: 10.1093/jmt/22.4.183

94. Russell JA. A circumplex model of affect. *J Pers Soc Psychol*. (1980) 39(6):1161–78. doi: 10.1037/h0077714

95. Downey LE, Blezat A, Nicholas J, Omar R, Golden HL, Mahoney CJ, et al. Mentalising music in frontotemporal dementia. *Cortex*. (2013) 49(7):1844–55. doi: 10.1016/j.cortex.2012.09.011

96. Lange EB, Zweck F, Sinn P. Microsaccade-Rate indicates absorption by music listening. *Conscious Cogn*. (2017) 55:59–78. doi: 10.1016/j.concog.2017.07.009

97. Hall SE, Schubert E, Wilson SJ. The role of trait and state absorption in the enjoyment of music. *PLOS ONE*. (2016) 11(11):e0164029. doi: 10.1371/journal.pone.0164029

98. Sandstrom GM, Russo FA. Absorption in music: development of a scale to identify individuals with strong emotional responses to music. *Psychol Music*. (2013) 41(2):216–28. doi: 10.1177/0305735611422508

99. Dvorak AL, Hernandez-Ruiz E. Comparison of music stimuli to support mindfulness meditation. *Psychol Music*. (2021) 49(3):498–512. doi: 10.1177/0305735619878497

100. Labbé A, Russo FA. Inventors; Method and System for Measuring, Calibrating and Training Psychological Absorption. United States of America (2022)).

101. Kaufer DI, Cummings JL, Ketchel P, Smith V, MacMillan A, Shelley T, et al. Validation of the npi-Q, a brief clinical form of the neuropsychiatric inventory. *J Neuropsychiatry Clin Neurosci*. (2000) 12(2):233–9. doi: 10.1176/jnp.12.2.233

102. Reisberg B, Borenstein J, Franssen E, Salob S, Steinberg G, Shulman E, et al. Behave-Ad: a clinical rating scale for the assessment of pharmacologically remediable behavioral symptomatology in Alzheimer's Disease. In: HJ Altman, editor. *Alzheimer's disease: problems, prospects, and perspectives*. Boston, MA: Springer US (1987). p. 1–16.

103. Grös DF, Antony MM, Simms LJ, McCabe RE. Psychometric properties of the state-trait inventory for cognitive and somatic anxiety (sticsa): comparison to the state-trait anxiety inventory (stai). *Psychol Assess*. (2007) 19(4):369. doi: 10.1037/1040-3590.19.4.369

104. Cohen-Mansfield J, Marx MS, Rosenthal AS. A description of agitation in a nursing home. *J Gerontol*. (1989) 44(3):M77–84. doi: 10.1093/geronj/44.3.M77

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 23 February 2023
DOI 10.3389/fdgth.2023.1043806

OPEN ACCESS

EDITED BY
Louis N. Awad,
Boston University, United States

REVIEWED BY
Margaret French,
Johns Hopkins University, United States
Anne Hickey,
Royal College of Surgeons in Ireland, Ireland

*CORRESPONDENCE
David J. Lin
✉ dlin7@mgh.harvard.edu

SPECIALTY SECTION
This article was submitted to Personalized Medicine, a section of the journal Frontiers in Digital Health

RECEIVED 14 September 2022
ACCEPTED 19 January 2023
PUBLISHED 23 February 2023

CITATION
DiCarlo JA, Erler KS, Petrilli M, Emerson K, Gochyyev P, Schwamm LH and Lin DJ (2023) SMS-text messaging for collecting outcome measures after acute stroke.
Front. Digit. Health 5:1043806.
doi: 10.3389/fdgth.2023.1043806

SMS-text messaging for collecting outcome measures after acute stroke

Julie A. DiCarlo[1], Kimberly S. Erler[2], Marina Petrilli[2], Kristi Emerson[1], Perman Gochyyev[2], Lee H. Schwamm[1,3] and David J. Lin[1,2]*

[1]Department of Neurology, Massachusetts General Hospital, Boston, MA, United States, [2]School of Health and Rehabilitation Sciences, MGH Institute of Health Professions, Boston, MA, United States, [3]Digital Enterprise Service, Mass General Brigham, Somerville, MA, United States

Introduction: Traditional methods for obtaining outcomes for patients after acute stroke are resource-intensive. This study aimed to examine the feasibility, reliability, cost, and acceptability of collecting outcomes after acute stroke with a short message service (SMS)-text messaging program.
Methods: Patients were enrolled in an SMS-text messaging program at acute stroke hospitalization discharge. Participants were prompted to complete assessments including the modified Rankin scale (mRS) and Patient-Reported Outcomes Measurement (PROM) Information System Global-10 at 30, 60, and 90 days postdischarge *via* SMS-text. Agreement and cost of SMS-text data collection were compared to those obtained from traditional follow-up methods (*via* phone or in the clinic). Participant satisfaction was surveyed upon program conclusion.
Results: Of the 350 patients who agreed to receive SMS texts, 40.5% responded to one or more assessments. Assessment responders were more likely to have English listed as their preferred language ($p = 0.009$), have a shorter length of hospital stay ($p = 0.01$), lower NIH stroke scale upon admission ($p < 0.001$), and be discharged home ($p < 0.001$) as compared to nonresponders. Weighted Cohen's kappa revealed that the agreement between SMS texting and traditional methods was almost perfect for dichotomized (good vs. poor) ($\kappa = 0.8$) and ordinal levels of the mRS score ($\kappa = 0.8$). Polychoric correlations revealed a significant association for PROM scores ($\rho = 0.4$, $p < 0.01$ and $\rho = 0.4$, $p < 0.01$). A cost equation showed that gathering outcomes *via* SMS texting would be less costly than phone follow-up for cohorts with more than 181 patients. Nearly all participants (91%) found the program acceptable and not burdensome (94%), and most (53%) felt it was helpful. Poststroke outcome data collection *via* SMS texting is feasible, reliable, low-cost, and acceptable. Reliability was higher for functional outcomes as compared to PROMs.
Conclusions: While further validation is required, our findings suggest that SMS texting is a feasible method for gathering outcomes after stroke at scale to evaluate the efficacy of acute stroke treatments.

KEYWORDS

stroke, digital health, SMS-text, outcomes, recovery

Introduction

Stroke is a leading cause of acquired adult disability worldwide (1). Recent substantial advances in acute stroke treatments (2, 3) and novel approaches to stroke rehabilitation (4, 5) have resulted in significant improvements in poststroke outcomes. To systematically evaluate the real-world benefit of such interventions, it is essential to reliably collect outcomes for patients after acute stroke discharge.

Current approaches to outcomes data collection face many logistical barriers. Follow-up care, during which outcomes are traditionally collected, requires patients to return to specialized

stroke centers and can be time-intensive, cost-prohibitive, and burdensome, relying on interaction with trained healthcare providers (6, 7). Phone calls to stroke patients to assess outcomes are also time-consuming and require dedicated and trained staff.

In recent years, there has been rapid adoption of digital and telehealth approaches in clinical care (8–10). Since mobile phones are one of the most popular forms of digital interaction (11) and are ubiquitous even among diverse demographic groups (11, 12), there is the potential to utilize short message service (SMS) texting for gathering assessments after stroke. SMS-texting programs have been used in a range of health conditions (13–15) for varying utilities, including intervention (16, 17), adherence (13), and data collection (18). Although app-based collection of outcomes after stroke has been explored (19), the feasibility of using SMS texting has not yet been examined in stroke. This study aimed to examine the feasibility, reliability, cost, and acceptability of an SMS-texting approach to gather health outcomes in the first 90 days after acute stroke.

Materials and methods

Our health system has articulated a goal of collecting functional outcomes after acute stroke discharge on all patients but has lacked the resources to accomplish this. As part of a clinical quality improvement initiative to assess barriers to success, we sought to increase the likelihood of data collection by leveraging an SMS-text messaging-based program on all discharged acute stroke patients for a several-month period. Using the services offered by a digital health technology company [Philips Patient Navigation Manager (formerly Medumo), Boston, MA, United States], we developed and launched an SMS-text follow-up program for patients discharged from the Massachusetts General Hospital (MGH) with a stroke ICD-10 code (I63, I60, I61, and G45) between June 8, 2020, and February 1, 2021. Patients were eligible to participate in the program if they had a valid mobile phone contact number in their medical chart. Patients who had not previously consented to receive SMS texts from their clinical care team at MGH received one consent SMS-text message at the time of acute hospital discharge, which remained active (i.e., giving the option to consent) for the duration of the program. If they did not consent, they did not receive any further messages. Patients had the option to decline participation in the program by responding "STOP" or simply not responding to the consent message.

Patients who consented to receiving SMS texts were enrolled in the program at the time of discharge regardless of their discharge destination (home or facility) and were provided instructions for unsubscribing (**Supplementary Table S1**). To familiarize patients with the SMS-texting method of communication and optimize patient engagement, patients also received weekly brain health educational tips developed by a multidisciplinary panel of clinicians, including neurologists, dietitians, and therapists (**Supplementary Table S1**).

Enrolled patients received an SMS text at 30, 60, and 90 days after hospital discharge, prompting them to complete the simplified modified Rankin scale (mRS) (20), a single-item, seven-level, ordinal measure of global disability, and then the Patient-Reported Outcomes Measurement Information System (PROMIS) Global-10, a 10-item measure of physical health, mental health, social health, pain, fatigue, and overall perceived quality of life. Individual items from each assessment were sent *via* SMS-text messages one at a time (i.e., each question of each assessment was one text message). Participants responded by directly texting the number corresponding to the answer of their selection. Participants had 1 week from receiving the prompt to complete the assessment at each time point. Responses were automatically saved in a secure database. Participants who completed all questions associated with the mRS at any given time point were considered responders, while those that did not were considered nonresponders. Participants who completed the mRS but did not complete all questions associated with the PROMIS Global-10 were still considered responders but were not scored on this assessment. The mRS can be dichotomized into good (score 0–2) and poor (score 3–6) outcomes (21). Global Physical Health (GPH) and Global Mental Health (GMH) *z*-scores (mean: 50; standard deviation (SD): 10) were derived from two 4-item summary scores extracted from the PROMIS Global-10 questions (22). At the conclusion of the program, enrolled patients received a satisfaction survey *via* an SMS text with six questions that had multiple-choice response options. Participants were counted as responders to the satisfaction survey if they responded to at least one question.

To evaluate the reliability of outcome measure scores obtained *via* SMS texting, mRS and PROMIS Global-10 scores closest in time to SMS-text responses were extracted from documented traditional follow-up encounters (clinic visit or follow-up phone call), when available. The clinically documented score was compared to the score from the closest SMS-text response in time.

To compare the yield and cost of the SMS texting approach to gather outcomes after acute stroke discharge with traditional methods of ascertaining outcomes, we added clinical staff and utilized a trained coordinator to call all consecutive patients discharged with stroke during a 3-month period to obtain their outcomes approximately 90 days after hospitalization discharge. The mRS and PROMIS Global-10 scores were also assessed *via* phone calls in the same order as SMS texts. Three attempts were made to reach each patient. Call attempts and time lengths were documented. The results of this intervention were then used to compare the cost between the two strategies (SMS texts vs. phone calls) for gathering poststroke outcomes.

Statistical analysis

Participant characteristics were examined with mean and SD, median and interquartile range (IQR), or *n* (%). Independent sample *t*-tests and chi-squared tests of independence were performed to compare clinical and demographic characteristics between those enrolled in the program and those not enrolled and between those who responded to the assessment prompts (responders) and those who did not (nonresponders).

Weighted Cohen's kappa (23) was calculated to assess the agreement between SMS texts and clinician- or coordinator-gathered responses for the mRS, and polychoric correlations were calculated to assess the agreement between GPH and GMH scores

(subscales of PROMIS Global-10). For the mRS, we examined agreement using the ordinal level (with quadratic weights, Supplementary Table S3) and the dichotomized level (good vs. poor) outcomes.

Descriptive statistics were used to characterize the time, yield, and cost comparison of SMS texting vs. phone call-based methods of gathering outcomes and to examine the results of the satisfaction survey.

The Massachusetts General Brigham Institutional Review Board (#2021P001342) approved this study, which was exempt from written informed consent as the data extracted for this study were gathered under the standard of care through a quality improvement initiative. Data will be made available from the corresponding author upon reasonable request.

Results

Of the 530 patients discharged from MGH with a stroke ICD-10 code between June 8, 2020, and February 1, 2021, 350 patients (66.04%) were enrolled in the program. Patients enrolled on average 6.8 ± 8.2 (mean ± SD) days after acute stroke hospital admission. Patients were not enrolled if they did not have valid contact information in the medical chart ($n = 151$) or declined to receive messages ($n = 29$) (Figure 1). Enrolled patients were more likely to have English listed as their preferred language ($\chi^2 = 5.44$, $p = 0.02$), have a lower NIH stroke scale upon admission ($t = 5.0$, $p \leq 0.001$), and have been discharged directly home from the hospital ($\chi^2 = 4.20$, $p = 0.04$) compared to those patients who did not enroll (Table 1).

Of those who enrolled, 40.5% ($n = 142$) responded to at least one SMS text to complete an assessment. The response rate at 30-day postdischarge was 28.6%, and the response rates at the 60- and 90-day time points were 24.3%. Of the responders, 30% ($n = 42$) responded at any two time points and 30% ($n = 43$) of responders responded at all three time points (Figure 2A). SMS-text response compliance is presented in Figure 2B. Responders to message prompts were more likely to have English as their preferred language ($\chi^2 = 9.33$, $p = 0.009$), have shorter acute hospital length of stay ($t = 2.5$, $p = 0.01$), have a lower NIH stroke scale upon admission ($t = 3.98$, $p < 0.001$), and have been discharged directly home ($\chi^2 = 24.94$, $p < 0.001$) (Table 1).

The median (IQR) modified Rankin scale score collected by SMS testing was 1 (0–3) at 30-, 60-, and 90-day postdischarge. The median (IQR) GPH scores were 44.9 (42.3–50.8), 47.7 (42.3–54.1), and 47.7 (41.1–54.1) and the median GMH scores were 45.8 (38.8–50.8), 43.5 (38.8–53.3), and 45.8 (36.9–50.8) at 30-, 60-, and 90-day postdischarge, respectively (Supplementary Table S2). The distributions of these outcomes gathered *via* SMS texting at 90 days are shown in Figure 3.

The mRS [median: 1 IQR: 1–3) from clinical follow-up encounters, within 13.0 ± 14.9 days of SMS-text responses across collection time points (Supplementary Figure S1), was available for 113 of the 142 patients who responded (Figure 4). Weighted Cohen's kappa between mRS scores obtained from SMS texting compared to follow-up encounters revealed almost perfect agreement ($\kappa = 0.8$, $p < 0.001$) for dichotomized (good vs. poor)

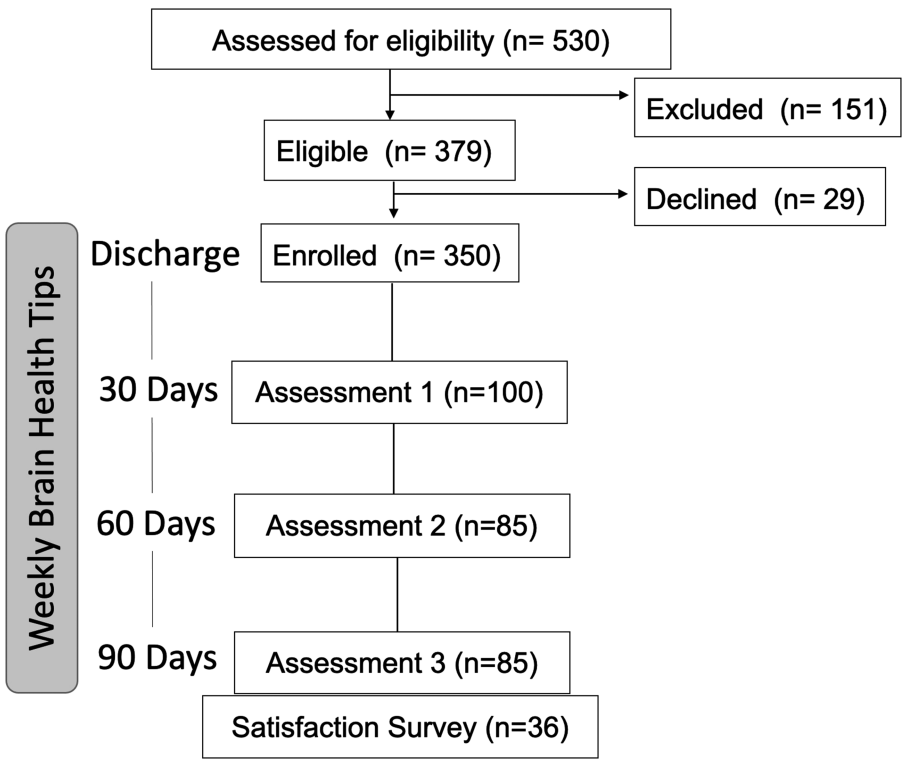


FIGURE 1
Study flow diagram. Cohorts at each assessment time point are unique.

TABLE 1 Cohort demographic and clinical characteristics.

	Enrolled (350)	Not enrolled (180)	*p*	Responders (142)	Nonresponders (208)	*p*
Age	66.3 ± 16.8	69.4 ± 16.7	0.5	65.2 ± 15.1	67.0 ± 17.7	0.06
Sex (male)	196 (56.0%)	92 (51.1%)	0.3	78 (54.9%)	118 (56.7%)	0.6
Hospital LOS	6.8 ± 8.2	8.1 ± 7.6	0.3	5.5 ± 6.4	7.6 ± 9.1	0.01*
NIH stroke scale[a]	3 [1–6]	5 [2–11]	<0.001*	3 [1–6]	8 [5–14]	<0.001*
Preferred language						
English	312 (89.1%)	154 (85.6%)	0.02*	133 (93.7%)	179 (86.1%)	0.009*
Other	38 (10.9%)	26 (14.4%)		9 (6.3%)	29 (13.9%)	
Discharge destination						
Home	200 (57%)	86 (48%)	0.04*	100 (70.4%)	100 (48.1%)	<0.001*
Facility	150 (43%)	94 (52%)		42 (29.6%)	108 (51.9%)	
Principal problem						
Ischemic	237 (67.7%)	124 (68.5%)	0.2	94 (66.2%)	143 (68.8%)	0.5
Hemorrhagic	90 (25.7%)	51 (28.3%)		37 (26.1%)	53 (25.5)	
TIA	23 (6.6%)	5 (2.8%)		11 (7.8%)	12 (5.8)	

LOS, length of stay; TIA, transient ischemic attack.
Scores reported as mean ± SD, median [interquartile range], or n (%).
*denotes statistical significance.
[a]NIH stroke scale scores at acute stroke hospital admission were available for 241/350 (68.9%) enrolled, 127/180 (70.6%) not enrolled, 100/142 (70.4%) responders, and 141/208 (67.8%) nonresponders.

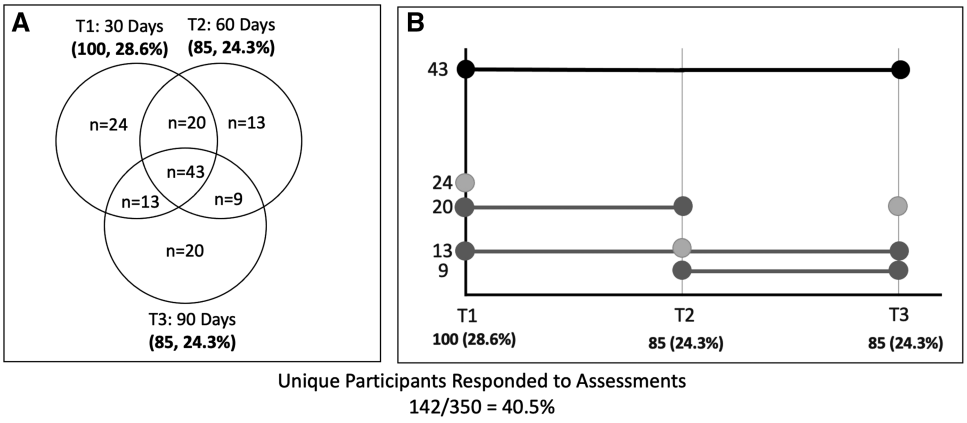


FIGURE 2
(A) Number of SMS-text assessment responses at different time points depicted in a Venn diagram (*N*, *N*%). *N* indicates unique participant responses. For example, *N* = 13 participants responded at 60 days poststroke, *N* = 20 answered at both 30 and 60 days, and *N* = 43 responded at all three study time points. (B) SMS-text assessment compliance over time. Shades of gray (dark to light) correspond to those that responded at all time points, those that responded at two timepoints (T_1 and T_2, T_1 and T_3, or T_2 and T_3), and those that responded at a single time point (T_1, T_2, or T_3). SMS, short message service.

and ordinal (with quadratic weights, Supplementary Table S3) levels ($\kappa = 0.8$, $p < 0.001$) of the mRS. The PROMIS Global-10 score was not routinely collected and so was only available for 19 patients from clinical encounters. There were significant associations between traditional methods and SMS texting of ascertaining PROMIS subscores ($\rho = 0.4$, $p < 0.01$, GPH, and $\rho = 0.4$, $p < 0.01$, GMH) (Figure 4).

To compare the yield of SMS texts with that of phone calls, we attempted to complete a 90-day phone call for all patients discharged within a 3-month time period. Of the 169 stroke patients discharged, we reached 104 (61.4%) by phone. Of those reached by phone, 59 (56.7%) were contacted on the first call attempt, 28 (26.9%) were contacted on the second attempt, and 17 (10.1%) were contacted on the third attempt. For every successful phone call, there were 2.5 unanswered calls. As compared to those who did not answer, patients who answered the phone calls were more likely to have English listed as their preferred language ($t = 3.9$, $p \leq 0.001$). Phone calls took 5.4 ± 2.9 minutes to complete. Unanswered calls took approximately 1.5 minutes.

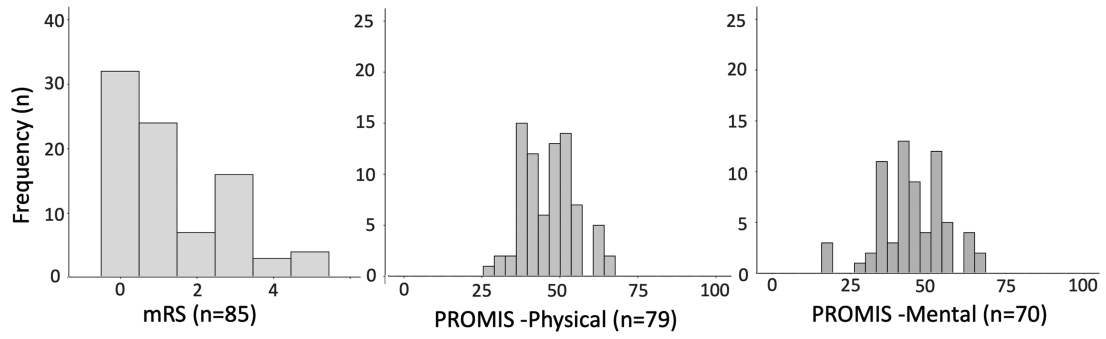


FIGURE 3
Distributions of 90-day outcomes (mRS, PROMIS Physical, and PROMIS Mental) gathered *via* SMS texting. SMS, short message service; mRS, modified Rankin scale; PROMIS Physical, Global Physical Health Score; PROMIS Mental, Global Mental Health Score.

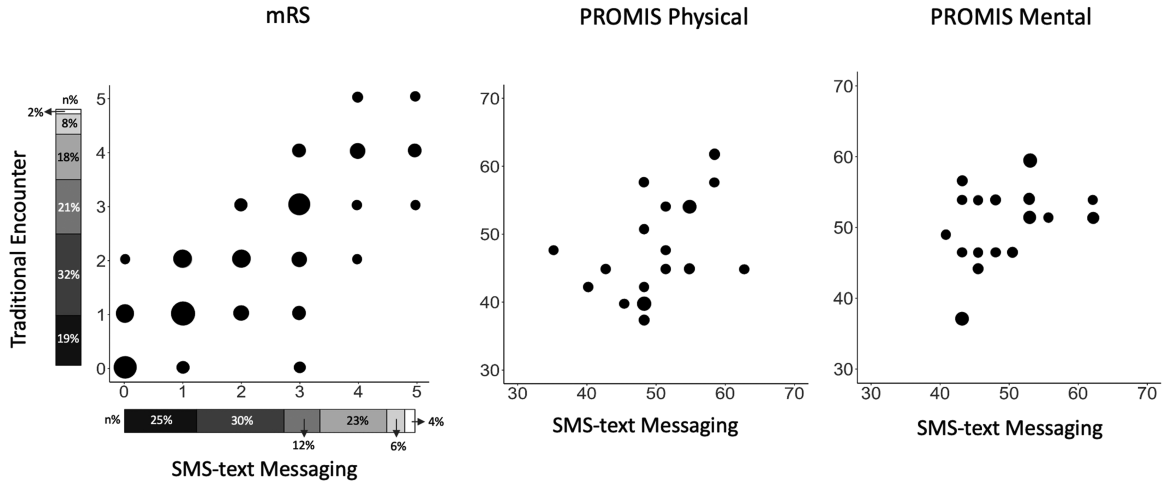


FIGURE 4
(**A**) mRS compared by modality (traditional encounter vs. SMS texting) across all collection time points (30, 60, and 90 days) for $N = 113$. Bubble plot to assess agreement between outcomes collected by traditional programs vs. SMS texting. Sizes of bubbles are directly proportional to the number of participants who provided answers *via* traditional methods and SMS texting. Subjects with data at multiple time points were compared at the latest available time point. The stacked bar near axes shows mRS distributions as collected by each modality with dark to light gradient representing scores 0–5, respectively. (**B**) PROMIS Physical and PROMIS Mental compared by modality for $N = 19$. Bubble plot to assess agreement between outcomes collected by traditional programs vs. SMS texting. The size of bubbles is directly proportional to the number of participants who provided answers *via* traditional methods and SMS texting. SMS, short message service; mRS, modified Rankin scale; PROMIS Physical, Global Physical Health Score; PROMIS Mental, Global Mental Health Score.

We estimated the cost per successful assessment achieved *via* SMS texts vs. phone calls. The total annual cost of the SMS-text program was \$12,500 to configure and run, with a cost per assessment defined as $\frac{\$12,500}{a}$ (where $a = \#$ of assessments). For phone calls, we estimated a fixed cost of 20% of a coordinator's time (for training and documentation) and added the average cost per phone call (both successful and unsuccessful) to each patient (n). In Boston, Massachusetts, coordinators make an average (avg.) of \$60,000 and thus \$0.48 per minute. Answered phone calls cost \$2.50 (5.4 min × \$0.48min), while unanswered phone calls cost \$0.70 (1.5 min × \$0.48) on average. This analysis yielded the following:

$$\frac{\begin{array}{c}[\text{fixed coordinator cost} + (\text{avg. cost of a successful call} \times \text{n}) \\ +(\text{avg. cost of an unsuccessful call} \\ \times \text{avg. number of unanswered calls} \times \text{n})]\end{array}}{\text{n}}$$

or

$$\frac{[\$12,000 + (2.5 \times \text{n} + \$0.70 \times 2.5 \times \text{n})]}{\text{number of subjects}},$$

suggesting that SMS texts become less costly than traditional phone calls if used for outcomes assessment in more than 181 stroke patients per year.

The results of the satisfaction survey revealed that the majority of patients found the educational material received *via* SMS texts to be helpful (53%), with only a small portion of patients finding information burdensome (6%). Most participants felt there was just the right number of tips (79%) and that the messages were clear and easy to understand (97%). Most felt that they were able to easily pick between the choices (91%) in the SMS texts that best

TABLE 2 Satisfaction survey results.

Question	*n* (%)
1: How helpful was it to receive a text with information, educational content, and tips about stroke?	
I did not find it helpful at all	2 (5.6)
Neutral	15 (41.7)
I found it very helpful	19 (52.7)
2: Did you find receiving information about stroke *via* text burdensome?	
I did not find the program burdensome	31 (93.9)
I found the program burdensome	2 (6.1)
3: What did you think of the number of tips and questions?	
Too few reminders	1 (3.0)
Just the right number of reminders	26 (78.8)
Too many reminders	6 (18.2)
4: The text messages about stroke were clear and easy to understand.	
Agree	32 (97.0)
Disagree	1 (3.0)
5: I was able to easily pick between the choices in the text messages that best described my degree of stroke recovery.	
Agree	30 (90.9)
Disagree	3 (9.1)
6: I was as comfortable answering questions by text as if I were answering in person or on the phone.	
Agree	30 (90.9)
Disagree	3 (9.1)

Satisfaction survey results are reported at n (%).

described their recovery. Most participants (91%) reported feeling as comfortable answering questions by SMS texts than by answering in person or *via* phone calls (**Table 2**).

Discussion

This is the first study to systematically examine the feasibility, reliability, cost, and acceptability of gathering outcomes after acute stroke *via* SMS texts. This novel method was found to be highly reliable for collecting the mRS scores and moderately reliable for collecting PROM scores. Even without specific program marketing or patient engagement campaigns, the SMS-text program yielded a 40.5% response rate. Compared to direct patient phone calls, SMS texting yielded fewer responses but is cost-saving for centers with annual stroke discharges exceeding 181 patients based on our costing equation. The experience of receiving text messages with assessments and brain health tips was overall very well received by stroke patients.

We found almost perfect agreement between mRS scores at both the dichotomized (good vs. poor outcome) and ordinal levels obtained *via* SMS texting compared to traditional methods. This high reliability provides the foundation for systematic evaluation of stroke survivors' outcomes at a large scale, which could help evaluate the efficacy of stroke treatments. We found moderate agreement with traditional methods for patient-reported measures. This suggests that certain stroke outcomes (i.e., ordinal ratings of global disability) may be more suitable for text-message programs. Prior studies for smoking cessation (24) and depression (25) have also found mixed reliability (fair to substantial) for self-report assessments. Digital means for collecting stroke outcomes may be limited by stroke-related impairments (i.e., language or cognitive deficits). Further exploration is required to examine the reliability of collecting different types of outcomes *via* SMS texting, particularly in stroke. Furthermore, modality-specific outcome measures, such as motor or language assessments after stroke, may require other types of digital or sensor-based approaches (26–28).

SMS texting had an overall lower yield (40.5%) in this study compared to targeted phone call follow-up method (61%, phone calls). The first attempt at reaching participants *via* SMS texts also yielded lower responses (28.6%) than the first attempt at reaching participants *via* phone calls (56.7%). This differs from a prior study that received more SMS-text data than paper diary data after birth control insertion, although notably this patient population was substantially younger as compared to ours (29). Stroke survivors who tend to be older and often suffer stroke-related deficits may be limited in their ability to use cell phones or read and write SMS texts. Survivors might also have limited access to their mobile devices when discharged to a facility, as a majority (70.4%) of responders were discharged home. Engaging caregivers to help with outcomes data collection *via* digital technology may be helpful in these cases. Low SMS-text response rates could also be attributed to participants feeling more comfortable declining to participate *via* SMS texts rather than directly to a care team member *via* phone calls. Another reason could be that the number of required questions needed to complete the assessments by SMS texts was too burdensome, and future research should determine the optimal number of questions to response ratio. In addition, future programs with a dedicated patient and caregiver outreach and SMS-text reminders (30) will likely yield higher response rates.

While the majority of stroke patients consented to SMS-text communication, there were differences between those who consented and those who did not. Individuals who consented to receive SMS texts were more likely to have English listed as their preferred language than another language, have a lower NIH stroke scale at admission, and be discharged home rather than to a postacute care facility. Similar differences were found between those who responded to SMS-text assessments vs. those who did not. Furthermore, outcomes gathered *via* SMS texts revealed that responders had predominantly mild disability (median mRS of 1, with an interquartile range of 0-3). Outcomes gathered *via* SMS texts may not be representative of all stroke survivors. A nutrition education program for low-income parents that used SMS texts for program evaluation also found limitations in those who could be reached (31). In stroke, different approaches may be required to reach non-English speakers (32) and those with more severe disabilities (33). If the use of SMS-text programs can diminish the burden of manual collection, outcome collection systems can work

in parallel to focus human resources on the more healthcare marginalized and disabled.

Due to the fixed cost of the SMS-text program for an unlimited number of participants, higher response rates would render a lower cost per participant. This contrasts with traditional methods such that outcomes collected during clinic visits or by phone calls require more resources for additional participants and thus cost increases per participant. Therefore, identifying the cost-to-response ratio that would favor SMS texting over traditional methods is essential for developing cost-saving programs. In our study, we show that gathering data *via* SMS texting would be cost-saving at scale for larger populations. Alternatively, it could also be cost-effective for smaller populations requiring outcome assessments at a greater frequency. For example, a previous study showed that SMS texting was more cost-effective for a weekly, two-question survey than the same paper-based survey (34) due to the frequency demand of the assessment. Future programs should consider the number of participants, data type, and sampling frequency and length in determining the cost-to-benefit ratio when using SMS texting.

Patients who participated in the program found it acceptable without additional burden. The majority found that the messages were clear, easy to understand, and easy to answer. The delivery of brain health tips was well received. These results are consistent with results from the acceptability of SMS texting in diverse clinical populations including individuals with depression (25), high blood pressure (35), and psychosis (36). Given the pervasive use of cell phones in modern society, cell phone-based outcome programs have significant promise for a wide range of patient populations including those with stroke. Although it is feasible to collect the mRS score *via* a mobile app (19), SMS texting leverages existing software without requiring a smartphone, additional download, or application knowledge.

Overall, our findings suggest that collecting functional outcomes *via* SMS texting during the first 3 months after stroke is feasible, acceptable, and reliable but that reach and cost-effectiveness should be further considered for broad clinical translation. Future programs should consider developing content in multiple languages and incorporating dedicated patient outreach materials. For example, educational material on how to view and respond to SMS-text communications delivered during the acute stroke inpatient stay could be helpful. Such content would help reach vulnerable populations. Increasing the yield of SMS-text outcome programs would decrease the cost per participant and make broad adoption across healthcare systems more feasible.

This study has several important limitations. The study was conducted at a single, urban, academic medical center in the northeastern United States with a predominantly White patient population. Findings may not be translatable to other hospitals in different locations with different patient populations. A limitation to communication *via* SMS texting is the chance that someone other than the intended recipient is receiving or interacting with the SMS texts. Moreover, the program could not gather information on the number of subjects who passed away during the study. In longitudinal data collection *via* SMS texting, subjects may see their responses from prior time points (in the SMS-text chain), which may lead to recall bias. Our sample of PROM data *via* both SMS texting and traditional methods was small ($n = 19$), and thus future studies with larger samples are needed to draw definitive conclusions. At last, the potential effects of our stroke and brain health educational program delivered *via* SMS texting were not systematically considered in our cost analysis.

Conclusion

Our study suggests that it is feasible, reliable, and acceptable to provide general stroke education and gather functional outcome measures *via* SMS-text messaging after acute stroke discharge. Replication of our results in an independent cohort and further validation of specific outcome types and assessment frequency and length are warranted. Our findings lay the foundation for using SMS texting to gather outcomes after stroke to better evaluate the real-world efficacy of stroke therapies.

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 Massachusetts General Brigham Institutional Review Board (#2021P001342) approved this study, which was exempt from written informed consent as the data extracted for this study was gathered under the standard of care through a quality improvement initiative.

Author contributions

All authors provided sufficient contributions to warrant authorship. JAD, KSE, MP, KE, PG, LHS, and DJL contributed to drafting and revising the manuscript for content. JAD, MP, KE, LHS, and DJL had a major role in acquiring data. JAD, LHS, and DJL contributed to conceptualizing and designing the study. JAD, KSE, PG, LHS, and DJL contributed to analyzing and interpreting the data. All authors contributed to the article and approved the submitted version.

Funding

This research was funded in part by a gift from Marriott Foundation.

Acknowledgments

The authors acknowledge Katrina Bennett, Philips Patient Navigation, for her assistance with product configuration and Kelly Sloane for her help developing the weekly brain health tips and coordinating with the multidisciplinary care teams.

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.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2023.1043806/full#supplementary-material.

References

1. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart disease and stroke statistics—2021 update. *Circulation*. (2021) 143:e254–743. doi: 10.1161/CIR.0000000000000950

2. Friedman HS. Tissue plasminogen activator for acute ischemic stroke. *N Engl J Med*. (1996) 334:1405. doi: 10.1056/NEJM199605233342114

3. Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med*. (2015) 372:1019–30. doi: 10.1056/NEJMoa1414905

4. Dawson J, Liu CY, Francisco GE, Cramer SC, Wolf SL, Dixit A, et al. Vagus nerve stimulation paired with rehabilitation for upper limb motor function after ischaemic stroke (VNS-REHAB): a randomised, blinded, pivotal, device trial. *Lancet*. (2021) 397:1545–53. doi: 10.1016/S0140-6736(21)00475-X

5. Dromerick AW, Geed S, Barth J, Brady K, Giannetti ML, Mitchell A, et al. Critical period after stroke study (CPASS): a phase II clinical trial testing an optimal time for motor recovery after stroke in humans. *Proc Natl Acad Sci U S A*. (2021) 118: e2026676118. doi: 10.1073/pnas.2026676118

6. Chiu C-C, Wang J-J, Hung C-M, Lin H-F, Hsien H-H, Hung K-W, et al. Impact of multidisciplinary stroke post-acute care on cost and functional status: a prospective study based on propensity score matching. *Brain Sci*. (2021) 11:161. doi: 10.3390/brainsci11020161

7. Tyagi S, Koh GC-H, Luo N, Tan KB, Hoenig H, Matchar DB, et al. Role of caregiver factors in outpatient medical follow-up post-stroke: observational study in Singapore. *BMC Fam Pract*. (2021) 22:74. doi: 10.1186/s12875-021-01405-z

8. Koonin LM, Hoots B, Tsang CA, et al. Trends in the use of telehealth during the emergence of the COVID-19 pandemic—United States, January–March 2020. *Morb Mortal Wkly Rep*. (2020) 69:1595–9. doi: 10.15585/mmwr.mm6943a3

9. Naito A, Wills A-M, Tropea TF, Ramirez-Zamora A, Hauser RA, Martino D, et al. Expediting telehealth use in clinical research studies: recommendations for overcoming barriers in North America. *NPJ Parkinsons Dis*. (2021) 7:34. doi: 10.1038/s41531-021-00177-8

10. Bhavnani SP, Narula J, Sengupta PP. Mobile technology and the digitization of healthcare. *Eur Heart J*. (2016) 37:1428–38. doi: 10.1093/eurheartj/ehv770

11. Pew-Research-Center. Mobile phone ownership over time. Mobile fact sheet (2021). Available from: https://www.pewresearch.org/internet/fact-sheet/mobile/

12. Schwamm LH. Telehealth: seven strategies to successfully implement disruptive technology and transform health care. *Health Aff*. (2014) 33:200–6. doi: 10.1377/hlthaff.2013.1021

13. Iribarren S, Beck S, Pearce PF, Chirico C, Etchevarria M, Cardinale D, et al. TextTB: a mixed method pilot study evaluating acceptance, feasibility, and exploring initial efficacy of a text messaging intervention to support TB treatment adherence. *Tuberc Res Treat*. (2013) 2013:349394.

14. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. *J Med Internet Res*. (2017) 19:e295. doi: 10.2196/jmir.7740

15. Mougalian SS, Gross CP, Hall EK. Text messaging in oncology: a review of the landscape. *JCO Clin Cancer Inform*. (2018) 2:1–9. doi: 10.1200/CCI.17.00162

16. Bobrow K, Brennan T, Springer D, Levitt NS, Rayner B, Namane M, et al. Efficacy of a text messaging (SMS) based intervention for adults with hypertension: protocol for the star (SMS text-message adherence support trial) randomised controlled trial. *BMC Public Health*. (2014) 14:28. doi: 10.1186/1471-2458-14-28

17. Naughton F, Prevost AT, Gilbert H, Sutton S. Randomized controlled trial evaluation of a tailored leaflet and SMS text message self-help intervention for pregnant smokers (MiQuit). *Nicotine Tob Res*. (2012) 14:569–77. doi: 10.1093/ntr/ntr254

18. Keoleian V, Polcin D, Galloway GP. Text messaging for addiction: a review. *J Psychoact Drugs*. (2015) 47:158–76. doi: 10.1080/02791072.2015.1009200

19. Cooray C, Matusevicius M, Wahlgren N, Ahmed N. Mobile phone–based questionnaire for assessing 3 months modified Rankin score after acute stroke. *Circ Cardiovasc Qual Outcomes*. (2015) 8:S125–30. doi: 10.1161/CIRCOUTCOMES.115.002055

20. Bruno A, Akinwuntan AE, Lin C, Close B, Davis K, Baute V, et al. Simplified modified Rankin scale questionnaire. *Stroke*. (2011) 42:2276–9. doi: 10.1161/STROKEAHA.111.613273

21. Sulter G, Steen C, De Keyser J. Use of the Barthel index and modified Rankin scale in acute stroke trials. *Stroke*. (1999) 30:1538–41. doi: 10.1161/01.STR.30.8.1538

22. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Qual Life Res*. (2009) 18:873–80. doi: 10.1007/s11136-009-9496-9

23. Warrens MJ. Cohen's weighted kappa with additive weights. *Adv Data Anal Classif*. (2013) 7:41–55. doi: 10.1007/s11634-013-0123-9

24. Thrul J, Mendel JA, Simmens SJ, Abroms LC. Collecting outcome data of a text messaging smoking cessation intervention with in-program text assessments: how reliable are the results? *Addict Behav*. (2018) 85:31–7. doi: 10.1016/j.addbeh.2018.05.012

25. Richmond SJ, Keding A, Hover M, Gabe R, Cross B, Torgerson D, et al. Feasibility, acceptability and validity of SMS text messaging for measuring change in depression during a randomised controlled trial. *BMC Psychiatry*. (2015) 15:68. doi: 10.1186/s12888-015-0456-3

26. Cramer SC, Koroshetz WJ, Finklestein SP. The case for modality-specific outcome measures in clinical trials of stroke recovery-promoting agents. *Stroke*. (2007) 38:1393–5. doi: 10.1161/01.STR.0000260087.67462.80

27. Erler KS, Wu R, DiCarlo JA, Petrilli MF, Gochyyev P, Hochberg LR, et al. Association of modified Rankin scale with recovery phenotypes in patients with upper extremity weakness after stroke. *Neurology*. (2022) 98(18):e1877–85. doi: 10.1212/WNL.0000000000200154

28. Kim GJ, Parnandi A, Eva S, Schambra H. The use of wearable sensors to assess and treat the upper extremity after stroke: a scoping review. *Disabil Rehabil*. (2021) 44:6119–38. doi: 10.1080/09638288.2021.1957027

29. Nippita S, Oviedo JD, Velasco MG, Westhoff CL, Davis AR, Castaño PM. A randomized controlled trial of daily text messages versus monthly paper diaries to collect bleeding data after intrauterine device insertion. *Contraception*. (2015) 92:578–84. doi: 10.1016/j.contraception.2015.09.004

30. Moran L, O'Loughlin K, Kelly BD. The effect of SMS (text message) reminders on attendance at a community adult mental health service clinic: do SMS reminders really increase attendance? *Ir J Med Sci*. (2018) 187:561–4. doi: 10.1007/s11845-017-1710-0

31. Grutzmacher SK, Munger AL, Speirs KE, Zemeir LA, Richard KC, Worthington L. Feasibility of bidirectional text messages in evaluating a text-based nutrition education program for low-income parents: results from the text2bhealthy program. *Eval Program Plann*. (2017) 64:90–4. doi: 10.1016/j.evalprogplan.2017.04.001

32. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of language barriers for healthcare: a systematic review. *Oman Med J*. (2020) 35:e122. doi: 10.5001/omj.2020.40

33. Lezzoni L. Eliminating health and health care disparities among the growing population of people with disabilities. *Health Aff*. (2011) 30:1947–54. doi: 10.1377/hlthaff.2011.0613

34. Johansen B, Wedderkopp N. Comparison between data obtained through real-time data capture by SMS and a retrospective telephone interview. *Chiropr Osteopat*. (2010) 18:10. doi: 10.1186/1746-1340-18-10

35. Leon N, Surender R, Bobrow K, Muller J, Farmer A. Improving treatment adherence for blood pressure lowering via mobile phone SMS-messages in South Africa: a qualitative evaluation of the SMS-text adherence support (STAR) trial. *BMC Fam Pract*. (2015) 16:80. doi: 10.1186/s12875-015-0289-7

36. D'Arcey J, Collaton J, Kozloff N, Voineskos AN, Kidd SA, Foussias G. The use of text messaging to improve clinical engagement for individuals with psychosis: systematic review. *JMIR Ment Health*. (2020) 7:e16993. doi: 10.2196/16993

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 11 April 2023
DOI 10.3389/fdgth.2023.1095110

OPEN ACCESS

EDITED BY
Kirsten Smayda,
MedRhythms, United States

REVIEWED BY
Kathleen M. M. Howland,
Berklee College of Music, United States
Tao Jiming,
Shanghai University of Traditional Chinese Medicine, China

*CORRESPONDENCE
Hantian Liu
htliu@bu.edu

SPECIALTY SECTION
This article was submitted to Personalized Medicine, a section of the journal Frontiers in Digital Health

RECEIVED 10 November 2022
ACCEPTED 22 March 2023
PUBLISHED 11 April 2023

CITATION
Liu H, Cordella C, Ishwar P, Betke M and Kiran S (2023) Consistent long-term practice leads to consistent improvement: Benefits of self-managed therapy for language and cognitive deficits using a digital therapeutic.
Front. Digit. Health 5:1095110.
doi: 10.3389/fdgth.2023.1095110

Consistent long-term practice leads to consistent improvement: Benefits of self-managed therapy for language and cognitive deficits using a digital therapeutic

Hantian Liu[1]*, Claire Cordella[2], Prakash Ishwar[3], Margrit Betke[1] and Swathi Kiran[2]

[1]Department of Computer Science, College of Arts and Sciences, Boston University, Boston, MA, United States, [2]Center for Brain Recovery, Boston University, Boston, MA, United States, [3]Department of Electrical and Computer Engineering, College of Engineering, Boston University, Boston, MA, United States

Background: Although speech-language therapy (SLT) is proven to be beneficial to recovery of post-stroke aphasia, delivering sufficiently high amounts of dosage remains a problem in real-world clinical practice. Self-managed SLT was introduced to solve the problem. Previous research showed in a 10-week period, increased dosage frequency could lead to better performance, however, it is uncertain if dosage still affects performance over a longer period of practice time and whether gains can be seen following practice over several months.
Objective: This study aims to evaluate data from a health app (Constant Therapy) to investigate the relationship between dosage amount and improvements following a 30-week treatment period. Two cohorts of users were analyzed. One was comprised of patients with a consistent average weekly dosage amount and the other cohort was comprised of users whose practice had higher variability.
Methods: We conducted two analyses with two cohorts of post-stroke patients who used Constant Therapy. The first cohort contains 537 "consistent" users, while the second cohort contains 2,159. The 30-week practice period was split into three consecutive 10-week practice windows to calculate average dosage amount. In each 10-week practice period, patients were grouped by their average dosage into low (0–15 min/week), medium (15–40 min/week) and moderate dosage (greater than 40 min/week) groups. Linear mixed-effects models were employed to evaluate if dosage amount was a significant factor affecting performance. Pairwise comparison was also applied to evaluate the slope difference between groups.
Results: For the consistent cohort, medium ($\beta = .002$, $t_{17,700} = 7.64$, $P < .001$) and moderate ($\beta = .003$, $t_{9,297} = 7.94$, $P < .001$) dosage groups showed significant improvement compared to the low dosage group. The moderate group also showed greater improvement compared to the medium group. For the variable cohort in analysis 2, the same trend was shown in the first two 10-week windows, however, in weeks 21–30, the difference was insignificant between low and medium groups ($\beta = .001$, $t = 1.76$, $P = .078$).
Conclusions: This study showed a higher dosage amount is related to greater therapy outcomes in over 6 months of digital self-managed therapy. It also showed that regardless of the exact pattern of practice, self-managed SLT leads to significant and sustained performance gains.

KEYWORDS

aphasia, stroke, technology, rehabilitation, dosage, therapy, data science

1. Introduction

Stroke is the most common disease that causes serious neurological disorders (1). Every year, over 795,000 people in the United States have a stroke, and aphasia or other communication disorders develop in approximately one-third of cases (2, 3). Compared to other patients, patients with aphasia are facing higher mortality and a higher degree of functional limitation, communication limitation, and social isolation (4, 5), making the need for effective rehabilitative approaches especially acute.

Previous research has shown that speech-language therapy (SLT) benefits functional language, language comprehension (listening and reading), and language production (speaking and writing) (6–14). Results also indicated that therapy at high intensity, high dosage, or over a longer period might be more beneficial compared to lower-intensity therapy (6). Moreover, high-intensity SLT over a short period appeared to help participants' language use in daily life and reduced the severity of their aphasia. However, high-intensity treatments might be less acceptable than less intensive therapy schedules for patients, as indicated by a significantly greater drop-out rate for higher-intensity regimens (6). Besides acceptability, there was also the problem of delivering sufficiently high therapy doses to patients in the real world, where practical realities (e.g., reimbursement caps, difficulties with mobility and travel, geographic isolation) placed severe limits on the amount of therapy actually received. National statistics available from the American Speech-Language-Hearing Association (ASHA) demonstrated a substantial reduction in the frequency and amount of SLT by the time patients had been discharged from acute or inpatient settings to community-based outpatient settings (15–17). A recent study of dosage amounts in a U.S.-based outpatient setting reported a median total therapy dosage of just 7.5 h for individuals with post-stroke aphasia (18). Similarly, another study of access to outpatient post-stroke rehabilitation services found that the average total dosage of outpatient SLT was 8 h total in the year following an individual's stroke (19). These average numbers were far from the number of hours of therapy recommended for high-intensity SLT. In fact, meta-analytic reviews have characterized high-intensity SLT protocols as providing total therapy dosages between 27 and 208 h, with positive effect studies tending to provide at least 50 total hours of therapy (6, 20).

Enabling patients to engage in in-home practice through computerized or app-based therapeutic programs could help patients to get more sufficient amounts of therapy and meet the dosage requirements of high-intensity SLT (13). Digital SLT interventions have been used as part of a treatment protocol in the form of smartphone, tablet, or computer-based programs. Some of these programs are entirely self-managed, meaning that patients can determine their own therapy schedule (14, 21–23). By giving patients the freedom to determine their practice schedule, researchers can access a wide range of practice frequencies, amounts and overall practice patterns from patient to patient. This variability provides a unique opportunity to probe practice-response relationships in SLT *via* dose articulation studies, which are a necessary first step toward the ultimate goal of establishing optimal dosage recommendations for SLT interventions (23, 24).

Recent efforts by Cordella et al. analyzed retrospectively collected data to evaluate the optimal dosage of interventions. In this study, the authors directly compared different dosage amounts of the same intervention in the context of self-managed digital therapies (23). This study focused on the relationship between the varied dosage frequency and the performance outcome across 13 different skill domains following a 10-week period of self-managed digital SLT. The results showed that higher dosage frequency groups (e.g., four or five times per week) achieved greater improvement vs. lower ones (e.g., once or twice per week) across all domains and also within a majority of individual subdomains. However, the definition of dosage in the Cordella et al. study is primarily the median number of days in a week patients practice, which is only one parameter to evaluate overall dosage (25). Other ways to calculate dosage have included session duration, total intervention duration, and total number of sessions administered (24–32). Moreover, it is not clear that 10 weeks is a sufficient duration of language therapy, especially in chronic survivors. Consequently, it is useful to evaluate improvements over a longer time period than 10 weeks, by which it would be possible to discover potentially more nuanced relationships between dosage and performance.

The goal of this study was to examine real-world therapy data to investigate the relationship between dosage amount, and midpoint and cumulative improvements following a 30-week treatment period using the Constant Therapy app. There were two main objectives of the current study. First, we investigated whether greater average weekly dosage—defined as number of minutes per week—led to greater performance gains over a 30-week period in a cohort of consistent users who practiced approximately the same average amount week to week. Second, in a larger cohort of more variable users we investigated the effect of weekly therapy dosage on performance outcomes across three consecutive 10-week intervals for a total of 30 weeks (i.e., 6 months). The two cohorts were denoted as consistent cohort and variable cohort. We hypothesized that in both analyses, greater practice amount would lead to better performance outcomes. Prior work has shown that during the first 10-week period of therapy, higher dosage frequency groups improved more compared to lower ones across all domains (23). Therefore, we hypothesized that such trend would persist in longer-term therapy that was practiced beyond 10 weeks.

2. Methods

2.1. Participants

Data used in this study are from patients who used the Constant Therapy app between March 2016 and July 2020. 30,129 unique users who reported having had a stroke with resultant speech, language, and cognitive deficits were included in the analyses with their consent to using exercise and

performance outcome data for research purposes. In order to evaluate the performance of longer-term therapy, a smaller number of users were filtered using criteria described in detail below. Overall, all users were engaged in the app for more than 10 weeks in order for their data to be included in the analyses. As described above, in the first cohort, 537 users practiced 30 weeks of consistent therapy (i.e., consistent cohort). In the second, variable cohort, the number of users differed among time periods. 2,159 patients are considered in the first 10 weeks, 1,314 in the second 10 weeks, and 812 in the last 10 weeks. The filtering procedure flowchart is shown in **Figure 1** to describe how we select the users from the whole population in the database. Note that all sessions we selected are self-managed sessions, which means no interference is made by any other individual including clinicians or Constant Therapy support team. Demographic details regarding participants are provided in **Table 1** after the filtering criteria are described.

2.2. Constant therapy program

Constant Therapy (CT) is an app-based, evidence-based digital therapeutic designed to improve multiple domains of language simultaneously using a self-managed approach (www.constanttherapy.com) (33). **Figure 2** depicts the CT therapy program using a tripartite schema (i.e., therapy target(s),

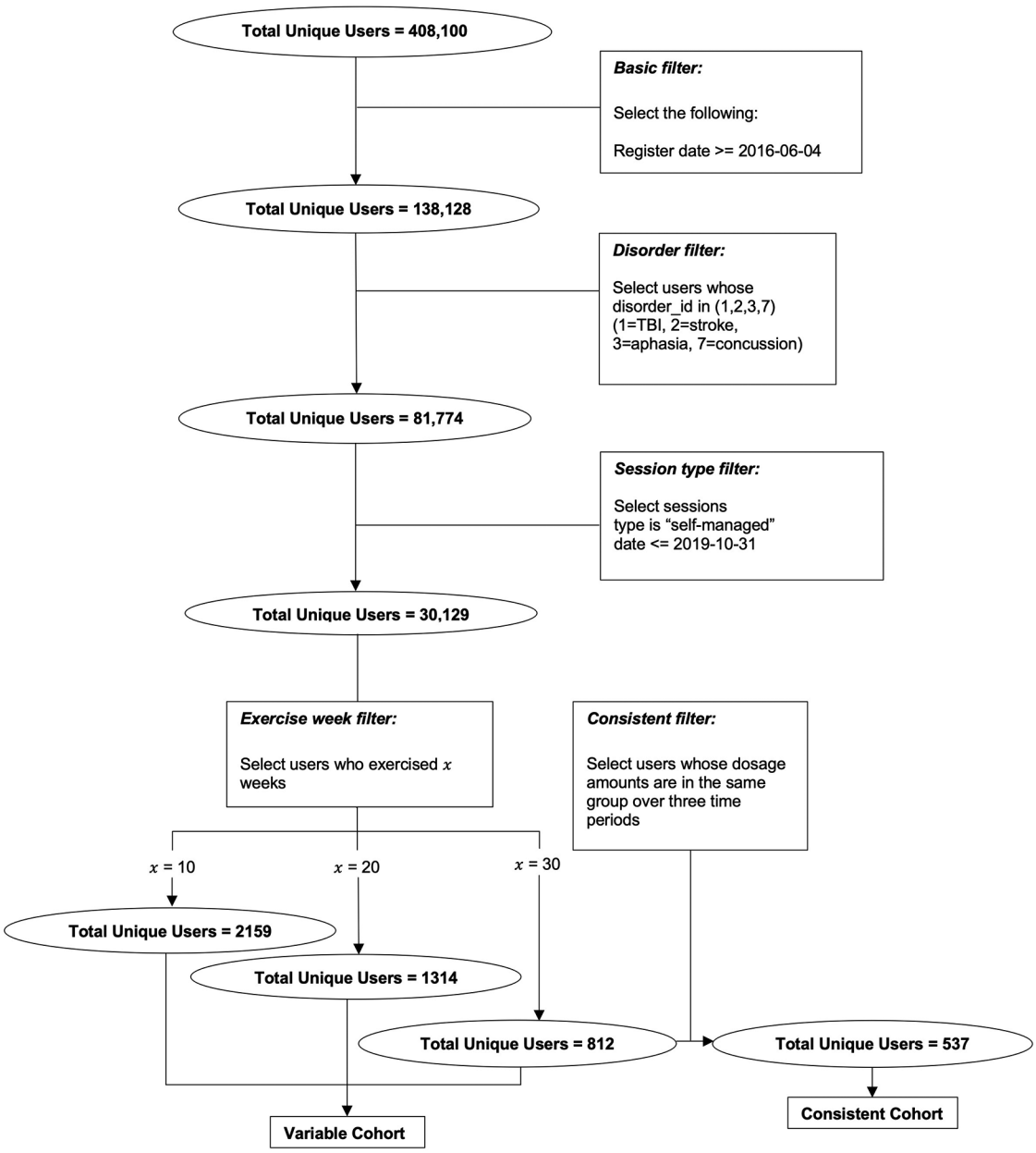


FIGURE 1
Flow chart of the data filtering procedure that results in the two cohorts for which analysis 1 and 2 are conducted, respectively.

ingredients, and mechanisms of action) following the Rehabilitation Treatment Specification System (RTSS) (34). There are several unique ingredients of the program, including (1) task variety with 266 different task types spanning speech, language and cognitive domains and functional daily activities that encompass them (e.g., listening to a voicemail, reading a map); (2) personalized goal setting enabling patients and their clinicians to identify high-priority, functionally relevant therapy goals across multiple domains; (3) adaptive difficulty that enables self-paced progression from easier to harder tasks within each targeted domain using an algorithm based on performance accuracy and consistency, allowing for therapy scaffolding in a way that mirrors in-person therapy techniques employed by skilled clinicians; (4) consistent feedback that is provided to the patient after every item, therapy goal and session; (5) ease of access that allows patients to log in and practice therapy at their convenience and progress at their own pace; and (6) the recommended therapy regimen that can be self-managed, reducing the need for regular face-to-face interaction with a clinician. Preliminary studies of CT have indicated that it is effective in inducing improvements in language outcomes in chronic post-stroke aphasia (22, 35, 36).

For this study, we aggregated data across 13 different skill domains: (1) analytical, (2) arithmetic, (3) attention, (4) auditory comprehension, (5) auditory memory, (6) naming, (7) phonological processing, (8) production, (9) quantitative, (10) reading, (11) visual memory, (12) visuospatial skills and (13) writing. When using the Constant Therapy program, users select skill domains they wish to improve and are assigned tasks based on that selection by the algorithm. Task difficulty is adjusted per individual user using an adaptive algorithm, with more difficult tasks assigned once patients have demonstrated mastery of prior tasks assigned with a high accuracy. The order in which more difficult tasks are assigned is according to a universal task progression order per domain. The progression order is thus a serial ranking of tasks from least to most difficult. Determination of each domain's progression order was based on research evidence in consultation with speech-language pathologists (37). Patient progress is subdomain specific, so improvement in one domain does not affect the progression order of other domains the patient is practicing simultaneously. In this way, during a session, patients practice tasks in order of subsequent increasing progression orders. Additionally, if a patient fails to improve at one progression order, a lower-level task will be assigned to the patient in addition to the original task. The Constant Therapy app records all data for each session for this study including accuracy per

TABLE 1 Summary statistics of the consistent cohort ($N = 1,448$).

Characteristics	Overall ($N = 1,448$)	0–15 min/week	15–40 min/week	>40 min/week
Age, mean (SD)	63.13 (13.68)	62.43 (13.98)	63.10 (13.86)	64.74 (12.69)
Baseline domain score, mean (SD)	0.32 (0.20)	0.33 (0.22)	0.30 (0.19)	0.33 (0.18)
Sex, *n* (%)				
Male	820 (56.6)	423 (56.0)	193 (55.0)	204 (59.8)
Female	628 (43.4)	333 (44.0)	158 (45.0)	137 (40.2)
Other	0 (0)	0 (0)	0 (0)	0 (0)
Chronicity, *n* (%)				
Acute (<6 months)	705 (48.7)	364 (48.1)	169 (48.1)	172 (50.4)
Chronic (>6 months)	743 (51.3)	392 (51.9)	182 (51.9)	169 (49.6)

N (patients) = 537.

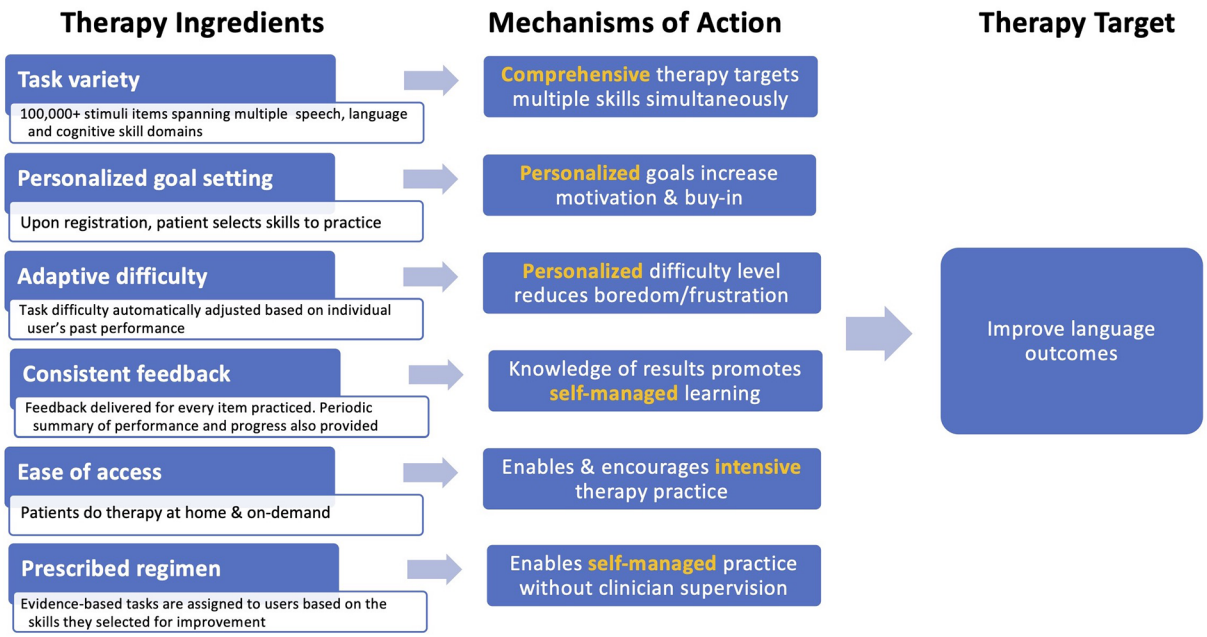


FIGURE 2
Ingredients, mechanisms, and targets of the constant therapy program, conceptualized within the rehabilitation treatment specification system (RTSS) framework.

trial, latency per trial, the progression order, timestamp, total exercises, and session duration.

Because users practice different task types at different levels of difficulty, it is not enough to evaluate the performance outcome using an accuracy metric alone. Instead, we derived a summative metric of performance accuracy that allows for comparison across different skill domains and task difficulty levels, called domain score. In a specific session, the highest progression order of the task passed or worked on and the lowest progression order of the task failed are recorded. Here passing a task indicates accuracy of the task is equal to or greater than 90%, working on means more than 40% and less than 90%, while failure means accuracy is lower than 40%. The domain score of the session is calculated by averaging the two progression numbers, which is an estimate of the session's difficulty level. After that, the domain score is normalized by dividing it by the total number of progression orders in the specific domain. Normalization is required because the numbers of progression orders vary from domain to domain, and the original number alone cannot be used to compare directly across different domains. More details of domain score and its calculation have been previously described (23). By averaging the domain score across sessions in a week (only if there are multiple sessions in a single week), it is possible to evaluate the improvement or deterioration of patients' performance over time in a single domain.

2.3. Determination of the different dosage groups

Prior to discussing the data analyses, it is important to describe the determination of the different dosage groups. For a specific patient, the term exercise week indicates a week in which the patient has exercise records; unless explicitly noted, *week* is defined as exercise week in this study. In an *n*-week time period, the average dosage amount is calculated by summing up the dosage amount in the *n* exercise weeks and dividing it by the total number of calendar weeks the patient spent to complete *n* weeks of practice, which may include some additional weeks that do not have exercise records. Patients were then binned into the following three groups based on their average dosage amount over a period spanning 10 exercise weeks: 0–15 min per week (low dosage group), 15–40 min per week (medium dosage group), and more than 40 min per week (moderate dosage groups). It should be noted here that users practicing greater than 40 min per week on average demonstrated a large dosage range (up to 1,736 min per week in a 10-week period).

We considered 30 (exercise) weeks of time in total to evaluate the relationship between dosage amount and performance outcome. The 30-week period was split into three 10-week periods, and dosage amounts were averaged separately in the three periods. Patients were considered consistent (Analysis 1) only if (1) they had at least 30 exercise weeks on record and (2) for each of the 10-week time periods, they stayed within the same dosage group. Since this dataset is relatively small and not reflective of the more variable practice patterns that characterize the majority of app users, we also wanted to include an analysis of patients with more variable usage habits (Analysis 2). In the three 10-week periods, patients were included if they had practice records in each of the 10 weeks. Crucially for this analysis, a specific patient could appear in different groups in different time periods (e.g., 0–15 min/week group in the first 10 weeks vs. 15–40 min/week group in the second 10 weeks), so it is not possible to compare the same dosage amount group across multiple 10-week periods, hence data in the three time periods were analyzed separately.

2.4. Statistical analyses

For all statistical analyses, the first week of the therapy within a 10-week period of exercises was indicated as the baseline week, and a comparison of domain scores between later weeks and the baseline week was made to address the performance outcome over this 10-week period. Because we were primarily interested in the effect of dosage amount on performance outcome, we began by grouping patients according to their average weekly dosage amount, measured by calculating the mean minutes per week of therapy. Patients were then binned into one of the three groups introduced above: 0–15 min per week, 15–40 min per week, and more than 40 min per week.

Linear mixed-effect models (LMMs) were run in order to examine domain score changes over time as a function of dosage amount group. The weekly domain score served as the dependent variable in the model, with fixed effects of time (week number), dosage amount group, cumulative practice amount (i.e., total hours spent completing therapy tasks), time × dosage amount group, and time × cumulative practice amount. Covariates of age, time since stroke (≤6 and >6 months), sex, and baseline domain scores were also included as fixed effects in the model. The model included random effects of patients and domains.

All statistical analyses were conducted in R (version 4.1.2; R Foundation for Statistical Computing) using *lme4*, *lmerTest*, *emmeans*, and *sjPlot* packages.

2.5. Ethics approval

This project was considered an institutional review board–exempt retrospective analysis by Pearl Institutional Review Board (#17-LNCO-101) under 45 Code of Federal Regulations 46.101 (b) category 2.

3. Results

3.1. Analysis 1: consistent users

A total of 537 patients and 1,448 records in different domains were selected as consistent practice patients by the criteria mentioned previously. As we are considering records of different domains from one specific patient separately, this can yield multiple records per patient. The statistical analysis is based on

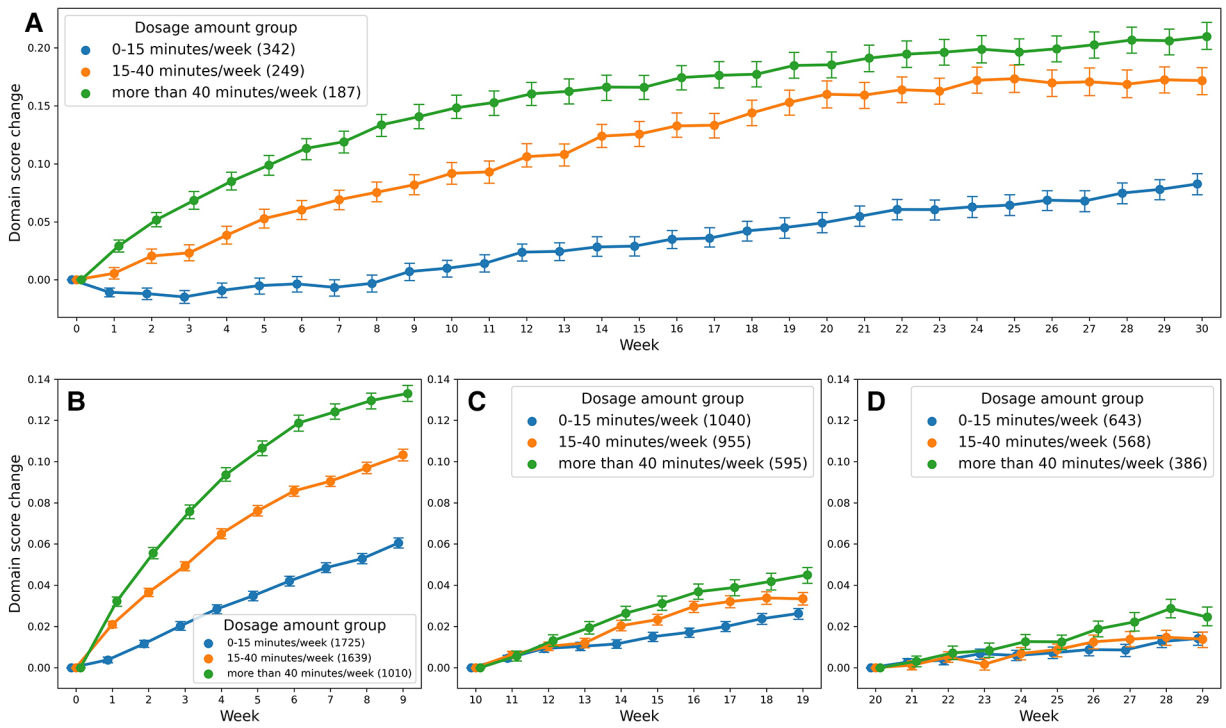


FIGURE 3
Change in domain score as a function of dosage amount group: (A) shows the score change for the consistent cohort, and (B–D) for the variable cohort.

the total number of 1,448 records. Among these records 820 are from male patients while 628 records are from female patients. The average age of patients is 63.13 (SD, 13.68) years old with 48.7% (705) in the acute recovery stage (less than 6 months prior to therapy initiation). The summary statistics for the entire cohort and for each dosage amount group are presented in Table 1. In general, age, sex, and chronicity did not differ among dosage groups ($P > .05$ in all comparisons).

Analysis 1 asked the question of whether greater average weekly dosage—defined as number of minutes per week—leads to greater performance gains over a 30-week period in a cohort of consistent users who practice approximately the same average amount week-to-week. The overall change in domain score (collapsed across domains) for the consistent group over 30 weeks is plotted in Figure 3A. The plot shows that, while all patients show improvements in the overall domain score, the 40+ min/week group shows greater changes in the domain score than the 0–15 min/week and 15–40 min/week groups over the 30-week time period. The statistical results for the consistent cohort are shown in Tables 2, 3. Specifically, a higher weekly domain score was associated with an increase in the number of weeks of therapy ($\beta = .004$; $t = 6.09$; $P < .001$), higher baseline domain score ($\beta = .378$; $t = 67.74$; $P < .001$), and greater practice amount (15–40 min/week: $\beta = .034$, $t = 7.49$, $P < .001$; 40+ min/week: $\beta = .091$, $t = 14.70$, $P < .001$). In addition, age ($\beta = -.001$; $t = -2.80$; $P = .005$) and time since stroke ($\beta = .019$; $t = 2.08$; $P = .038$) were also significant predictors of domain score, with younger age and acute chronicity associated with a higher weekly domain score. Sex was not a significant predictor of domain score.

TABLE 2 Final linear mixed-effects model results summary of consistent cohort (fixed effects).

Predictors	Estimates (SE)	*t* test (df)	*P* value
Fixed effects			
Intercept	2.51×10^{-1} (2.94×10^{-2})	8.56 (7.60×10^{1})	***
Week	3.88×10^{-3} (6.37×10^{-4})	6.09 (2.15×10^{1})	***
Dosage group (15–40 min/week)	3.43×10^{-2} (4.58×10^{-3})	7.49 (1.76×10^{4})	***
Dosage group (>40 min/week)	9.06×10^{-2} (6.17×10^{-3})	14.70 (9.37×10^{3})	***
Domain score baseline	3.78×10^{-1} (5.58×10^{-3})	67.74 (4.26×10^{4})	***
Age	-9.42×10^{-4} (3.37×10^{-4})	-2.80 (5.03×10^{2})	**
Sex (male)	3.94×10^{-4} (9.04×10^{-3})	0.04 (5.04×10^{2})	
Chronicity (acute)	1.86×10^{-2} (8.94×10^{-3})	2.08 (5.04×10^{2})	*
Week: Dosage group (15–40 min/week)	2.01×10^{-3} (2.63×10^{-4})	7.64 (1.77×10^{4})	***
Week: Dosage group (>40 min/week)	2.81×10^{-3} (3.54×10^{-4})	7.94 (9.30×10^{3})	***

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1.
Model equation: domain score (weekly average) ~ 1 + week * dosage group + domain score baseline + age + sex + chronicity + (1 + week:patient) + (1 + week: domain).

Most importantly given our study objectives, the time × dosage amount group interaction was significant ($F = 38.78$, $P < .001$). From this result, we note that although all groups of consistent app users improved over the 30-week therapy period, the rate of improvement was driven by the weekly dosage amount. Compared to the group practicing 0–15 min per week, the 15–40 min per week group ($\beta = .002$, $t_{17,700} = 7.64$, $P < .001$) and the group practicing more than 40 min per week ($\beta = .003$, $t_{9,297} = 7.94$, $P < .001$) showed significantly higher weekly domain scores

TABLE 3 Final linear mixed-effects model results summary of consistent cohort (random effects).

Predictors	Variance (SD)	Correlation
Random effects		
Residual	1.65×10^{-2} (1.28×10^{-1})	N/A
Patient (intercept)	1.03×10^{-2} (1.02×10^{-1})	N/A
Domain (intercept)	4.46×10^{-3} (6.68×10^{-2})	N/A
Week:patient (slope)	3.78×10^{-5} (6.15×10^{-3})	−0.20
Week:domain (slope)	3.99×10^{-6} (2.00×10^{-3})	−0.34

TABLE 4 Pairwise comparisons of slopes by dosage amount group (consistent cohort).

Contrast	Estimate (SE)	*t* test	*P* value
0–15 min/week vs. 15–40 min/week	-2.01×10^{-3} (2.63×10^{-4})	−7.64	***
0–15 min/week vs. >40 min/week	-2.81×10^{-3} (3.54×10^{-4})	−7.94	***
15–40 min/week vs. >40 min/week	-7.99×10^{-4} (2.94×10^{-4})	−2.69	*

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1.

over time. A *post hoc* comparison of slopes across each of the three dosage groups revealed a significantly greater rate of improvement for the moderate dosage (40+ min/week) group compared to the medium dosage (15–40 min/week) group and the low dosage (0–15 min/week). This reinforces the notion that incremental increases in weekly therapy dosage (i.e., 0–15 vs. 15–40 vs. 40 + min/week) yield significantly greater gains in improvements over a 30-week period for this cohort of consistent app users (Table 4).

Analysis 1 took a conservative approach to evaluate the effects of practicing long-term therapy, only users that consistently practiced for 30 weeks were included in the analyses. Consequently, the number of such users was relatively low, with only 537 individual users. A perusal of the database of users indicated that users were more likely to be variable in their practice patterns, sometimes practicing more often and sometimes practicing less often. To evaluate whether this variable practice pattern influenced the extent of domain score change, we conducted Analysis 2.

3.2. Analysis 2: variable users

In Analysis 2, in each 10-week period, the numbers of patients in this cohort are subject to change and vary from period to period. Demographic information about users in each of the three time periods is listed in Table 5. Age and sex are distributed evenly across the three time periods and the three groups in each time slot. However, the average baseline

TABLE 5 Summary statistics of the variable cohort.

Time period	Characteristics	Overall	0–15 min/week	15–40 min/week	>40 min/week
Week 1–10 ($N = 12{,}112$)	Age, mean (SD)	63.36 (13.51)	62.52 (13.77)	63.85 (13.71)	64.45 (12.35)
	Baseline domain score, mean (SD)	0.33 (0.20)	0.32 (0.21)	0.33 (0.19)	0.36 (0.18)
	Sex, *n* (%)				
	Male	6,902 (57.0)	3,061 (55.5)	2,488 (58.3)	1,353 (58.0)
	Female	5,151 (42.5)	2,423 (43.9)	1,764 (41.3)	964 (41.3)
	Other	59 (0.5)	30 (0.6)	12 (0.4)	17 (0.7)
	Chronicity, *n* (%)				
	Acute (<6 months)	6,899 (57.0)	3,080 (55.9)	2,468 (57.9)	1,351 (57.9)
	Chronic (>6 months)	5,213 (43.0)	2,434 (44.1)	1,796 (42.1)	983 (42.1)
Week 11–20 ($N = 6{,}888$)	Age, mean (SD)	63.39 (13.53)	62.59 (14.18)	63.99 (13.05)	64.45 (12.46)
	Baseline domain score, mean (SD)	0.41 (0.22)	0.37 (0.22)	0.43 (0.22)	0.47 (0.20)
	Sex, *n* (%)				
	Male	3,906 (56.7)	1,931 (57.4)	1,237 (55.4)	738 (57.3)
	Female	2,975 (43.2)	1,435 (42.6)	993 (44.5)	547 (42.5)
	Other	7 (0.1)	1 (0.0)	3 (0.1)	3 (0.2)
	Chronicity, *n* (%)				
	Acute (<6 months)	3,656 (53.1)	1,781 (52.9)	1,205 (54.0)	670 (52.0)
	Chronic (>6 months)	3,232 (46.9)	1,586 (47.1)	1,028 (46.0)	618 (48.0)
Week 21–30 ($N = 4{,}162$)	Age, mean (SD)	63.66 (13.06)	63.00 (13.31)	64.07 (13.05)	64.61 (12.38)
	Baseline domain score, mean (SD)	0.43 (0.23)	0.39 (0.23)	0.45 (0.23)	0.50 (0.21)
	Sex, *n* (%)				
	Male	2,398 (57.6)	1,151 (57.0)	746 (56.8)	501 (60.4)
	Female	1,762 (42.3)	867 (42.9)	567 (43.1)	328 (39.6)
	Other	2 (0.1)	1 (0.1)	1 (0.1)	0 (0)
	Chronicity, *n* (%)				
	Acute (<6 months)	2,011 (48.3)	989 (49.0)	622 (47.3)	400 (48.3)
	Chronic (>6 months)	2,151 (51.7)	1,030 (51.0)	692 (52.7)	429 (51.7)

N (patient 0–15 min/week) = 2,159.
N (patient 15–40 min/week) = 1,314.
N (patient >40 min/week) = 812.

TABLE 6 Final linear mixed-effects model results summary of the variable cohort (fixed effects).

Time period	Predictors	Estimates (SE)	*t* test (*df*)	*P* value
Fixed effects				
Week 1–10	Intercept	1.51×10^{-1} (1.31×10^{-2})	11.49 (2.27×10^{1})	***
	Week	9.00×10^{-3} (9.96×10^{-4})	9.04 (1.57×10^{1})	***
	Dosage group (15–40 min/week)	2.21×10^{-2} (1.69×10^{-3})	13.07 (3.93×10^{4})	***
	Dosage group (>40 min/week)	4.16×10^{-2} (2.27×10^{-3})	18.31 (1.90×10^{4})	***
	Domain score baseline	6.05×10^{-1} (2.58×10^{-3})	234.45 (1.14×10^{5})	***
	Age	-2.70×10^{-4} (9.62×10^{-5})	-2.81 (1.72×10^{3})	**
	Sex (male)	4.49×10^{-4} (2.63×10^{-3})	0.17 (1.71×10^{3})	
	Sex (not specified)	2.12×10^{-2} (1.77×10^{-2})	1.20 (1.84×10^{3})	
	Chronicity (acute)	9.29×10^{-3} (2.64×10^{-3})	3.52 (1.70×10^{3})	***
	Week: Dosage group (15–40 min/week)	2.74×10^{-3} (3.18×10^{-4})	8.60 (4.18×10^{4})	***
	Week: Dosage group (>40 min/week)	5.58×10^{-3} (4.28×10^{-4})	13.02 (2.07×10^{4})	***
Week 11–20	Intercept	7.14×10^{-2} (1.01×10^{-2})	7.05 (4.45×10^{1})	***
	Week	2.94×10^{-3} (5.29×10^{-4})	5.56 (2.56×10^{1})	***
	Dosage group (15–40 min/week)	1.59×10^{-3} (5.12×10^{-3})	0.31 (1.26×10^{4})	
	Dosage group (>40 min/week)	-5.01×10^{-3} (6.63×10^{-3})	-0.76 (6.24×10^{3})	
	Domain score baseline	7.79×10^{-1} (2.58×10^{-3})	301.62 (4.94×10^{4})	***
	Age	-2.80×10^{-4} (8.98×10^{-5})	-3.11 (9.26×10^{2})	**
	Sex (male)	-1.03×10^{-4} (2.44×10^{-3})	-0.42 (9.28×10^{2})	
	Sex (not specified)	-1.90×10^{-2} (3.25×10^{-2})	-0.59 (1.23×10^{3})	
	Chronicity (acute)	1.48×10^{-2} (2.43×10^{-3})	6.10 (9.28×10^{2})	***
	Week: Dosage group (15–40 min/week)	1.57×10^{-3} (3.51×10^{-4})	4.48 (1.38×10^{4})	***
	Week: Dosage group (>40 min/week)	3.33×10^{-3} (4.58×10^{-4})	7.27 (7.14×10^{3})	***
Week 21–30	Intercept	6.78×10^{-2} (1.17×10^{-2})	5.82 (1.02×10^{2})	***
	Week	1.63×10^{-3} (4.53×10^{-4})	3.60 (4.09×10^{1})	***
	Dosage group (15–40 min/week)	-4.00×10^{-3} (1.01×10^{-2})	-0.40 (3.02×10^{3})	
	Dosage group (>40 min/week)	-2.64×10^{-2} (1.25×10^{-2})	-2.11 (1.41×10^{3})	*
	Domain score baseline	8.13×10^{-1} (3.01×10^{-3})	270.21 (2.69×10^{4})	***
	Age	-3.99×10^{-4} (1.04×10^{-4})	-3.84 (5.94×10^{2})	***
	Sex (male)	-2.89×10^{-4} (2.77×10^{-3})	-0.11 (5.74×10^{2})	
	Sex (not specified)	-2.07×10^{-2} (4.64×10^{-2})	-0.45 (1.09×10^{3})	
	Chronicity (acute)	1.03×10^{-2} (2.75×10^{-3})	3.73 (5.73×10^{2})	***
	Week: Dosage group (15–40 min/week)	7.29×10^{-4} (4.14×10^{-4})	1.76 (3.11×10^{3})	.
	Week: Dosage group (>40 min/week)	2.49×10^{-3} (5.14×10^{-4})	4.85 (1.48×10^{3})	***

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1.
Model equation: domain score (weekly average) ~ 1 + week * dosage group + domain score baseline + age + sex + chronicity + (1 + week:patient) + (1 + week:domain).

domain score increased over time (week 1–10: 0.33, week 11–20: 0.41, week 21–30: 0.43), which indicates that patients were improving as part of the continued therapy process. Another factor to note is that as time progressed, the portion of acute patients decreased (week 1–10: 57.0%, week 11–20: 53.1%, week 21–30: 48.3%).

Analysis 2 asked the question of whether greater amounts of weekly therapy led to greater performance gains across three consecutive 10-week intervals (for a total of 30 weeks), in a larger cohort of more variable users. As shown in **Tables 6**, **7**, similar to the consistent cohort, time (week 1–10: $\beta = .009$, $t = 9.04$, $P < .001$, week 11–20: $\beta = .003$, $t = 5.56$, $P < .001$, week 21–30: $\beta = .002$, $t = 3.60$, $P < .001$), acute condition (week 1–10: $\beta = .009$, $t = 3.52$, $P < .001$, week 11–20: $\beta = .015$, $t = 6.10$, $P < .001$, week 21–30: $\beta = .010$, $t = 3.73$, $P < .001$) and greater baseline domain score (week 1–10: $\beta = .604$, $t = 234.45$, $P < .001$, week 11–20: $\beta = .779$, $t = 301.62$, $P < .001$, week 21–30: $\beta = .813$, $t = 270.21$, $P < .001$) were also associated with greater weekly domain score within each of the 10 week analysis periods. **Figures 3B–D** shows the change in domain score of this cohort in three different time periods.

Crucial to our question of interest, the interaction of time × dosage amount group was significant across each of the three 10 week analysis periods. Compared to the 0–15 min/week group, the 15–40 min/week (week 1–10: $\beta = .003$, $t = 8.60$, $P < .001$, week 11–20: $\beta = .002$, $t = 4.48$, $P < .001$) and 40 + min/week groups (week 1–10: $\beta = .006$, $t = 13.02$, $P < .001$, week 11–20: $\beta = .3$, $t = 7.27$, $P < .001$) showed greater rates of performance improvement in the first and second 10-week analysis intervals. Post hoc comparisons of slopes (**Table 8**) demonstrated a significantly greater rate of improvement also for the 40 + min/week compared to the 15–40 min/week in both the first and second 10-week intervals. For the final 10-week analysis interval (i.e., weeks 20–29 of therapy), a similar pattern of significance emerged, with the rate of improvement being significantly greater for 40+ min/week vs. 0–15 min/week group ($\beta = .002$, $t = 4.85$, $P < .001$), but with no significant difference in rates of improvement for the 15–40 and 0–15 min/week ($\beta = .001$, $t = 1.76$, $P = .078$). Post hoc tests revealed there was also a significantly greater rate of improvement for the moderate vs. medium dosage group.

TABLE 7 Final linear mixed-effects model results summary of variable cohort (random effects).

Time period	Predictors	Variance (SD)	Correlation
Random effects			
Week 1–10	Residual	1.45×10^{-2} (1.21×10^{-1})	N/A
	Patient (intercept)	2.57×10^{-3} (5.07×10^{-2})	N/A
	Domain (intercept)	1.70×10^{-3} (4.13×10^{-2})	N/A
	Week: patient (slope)	1.36×10^{-4} (1.16×10^{-2})	0.43
	Week: domain (slope)	1.14×10^{-5} (3.38×10^{-3})	−0.15
Week 11–20	Residual	1.02×10^{-2} (1.01×10^{-1})	N/A
	Patient (intercept)	5.58×10^{-3} (7.47×10^{-2})	N/A
	Domain (intercept)	6.69×10^{-4} (2.59×10^{-2})	N/A
	Week: patient (slope)	6.42×10^{-5} (8.01×10^{-3})	−0.91
	Week: domain (slope)	2.29×10^{-6} (1.51×10^{-3})	−0.50
Week 21–30	Residual	9.08×10^{-3} (9.53×10^{-2})	N/A
	Patient (intercept)	1.62×10^{-2} (1.27×10^{-1})	N/A
	Domain (intercept)	3.29×10^{-4} (1.81×10^{-2})	N/A
	Week: patient (slope)	4.72×10^{-5} (6.87×10^{-3})	−0.98
	Week: domain (slope)	9.73×10^{-7} (9.87×10^{-4})	−0.56

Model equation: domain score (weekly average) ~ 1 + week * dosage group + domain score baseline + age + sex + chronicity + (1 + week:patient) + (1 + week: domain).

TABLE 8 Pairwise comparisons of slopes by dosage amount group (variable cohort).

Time period	Contrast	Estimate (SE)	*t* test	*P* value
Week 1–10	0–15 min/week vs. 15–40 min/week	-2.74×10^{-3} (3.18×10^{-4})	−8.60	***
	0–15 min/week vs. >40 min/week	-5.58×10^{-3} (4.28×10^{-4})	−13.02	***
	15–40 min/week vs. >40 min/week	-2.84×10^{-3} (3.89×10^{-4})	−7.30	***
Week 11–20	0–15 min/week vs. 15–40 min/week	-1.57×10^{-3} (3.51×10^{-4})	−4.48	***
	0–15 min/week vs. >40 min/week	-3.33×10^{-3} (4.58×10^{-4})	−7.27	***
	15–40 min/week vs. >40 min/week	-1.75×10^{-3} (4.32×10^{-4})	−4.06	***
Week 21–30	0–15 min/week vs. 15–40 min/week	-7.29×10^{-4} (4.14×10^{-4})	−1.76	
	0–15 min/week vs. >40 min/week	-2.49×10^{-3} (5.14×10^{-4})	−4.85	***
	15–40 min/week vs. >40 min/week	-1.76×10^{-3} (5.07×10^{-4})	−3.48	**

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1.

4. Discussion

This study aimed to examine if self-managed therapy could be sustained over a long period of time, and if greater average amounts of therapy were associated with greater therapy outcomes. To address these broad questions, therapy practice over a course of a 30-week treatment period (i.e., 6 months) was evaluated for different dosage amounts. Specifically, we evaluated whether greater average weekly dosage—defined as number of minutes per week—led to greater performance gains over a 30-week period in a cohort of consistent users who practice approximately the same average amount week to week. A second analysis examined a larger cohort of variable users, also over a course of 30-week period, to see if performance outcomes at each 10-week period showed relative greater gains for high practice frequency than lower practice frequencies.

There were several main conclusions to be drawn from our study results. Firstly, patients were able to practice consistently for 30 weeks of self-managed therapy and this practice was associated with concurrent improvements in domain scores. Not surprisingly, in this context, users who practiced more than 40 min per week showed greater improvements in the average domain score than users who practiced less than 15 min per week. These results suggest that consistent and sustained practice can result in therapy improvements and that these gains are maintained 20–30 weeks from the therapy onset time. Notably, patients who practiced more variably over a 30-week treatment period likewise demonstrated that greater weekly average dosage amounts were associated with greater improvements in overall domain score. In particular, users who practiced more than 40 min per week showed significantly greater performance gains than users who followed a medium (15–40 min) or low (0–15 min) dosage practice regimen. This was the case in each of the three 10-week intervals of interest, demonstrating that dosage amount matters for therapy outcomes not just in the beginning but also throughout the course of treatment. It should also be noted, as Figure 4 shows, users who practiced more than 40 min per week (Figure 4C) also tended to practice more frequently, with a portion of 65.1% practicing more than 5 days per week and 27.8% practiced every day, compared to less frequent, more massed practice patterns in the medium (15–40 min) (Figure 4B) and low (0–15 min) (Figure 4A) dosage groups.

One notable observation is that by the 21–30 week period, the proportion of chronic patients (greater than 6 months post injury) was higher than in the first 1–10 week period, where they were more acute patients (less than 6 months post injury). This observation was true for both the consistent and variable group analyses. These results suggest that chronic survivors are able to sustain practice over long periods of time (>20 weeks) and demonstrate noticeable improvements on the domain score within the Constant Therapy program.

Results from both consistent and variable user cohorts demonstrated significant gains in domain score across the entire 30-week period of interest in our analyses. In both cohorts, the greatest rates of improvement occurred in the early weeks of therapy but crucially, all users were able to maintain performance gains during later weeks of therapy (e.g., weeks 10–20; 20–30). Moreover, for users following a relatively higher dosage practice regimen, these additional weeks of therapy resulted not only in maintenance of initial gains but in significant additional gains. This result underscores the promise of higher dose therapy to induce gains over a much longer time period than has previously been shown.

In line with prior research, our results show that relatively higher dosage therapy regimens are associated with greater gains in performance as compared to medium or low dosage regimens (6, 38). This study is among a relatively small number of studies to directly compare the effects of varied dose of the same behavioral intervention. The small number of these dose articulation studies has been identified as a major barrier to the

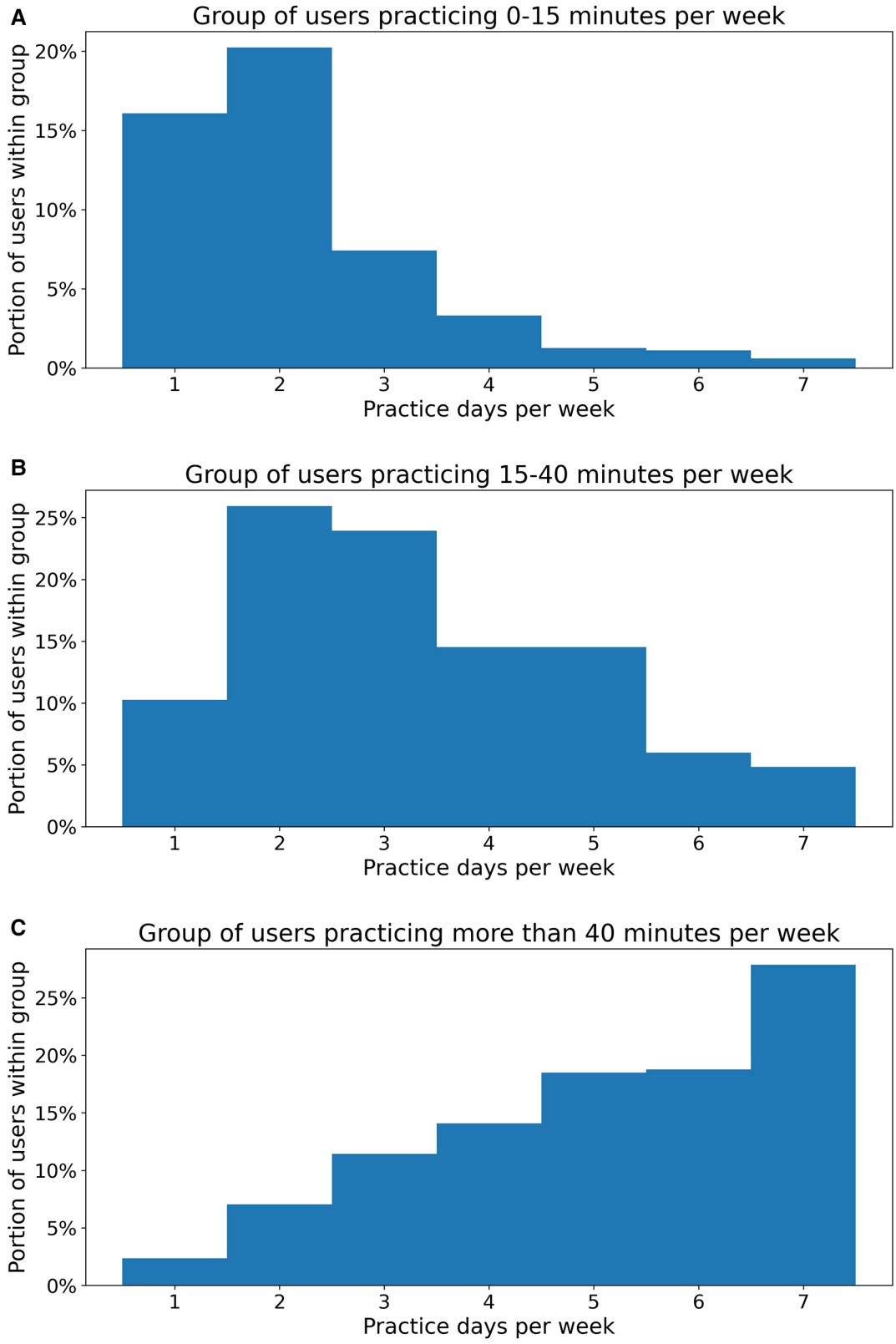


FIGURE 4
Practice frequency distribution of users in different dosage amount groups.

development of optimal dosage guidelines for speech-language pathologists. A recent systematic review found just six studies that reported direct dosage comparisons, and all of these focused on traditional clinician-mediated interventions (39). To our knowledge, only one prior study has investigated the effect of varied dosage on treatment outcomes for a self-managed digital

therapy (23). The current study extends on this finding in critical ways by demonstrating that (1) high-intensity, self-managed SLT leads to significant performance gains over a much more extended therapy time than previously shown (30 vs. 10 weeks) and (2) performance gains are greater for users who practice a greater average amount, regardless of whether they are consistent or more variable in their usage pattern.

The current study also contributes to existing literature through its use of a real-world, ecologically valid dataset. Although the efficacy of high dose speech-language therapy has been established in the literature, there is a gap in translating these research findings to clinical practice. Translation of research findings is complicated by several barriers that include, among others, a large discrepancy in the amount of therapy recommended in research compared to the amount of therapy that is realistically attainable in the clinical setting (15, 18). By analyzing data from two cohorts of patient users who showed natural divergence in the pattern and amount of app-based practice logged over the 30-week time period, we were able to investigate effects of different dosage amounts taking into account the actual amount and types of practice of a large number of real-world users. This ensures greater generalizability of results to the clinical and real-world settings. Our results are encouraging because they not only show that higher intensity (40 min or more per week in our study) is feasible for a sizable group of real-world users but they also show that regardless of your exact pattern of practice (consistent vs. variable; moderate vs. medium vs. low), self-managed SLT leads to significant and sustained performance gains. Our results also demonstrate that higher-intensity therapy may look different in self-managed settings compared to highly controlled laboratory or clinical trial settings. In the latter, weekly dosage prescriptions are very high but total intervention duration is relatively short, whereas in our data, weekly dosage amounts are comparatively more modest but users instead practice for many more weeks (30+ weeks), resulting in cumulative dosage amounts that are comparable to high-intensity regimens as reported in the literature (38). Also important to consider is that CT or other app-based, at-home therapy can be used as an adjuvant to other SLT within the context of patients' longer-term trajectory of recovery; patients may for instance receive direct SLT in early post-acute recovery stages but turn to use of at-home, self-managed therapy after exhausting options for insurance-covered direct SLT. Finally, we note that the data analyzed in this study is the result of entirely self-managed practice, meaning that users were not given explicit instructions on the amount or frequency with which to practice. It is likely that dosage amounts—and possibly also the resultant therapy gains—could be augmented if users were advised on a specific practice regimen.

Importantly, the current study focused on measuring improvement *via* an in-app task improvement measure (i.e., domain score). Though outside the scope of the current study, it will be essential in future work to evaluate the generalizability of in-app domain score improvements to standardized measures of global language severity (e.g., WAB-R Aphasia Quotient), to real-world communication settings and conducted with large numbers of users. Prior clinical studies of the Constant Therapy app have reported clinically significant gains in both global language severity measures and quality of life scores following in-app practice (35, 36). Des Roches et al. found significant pre-post improvements on the WAB-R Aphasia Quotient and composite severity score on the Cognitive Linguistic Quick Test among an experimental group of patients using the CT program as an adjuvant to traditional SLT; no such changes were seen among control participants receiving only traditional SLT (36). Most recently, Braley and colleagues conducted a randomized clinical trial comparing language-based outcomes following digital-only CT therapy compared to traditional SLT. Participants receiving digital-only CT therapy improved 6.75 points on the WAB-R AQ and also demonstrated significant improvement in overall quality of life, as measured by the Stroke and Aphasia Quality of Life Scale 39 (SAQOL-39) (35). Taken together, these findings lend encouraging evidence in support of treatment generalization for CT app users. It is also worthwhile to note that unlike rote paper-and-pencil therapy exercises, CT tasks are functional in nature (e.g., reading a museum map to determine where a given exhibit is), which may make it more likely for in-app improvements to generalize to out-of-app settings.

4.1. Limitations

We note several limitations of the current study. First is the lack of standardized performance metrics to characterize baseline severity and relatedly, the reliance on patient self-report for reporting of demographic and etiological details. The Constant Therapy app makes it possible to collect a large amount of real-world data about users and their daily performance patterns but because it is entirely self-managed, our dataset did not include standardized assessment metrics that might typically be collected in a clinic setting (e.g., Western Aphasia Battery-Revised aphasia quotient). Likewise, we did not have access to detailed information about concurrent medical and cognitive comorbidities, motivation levels, or personality types, all of which have the potential to influence therapy outcomes. Our analysis models do take into account basic demographic information such as age, sex, and chronicity and we also include random effects of patients in all analysis models. For baseline severity, we use the baseline domain score as a proxy measure. Nonetheless, future models with more detailed patient factors would likely lead to more robust and generalizable results.

A second limitation of the current study relates to the way in which users were assigned to their respective dosage groups and the way in which we chose to bin these groups. Users were binned into one of the three dosage amount groups according to their usage patterns and not by random assignment, leading to the possibility for some degree of self-selection into these groups (e.g., more severe users practicing less). To account for this potential effect of severity on results, we included baseline domain score—our proxy for starting severity—as a covariate in all statistical models. We also acknowledge that the current study employs data-informed but clinically

arbitrary cutoffs to determine grouping into low, medium and moderate dosage groups. We therefore are careful to interpret results as providing support for higher vs. lower dose therapy rather than for a specific therapy prescription in minutes (e.g., 40 or min/week).

A final limitation is that there is insufficient information available on whether users had access to other direct therapy services. It is possible that some users may have used the app-based regimen in combination with traditional, in-person SLT, while others may have solely relied on the app. Differences in the amount of outside (i.e., non-app-based) therapy received by users across the dosage groups could potentially affect the results. High dosage app users may also be receiving more outside therapy, making it difficult to attribute any improvement in performance solely to increased in-app practice.

5. Conclusion

This study explored the relationship between the weekly dosage amount that stroke patients practiced in an app-based, self-managed therapeutic program and their performance improvement over a 30-week period. The results showed that across all users, the moderate dosage group (more than 40 min per week) achieved greater performance gains compared to medium (15–40 min per week) and low (0–15 min per week) dosage groups. A similar trend was noted between the medium and low dosage groups. Thus, our results show that performance gains are greater for users who practice a greater average amount. One possible further research direction could be suggesting a new evaluation metric to link in-app performance gains with real-world improvement.

Data availability statement

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

Ethics statement

This project was considered an institutional review board–exempt retrospective analysis by Pearl Institutional Review Board (#17-LNCO-101) under 45 Code of Federal Regulations 46.101 (b) category 2.

Author contributions

All authors contributed to the methodology design. HL performed data filtering and analysis. CC provided some source code for data analysis. CC and SK contributed to data interpretation. PI and MB provided significant help with the overall direction of the research. HL, CC and SK wrote the original draft of the paper while all authors contributed to reviewing the following versions. All authors contributed to the article and approved the submitted version.

Funding

This study is funded by Rafik B. Hariri Institute for Computing and Computational Science and Engineering, Boston University.

Conflict of interest

SK owns ownership stock in Constant Therapy Health, which is the software program described in this study.

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.

Publisher's note

References

1. Berthier ML. Poststroke aphasia. *Drugs Aging*. (2005) 22:163–82. doi: 10.2165/00002512-200522020-00006

2. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart disease and stroke statistics-2022 update: a report from the American heart association. *Circulation*. (2022) 145:153–639. doi: 10.1161/CIR.0000000000001052

3. Lam JMC, Wodchis WP. The relationship of 60 disease diagnoses and 15 conditions to preference-based health-related quality of life in Ontario hospital-based long-term care residents. *Med Care*. (2010) 48:380–7. doi: 10.1097/MLR.0b013e3181ca2647

4. Laska AC, Hellblom A, Murray V, Kahan T, Von Arbin M. Aphasia in acute stroke and relation to outcome. *J Intern Med*. (2001) 249:413–22. doi: 10.1046/j.1365-2796.2001.00812.x

5. Bullier B, Cassoudesalle H, Villain M, Cogné M, Mollo C, De Gabory I, et al. New factors that affect quality of life in patients with aphasia. *Ann Phys Rehabil Med*. (2020) 63:33–7. doi: 10.1016/j.rehab.2019.06.015

6. Brady MC, Kelly H, Godwin J, Enderby P, Campbell P. Speech and language therapy for aphasia following stroke. *Cochrane Database Syst Rev*. (2016) 2016:1–309. doi: 10.1002/14651858.CD000425.pub4

7. Weidner K, Lowman J. Telepractice for adult speech-language pathology services: a systematic review. *Perspect ASHA Spec Interest Groups*. (2020) 5:326–38. doi: 10.1044/2019_persp-19-00146

8. Merlino S. Making sounds visible in speech-language therapy for aphasia. *Soc Interact Video-Based Stud Hum Sociality*. (2021) 4. doi: 10.7146/si.v4i3.128151

9. Beeke S, Beckley F, Johnson F, Heilemann C, Edwards S, Maxim J, et al. Conversation focused aphasia therapy: investigating the adoption of strategies by

people with agrammatism. *Aphasiology*. (2015) 29:355–77. doi: 10.1080/02687038.2014.881459

10. Ferguson A, Armstrong E. Reflections on speech-language therapists' talk: implications for clinical practice and education. *Int J Lang Commun Disord*. (2004) 39:469–507. doi: 10.1080/1368282042000226879

11. Horton S, Byng S. Examining interaction in language therapy. *Int J Lang Commun Disord*. (2000) 35:355–75. doi: 10.1080/136828200410627

12. Laakso M. Collaborative participation in aphasic word searching: comparison between significant others and speech and language therapists. *Aphasiology*. (2015) 29:269–90. doi: 10.1080/02687038.2013.878450

13. Des Roches CA, Kiran S. Technology-based rehabilitation to improve communication after acquired brain injury. *Front Neurosci*. (2017) 11:1–17. doi: 10.3389/fnins.2017.00382

14. Palmer R, Dimairo M, Cooper C, Enderby P, Brady M, Bowen A, et al. Self-managed, computerised speech and language therapy for patients with chronic aphasia post-stroke compared with usual care or attention control (Big CACTUS): a multicentre, single-blinded, randomised controlled trial. *Lancet Neurol*. (2019) 18:821–33. doi: 10.1016/S1474-4422(19)30192-9

15. Adults in Healthcare: Outpatient National Data Report 2012–2016 (2019), p. 51.

16. Adults in Healthcare: Inpatient Rehab National Data Report 2012–2016 (2019), p. 53.

17. National Outcomes Measurement System: Adults in Healthcare: Acute Hospital National Data Report 2012–2016. American Speech-Language-Hearing Association (2019).

18. Cavanaugh R, Kravetz C, Jarold L, Quique Y, Turner R, Evans WS. Is there a research–practice dosage gap in aphasia rehabilitation? *Am J Speech Lang Pathol*. (2021) 30:2115–29. doi: 10.1044/2021_AJSLP-20-00257

19. Skolarus LE, Feng C, Burke JF. No racial difference in rehabilitation therapy across all post-acute care settings in the year following a stroke. *Stroke*. (2017) 48:3329–35. doi: 10.1161/STROKEAHA.117.017290

20. Bhogal SK, Teasell R, Speechley M. Intensity of aphasia therapy, impact on recovery. *Stroke*. (2003) 34:987–93. doi: 10.1161/01.STR.0000062343.64383.D0

21. Kurland J, Liu A, Stokes P. Effects of a tablet-based home practice program with telepractice on treatment outcomes in chronic aphasia. *J Speech Lang Hear Res*. (2018) 61:1140–56. doi: 10.1044/2018_JSLHR-L-17-0277

22. Godlove J, Anantha V, Advani M, Des Roches C, Kiran S. Comparison of therapy practice at home and in the clinic: a retrospective analysis of the constant therapy platform data set. *Front Neurol*. (2019) 10:1–10. doi: 10.3389/fneur.2019.00140

23. Cordella C, Munsell M, Godlove J, Anantha V, Advani M, Kiran S. Dosage frequency effects on treatment outcomes following self-managed digital therapy: retrospective cohort study. *J Med Internet Res*. (2022) 24:1–14. doi: 10.2196/36135

24. Warren SF, Fey ME, Yoder PJ. Differential treatment intensity research: a missing link to creating optimally effective communication interventions. *Ment Retard Dev Disabil Res Rev*. (2007) 13:70–7. doi: 10.1002/mrdd.20139

25. Cherney LR. Aphasia treatment: intensity, dose parameters, and script training. *Int J Speech Lang Pathol*. (2012) 14:424–31. doi: 10.3109/17549507.2012.686629

26. Raymer AM, Kohen FP, Saffell D. Computerised training for impairments of word comprehension and retrieval in aphasia. *Aphasiology*. (2006) 20:257–68. doi: 10.1080/02687030500473312

27. Sage K, Snell C, Lambon Ralph MA. How intensive does anomia therapy for people with aphasia need to be? *Neuropsychol Rehabil*. (2011) 21:26–41. doi: 10.1080/09602011.2010.528966

28. Bakheit AMO, Shaw S, Barrett L, Wood J, Carrington S, Griffiths S, et al. A prospective, randomized, parallel group, controlled study of the effect of intensity of speech and language therapy on early recovery from poststroke aphasia. *Clin Rehabil*. (2007) 21:885–94. doi: 10.1177/0269215507078486

29. Basso A, Caporali A. Aphasia therapy or the importance of being earnest. *Aphasiology*. (2001) 15:307–32. doi: 10.1080/02687040042000304

30. Denes G, Perazzolo C, Piani A, Piccione F. Intensive versus regular speech therapy in global aphasia: a controlled study. *Aphasiology*. (1996) 10:385–94. doi: 10.1080/02687039608248418

31. Baker E. Optimal intervention intensity. *Int J Speech Lang Pathol*. (2012) 14:401–9. doi: 10.3109/17549507.2012.700323

32. Harnish SM, Neils-Strunjas J, Lamy M, Eliassen J. Use of fMRI in the study of chronic aphasia recovery after therapy: a case study. *Top Stroke Rehabil*. (2008) 15:468–83. doi: 10.1310/tsr1505-468

33. Kiran S, Roches CD, Balachandran I, Ascenso E. Development of an impairment-based individualized treatment workflow using an iPad-based software platform. *Semin Speech Lang*. (2014) 35:038–50. doi: 10.1055/s-0033-1362995

34. Fridriksson J, Basilakos A, Boyle M, Cherney LR, DeDe G, Gordon JK, et al. Demystifying the complexity of aphasia treatment: application of the rehabilitation treatment specification systemx. *Arch Phys Med Rehabil*. (2022) 103:574–80. doi: 10.1016/j.apmr.2021.08.025

35. Braley M, Pierce JS, Saxena S, De Oliveira E, Taraboanta L, Anantha V, et al. A virtual, randomized, control trial of a digital therapeutic for speech, language, and cognitive intervention in post-stroke persons with aphasia. *Front Neurol*. (2021) 12:1–16. doi: 10.3389/fneur.2021.626780

36. Des Roches CA, Balachandran I, Ascenso EM, Tripodis Y, Kiran S. Effectiveness of an impairment-based individualized rehabilitation program using an iPad-based software platform. *Front Hum Neurosci*. (2015) 8:1–29. doi: 10.3389/fnhum.2014.01015

37. Kiran S, Gerst K, Dubas E. Abstract TMP48: understanding optimal dosage frequency and patient engagement on improving outcomes using digital therapy. *Stroke*. (2019) 50. doi: 10.1161/str.50.suppl_1.TMP48

38. The Rehabilitation and recovery of People with Aphasia after StrokE (RELEASE) Collaborators, Brady MC, Ali M, VandenBerg K, Williams LJ, Williams LR, Abo M, et al. Dosage, intensity, and frequency of language therapy for aphasia: a systematic review–based, individual participant data network meta-analysis. *Stroke*. (2022) 53:956–67. doi: 10.1161/STROKEAHA.121.035216

39. Harvey SR, Carragher M, Dickey MW, Pierce JE, Rose ML. Treatment dose in post-stroke aphasia: a systematic scoping review. *Neuropsychol Rehabil*. (2021) 31:1629–60. doi: 10.1080/09602011.2020.1786412

frontiers | Frontiers in Digital Health

TYPE Perspective
PUBLISHED 17 April 2023
DOI 10.3389/fdgth.2023.1086219

OPEN ACCESS

EDITED BY
Sabrina R. Taylor,
MedRhythms, Inc., United States

REVIEWED BY
Christina Brezing,
Columbia University, United States

*CORRESPONDENCE
Anthony Watson
anthony.watson@peartherapeutics.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Personalized Medicine, a section of the journal Frontiers in Digital Health

RECEIVED 01 November 2022
ACCEPTED 23 March 2023
PUBLISHED 17 April 2023

CITATION
Watson A, Chapman R, Shafai G and Maricich YA (2023) FDA regulations and prescription digital therapeutics: Evolving with the technologies they regulate.
Front. Digit. Health 5:1086219.
doi: 10.3389/fdgth.2023.1086219

FDA regulations and prescription digital therapeutics: Evolving with the technologies they regulate

Anthony Watson[1*†], Richard Chapman[2†], Gigi Shafai[3] and Yuri A. Maricich[3]

[1]Regulatory Affairs, Pear Therapeutics (US), Inc., Boston, MA, United States, [2]Global Regulatory Affairs – Devices, Sanofi US, Cambridge, MA, United States, [3]Medical Affairs, Pear Therapeutics (US), Inc., Boston, MA, United States

Technological progress in digital therapeutics—and, in particular prescription digital therapeutics (PDTs)—has outpaced the processes that the Food and Drug Administration (FDA) uses to regulate such products. Digital therapeutics have entered the health care ecosystem so rapidly that substantial misunderstandings exist about how they are evaluated and regulated by the FDA. This review briefly explains the relevant regulatory history of software as medical devices (SaMDs) and reviews the current regulatory landscape in which prescription and non-prescription digital therapeutics are developed and approved for use. These are important issues because PDTs, and digital therapeutics in general, are an explosively growing field in medicine and offer many advantages over conventional face-to-face treatments for the behavioral dimensions of a wide range of conditions and disease states. By allowing access to evidence-based therapies remotely and privately, digital therapeutics can reduce existing disparities in care and improve health equity. But clinicians, payers, and other healthcare stakeholders must appreciate the rigor of the regulatory frameworks within which PDTs are approved for use.

KEYWORDS

digital therapeutics, prescription digital therapeutics, food and drug administration, regulatory reform, FDA, software as a medical device (SaMD)

Introduction

As in practically every other sphere of life, medicine in the past two decades has become ever more digitized. At every level, from the conduct of wholly remote digital clinical trials to the use of digital diagnostic tools and, increasingly, the use of digital therapeutics (whether prescription or non-prescription) to treat serious disease states, digital technologies are transforming healthcare. The expanded use of digital therapeutics has been fueled by the ever-increasing prevalence of chronic or difficult-to-treat conditions such as mental health conditions, substance use disorders, insomnia, and lower back pain as well as by acute shortages of providers who are skilled in delivering the behavioral therapy components of care that are so often critical to patient recovery.

But, in part because of the speed with which prescription digital therapeutics (PDTs) have entered the health care ecosystem, some uncertainty exists about how rigorously these devices are evaluated and exactly how they are currently regulated by the Food and Drug Administration (FDA). Torous et al., recently outlined how regulatory approaches can improve innovation in digital devices (1). This perspective expands on this theme, explaining the history and current regulatory landscape in which digital therapeutics are approved for use and will examine the ways the FDA is adapting to this new paradigm.

The evolution of software regulation

The 1976 Medical Device Amendments to the Food, Drug, and Cosmetic Act requires that FDA regulate products intended to diagnose, treat, and/or manage disease. After 1987, FDA was fully aware of the potential role of computers and software in healthcare when it published its "Draft Policy on the Regulation of Computer Products." This document provided guidelines about which software products were regulated as medical devices and which were exempt from regulatory controls such as premarket notifications. Specifically, the guidance stated that the following software was not subject to registration, listing and premarket notification (i.e., FDA authorization): (1) general purpose articles, (2) computer products manufactured by licensed practitioners for use in their practice (3) computer products used in teaching and non-clinical research, and (4) computer products which provide opportunity for competent human intervention. The guidance further stated that the following computer products would require notification to FDA prior to marketing: (1) computer products excluding competent human intervention and (2) substantially equivalent computer products. Computer products that do not meet any of the other criteria would be subject to premarket approval (see section "FDA pathways for digital technologies"). A 1989 draft policy statement, "FDA Policy for the Regulation of Computer Products," reiterated the 1987 draft and was the agency's operational policy for almost 20 years (2).

In the years since, and particularly in the last decade, there has been a proliferation of consumer industry healthcare apps that were never considered in the original software policies. In addition, the sheer volume and diversity of the products and manufacturers has seemingly been daunting to FDA. FDA attempted to clarify the kinds of software it would regulate in the draft Mobile Medical Applications (MMA) guidance document published in 2013 and updated twice in 2019 (3). The 2019 guidance was updated to reflect the issuance of the final rule, "Medical Devices; Medical Device Classification Regulations To Conform to Medical Software Provisions in the 21st Century Cures Act" (86 FR 20278) and the guidance "Clinical Decision Support Software" (referred to as CDS guidance throughout the rest of this document) issued on September 28, 2022. This guidance excluded from regulation certain low-risk software that met the definition of a medical device, although certain quality-related activities were recommended for manufacturers.

FDA and industry seemed to be moving towards a common ground of using a risk-based approach to regulating software. Meanwhile, Congress was watching this evolution closely, and in December 2016, Congress passed the 21st Century Cures Act (4). Among other changes, the Cures Act redefined "medical device" to exclude certain types of software such as medical device data systems (e.g., a device intended to transmit, receive, display, or convert without changing, data from a medical device). Many of the software types excluded from regulation were consistent with FDA's present risk-based approach and aligned closely with the MMA Guidance. The Cures Act also clarified that FDA was not permitted to review parts of a software system that were not regulated, although the boundaries of this system would be dependent on the manufacturer's risk assessment. FDA has since published a flurry of guidance documents attempting to clarify their evolving interpretation of the Cures Act (3, 5, 6). Specifically, these guidance documents attempted to clarify (1) what types of clinical decision support the FDA would and would not regulate, (2) what types of medical device functions FDA would and would not regulate, and (3) what information FDA would expect to see in a submission for regulated medical device functions.

FDA has also been examining the way in which it works with digital health companies. The Digital Health Center of Excellence (DHCoE) was created within the Center for Devices and Radiological Health (CDRH) to lead efforts to catch up with the digital revolution (7). The DHCoE takes a strategic view of digital health devices, i.e., by working broadly with the FDA, other agencies, and external stakeholders to address regulatory approaches as opposed to simply producing guidance documents regarding specific technologies or processes related to approval or product-specific efforts.

The FDA currently regulates digitally-delivered treatments that meet the definition of software as a medical device (SaMD) (8). SaMD is defined by the International Medical Device Regulators Forum as "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device" (9). The "medical purposes" include the diagnosis, mitigation, treatment, or prevention of disease. In the United States, SaMD products are primarily regulated through the traditional approaches used to approve low-to-moderate-risk medical devices (i.e., devices that pose a low-to-moderate risk of harm to patients as a result of using a device).

FDA pathways for digital technologies

SaMD products, like all medical devices, are evaluated for their perceived potential risk to patients and are assigned to one of three classes: Class I (low risk); Class II (moderate risk); and Class III (high risk) (10). Class II devices require general regulatory controls (i.e., broad requirements for provision of information to users), and often special regulatory controls, such as a requirement for clinical data specific to a product in order to provide reasonable assurance of safety and effectiveness or to demonstrate substantial equivalence to a predicate device (11). Class III devices require general controls as well as premarket approval.

Although the first FDA-authorized PDTs were authorized as Class II devices based on their indications (requiring special controls) (11), different digital therapeutics may end up in different risk classes based on their area of treatment (12–14). The Code of Federal Regulations (CFR) lists a variety of regulations regarding computerized therapies that are unique to the diseases that a particular SaMD product is designed to treat. The regulations, therefore, are "fit for purpose." As an example, SaMDs developed for psychiatric disorders follow the requirements for Computerized Behavioral Therapy device for psychiatric disorders (21 CFR 882.5801) (15), while SaMDs for

gastrointestinal conditions are categorized and follow the requirements for Computerized Behavioral Therapy device for treating symptoms of gastrointestinal conditions (21 CFR 876.5960) (16). FDA looks not only at past decisions but considers the specific circumstances involved in each approval.

Importantly, Class II devices generally include special controls. For example, FDA might specify requirements around labeling or clinical data to satisfy questions of safety and effectiveness (11). Some existing PDT authorizations specify the requirement for subsequent products to include clinical data, which are necessary to provide reasonable assurance of safety and effectiveness (12, 13, 17–19).

Once any kind of digital treatment has been evaluated in one or more clinical studies (e.g., randomized controlled trials) the data and formal requests for authorization are submitted via one of two FDA pathways, each with regulatory and evidence-based requirements:

- The *de novo* pathway, which requires clinical data demonstrating the safety and effectiveness of the device (20). Devices authorized via this pathway can then serve as "predicates" for other devices.
- The 510(k) clearance pathway, which requires the submission of clinical data demonstrating substantial equivalence in terms of safety and effectiveness to a predicate product authorized either via the *de novo* or another 510(k) pathway (21).

Both pathways involve the submission of detailed data reports and product descriptions that inform the creation of patient and clinician labeling/instructions for use if the product is authorized.

Work is underway to create a dedicated FDA regulatory framework for SaMD products such as PDTs that reflects the unique attributes of these devices. For example, unlike pharmaceuticals, SaMD products can be frequently updated following FDA authorization, and products relying on artificial intelligence as a component of treatment may "learn" or change how their algorithms perform over time.

In 2017, the FDA announced the Software Precertification Pilot Program, which, it is hoped, will provide more streamlined and efficient regulatory oversight of software-based medical devices developed by manufacturers who have demonstrated a robust culture of quality, organizational excellence, and willingness to monitor their products once they reach the market. Nine companies have participated in the pilot program and have committed to reviewing real-world performance of their products to ensure patient safety and product quality (22).

The proposed approach looks first at the digital health technology developer, rather than solely at the product, which is the current focus of traditional medical device regulations. The new processes seek to accommodate the rapidity with which software products can respond to glitches, adverse events, and other safety concerns. In the Pre-Cert program, the FDA is proposing that software products from authorized companies would continue to meet the same safety and effectiveness standard that the agency expects for products that have followed the traditional path to market. FDA released a final report on this program in September 2022 (23). The report concluded that FDA could implement some changes under present authorities but would need legislative changes to implement others.

Prescription digital therapeutics

Digital treatments, like other therapeutic products, may be prescription or non-prescription. Prescription products require initiation by a licensed healthcare professional, as governed by state-level health authorities. The stipulation for prescription is based on review of the product and a variety of factors by FDA. Prescriptions may be required for the treatment of serious disease, the use of higher-risk devices, the need for a secure diagnosis by a trained clinician, monitoring and follow-up to determine appropriate response, and/or to compare treatment options to determine optimal treatment approaches.

PDTs are software-based treatments delivered on smartphones or tablets that address the behavioral dimensions of many diseases and conditions (8). The first FDA-authorized PDT to make treatment claims was reSET® (to treat patients with substance use disorders) in 2017 (17, 24). This new class of therapy is expanding rapidly, in terms of coverage by payors and overall market size. In January 2022, a Research and Markets analysis valued the 2021 global market for digital therapeutics at $3.35 billion and estimated it would reach $12.1 billion by 2026 (25, 26).

While the first software-based therapeutics were PDTs, non-prescription digital treatments are similar and some of these have received FDA market authorization (14, 24). For example, the non-prescription Natural Cycles (27) software application that lets women track their menstrual cycles was approved via the *de novo* pathway as a Class II device, while another non-prescription menstruation tracker, Clue, was authorized via the 510(k) pathway using Natural Cycles as a predicate device (28).

Unlike health and wellness apps, PDTs specifically treat diseases and, therefore, are regulated by FDA and categorized as Class II devices. Although PDTs, and digital therapeutics in general, are technologically different from traditional medical devices, they are currently reviewed and authorized by CDRH using regulatory pathways and processes that have not always been aligned with the rapid, dynamic, and iterative nature of treatments delivered as software.

In some cases, PDTs may be intended to be used alongside standard of care pharmacotherapy. In such cases, such as reSET-O®, which is intended to be used alongside the pharmacotherapy buprenorphine, both FDA's CDRH *and* the Center for Drug Evaluation and Research (CDER) review and provide input, even if CDRH was the primary review center (12). We are already seeing expansion of drug/software combination products that may be regulated as drugs with CDER as the primary review center and CDRH as the consulting center.

Patient safety, trust, and transparency for public health

FDA has recognized for decades that software is not risk free. Software can result in adverse events, mistreatment, lack of treatment, or other errors across many disease areas (29, 30). It is appropriate, therefore, that FDA regulates software that carries

risk under their risk-based framework to protect public health. Organizations and stakeholders including payers, provider organizations, clinicians, and developers, have a responsibility to their patients to use products that are safe and effective. Maintaining trust and transparency is critical for patients and public health. Developers' compliance with FDA regulations and best practices is critical to maintain trust and transparency, and reduce the risk of harm. The vast majority of consumer medical apps are not regulated by FDA because these products are, presumptively, only intended to help individuals maintain general fitness, health, or wellness, and do not meet the definition of a medical device as defined above. In the authors' opinion, it is important that FDA continue to enforce the line between regulated and unregulated products to protect patients and maintain trust for the benefit of public health (29, 30).

Discussion

The Pre-Cert program and the Digital Health Center of Excellence mentioned previously are examples of the kinds of changes that can create FDA regulatory frameworks aligned with different product types to improve transparency, clinical responsibility, authorization efficiency, and clear labeling for stakeholders. The rise of FDA-regulated digital therapeutics has spurred regulatory evolution and provides experience to support further refinement and richness in FDA regulatory frameworks that balance risk and speed of bringing effective treatments to market, while maintaining public health. FDA, policymakers, research experts, and developers, must work together to make FDA policies related to digital therapeutics as nimble, flexible, and dynamic as the technologies themselves. However, reasonable and flexible regulation only works with responsible enforcement by FDA and compliance by the industry.

Author contributions

RC and AW wrote the first draft of the manuscript, reviewed subsequent versions, and approved the manuscript for submission. GS commissioned the project and outlined the scope, reviewed the manuscript and contributed to substantive revisions. YM provided direction on the manuscript, discussed content with co-authors, and reviewed and approved of manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors thank Stephen Braun, Medical Editor at Pear Therapeutics (US), Inc., for editorial assistance in the preparation of this manuscript.

Conflict of interest

This manuscript was funded entirely by Pear Therapeutics, (US), Inc. The funder had the following involvement with this manuscript: salaries for the employees of Pear Therapeutics (US), Inc. who were involved in this manuscript and payment for all fees associated with submission and/or publication. AW, GS, and YM are employees of Pear Therapeutics (US), Inc. RC is an employee of Sanofi U.S. AW and RC are former FDA employees. Both RC and AW were leaders in the development of medical device software regulatory policy within the FDA and the digital health field.

Publisher's note

References

1. Torous J, Stern AD, Bourgeois FT. Regulatory considerations to keep pace with innovation in digital health products. *NPJ Digit Med.* (2022) 5(1):121. doi: 10.1038/s41746-022-00668-9

2. Federal Register. Medical devices; medical software devices; Notice of public workshop. Published July 15, 1996 Vol. 61, No. 136. (Accessed July 11, 2022).

3. Food and Drug Administration. Policy for device software functions and mobile medical applications; guidance for industry and food and drug administration staff. Available at: https://www.fda.gov/media/80958/download (Accessed July 11, 2022).

4. Food and Drug Administration. 21st century cures act. Available at: https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act (Accessed October 31, 2022).

5. Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. In (2019).

6. Food and Drug Administration. Policy for device software functions and mobile medical applications (2019).

7. Food and Drug Administration. About the digital health center of excellence. Available at: https://www.fda.gov/medical-devices/digital-health-center-excellence/about-digital-health-center-excellence (Accessed August 17, 2022).

8. Food and Drug Administration. Software as a medical device (SaMD). Available at: https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd (Accessed February 16, 2022).

9. International Medical Device Regulators Forum. Available at: https://www.imdrf.org/ (Accessed March 14, 2023).

10. Congressional Research Service. FDA regulation of medical devices. Washington DC (2023).

11. Food and Drug Administration. Class II special controls documents. Available at: https://www.fda.gov/medical-devices/guidance-documents-medical-devices-and-

radiation-emitting-products/class-ii-special-controls-documents (Accessed October 25, 2022).

12. Food and Drug Administration. reSET-O 510k summary. Available at: https://www.accessdata.fda.gov/cdrh_docs/pdf17/K173681.pdf (Accessed February 24, 2021).

13. Food and Drug Administration. Section 510(k) approval letter for somryst, computerized behavioral therapy device for pschiatric disorders, class II. Available at: https://www.accessdata.fda.gov/cdrh_docs/pdf19/K191716.pdf (Accessed July 14, 2020).

14. Food and Drug Administration. DEN 160018 FDA decision summary for reSET. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiqn_C3h9XqAhWhhHIEHUCPCSgQFjAAegQIARAB&url=https%3A%2F%2Fwww.accessdata.fda.gov%2Fcdrh_docs%2Freviews%2FDEN160018.pdf&usg=AOvVaw2PF-Bfdh76gRPUUMIsoxQs (Accessed July 17, 2020).

15. Food and Drug Administration. Computerized behavioral therapy device for psychiatric disorders. Available at: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=882.5801 (Accessed October 27, 2022).

16. Food and Drug Administration. Mahana parallel digital cognitive behavioral therapy (CBT) mobile application for irritable bowel syndrome (IBS) chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/. Available at: https://www.accessdata.fda.gov/cdrh_docs/pdf21/K211372.pdf (Accessed October 27, 2022).

17. Food and Drug Administration, Center for Devices and Radiological Health. De novo classification request for reSET (DEN160018). Washington, DC (2017).

18. Food and Drug Administration. FDA permits marketing of first game-based digital theraeutic to improve attention fuction in children with ADHD). Available at: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-first-game-based-digital-therapeutic-improve-attention-function-children-adhd (Accessed October 27, 2022).

19. Mahana therapeutics obtains FDA marketing authorization for the first prescription digital therapeutic to teat irritable bowel syndrome. Available at: https://www.mahana.com/press/parallel-fda-authorization (Accessed October 27, 2022).

20. Food and Drug Administration. Denovo classification request. Available at: https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/de-novo-classification-request (Accessed March 14, 2023).

21. Food and Drug Administration. Premarket notification 510(k). Available at: https://www.fda.gov/medical-devices/premarket-submissions/premarket-notification-510k.

22. Food and Drug Administration. Digital health software precertification (Pre-Cert) program. Available at: https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program (Accessed August 24, 2022).

23. Food and Drug Administration. The software precertification (pre-cert) pilot program: tailored total prodct lifecycle approaches and key findings (2022). Available at: https://www.fda.gov/media/161815/download (Accessed October 24, 2022).

24. Health Advances Blog. Who's really first in FDA cleared digital therapeutics?. Available at: https://healthadvancesblog.com/2017/11/13/whos-really-first-in-fda-cleared-digital-therapeutics/ (Accessed October 27, 2022).

25. Globe Newswire. Global digital therapeutics market report 2022: analysis and forecasts 2020-2026. Available at: https://www.globenewswire.com/news-release/2022/01/31/2375479/28124/en/Global-Digital-Therapeutics-Market-Report-2022-Analysis-Forecasts-2020-2026-Market-to-Reach-12-1-Billion-by-2026.html (Accessed August 13, 2022).

26. Aungst TD, Franzese C, Yoona K. Digital health implications for clinical pharmacists services: a primer on the current landscape and future concerns. *J Am Coll Clin Pharm*. (2020) 4(4):514–24. doi: 10.1002/jac5.1382

27. Food and Drug Administration. De novo classification request for natural cycles. Available at: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170052.pdf (Accessed October 27, 2022).

28. Food and Drug Administration. Clue birth control 510(k) premarket notification. Available at: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K193330 (Accessed October 27, 2022).

29. Huckvale K, Adomaviciute S, Prieto JT, Leow MK, Car J. Smartphone apps for calculating insulin dose: a systematic assessment. *BMC Med.* (2015) 13:106. doi: 10.1186/s12916-015-0314-7

30. Knox R, Tenenbaum C. Regulating digital health apps needs user-centered reform. STAT (2021). Available at: https://www.statnews.com/2021/08/03/refor-regulatory-landscape-digital-health-applications/ (Accessed October 27, 2022).

frontiers | Frontiers in Digital Health

TYPE Clinical Trial
PUBLISHED 23 May 2023
DOI 10.3389/fdgth.2023.1062471

OPEN ACCESS

EDITED BY
Tim Campellone,
Woebot Labs Inc., United States

REVIEWED BY
Katrina Prior,
The University of Sydney, Australia
Arkers Kwan Ching Wong,
Hong Kong Polytechnic University, Hong Kong SAR, China

*CORRESPONDENCE
Shana A. Hall
✉ shana@limbix.com
Jessica I. Lake
✉ jess@limbix.com

†These authors share first authorship

‡These authors share last authorship

RECEIVED 05 October 2022
ACCEPTED 27 April 2023
PUBLISHED 23 May 2023

CITATION
Kulikov VN, Crosthwaite PC, Hall SA, Flannery JE, Strauss GS, Vierra EM, Koepsell XL, Lake JI and Padmanabhan A (2023) A CBT-based mobile intervention as an adjunct treatment for adolescents with symptoms of depression: a virtual randomized controlled feasibility trial.
Front. Digit. Health 5:1062471.
doi: 10.3389/fdgth.2023.1062471

A CBT-based mobile intervention as an adjunct treatment for adolescents with symptoms of depression: a virtual randomized controlled feasibility trial

Vera N. Kulikov[1†], Phoebe C. Crosthwaite[1†], Shana A. Hall[1*], Jessica E. Flannery[2], Gabriel S. Strauss[3], Elise M. Vierra[4], Xin L. Koepsell[4], Jessica I. Lake[2*‡] and Aarthi Padmanabhan[1‡]

[1]Research Department, Limbix Health, San Francisco, CA, United States, [2]Science Department, Limbix Health, San Francisco, CA, United States, [3]Product Department, Limbix Health, San Francisco, CA, United States, [4]Content Department, Limbix Health, San Francisco, CA, United States

Background: High rates of adolescent depression demand for more effective, accessible treatment options. A virtual randomized controlled trial was used to assess the feasibility and acceptability of a 5-week, self-guided, cognitive behavioral therapy (CBT)-based mobile application, Spark, compared to a psychoeducational mobile application (Active Control) as an adjunct treatment for adolescents with depression during the COVID-19 pandemic.
Methods: A community sample aged 13–21, with self-reported symptoms of depression, was recruited nationwide. Participants were randomly assigned to use either Spark or Active Control ($N_{Spark} = 35$; $N_{Active\ Control} = 25$). Questionnaires, including the PHQ-8 measuring depression symptoms, completed before, during, and immediately following completion of the intervention, evaluated depressive symptoms, usability, engagement, and participant safety. App engagement data were also analyzed.
Results: 60 eligible adolescents (female = 47) were enrolled in 2 months. 35.6% of those expressing interest were consented and all enrolled. Study retention was high (85%). Spark users rated the app as usable (System Usability $Scale_{mean} = 80.67$) and engaging (User Engagement Scale-Short $Form_{mean} = 3.62$). Median daily use was 29%, and 23% completed all levels. There was a significant negative relationship between behavioral activations completed and change in PHQ-8. Efficacy analyses revealed a significant main effect of time, $F = 40.60$, $p < .001$, associated with decreased PHQ-8 scores over time. There was no significant Group × Time interaction ($F = 0.13$, $p = .72$) though the numeric decrease in PHQ-8 was greater for Spark (4.69 vs. 3.56). No serious adverse events or adverse device effects were reported for Spark users. Two serious adverse events reported in the Active Control group were addressed per our safety protocol.
Conclusion: Recruitment, enrollment, and retention rates demonstrated study feasibility by being comparable or better than other mental health apps. Spark was highly acceptable relative to published norms. The study's novel safety protocol efficiently detected and managed adverse events. The lack of significant difference in depression symptom reduction between Spark and

Active Control may be explained by study design and study design factors. Procedures established during this feasibility study will be leveraged for subsequent powered clinical trials evaluating app efficacy and safety.
Clinical Trial Registration: https://clinicaltrials.gov/ct2/show/NCT04524598

KEYWORDS

cognitive behavioral therapy, digital therapeutics, adolescent depression, feasibility, mHealth, mental health

1. Introduction

Depression, a highly prevalent mental health disorder among adolescents, is a growing crisis within the US (1, 2). Depressive episodes and symptoms affect up to 26% of adolescents annually, with depression and suicide rates rising sharply in recent years (1). Adolescent depression has far-reaching consequences including impairments in academic and work performance and social and family relationships, substance use, and exacerbation of other health conditions (3–6). Adolescent depression places significant economic burdens on the US healthcare system, with higher medical costs than those of almost any other adolescent mental health condition (7, 8). The COVID-19 pandemic disrupted the daily lives of adolescents around the globe, and it is estimated that global prevalence of depression symptoms amongst adolescents doubled as a result (9). With the demand for mental healthcare likely to continue increasing in coming years, the development of effective and accessible treatment options, such as digital interventions, is critical to reducing youth depression.

Despite high prevalence rates of depression, up to 80% of adolescents do not receive mental health treatment when necessary (10, 11). There are many reasons that adolescents do not receive adequate mental health care in times of need. First, social stigma surrounding mental healthcare causes adolescents to be hesitant to seek treatment (12). Additionally, limited access to effective mental health care means that those who do seek treatment are often unable to access it in times of need; because there is a nationwide lack of availability of speciality-trained clinicians, especially in rural areas, and mental health providers often get referrals from a variety of sources (primary care physicians, schools, self-referral) (13–15). Cost is also a barrier, with 11% of the population not seeking therapy because it is not covered by insurance, and an even bigger barrier for low-income individuals, with 30% of Medicaid patients reporting cost as an obstacle (16, 17). Finally, individuals who can afford treatment often do not have the time or ability to devote to weekly therapy, due to caregivers' employment commitments, school and after-school activities, or other responsibilities (18).

Digitally-delivered health interventions for mental illness address these barriers by providing private, accessible, cost-effective, and convenient means of treatment that can also increase engagement and self-disclosure due to lessened stigmatization (19–22). Critically, such interventions can serve as a first line of defense for treatment, eliminating wait times to access treatment and reducing high economic costs associated with traditional in-person psychotherapy. They are also available on demand so intervention sessions can be completed at the adolescent's convenience, and can be split into smaller sections of time, which may allow them to more readily fit into a daily routine. Digital treatments *via* mobile application hold particular promise as a widely-accessible treatment for adolescent mental illness– as adolescent smartphone ownership in the United States increased to 95% in 2018 (23). 45% of teens describe their internet use as "near constant" with around 9 in 10 teens reporting that they go online multiple times per day (23). The nearly universal use of smartphones within the U.S., which persists regardless of gender, race, ethnicity, and socioeconomic background, makes it a powerful tool to increase accessibility to mental health interventions (24). Therefore, digital technologies, such as mobile applications, could be leveraged to fill the depression treatment gap.

Cognitive-behavioral therapy (CBT) is a therapeutic approach that can be implemented in the context of digital therapeutics, which "deliver evidence-based therapeutic interventions that are driven by high quality software programs to prevent, manage, or treat a medical disorder or disease"(25). It is used for the prevention and treatment of depression in children and adolescents and is a recommended form of treatment by the American Academy of Pediatrics (26). Digital forms of CBT have been shown to be effective in the treatment of anxiety and depression in youth (27). Behavioral activation (BA), a core CBT skill that has been shown to be effective in conjunction with other CBT skills, like cognitive restructuring, or as a standalone treatment, is an activity performed so that the patient 1) increases engagement with adaptive and contextually relevant activities that induce feelings of mastery or pleasure, 2) advances their personal goals using a combination of motivational strategies, reward-seeking, natural reinforcers, and self-monitoring, and 3) reduces harmful and avoidant behaviors that often manifest during depressive episodes (28). BA-specific therapy is a successful method across multiple durations of treatment for treating depression in adolescents (29). Given that BA is individually paced, self-driven, and self-monitored, it can be easily delivered digitally, which may be appealing to depressed youth who have limited access to or lack of interest in traditional care. Recent evidence suggests that behavioral aspects of CBT are as effective as cognitive approaches in reducing depressive symptoms in youth and may mechanistically drive symptomatic reduction in CBT (30–32). A digital BA program for adolescent depression represents an exciting new direction for treatment. BA is a component of CBT treatment that emphasizes the connection between mood and behaviors. It has been shown to be successful when used in conjunction with other CBT skills, such as cognitive restructuring, but also when used as its own treatment, particularly for adolescents (33–36).

Digital applications of CBT are well supported as a comparable and effective alternative to traditional CBT (37). Computer-based CBT has been associated with significant effects on symptoms of depression in adolescents and growing evidence supports self-guided, smartphone based-apps as a promising treatment option for depression (38). While digital mental health interventions are an effective way to increase accessibility to proper mental health care, there remains a lack of digital treatment options for adolescents. To our knowledge, there are no digital therapeutics designed to treat adolescent depression approved by the FDA and the current study is the first feasibility trial for a digital therapeutic in adolescents. This digital BA program was designed to address the need for both accessible and evidence-based treatment for adolescents amidst a growing mental health crisis. The current research aimed to investigate the feasibility of a novel CBT-based mobile-app to treat adolescent depression.

This feasibility study was initiated during the COVID-19 pandemic as a means to provide accessible mental health resources to adolescents. The purpose of this randomized controlled trial (RCT) was to assess the feasibility and acceptability of a 5-week, self guided CBT-based mobile app program primarily focused on BA (Spark v2.0, hereafter referred to as Spark), compared to an active psychoeducational control condition (Active Control) for an adjunct treatment of adolescents with symptoms of depression. Study's primary aims included evaluating (1) study feasibility, based on recruitment rate, enrollment rate, and retention rate of participants, (2) acceptability of the app for the target population, based on usability (as evaluated by Systems Usability Scale [SUS] and post-intervention questionnaire responses) and engagement (as evaluated by the User Engagement Scale—Short Form [UES-SF]) and (3) the feasibility of a novel protocol for monitoring participant safety during a fully decentralized virtual clinical trial of a digital intervention, based on the rate of total number of clinical concerns identified in each group. A fourth (4) aim, considered a secondary aim, was to evaluate the preliminary evidence of clinical efficacy, exploring the differences in PHQ-8 score for each group over time, differences between groups in additional aspects of mood and health (Mood and Feelings Questionnaire [MFQ], Patient Reported Outcomes Measurement Information System—Pediatric [PROMIS—Pediatric], General Anxiety Disorder -7 [GAD-7] and Brief Resilience Scale [BRS]), and safety, determined by measuring the number of ADEs, SAEs, and UADEs identified in each group. The current study hypothesized that leveraging engaging mobile technologies would result in high treatment engagement, and preliminary evidence of clinical efficacy.

2. Materials and methods

2.1. Eligibility

Participants were eligible for the study if they 1) were between the ages of 13 and 21; 2) had self-reported symptoms of depression; 3) were residing in the USA for the duration of the 5-week study; 4) were under the care of a US-based primary care and/or licensed mental healthcare provider and willing to provide their provider's contact information (to contact them in case of a concern for participant safety); 5) were fluent and literate in English and had a legal guardian (if under 18 years of age) who was fluent and literate in English; 6) had access to an eligible smartphone (ie. one capable of downloading and running the digital therapeutic, meaning a iPhone 5s or later or running Android 4.4 KitKat or later); 7) had regular internet access (i.e., access to internet either within their home, school environment or other locations on a daily basis, with no planned time without regular internet access during the intervention period); and 8) were willing to provide informed e-consent/assent and had a legal guardian willing to provide informed e-consent (if under 18 years of age). The criteria that required participants to be under the care of a US based primary care and/or licensed mental healthcare provider was included to 1) evaluate the feasibility of the Spark app as an adjunct treatment for depression, and 2) to manage participant safety.

Participants were ineligible if they self-reported 1) a lifetime suicide attempt, 2) active self-harm, 3) active suicidal ideation with intent, or 4) a prior diagnosis by a clinician of bipolar disorder, substance use disorder, or any psychotic disorder including schizophrenia, or 5) if they were incapable of understanding or completing the study procedures or the digital intervention as determined by the participant, legal guardian, healthcare provider, or the clinical research team.

If participants were under the age of 18 and not determined to be legally emancipated, legal guardians were required to be involved in study procedures, including taking part in the initial onboarding session, providing consent, completing weekly questionnaires and receiving study correspondence when necessary.

Of note, the age range of 13–21 for study recruitment presents the variable adolescent period across individuals and is generally thought to extend through the second decade and into the third decade of life, roughly defined by the onset and completion of pubertal maturation as well as other psychosocial, socio-emotional, and cultural factors (39, 40). In the context of medical devices, including digital therapeutics, the US Food and Drug Administration (FDA) defines adolescence as between the ages of 12 and 21. Depression is also highly prevalent across this entire age range (41). As such, the goal of the current study was to assess feasibility of Spark as a digital therapeutic adjunct treatment for adolescent depression symptoms in this age range. We did not include those who were 12 years old due to Children's Online Privacy Protection Act (COPPA) restrictions for mobile applications in children under the age of 13.

2.2. Procedures

2.2.1. Participant recruitment

Participants were recruited *via* online paid advertising on social media platforms, such Facebook and Instagram, and word of mouth. Paid advertisement campaigns were targeted towards 13–21 year-olds and the legal guardians of 13–17 year olds who

were located within the US and English-speaking. After seeing and clicking on an advertisement, participants and/or legal guardians were directed to a landing page where they received an overview of the study and reviewed the presented eligibility criteria. If they determined themselves or their child eligible, they clicked on a link to schedule a consent appointment.

No formal power calculations were conducted to determine sample size. A target sample size of sixty was determined to be sufficient to evaluate feasibility, usability, and preliminary evidence of efficacy (42). This target sample size accounted for a predicted attrition rate of 20%–30% based on previous studies of digital CBT-based interventions for adolescent mental health (43–45). Recruitment was completed in two months, beginning July 23 2020, and ending on September 29 2020.

2.2.2. Consent and Pre-intervention

This study was reviewed and approved by the Western Copernicus Group (WCG) Institutional Review Board (IRB) (ethical approval ID: WIRB® Protocol #20201686) with an abbreviated investigational device exemption for non-significant risk devices and was registered on clinicaltrials.gov (NCT04524598). This study was Phase I in two phases of clinical testing. In Phase II, a larger-scale RCT was conducted to evaluate the efficacy and safety of Spark, following product updates made as a result of Phase I study findings. These results will be reported elsewhere. The consent and onboarding process was completed *via* video conferencing, using the HIPAA-compliant Google Meet video-communication service, between a clinical research coordinator, the participant, and the participant's consenting guardian (if under 18). All participants provided written electronic informed consent, if over the age of 18, or assent, if under the age of 18. Written guardian informed consent was obtained from those under 18 years old.

After providing informed consent, participants and legal guardians were screened for eligibility, which involved the coordinator reviewing the criteria and the participant verbally confirming their eligibility. If the participant was under 18 years old, legal guardians were asked to leave the room while participants confirmed eligibility in order to provide the participant with a private setting to discuss sensitive topics, including self-harm and suicide/suicidal ideation. Afterwards, legal guardians returned to confirm their child's eligibility. Following the standard practice for health care providers, the research coordinators informed all participants about the limits of confidentiality, including the circumstances in which information related to safety risk would be shared with others. In clinical work with minors under the age of 18, these discussions involve what information will be shared with legal guardians. It is expected that information related to potential safety risk of minors would be shared with legal guardians so that appropriate services could be sought. We therefore expect a similar level of accuracy in reporting self-harm or suicide/suicidal ideation as what would occur in standard practice. Participants that met eligibility criteria during the onboarding session then used a web portal to fill out baseline questionnaires, including the Patient Health Questionnaire-8 (PHQ-8) (46), which measures symptoms of depression (see Questionnaires below). Baseline questionnaires took approximately 10–20 min to complete. Participants that met eligibility criteria were randomly assigned to the Spark or the Active Control group with a 1:1 ratio, using a fully random algorithm for randomization. Participants were guided by the coordinator to download the app and create an account. Once the participant logged in, they saw whether they had been randomized to Spark or the Active Control. Neither participants nor study staff were blinded to the assigned study condition. Participants and legal guardians were also provided with mental health resources and a safety plan (47) that could be completed in their own time.

2.2.3. Five week intervention

Participants in both Spark and Active Control groups had access to their assigned app for a 5 week intervention period. All participants completed two weekly questionnaires in the app: 1) the PHQ-8 about their depression symptoms, and 2) an adverse events questionnaire (AEQ) about their safety (see Questionnaires below). These questionnaires took approximately 10–20 min to complete. Automated app notification reminders to complete these questionnaires were sent to participants. Legal guardians completed an AEQ on a weekly basis *via* a web portal. Both participants and legal guardians had access to their weekly questionnaires for seven days. Reminders were sent the day after the participant or legal guardian did not complete a weekly set of questionnaires, with a warning that participants would be withdrawn if they did not complete the AEQ questionnaire due to being unable to monitor their safety. If a participant did not complete the weekly questionnaires two weeks in a row, they were emailed that they will be withdrawn from the study. Both emails were templated.

2.2.3.1. Spark group

The treatment intervention, Spark (v2.0), was a 5-level, interactive program. Our program was modeled on evidenced based treatment (EBT) protocols for behavioral activation (35, 48–52), particularly for adolescents. Following those EBTs, we retained the same therapeutic ingredients: 1) an introduction to the BA model 2) getting active and charting progress (including focus on BAs, tracking mood and behavior, and identifying activities that align with users values), 3) skill building and addressing barriers and avoidance (includes sessions on problem solving, goal setting, and identifying barriers that can get in the way of accomplishing goals), 4) practice (includes practice and consolidation of skills), 5) moving forward/planning for continued activation (includes review of treatment gains, and relapse prevention strategies). This version of our intervention built upon the previous version of the app called Spark (v1.0) (53). User experience data from post-study interviews, from a previous study of an earlier version of Spark, was used to inform the design of the version of the Spark app used in this study. Levels in the app progress in a linear fashion; participants had to complete each task before they could progress onto the next task. Each level was designed to take less than 60 min and participants were recommended to complete one level per week, though they could progress at their own

pace. Participants were guided through the program by a character called "Limbot." This character encourages the user to complete the program and provides personal examples of how they have undertaken behavioral activation therapy. In level 1, participants completed onboarding and learning tasks. During onboarding, they received a tutorial on the app interface and a description of the BA program. The first learning task included information about the behavioral (BA) model of depression, focusing on the relationship between mood and behavior, and how it can lead to a downward cycle of depression. Next, participants learned about breaking the cycle of depression by changing behavior. They received information about how completing activities that align with their values can help the activities be more effective at improving their mood. Participants identified values that were important to them (54). At the end of lesson 1, participants were taught how to schedule activities centered around their previously identified values and were given a walkthrough tutorial of the activities tab. Level 2 through Level 5 focused on activity scheduling and review. Participants were asked to schedule activities within the app and then complete those

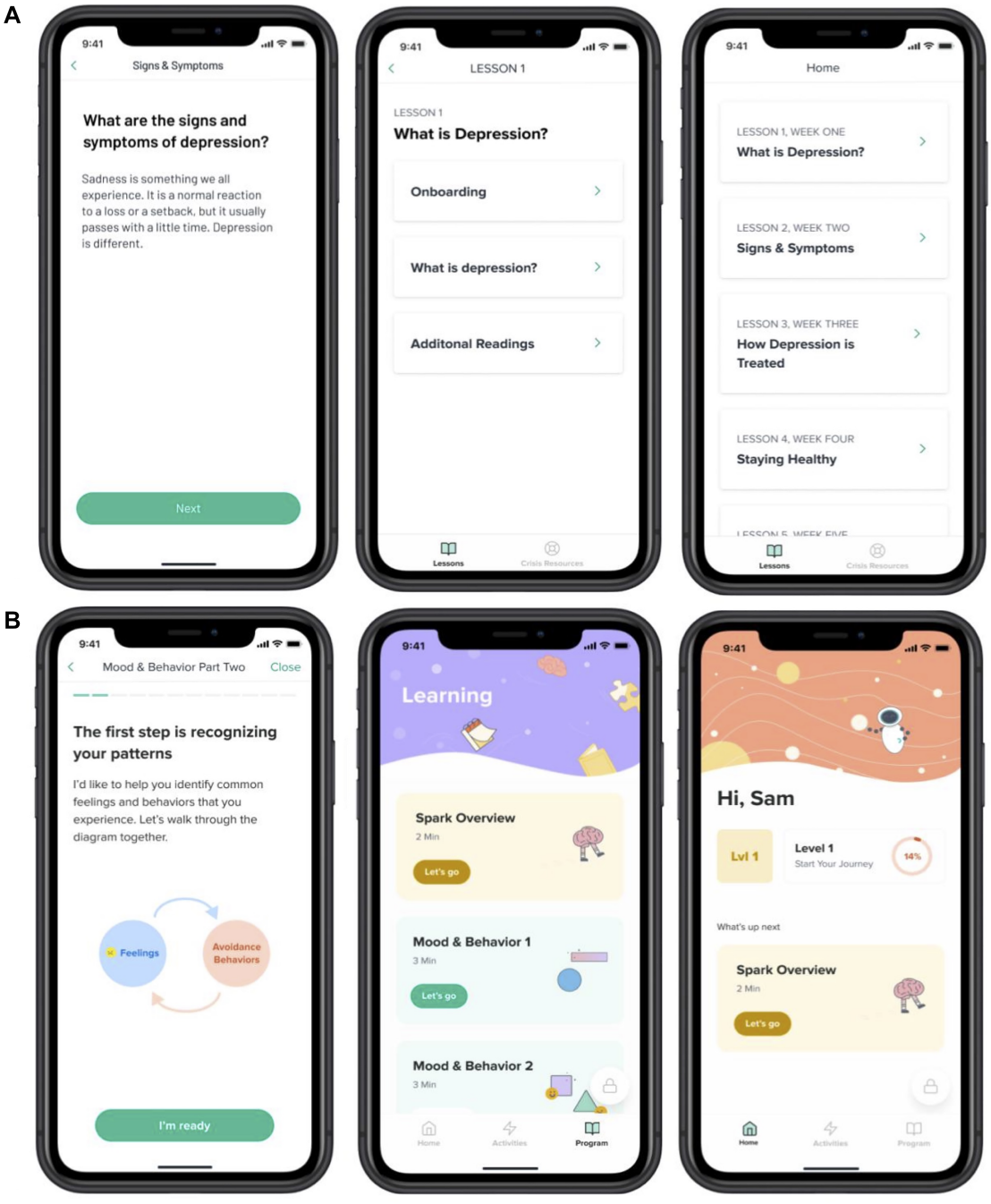


FIGURE 1
Examples of screens from the Active Control (A) and Spark (B) apps.

activities outside of the app. Participants were encouraged to log into the app and reflect on the activity that they completed, answering questions about how the activity aligned with their selected values (Lesson 1) and how it made them feel. If participants did not complete their scheduled activity, they were asked questions that encouraged them to reflect upon the roadblocks they encountered and how they can combat them in the future. At the end of each level, participants received acknowledgement from the Limbot character and learned about the goal for the next level. Crisis resources could be accessed in the app at any time. See **Figure 1** for an illustration of the app interface.

2.2.3.2. Active control group

The Active Control was an app containing educational content related to symptoms and treatments for depression, healthy habits and resources. The content was largely based on the NIMH Teenage Depression ebook (55). It did not include CBT or BA components. Participants did not have the ability to enter free form text in the app. The Active Control was designed to be similar to Spark in duration, and modality of delivery and contained five lessons. Content in the Active Control app was not gated; it was possible to access later lessons without having reviewed earlier lessons. See **Figure 1** for an illustration of the interface.

2.2.4. Post-Intervention

After the 5-week intervention period, participants and their legal guardian were emailed links to complete post-intervention self-report assessments, which took approximately 10–20 min to complete. Participants and their legal guardian received reminders to complete their assessments if they did not complete the questionnaires after one week of being granted access. These emails were templated. Participants who did not complete the post-intervention assessments within 4 weeks from the end of the intervention period lost access to their assessments at that time and were considered lost to follow up. Participants were compensated $25 in the form of an electronic gift card for completing the post-intervention assessments regardless of app usage.

2.2.5. Post-Intervention interviews

Select participants and legal guardians were invited to participate in 1 hour interviews for product feedback. Participants were selected to take part in these interviews based on different factors including age, geographic location, and level of app engagement. Participants were compensated $25 in the form of an electronic gift card for participating. These data are out of scope for this manuscript and are not discussed further.

2.3. Safety protocol

During the study period, trained study staff followed a rigorous safety protocol with study PI and clinician oversight. Clinical concerns that arose at any time during the study were logged. Clinical concerns were defined as any potentially concerning information reported during the trial that indicated a potential risk to health in the past, present, or future, or that signaled abuse. Clinical concerns were identified through four channels:

- Text entered within the Spark app identified by a research coordinator as concerning (defined by the safety protocol)
- Deterioration of symptoms of depression, defined as a PHQ-8 score ≥ 15 (moderately severe or higher) (46) and a ≥ 5 point increase from baseline (56)
- Text in any questionnaire identified by a research coordinator as concerning
- Spontaneously reported harm by participants or legal guardians, including self-harm or abuse, during direct communication with study staff or *via* email

Any clinical concern identified during the study triggered the safety protocol, regardless of severity. The safety protocol dictated that, during the onboarding session, if a participant indicated that they were in immediate distress or danger, the study coordinator would direct them towards emergency services (e.g., the nearest emergency room or calling 911). Otherwise if a clinical concern was identified in an asynchronous context, or during the onboarding session but did not require immediate referral to emergency services, it was escalated to the study investigator. Study investigators reviewed mild concerns weekly and moderate concerns within 24 h, along with any other relevant information or safety data. The study investigator would determine whether the clinical concern required escalation to the study clinician based on criteria established in the safety protocol and within 48 h the study investigator would determine whether the participant was safe and eligible to continue with the study, consulting with the study clinician as needed. If the safety concern was related to suicidality, the study investigator or clinician was trained to administer the Ask Suicide-Screening Questions (ASQ) toolkit (57). If the study clinician determined that the participant was no longer eligible to continue with the study, or if the clinician could not monitor safety due to not being able to reach the participant or other listed contacts, the participant would be informed, withdrawn from the study, and sent mental health resources. Participants were also withdrawn from the study if they did not complete the weekly Adverse Event Questionnaire for two consecutive weeks. (Note: this procedure was implemented in the second month of enrollment, as during this virtual and decentralized RCT we were otherwise unable to determine participant safety).

After study completion, an internal clinician who was not otherwise involved in the study, reviewed all clinical concern data. Those that the clinician judged to be potential adverse events were sent to an external clinician. These clinical concerns, along with accompanying relevant safety data, were classified as relevant as adverse events (AE), adverse device effects (ADE), serious adverse events (SAE), and unanticipated adverse device

TABLE 1 Definitions for external clinician categorization of adverse events (AEs).

Adverse Event	**An adverse event (AE)** is an untoward medical occurrence, unintended disease or injury, or untoward clinical signs (including abnormal laboratory findings) in subjects (3.50), users or other persons, whether or not related to the investigational medical device (3.29) and whether anticipated or unanticipated. Note 1 to entry: This definition includes events related to the investigational medical device or the comparator (3.12). Note 2 to entry: This definition includes events related to the procedures involved.
Serious Adverse Event	**Serious Adverse Events/Serious Adverse Device Effects**: An adverse event or adverse device effect is considered serious if it meets any of the following criteria: • Is fatal; • Is life-threatening, meaning, the participant was, in the view of the investigator, at immediate risk of death from the reaction as it occurred; • Leads to persistent or significant disability/ incapacity, i.e., the event causes a substantial disruption of a person's ability to conduct normal life functions; • Requires or prolongs inpatient hospitalization; • Is an important medical event, based on appropriate medical judgment, that may jeopardize the participant, or the participant may require medical or surgical intervention to prevent one of the other outcomes above. Note 1: Planned hospitalization for a pre-existing condition, or a procedure required by the CIP (3.9), without serious deterioration in health, is not considered a serious adverse event. Note 2: Serious adverse device effect (SADE): adverse device effect that has resulted in any of the consequences characteristic of a serious adverse event.
Adverse Device Effect	**An adverse device effect (ADE)** is an adverse event related to the use of an investigational medical device. This includes any adverse event resulting from insufficiencies or inadequacies in the instructions for use, the deployment, the implantation, the installation, the operation, or any malfunction of the investigational medical device. This also includes any event that is a result of a user error or intentional misuse. Note: For this study, ADEs may occur in either the Spark or Active Control arms.
Unanticipated Adverse Device Effect	(UADEs, as defined in 21 CFR 812.3, also referred to as "Unanticipated Problems"): Any serious adverse effect on health or safety or any life-threatening problem or death caused by, or associated with, a device, if that effect, problem, or death was not previously identified in nature, severity, or degree of incidence in the investigational plan or application; OR Any other unanticipated serious problem associated with a device that relates to the rights, safety, or welfare of subjects.

effects (UADE) (58–60). Definitions used for adverse events classification can be found in **Table 1**.

2.4. Questionnaires

Different measures were used to assess the characteristics of the study population, general mood, depression and anxiety symptoms, and overall health. All questionnaires were delivered to both Spark and Active Control users. The schedule of assessments can be referenced in **Table 2**.

TABLE 2 Baseline and post-intervention assessments for participants and legal guardians were completed *via* a secure web portal. Weekly participant assessments were completed in the mobile app. Weekly parent assessments were completed *via* a secure web portal.

	Baseline	Weekly during the 5-week intervention	Post-intervention
Patient Health Questionnaire (PHQ-8)*	X	X	X
Baseline Questionnaire-Participant*	X		
Baseline Questionnaire-Parent*	X		
Brief Resilience Scale (BRS)	X		
Generalized Anxiety Disorder (GAD-7)*	X		X
PROMIS Pediatric Global Health Scale*	X		X
PROMIS Parent Proxy Global Health Scale	X		X
Mood and Feelings Questionnaire (Short Parent Version)*	X		X
Adverse Events Questionnaire-Participant*		X	X
Adverse Events Questionnaire- Parent*		X	X
Post-intervention Questionnaire-Participant*			X
Post-intervention Questionnaire- Parent*			X
System Usability Scale*			X
User Engagement Scale —Short Form*			X

*Indicates Questionnaires that were reported in this manuscript.

2.4.1. Baseline demographics questionnaire

The Baseline Demographics Questionnaire was an internally developed questionnaire that included demographic questions in regards to the adolescent participant's gender (i.e., male, female, or gender non-binary), ethnicity, race, and age, questions about prior and current treatment for depression and other mental health disorders. Choice questions, with answer choices of "yes" or "no" were used to evaluate whether the participant had been diagnosed with depression or any other mental health, cognitive, or developmental disorder, followed by a free-form text field asking for details about any disorder, besides depression, with which they had been diagnosed. A multi-select choice question was used to evaluate previous or concurrent treatment for depression, with a free-form text field provided if the participant selected "Other" for forms of treatment. A free-form text field was also provided, asking the participant to list all medication they were taking when beginning the intervention. Separate versions of the baseline demographics questionnaire were completed by participants and legal guardians, where legal

guardians completed questions about their education level, and their child's demographics, diagnosis and treatment.

2.4.2. Patient health questionnaire (PHQ-8)

The PHQ-8 consists of eight descriptive phrases of depressive symptoms (61). Participants rated how often they were bothered by any of those symptoms over the last fortnight; (0) Not at All; (1) Several Days; (2) More than Half the Days; (3) Nearly Every Day. Possible scores ranged from 0 to 24, with a higher score indicating more severe depressive symptoms. This assessment was delivered at baseline, weekly during the 5-week intervention and post-intervention. Only the participant completed the PHQ-8. Participants had a full week to complete each weekly PHQ-8 in app after the baseline PHQ-8. Participants had one month to complete the post-intervention PHQ-8. The PHQ-8 is a well established measure to both diagnose and assess the severity of depressive disorders (62). Evidence supports the high internal reliability of the PHQ-8 (Cronbach's $\alpha = .89$) and its high construct validity, with the PHQ-8 score correlating strongly with patient mental health (.73) (46).

2.4.3. Adverse event questionnaire (AEQ)

The AEQ was an internally developed questionnaire that assessed consenting guardian- and participant-reported clinical concerns. Participants and legal guardians were asked to rate clinical concerns in terms of severity, on a scale of (0) Not at all to (4) Extremely, to provide the start and stop date (if applicable), and to indicate whether they believed the reported concern was related to study intervention. This assessment was delivered during the 5-week intervention and at post-intervention. Separate versions of the AEQ were completed by the participant and legal guardian.

2.4.4. Post-intervention questionnaire

The post-intervention questionnaire was developed internally and administered at post-intervention including questions about current treatment for depression and other mental health disorders and any changes in treatment since baseline. The questionnaire also asked whether participants and legal guardians thought the program helped them, and questions evaluating participant experience using the program as a whole. Mood improvement was captured through the following question for participants: "How much do you feel like this mobile app improved your symptoms of depression?" and for parents: "How much do you feel like this mobile app improved your child's symptoms of depression?". Respondents indicated their response using a 10 point scale (0 = Didn't improve at all, 5 = Moderately Improved, 10 = Improved Completely). Participants and legal guardians completed different versions of the post-intervention questionnaire.

2.4.5. The system usability scale (SUS)

The SUS is a validated scale used to assess the usability of a system originally developed by Brooke (63). It was modified for use in this study to evaluate app usability at post-intervention. It consisted of 10 questions about how easy it was to use the app (63, 64). Responses are given on a 5-point Likert scale from (0) Strongly Disagree to (4) Strongly Agree. Item responses are summed and multiplied by 2.5 such that final scores range from 0 to 100. A score above 68 is considered above average. Only the participant completed the SUS. The SUS is supported as an easy to administer yet highly reliable method (Cronbach's $\alpha = 0.911$) for measuring the usability of a product (65).

2.4.6. The user engagement scale short form (UES-Sf)

The UES-SF has 12 questions about how engaging participants found the app (66) and was delivered post-intervention. Responses are given on a 5-point Likert scale from (1) Strongly Disagree to (5) Strongly Agree. Item responses are averaged across all questions to generate a general engagement score ranging from 1 to 5. Only the participant completed the UES-SF. Data supports the UES-SF as a statistically reliable scale that can effectively estimate full UES scores (66).

2.4.7. Generalized anxiety disorder 7-item scale (GAD-7)

The GAD-7 is a brief seven-item self-report measure of anxiety. The scale has been found to be reliable and valid (67), and was used to evaluate changes in anxiety given the high comorbidity between anxiety and depression. The GAD-7 scale was delivered at baseline and post-intervention. This assessment was delivered at baseline and post-intervention. Only the participant completed the GAD-7.

2.4.8. PROMIS pediatric global health scale & PROMIS parent proxy global health scale

These are 9-item measures that produce essentially a unidimensional measure of global health perception/well-being[3]. The PROMIS Parent Proxy Global Health Scale was written parallel to the PROMIS Pediatric Global Health Scale to allow consenting guardians to report on the perceived global health/well-being of their child. These scales are supported as a brief and reliable method to measure the global health status of children (68, 69). Both scales start with 4 descriptive phrases paired with scale of 5–1, asking the user to evaluate different aspects of their global health perception/well-being; (5) Excellent, (4) Very Good, (3) Good, (2) Fair, (1) Poor, followed by 3 questions with descriptive phrases paired with a scale of 5–1; (5) Always, (4) Often, (3) Sometimes, (2) Rarely, (1) Never; and two final phrases paired with a scale of 1–5, (1) Never, (2) Almost Never, (3) Sometimes, (4) Often, (5) Almost Always. Possible scores ranged from 0 to 24, with a higher score indicating a lower quality of life. The PROMIS scales were delivered at baseline and post-intervention. The consenting guardian completed the PROMIS Parent Proxy Global Health Scale 7 + 2 and the participant completed the PROMIS Pediatric Global Health Scale.

2.4.9. Mood and feelings questionnaire short parent version (MFQ-Ps)

The MFQ-PS was used to record change in parent-reported depressive symptoms. The MFQ consists of 13 descriptive phrases paired with scales rated 0–2; (0) True, (1) Sometimes, (2)

Not True. Possible scores range from 0 to 26, with a higher the score indicating the higher the likelihood the child is suffering from depression, as reported by a consenting guardian. The MFQ-PS was delivered at the baseline and post-intervention. Only the consenting guardian completed the MFQ-PS. This scale is supported as a brief and reliable method of evaluating depressive symptoms (70).

2.4.10. Brief resilience scale (BRS)

The BRS is a 6 item self-report measure for assessing the ability to "bounce back" or recover from stress. It has been shown to be reliable and to measure a unitary construct (71). The BRS was delivered at the baseline. Only the participant completed the BRS.

A description of an additional exploratory questionnaire (COVID questionnaire) administered during the study can be found in the **Supplementary Materials**.

2.5. Analysis

2.5.1. Participant characteristics and feasibility outcomes

Participant characteristics were evaluated per study arm and for the full study sample. Chi-squared tests and two-sample *t*-tests were used to evaluate significance of any group differences, as appropriate. Study feasibility was evaluated as 1) recruitment rate: the proportion of those who scheduled an onboarding session out of those who expressed interest in the study, 2) enrollment rate: the proportion of participants enrolled in the study out of those who scheduled an onboarding session and 3) retention rate: the proportion of those who completed the post-intervention survey out of those who enrolled in the study. Microsoft Excel was used to analyze these data.

2.5.2. App acceptability: usability and engagement

App acceptability consists of app usability and engagement. Usability was collected *via* the SUS and post-intervention questionnaire. Exploratory comparisons of SUS, post-intervention questionnaire, and UES-SF scores were conducted between the Spark and Active Control groups using two-sample *t*-tests. Spark app engagement was collected *via* self-report, the UES-SF, and app usage data. App usage data included: (1) the percent of daily active users who used Spark on each intervention day, along with the median percent of daily active users across the full intervention period; and (2) the percent of Spark participants who completed each of the five levels of Spark, along with the percent of participants who completed behavior activation activities. Daily active use was defined as opening the app for any duration. Descriptive statistics are reported for app usage. Finally, a correlation was run to examine the relationship between post-intervention and baseline PHQ-8 scores and the number of behavioral activation activities completed. Microsoft Excel was used to analyze these data.

2.5.3. Study safety protocol feasibility

The number of total clinical concerns identified in each group was evaluated. We used free-form text to identify clinical concerns in the Spark group. We note that the Active Control group did not have the ability to enter free-form text into the app. Therefore, we report descriptive statistics about the total number of clinical concerns captured for each group without direct comparison. We report the sources of clinical concerns, the number of participants that had clinical concerns escalated to the study clinician, and the number of participants that had clinical concerns that elicited clinician reachout to the participant. The feasibility of capturing clinical concerns through a variety of sources and of managing safety concerns in a fully virtual setting was evaluated. Microsoft Excel was used to analyze these data.

2.5.4. App efficacy and safety

Differences in PHQ-8 scores for each group over time were explored. Multiple imputation was used to account for missing data points, excluding participants with only baseline scores. First, analyses were conducted to determine if data were missing completely at random and whether patterns of missing data differed between groups. Little's test (72) was used to determine whether data were missing completely at random and a chi-square test was conducted to identify whether there were significant differences between groups in the proportion of missing data across weeks. Because participants had seven days to complete each weekly PHQ-8, the assumption that spacing between the six timepoints was consistent across time and groups was evaluated using a generalized linear mixed-effect model (GLMM) with a 2-level PAN method (73) with numbers of days since baseline PHQ-8 completion as the dependent variable. Main effects of Group (Treatment vs. Active Control) and Week (six timepoints) were analyzed along with the Group × Week interaction. Finally, to test for group differences in the change in PHQ-8 scores over time,an exploratory GLMM was conducted using a 2-level PAN method and examined the main effects of Group, Week, and the Group × Week interaction. Days between successive PHQ-8 completions was included as a random-effect for the slope at the individual level to control for irregular spacing between questionnaire completion timepoints. Random effects also included a participant-level intercept. As the primary objective of this study was not to evaluate efficacy, this analysis was not powered to detect significant group differences in PHQ-8 scores. An exploratory analysis measured the change in PHQ-8 scores between baseline and post-intervention for individuals with a baseline PHQ-8 score ≥ 10, consistent with moderate symptoms of depression in both groups. Descriptive statistics are presented for this analysis. R version 4.1.1 (2021–08–10) was used to complete these analyses, and included using self-written code and the following packages: Rmisc, reshape2, stringr, ggplot2 and lmerTest.

The standardized mean-difference effect size and 95% confidence intervals were calculated for the MFQ, PROMIS Pediatric, GAD-7, and BRS measures using the Practical Meta-Analysis Effect Size Calculator created by W. Lipsey and David B. Wilson, 2001.

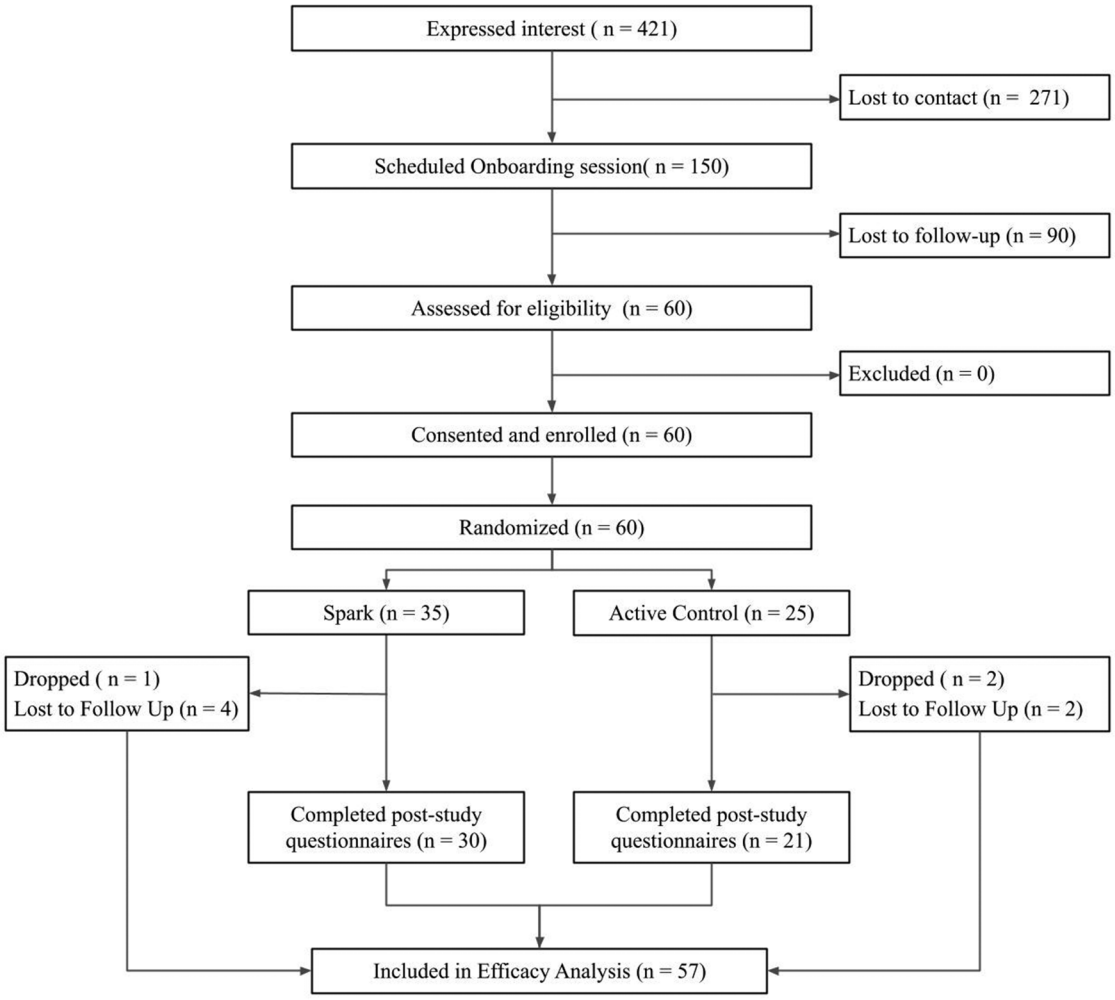


FIGURE 2
The flow of participants through the study procedures, from expression of interest to efficacy analysis.

App safety was determined by measuring the number of ADEs, SAEs, and UADEs identified in each group. Descriptive statistics about the number of AEs, ADEs, and UADEs captured for each group are reported. Microsoft Excel was used to analyze these data.

3. Results

3.1. Participant characteristics & feasibility outcomes

Over two months, sixty eligible participants were enrolled in the study. See **Figure 2** for the CONSORT diagram. 421 participants expressed interest in the study *via* a web form, of which 150 scheduled an onboarding session, representing a 35.6% recruitment rate. Of the 150 who scheduled an onboarded session, 60 attended their onboarding session, were determined to be eligible to participate, consented/assented and were enrolled, representing a 40% enrollment rate. Of these 60 participants, 35 were randomized to the Spark arm and 25 to the Active Control arm. 51 participants completed the study ($n_{Spark} = 30$, $n_{Active\ Control} = 21$), representing a 85% retention rate by post-intervention. Of those that did not complete the study, 3 participants ($n_{Spark} = 1$, $n_{Active\ Control} = 2$) were withdrawn per the safety protocol, due to missing two consecutive weekly questionnaires or safety events, and 6 participants were considered lost-to-follow up ($n_{Spark} = 4$, $n_{Active\ Control} = 2$) due to not completing post-intervention questionnaires.

See **Table 3** for participant characteristics. The sample recruited, consisting of 13–21 year olds ($n_{Spark} = 17.91$ [2.36]; $n_{Active\ Control} = 16.96$ [2.57]),was 78% female, which is consistent with higher rates of depression in adolescent girls (74, 75). The average PHQ score at baseline was 13.82, which is considered moderate severity (46). The majority of participants ($n = 32$, 53%) reported a depression diagnosis and 28 participants (46.6%) reported that they were currently receiving treatment specifically for depression at baseline. The majority of participants ($n = 37$, 62%) were over 18 years old in both conditions ($n_{Spark} = 19$; $n_{Active\ Control} = 18$). Additionally, 29 legal guardians ($n_{Spark} = 16$; $n_{Active\ Control} = 13$) were enrolled.

TABLE 3 Baseline characteristics of adolescent participants and legal guardians enrolled within the study.

Adolescent Participants			
	Spark ($N = 35$)	Active Control ($N = 25$)	Test Statistic
Age, M (SD)	17.91 (2.36)	16.96 (2.57)	$t(58) = 2.00$, $p = .14$
Gender, *N* (%)			χ^2 (2) = .93, $p = .62$
Male	6 (17.14%)	5 (20.00%)	
Female	28 (80.00%)	19 (76.00%)	
Non-binary	1 (2.86%)	1 (4.00%)	
Race, *N* (%)			χ^2 (5) = .59, $p = .99$
American Indian/Alaska Native	1 (2.86%)	0 (0.00%)	
Asian	7 (20.00%)	4 (16.00%)	
Black or African American	2 (5.71%)	3 (12.00%)	
Native Hawaiian or Other Pacific Islander	0 (0.00%)	0 (0.00%)	
Unknown	2 (5.71%)	0 (0.00%)	
White	20 (57.14%)	17 (68.00%)	
Mixed Race	3 (8.57%)	1 (4.00%)	
Ethnicity, *N* (%)			χ^2 (1) = .91, $p = .34$
Hispanic/Latino	6 (17.14%)	4 (16.00%)	
Not Hispanic/Latino	29 (82.85%)	21 (84.00%)	
Baseline PHQ-8, M (SD)	13.74 (6.02)	13.92 (5.32)	$t(58) = 2.00$, $p = .90$
Severity, *N* (%)			χ^2 (1) = .86, $p = .35$
mild-moderate (up to 15)	23 (65.71%)	16 (64.00%)	
moderate to severe (above 15)	12 (34.29%)	9 (36.00%)	
Depression Diagnosis, *N* (%)	18 (51.43%)	14 (56.00%)	χ^2 (1) = .73, $p = .39$
Concurrent treatment for depression, *N* (%)			χ^2 (5) = .57, $p = .99$
Medication only	5 (14.29%)	8 (32.00%)	
None	19 (54.29%)	12 (48.00%)	
Other	1 (2.86%)	0 (0.00%)	
Psychotherapy only	4 (11.43%)	2 (8.00%)	
Medication and Psychotherapy	5 (14.28%)	3 (12.00%)	
Unknown	1 (2.86%)	0 (0.00%)	
Legal Guardians			
	Spark ($N = 16$)	Active Control ($N = 13$)	
Education Level, *N* (%)			χ^2 (5) = .50, $p = 0.99$
Middle school	3 (18.75%)	1 (7.69%)	
High school/GED	1 (6.25%)	0 (0.00%)	
Some college	1 (6.25%)	3 (23.07%)	
Associate's and/or Bachelor's degree	9 (56.25%)	6 (46.15%)	
Master's degree	2 (12.50%)	2 (15.38%)	
Doctoral or Professional degree	0 (0.00%)	1 (7.69%)	

3.2. App acceptability: engagement & usability

As seen in Table 4, participants reported using Spark to be a more engaging experience than using the Active Control on the UES-SF ($t(49) = 3.46$, $p < .005$). Both apps were rated as having above-average usability, as indicated by a score of 68 or higher on the SUS scale. Exploratory between-group analyses were conducted. No differences were found as measured by the SUS mean scores in each condition ($t(49) = 1.50$, $p > .1$). Additionally, participants that used Spark reported a higher average improvement in symptoms of depression than participants that used the Active Control ($t(49) = 4.96$, $p < .001$). Legal guardians of participants who used either Spark or the Active control did not indicate a difference in subjective reports of symptom improvement between the two apps ($t(16) = 0.83$, $p > .1$). Both participants who used Spark and the legal guardians of these participants reported higher enjoyability ratings of the app compared to the Active Control users (participants: $t(49) = 4.55$, $p < .001$) and their legal guardians: $t(16) = 2.77$, $p < .05$). See Table 5 for more detail.

TABLE 4 The mean SUS and UES-SF scores for the two conditions. The mean usability and engagement for Spark users was higher than for the Active Control.

	Spark ($N = 30$)	Active Control ($N = 21$)	Test Statistic
SUS, M (SD)	80.67 (11.91)	75.83 (10.50)	$t(49) = 1.50$, $p = .14$
UES-SF, M (SD)	3.62 (0.52)	3.10 (0.54)	$t(49) = 3.46$, $p = .001$

We also investigated app engagement metrics. The median number of daily active users on a given day across the 5-week intervention period was 29%, and the 35-day retention rate was 26% (Figure 3). 94% of participants who received Spark completed level 1, with decreases in level completion in subsequent levels to 23% completing level 5 (Figure 4). Only levels 2–5 consisted of completing behavioral activations. 60% of the participants completed at least 5 behavioral activations (Figure 4). Furthermore, we found a significant negative relationship between the magnitude of change in PHQ-8 scores from the post-intervention and baseline timepoints, and the number of BAs that were completed ($r(32) = -0.38$, $p = 0.03$; Figure 5).

3.3. Study safety protocol feasibility

During the 5-week intervention period, 56 potential clinical concerns were logged and evaluated by study investigators ($n_{Spark} = 16$, $n_{Active\ Control} = 11$; see Figure 6). Any text that mentioned symptoms of depression from more serious (e.g., suicidal ideation) to less serious (e.g., cried all day) was logged for review. Of the 40 potential clinical concerns identified in the Spark group, 13 were identified from free-form text entries in Spark and the remaining 27 were identified in the adverse event questionnaire (AEQ), which prompted participants to indicate worsening, frequency, and intensity of mood. Following guidelines listed in the safety protocol, 35/40 logged events did not meet criteria for a potential safety concern and were consistent with expected day-to-day events or expected symptoms of depression, without an indication of worsening in intensity, frequency, or duration. Therefore, the study investigators consulted with the study clinician regarding five of

TABLE 5 Post-intervention questionnaire app feedback question ratings.

Question	Participants ($n = 51$)			Parents ($n = 18$)		
Question (on a scale of 0–10)	Spark ($n = 30$), Mean (SD)	Active Control ($n = 21$), Mean (SD)	*t*-test	Spark ($n = 10$), Mean (SD)	Active Control ($n = 8$), Mean (SD)	*t*-test
Mood improvement	5.07 (2.30)	1.90 (2.17)	$t(49) = 4.96$, $p < .001$	4.90 (1.91)	2.88 (3.27)	$t(16) = 0.83$, $p > .1$
Enjoyableness of mobile app	6.83 (2.05)	3.95 (2.46)	$t(49) = 4.55$, $p < .001$	6.10 (1.85)	2.75 (3.24)	$t(16) = 2.77$, $p < .05$

these participants' clinical concerns. The study clinician used the study safety protocol and their clinical judgment to determine whether clinician outreach was required. The study clinician decided that two out of these five participants were at sufficient risk and contacted them to confirm their safety. Out of the total 16 potential clinical concerns in the Active Control group, one was from a clinical deterioration in depression symptoms (as measured by the PHQ-8), 13 were reported in the AEQ, one was from text entered by a parent in the post-intervention questionnaire, and one was reported in an email response from a parent. Following the same safety protocol, 6/13 logged events did not meet criteria for a potential safety concern; therefore, the study investigators reported seven participants' clinical concerns in the Active Control group to the study clinician. The study clinician decided that one of these participants was at sufficient risk and contacted them and their legal guardian to confirm safety. In summary, 16 out of 35 participants in the treatment group and 11 out of 25 participants in the control group had potential clinical concerns logged, with some individuals in each group having multiple logs, resulting in higher total log counts than the number of participants. Five participants from the treatment group and seven participants from the control group had potential clinical concerns that were escalated to clinicians for safety evaluation. This resulted in 0 AE/SAE classifications for the treatment group and 2 SAEs for the control group.

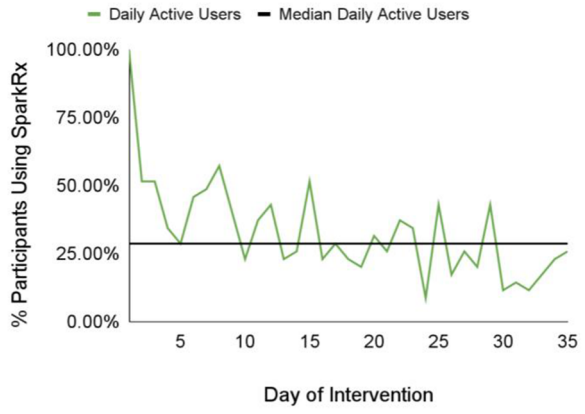


FIGURE 3
The median value and daily number of daily active users on a given day across the 5-week intervention period.

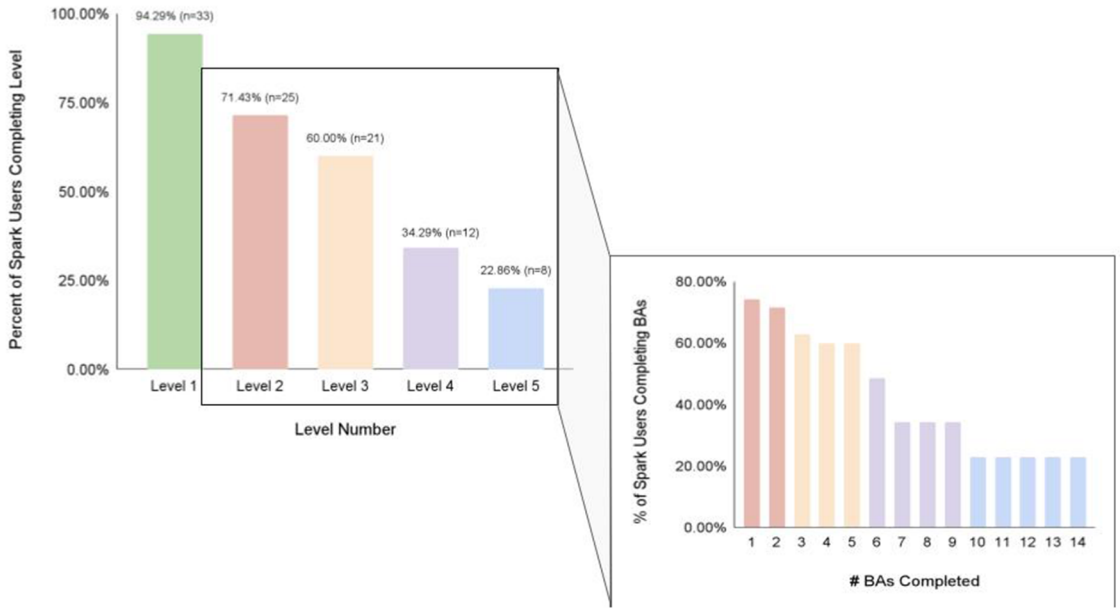


FIGURE 4
The percent of Spark users completing each level of the Spark intervention, along with the number of behavioral activations completed by a certain percentage of Spark users.

3.4. App efficacy and safety

Three participants were excluded from efficacy analyses due to having completed only the baseline PHQ-8 ($n_{Spark} = 1$, $n_{Active\ Control} = 2$), wwhich did not allow for imputation of missing data.

Within weekly PHQ-8s, 6.1% were missing. No item-level data were missing. Little's test suggested that data were not missing at random ($\chi^2(26) = 52.886$, $p = .0014$). There were no group differences in missing data ($\chi^2(5) = 0.99$, $p = 1.00$).

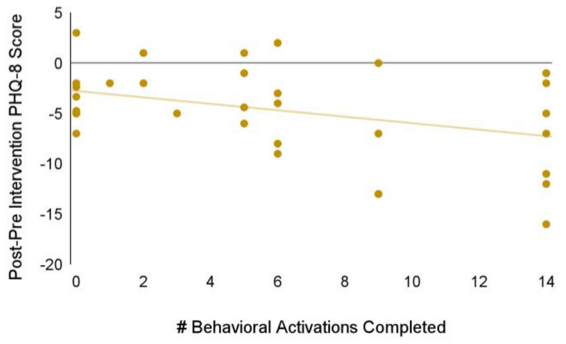


FIGURE 5
Relationship between the magnitude of change in PHQ-8 scores between the baseline and post-intervention timepoints, and the number of BAs that were completed.

Analyses investigating differences across Group and Week in the number of days between the completion of the baseline PHQ-8 and each weekly PHQ-8 showed a significant effect of Week, $F =$ 2,470.35, $p < .001$, as the number of days since baseline increased for each successive weekly PHQ-8. There was no main effect of Group, $F = 1.96$, $p = 0.16$, nor was there an interaction between Group and Week, $F = 1.158$, $p = .33$, indicating that differences in the timing of completion of PHQ-8s by week did not differ between the two groups.

The GLMM exploring PHQ-8 scores as a function of Group and Week showed a significant main effect of Week, $F = 40.600$, $p < .001$, demonstrating that depression symptoms declined over time. However, no main effect of Group, $F = 0.004$, $p = .95$, nor Group × Week interaction, $F = 0.125$, $p = .72$, was observed (Figure 7). The lack of a Group × Week interaction appears to have been driven by a larger than expected reduction in symptoms in the Active Control arm, ΔPHQ-$8_{Active\ Control} = 3.56$, as the average reduction in symptoms in the Spark group, ΔPHQ-$8_{Spark} = 4.69$, was close to reaching a clinically meaningful change (defined as ΔPHQ-$8 \geq 5$; see Table 6) (46, 76, 77). However, an exploratory analysis showed that Spark users with moderate or higher levels of depression (PHQ-$8 \geq 10$) demonstrated, on average, a clinically meaningful reduction in depressive symptoms, while Active control users did not (ΔPHQ-$8_{Spark} = 5.62$ (4.68), $n_{Spark} = 26$; ΔPHQ-$8_{Active\ Control} =$ 3.72 (5.01), $n_{Active\ Control} = 19$).

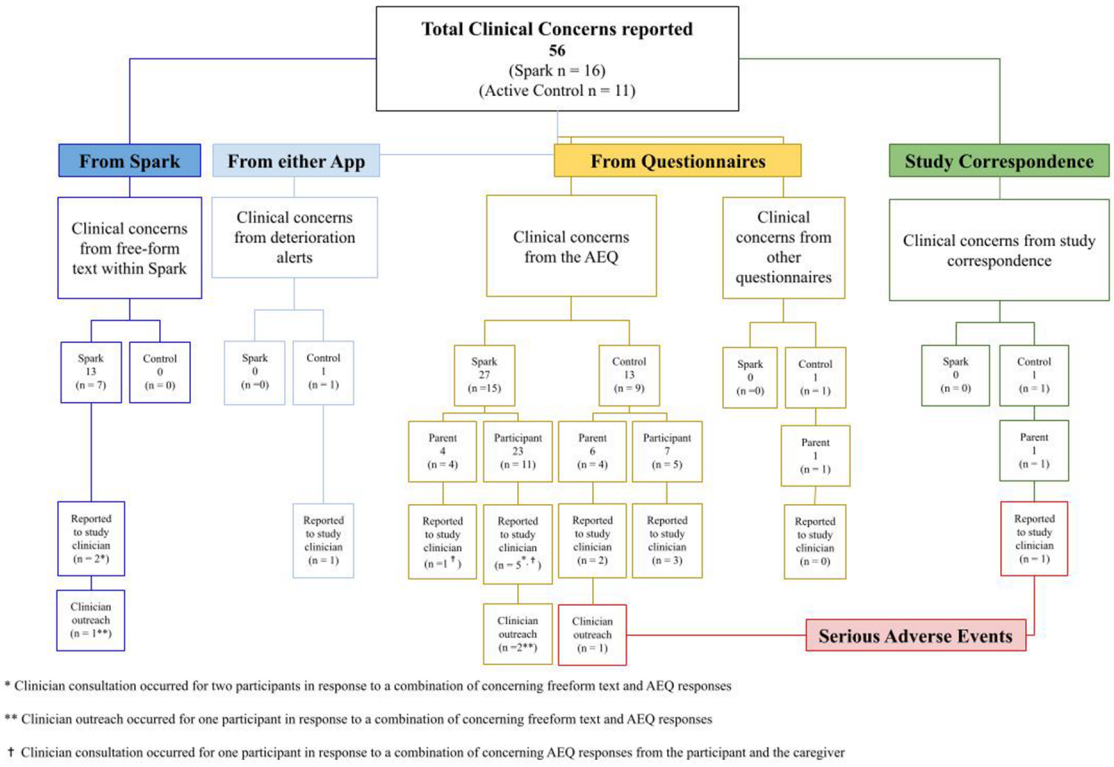


FIGURE 6
Summary chart of clinical concerns including the sources of clinical concerns, the number of clinical concerns escalated to the study clinician, and the number of clinical concerns that elicited clinician reachout to the participant.

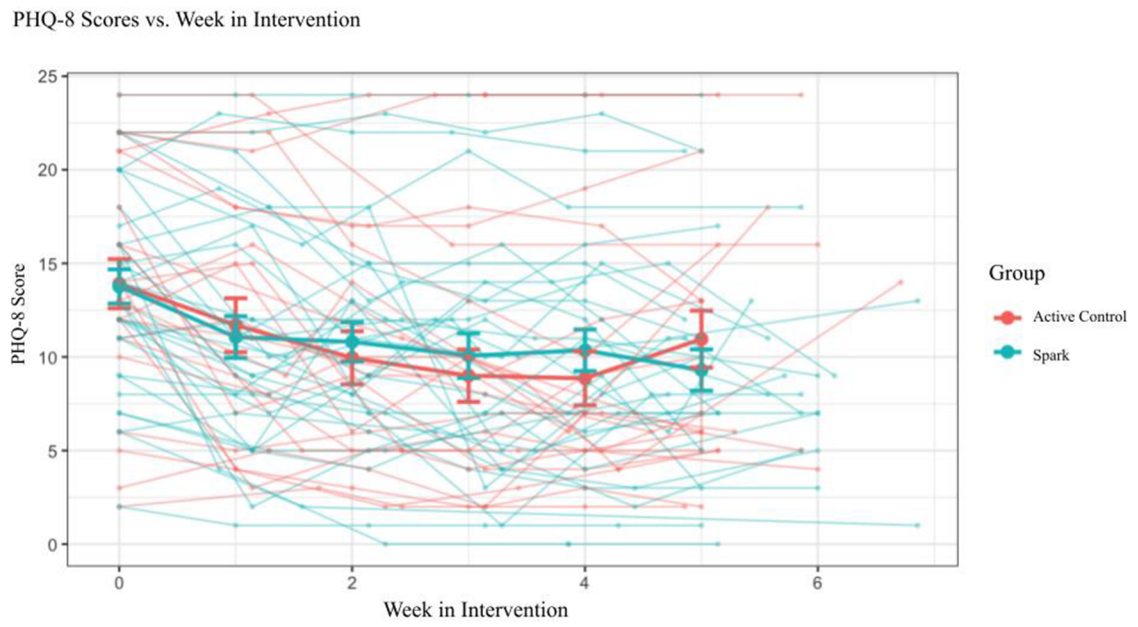


FIGURE 7
Imputed PHQ-8 scores for participants that completed two or more PHQ-8 questionnaires (n = 57) and separated by condition.

TABLE 6 Change in depressive symptoms at baseline vs. Post-Intervention by group as evaluated by the PHQ-8.

	Baseline	Post-intervention	Mean difference
Spark, M (SD)	13.76 (5.31)	9.06 (5.76)	4.69 (4.53)
Active Control, M (SD)	13.91 (6.30)	10.36 (6.98)	3.56 (5.03)

TABLE 7 Change in GAD-7, MFQ, and PROMIS pediatric at baseline and post-intervention, mean difference and Cohen's D.

GAD-7				
	Baseline	Post-intervention	Mean Difference	Effect size
Spark, M (SD)	11.26, 35 (4.85)	8.77, 30 (5.98)	−2.49	d = −.18 95% CI [−.58,.18]
Active Control, M (SD)	12.08, 25 (5.20)	10.10, 21 (5.96)	−1.98	
MFQ				
	Baseline	Post-intervention	Mean Difference	Effect size
Spark, M (SD)	18.63, 16 (4.32)	9.00, 10 (6.57)	−4.31	d = .25 95% CI [−.35,.85]
Active Control, M (SD)	12.08, 13 (5.78)	8.00, 8 (4.63)	−4.08	
PROMIS Pediatric (Global Health)				
	Baseline	Post-intervention	Mean Difference	Effect size
Spark, M (SD)	35.88, 35 (6.62)	37.97, 30 (7.86)	2.09	d = .06 95% CI [−.68,.82]
Active Control, M (SD)	34.83, 25 (6.27)	35.50, 21 (6.77)	.67	

In relation to app safety, there were a total of 2 SAEs, which both occurred in the Active Control group. One SAE was reported in the weekly AEQ; a parent reported that their child was hospitalized due to depressive symptoms. The clinician contacted the participant and parent and confirmed the participant was safe and eligible to continue with the study. The second SAE was reported *via* email; a parent wrote that their child had been hospitalized for a suicide attempt. Since the individual was receiving care at the hospital, there was no study clinician reachout. This participant was also withdrawn from the study due to our inability to accurately monitor their safety during the intervention period as they did not complete the AEQ questionnaire over two consecutive weeks during the 5-week intervention period). There were no ADEs or UADEs reported in either group.

No significant effect was determined when comparing baseline and post-intervention mean scores across groups was determined for any other measure (GAD-7, MFQ, PROMIS Pediatric Global Health), except for the MFQ (see Table 7).

4. Discussion

The results of this study determined that 1) it was feasible to evaluate a 5-week, self-guided CBT-based mobile app program compared to an active educational control app for an adjunct treatment of adolescents with symptoms of depression in a nationwide virtual and decentralized RCT, 2) adolescents found the app acceptable, and 3) our safety protocols were robust for monitoring participant safety. Additionally, there was a promising reduction in depression symptoms for participants who received Spark, though the difference in symptom reduction between Spark and Active Control was not statistically significant. Finally, there were 0 serious adverse events in the

Spark group and 2 serious adverse events in the control group. This suggests that participants in the Spark group were not at greater risk of a serious adverse event than participants in the active control group.

4.1. Study feasibility

The enrolled sample successfully represented a range in age, gender, race, ethnicity, and depression symptom severity. Though females were more heavily represented, this is consistent with the etiology of depression in adolescents (78). The recruited sample was racially diverse compared to other feasibility studies, which may have been a benefit of our decentralized approach to virtually recruiting participants nationwide (79). The racial and ethnic background of participants in the study was in line with national racial and ethnic census data and with the demographic distribution of depression among adolescents (80–82). The diversity reflected in the study sample is a strength and may allow for greater generalizability of feasibility, engagement, and usability findings to the wider population of adolescents with depression.

Target enrollment was reached in two months for this trial, demonstrating the success of our online recruitment strategy and the perceived feasibility of our enrollment procedures. This recruitment speed may also underscore the demand for mental health resources in this population and during the COVID-19 pandemic, as well as reflect an interest in and receptivity to digital health solutions. Additionally, our recruitment, enrollment and retention rates were high compared to other feasibility studies that enrolled similar populations (those with depression (83) and/or adolescents (84) through online recruitment for remote interventions (83–85). For example, our enrollment rate was double a feasibility trial evaluating the effectiveness of clinical trials conducted in a virtual setting, or 21% (205 out of 958) vs. 40% (60 out of 150) of participants screened vs. those that enrolled (85). Despite this success, a few areas of improvement were identified. Improvements to increase retention could include sending more regular reminders to participants to remind them to complete questionnaires and additional modalities for reminders, such as text and email notifications. Additionally, tailoring availability of onboarding sessions to later hours in the day or weekends could allow faster enrollment, especially for participants under the age of 18, given the required involvement of legal guardians and scheduling constraints around school hours.

4.2. App acceptability

Participants that used Spark rated it as more usable than those that used the Active Control app in terms of enjoyment and in terms of its impact on improving their symptoms of depression. Furthermore, both users of Spark and the Active control rated the app as well above average usability (64). While there was no significant difference in the ratings of usability of the two apps, Spark users rated it, on average, as more usable on the SUS scale than Active Control users, suggesting that its interactive features are easy to use. Engagement was also high for the Spark group: with an engagement rating above 3.5 (out of 5), this is comparable to similar studies (86, 87). All users except one gave Spark an engagement rating above 3 and Spark was rated as significantly more engaging than the Active Control app. Together, this suggests that Spark is highly acceptable to study participants.

App engagement metrics are as good or better than other depression apps on the market. Baumel and colleagues report that the median daily open rate for real-world usage of depression apps is 4.8% (88), and is 4.06 times higher for research studies (88, 89), which is lower than the median daily active use we found. They also found that the 30-day retention rate is 3.3% for real-world usage of mental health apps (88). Even a 4.06 fold increase in average engagement for apps in research studies (89) would put our 35-day retention rate of 26% above the average. Though adherence (completion of all levels in the app) was only at 23%, engagement in digital therapeutics for mental health is a challenge across the field (90). This low adherence may be contributing to the non-significant difference in changes in PHQ between groups. Interestingly, the relationship between the number of behavioral activations completed and the reduction in PHQ-8 scores is similar to or stronger than other studies that report little or no relationship between app dose and treatment response (91–93). This suggests that if engagement increases, this may facilitate even greater improvements in depression symptoms.

One reason Spark may have had high engagement is because of its reliance on BA, which is inherently self-paced and may appeal to self-motivated adolescents. A 2021 meta-analysis of digital intervention studies showed that flexibility was a component often used to increase adherence and engagement (36). Furthermore, users of apps that help treat depression have stated a desire to have space for positive emotions within digital mental health products they are using (94), a quality inherent to BAs. However, for individuals who may not feel self-motivated, it is important to incorporate additional features to enhance engagement, like reminders. The therapeutic qualities of BA can be further enhanced in the digital setting with the inclusion of additional features allowing for increased personalization, gamification, and ease of use (36), which will be important for future versions of Spark.

It is worth noting that operationalizing and measuring meaningful engagement is a challenge in the field of digital therapeutics and is critical for understanding how adherence and engagement impacts therapeutic outcomes (94). This is an area in which Limbix is actively working (90). In future versions of Spark, a focus on improvement engagement, like including mood-tracking activities, mindfulness, psychoeducation, and relapse prevention in addition to the behavioral activation activity scheduling that was included here may help to improve outcomes. Furthermore, though each level could have taken up to 60 min to complete, which may seem like too long for adolescents to be able to engage, we do not believe that this was actually a barrier to engagement. This time was purposely overestimated so that teens would not feel discouraged if it took longer to complete a module than anticipated. This estimate also included time to do BA activities outside of the app, and additionally, adolescents could go at their own pace, using the app for only a few minutes per day,

and still complete each module. We felt it was important to keep this amount of content in the treatment so that we could retain essential clinical components to improve outcomes; having an evidence-based treatment is rated as one of the five critical features of evaluating mental health apps according to the American Psychiatric Association (96) and is viewed as an increasingly necessary feature of digital health solutions (97). Therefore, we believe the primary goal is to modify the app to make the material more engaging while still maintaining a high standard for clinical quality. Though these are preliminary analyses, these results suggest promising directions for future work.

4.3. Study safety protocol feasibility

A third aim of this study was to develop and test the feasibility of using a detailed, thorough method for monitoring safety in a decentralized, virtual trial of a mobile application. Typically, safety protocols for studies of digital interventions are either not reported (95, 96) or consist of unstructured monitoring with safety intervention at the investigator's discretion (97). Nevertheless, a thorough approach as implemented here may be especially critical for ensuring safety of study participants within the context of a completely virtual and decentralized trial. Additionally, the use of mobile technology affords the opportunity to standardize data collection around safety rather than relying exclusively on spontaneous reporting. The safety protocol was successful in ensuring participant safety throughout the study period. It provided a standardized and rigorous method to track participant and guardian reported clinical concerns in both study arms. This protocol allowed study investigators to determine which clinical concerns met criteria to be considered adverse events as well as the severity of such events. The clinician outreach approach outlined in the protocol was feasible and effective for determining relatedness of adverse events to the study apps and assuring participant safety. Opportunities for refining the safety protocol in the future could include increasing automation in identifying potential clinical concerns to reduce the potential for human error or oversight.

4.4. Preliminary App efficacy & safety

The preliminary clinical efficacy and safety of Spark was evaluated compared to an active control condition. The lack of serious adverse events in the Spark group, compared to two in the Active Control group, suggests that Spark does not pose any additional risk to users. Efficacy was measured by a reduction in depressive symptoms as measured by the PHQ-8. There was a significant main effect of Time, indicating that both groups reported improvements in symptoms of depression over the intervention period. While we did not observe a statistically significant difference in symptom reduction between groups, Spark users experienced a greater numeric decrease in PHQ-8 scores compared to Active Control users. The reduction of depression symptoms in the Spark group was promising, as the average reduction in depression symptoms approached a clinically meaningful change. Symptom reduction in the Spark group may have been limited by a floor effect introduced by the inclusion of participants with all levels of baseline symptom severity. This possibility was supported by a *post hoc* analysis of only participants with at least moderate baseline symptom severity that showed a clinically meaningful reduction in symptom severity at post-intervention. In fact, recent evidence suggests that digital interventions may be most effective for more severe forms of depression (98).

The lack of statistical significance in symptom reduction between groups is not surprising, given that this trial was not designed or powered to detect statistical differences in symptom reduction between Spark and the Active Control. Notably, this finding seemed to have been driven at least in part by a larger than expected reduction in symptoms in the Active Control group (26, 99–101), which might be explained by a number of study considerations. First, the study design did not control for participants beginning new treatment or changing treatment for a mental health condition immediately prior to or during the study intervention period. Additionally, the psychoeducational material in the Active Control app may have had therapeutic impact, as psychoeducation is used as a form of treatment (102) and is considered a therapeutic element of CBT. Finally, changing impacts of the pandemic may have played a role, as changes to federal, state, and regional policies occurred, including those related to remote schooling, during the conduct of the trial. Future studies powered to detect statistical differences between groups will be necessary to evaluate the efficacy of Spark relative to an active control condition.

4.5. Limitations and future directions

Though these data support study feasibility and the acceptability of the Spark app, limitations remain. The recruited sample was predominantly female (78, 103), which is consistent with prevalence rates of depression in adolescence (104). However, a limitation is that these results are not generalizable to males and gender non-binary individuals. Future studies should consider alternative sampling methods that result in a more equal sampling to better understand the effects in non-female populations. In addition to this, our eligibility criteria required that participants were under the care of a US-based primary care and/or licensed mental healthcare provider. This criteria was included to; 1) evaluate the feasibility of the Spark app as an adjunct treatment for depression and 2) manage participant safety. We acknowledge that many adolescents are not under the care of a US-based primary care and/or licensed mental healthcare provider and as a result, our sample may not be generalizable to the adolescent population in the US. Participants and their legal guardians were required to be fluent in English in order to enroll in the study and use the study apps, in turn limiting access for those who are not English-speaking. While for this study no participants were determined ineligible due to this criteria, individuals from minority populations who do not speak English are in need of mental health services and future work will be needed to determine whether it is feasible to use

Spark in such populations. Also, we included participants that were receiving other forms of treatment at baseline. Though excluding such participants may have increased the efficacy of Spark, this choice was made because Spark is intended to be used as an adjunct treatment and we wanted to make Spark widely available to those who were looking for additional resources during the COVID-19 pandemic. This likely increased the ecological validity of this study given Spark's intended use. Because efficacy analyses were preliminary, we did not statistically control for changes in treatment for mental health conditions prior to, or during the study intervention period or stratify this variable between groups. As a result, reductions in depression severity or lack of group differences could be attributed to changes in concomitant treatments that participants were receiving. Future studies should ensure stability on concurrent treatments and control for changes in treatment during the study intervention period. Engagement analyses were limited to subjective measures, whereas objective measures of app use analytics would provide a more complete picture of engagement. Additionally, while the study's safety protocol was supported based on AE ratings and clinical concern rates, improvements can be made. In this study's safety protocol, we withdrew participants from the study if they did not complete two weekly questionnaires in a row. This criterion was implemented in order to motivate participant completion of questionnaires, including the AEQ, which would allow better monitoring of participant safety. For future studies, it would be preferable to maintain participant involvement in the study and remove this criteria for withdrawing participants, in order to not miss potential data from these withdrawn participants. An additional limitation of study procedures was that suicidality and comorbidities were not assessed using standardized measures in every participant to confirm eligibility. While thorough screening measures were taken to provide the participants with a self-reported confirmation of eligibility, in future studies, we may implement standardized screenings.

Lastly, we recognize that this study was not powered to detect statistical differences between groups and all statistical analyses are considered exploratory. Future studies will be required to evaluate efficacy, safety, and engagement of Spark relative to an active control condition or other digital therapeutics.

4.6. Conclusion

This feasibility study demonstrated the robustness of online recruitment techniques, strong engagement with and potential therapeutic benefit of Spark, and the effectiveness of the novel safety protocol to monitor and ensure patient safety. These findings will be used to inform and direct future product development as well as a powered RCT to evaluate app efficacy. The results of this feasibility trial provide preliminary support for the use of Spark as a novel digital treatment for adolescent depression and may point to the utility of digital therapeutics in addressing existing barriers in access to effective mental health care.

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 Western Copernicus Group (WCG) Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

JIL and AP designed the trial and oversaw study operations. VNK, PCC, and SAH conducted study analysis. JIL, JEF, VNK, PCC, SAH, and AP wrote the paper. VNK and PCC prepared the figures under the supervision of SAH JIL, JEF, SAH, AP, GSS, EMV, and XLK revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

Product development of Spark was funded in part by NIH grant R44MH125636.

Acknowledgments

We would like to thank Lauren Smith and Stella Kim for contributing to data collection for this study. We also thank Lang Chen for data analysis support. We also thank Isabel Enriquez for her contributions to finalizing manuscript wording throughout the text.

Conflict of interest

Authors VNK, PCC, SAH, JEF, GSS, EMV, XLK, JIL, and AP are employed in Limbix Health, Inc. and are stakeholders in Limbix Health, Inc.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2023.1062471/full#supplementary-material.

References

1. Patton GC, Sawyer SM, Santelli JS, Ross DA, Afifi R, Allen NB, et al. Our future: a lancet commission on adolescent health and wellbeing. *Lancet*. (2016) 387:2423–78. doi: 10.1016/S0140-6736(16)00579-1

2. Keyes KM, Gary D, O'Malley PM, Hamilton A, Schulenberg J. Recent increases in depressive symptoms among US adolescents: trends from 1991 to 2018. *Soc Psychiatry Psychiatr Epidemiol*. (2019) 54:987–96. doi: 10.1007/s00127-019-01697-8

3. Galaif ER, Sussman S, Newcomb MD, Locke TF. Suicidality, depression, and alcohol use among adolescents: a review of empirical findings. *Int J Adolesc Med Health*. (2007) 19:27–35. doi: 10.1515/IJAMH.2007.19.1.27

4. Verboom CE, Sijtsema JJ, Verhulst FC, Penninx BWJH, Ormel J. Longitudinal associations between depressive problems, academic performance, and social functioning in adolescent boys and girls. *Dev Psychol*. (2014) 50:247–57. doi: 10.1037/a0032547

5. Jaycox LH, Stein BD, Paddock S, Miles JNV, Chandra A, Meredith LS, et al. Impact of teen depression on academic, social, and physical functioning. *Pediatrics*. (2009) 124:e596–605. doi: 10.1542/peds.2008-3348

6. Katon WJ. Clinical and health services relationships between major depression, depressive symptoms, and general medical illness. *Biol Psychiatry*. (2003) 54:216–26. doi: 10.1016/S0006-3223(03)00273-7

7. Rao U, Chen L-A. Characteristics, correlates, and outcomes of childhood and adolescent depressive disorders. *Dialogues Clin Neurosci*. (2009) 11:45–62. doi: 10.31887/DCNS.2009.11.1/urao

8. Torio CM, Encinosa W, Berdahl T, McCormick MC, Simpson LA. Annual report on health care for children and youth in the United States: national estimates of cost, utilization and expenditures for children with mental health conditions. *Acad Pediatr*. (2015) 15:19–35. doi: 10.1016/j.acap.2014.07.007

9. Racine N, McArthur BA, Cooke JE, Eirich R, Zhu J, Madigan S. Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: a meta-analysis. *JAMA Pediatr*. (2021) 175(11):1142–50. doi: 10.1001/jamapediatrics.2021.2482

10. Bose J. Key substance use and mental health indicators in the United States: results from the 2017 national survey on drug use and health (2018) 124.

11. Kataoka SH, Zhang L, Wells KB. Unmet need for mental health care among U.S. Children: variation by ethnicity and insurance status. *Am J Psychiatry*. (2002) 159:1548–55. doi: 10.1176/appi.ajp.159.9.1548

12. Telesia L, Kaushik A, Kyriakopoulos M. The role of stigma in children and adolescents with mental health difficulties. *Curr Opin Psychiatry*. (2020) 33:571. doi: 10.1097/YCO.0000000000000644

13. Cuddy E, Currie J. Treatment of mental illness in American adolescents varies widely within and across areas. *Proc Natl Acad Sci U S A*. (2020) 117:24039–46. doi: 10.1073/pnas.2007484117

14. Douglas D, Diehl S, Honberg R, Kimball A. *The unfulfilled promise of parity*. Arlington, VA: National Alliance on Mental Illness (2016). 14. https://www.nami.org/Support-Education/Publications-Reports/Public-Policy-Reports/Out-of-Network-Out-of-Pocket-Out-of-Options-The/Mental_Health_Parity2016.pdf

15. Carbonell Á, Navarro-Pérez J-J, Mestre M-V. Challenges and barriers in mental healthcare systems and their impact on the family: a systematic integrative review. *Health Soc Care Community*. (2020) 28:1366–79. doi: 10.1111/hsc.12968

16. Meredith LS, Stein BD, Paddock SM, Jaycox LH, Quinn VP, Chandra A, et al. Perceived barriers to treatment for adolescent depression. *Med Care*. (2009) 47:677–85. doi: 10.1097/MLR.0b013e318190d46b

17. *The doctor is out*. Arlington, VA: National Alliance on Mental Illness (2017). 15. https://www.nami.org/Support-Education/Publications-Reports/Public-Policy-Reports/The-Doctor-is-Out/DoctorIsOut

18. Aguirre Velasco A, Cruz ISS, Billings J, Jimenez M, Rowe S. What are the barriers, facilitators and interventions targeting help-seeking behaviours for common mental health problems in adolescents? A systematic review. *BMC Psychiatry*. (2020) 20:293. doi: 10.1186/s12888-020-02659-0

19. Fein JA, Pailler ME, Barg FK, Wintersteen MB, Hayes K, Tien AY, et al. Feasibility and effects of a web-based adolescent psychiatric assessment administered by clinical staff in the pediatric emergency department. *Arch Pediatr Adolesc Med*. (2010) 164:1112–7. doi: 10.1001/archpediatrics.2010.213

20. Gardner W, Klima J, Chisolm D, Feehan H, Bridge J, Campo J, et al. Screening, triage, and referral of patients who report suicidal thought during a primary care visit. *Pediatrics*. (2010) 125:945–52. doi: 10.1542/peds.2009-1964

21. Bradford S, Rickwood D. Acceptability and utility of an electronic psychosocial assessment (myAssessment) to increase self-disclosure in youth mental healthcare: a quasi-experimental study. *BMC Psychiatry*. (2015) 15:305. doi: 10.1186/s12888-015-0694-4

22. Scott MA, Wilcox HC, Schonfeld IS, Davies M, Hicks RC, Turner JB, et al. School-Based screening to identify at-risk students not already known to school professionals: the Columbia suicide screen. *Am J Public Health*. (2009) 99:334–9. doi: 10.2105/AJPH.2007.127928

23. Teens, Social Media & Technology 2018. *Pew research center: Internet*. DC: Science & Tech (2018). https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/(Accessed December 22, 2021)

24. Anderson M, Jiang J. *Teens, social Media, and technology 2018*. DC: Pew Research Center (2018). 20. https://www.pewinternet.org/wp-content/uploads/sites/9/2018/05/PI_2018.05.31_TeensTech_FINAL.pdf

25. Digital therapeutics definition and core principles. (2019). https://dtxalliance.org/wp-content/uploads/2021/01/DTA_DTx-Definition-and-Core-Principles.pdf

26. Clarke G, DeBar LL, Pearson JA, Dickerson JF, Lynch FL, Gullion CM, et al. Cognitive behavioral therapy in primary care for youth declining antidepressants: a randomized trial. *Pediatrics*. (2016) 137:e20151851. doi: 10.1542/peds.2015-1851

27. Ebert DD, Zarski A-C, Christensen H, Stikkelbroek Y, Cuijpers P, Berking M, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. *PLoS One*. (2015) 10:e0119895. doi: 10.1371/journal.pone.0119895

28. Hopko DR, Lejuez CW, Ruggiero KJ, Eifert GH. Contemporary behavioral activation treatments for depression: procedures, principles, and progress. *Clin Psychol Rev*. (2003) 23:699–717. doi: 10.1016/S0272-7358(03)00070-9

29. Tindall L, Mikocka-Walus A, McMillan D, Wright B, Hewitt C, Gascoyne S. Is behavioural activation effective in the treatment of depression in young people? A systematic review and meta-analysis. *Psychol Psychother*. (2017) 90:770–96. doi: 10.1111/papt.12121

30. Huguet A, Rao S, McGrath PJ, Wozney L, Wheaton M, Conrod J, et al. A systematic review of cognitive behavioral therapy and behavioral activation apps for depression. *PLoS One*. (2016) 11:e0154248. doi: 10.1371/journal.pone.0154248

31. Pass L, Lejuez CW, Reynolds S. Brief behavioural activation (brief BA) for adolescent depression: a pilot study. *Behav Cogn Psychother*. (2018) 46:182–94. doi: 10.1017/S1352465817000443

32. Ritschel LA, Ramirez CL, Jones M, Craighead WE. Behavioral activation for depressed teens: a pilot study. *Cogn Behav Pract*. (2011) 18:281–99. doi: 10.1016/j.cbpra.2010.07.002

33. Jacobson NS, Dobson KS, Truax PA, Addis ME, Koerner K, Gollan JK, et al. A component analysis of cognitive-behavioral treatment for depression. *J Consult Clin Psychol*. (1996) 64:295–304. doi: 10.1037/0022-006X.64.2.295

34. Dobson KS, Hollon SD, Dimidjian S, Schmaling KB, Kohlenberg RJ, Gallop RJ, et al. Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the prevention of relapse and recurrence in major depression. *J Consult Clin Psychol*. (2008) 76:468–77. doi: 10.1037/0022-006X.76.3.468

35. McCauley E, Gudmundsen G, Schloredt K, Martell C, Rhew I, Hubley S, et al. The adolescent behavioral activation program: adapting behavioral activation as a treatment for depression in adolescence. *J Clin Child Adolesc Psychol*. (2016) 45:291–304. doi: 10.1080/15374416.2014.979933

36. Martin F, Oliver T. Behavioral activation for children and adolescents: a systematic review of progress and promise. *Eur Child Adolesc Psychiatry*. (2019) 28:427–41. doi: 10.1007/s00787-018-1126-z

37. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. *JMIR Ment Health*. (2021) 8:e30000. doi: 10.2196/30000

38. Himle JA, Weaver A, Zhang A, Xiang X. Digital mental health interventions for depression. *Cogn Behav Pract*. (2022) 29:50–9. doi: 10.1016/j.cbpra.2020.12.009

39. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. *Lancet Child Adolesc Health*. (2018) 2:223–8. doi: 10.1016/S2352-4642(18)30022-1

40. Spear LP. The adolescent brain and age-related behavioral manifestations. *Neurosci Biobehav Rev*. (2000) 24:417–63. doi: 10.1016/S0149-7634(00)00014-2

41. Twenge JM, Cooper AB, Joiner TE, Duffy ME, Binau SG. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005-2017. *J Abnorm Psychol*. (2019) 128:185–99. doi: 10.1037/abn0000410

42. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharm Stat*. (2005) 4:287–91. doi: 10.1002/pst.185

43. Wright B, Tindall L, Littlewood E, Allgar V, Abeles P, Trépel D, et al. Computerised cognitive–behavioural therapy for depression in adolescents: feasibility results and 4-month outcomes of a UK randomised controlled trial. *BMJ Open*. (2017) 7:e012834. doi: 10.1136/bmjopen-2016-012834

44. Calear AL, Christensen H, Mackinnon A, Griffiths KM, O'Kearney R. The YouthMood project: a cluster randomized controlled trial of an online cognitive behavioral program with adolescents. *J Consult Clin Psychol*. (2009) 77:1021–32. doi: 10.1037/a0017391

45. Abeles P, Verduyn C, Robinson A, Smith P, Yule W, Proudfoot J. Computerized CBT for adolescent depression ("stressbusters") and its initial evaluation through an extended case series. *Behav Cogn Psychother*. (2009) 37:151–65. doi: 10.1017/S1352465808005067

46. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord*. (2009) 114:163–73. doi: 10.1016/j.jad.2008.06.026

47. Stanley-Brown Safety Planning Intervention. Stanley-Brown Safety Planning Intervention (2018) https://suicidesafetyplan.com/(Accessed October 3, 2022)

48. Lejuez CW, Hopko DR, Acierno R, Daughters SB, Pagoto SL. Ten year revision of the brief behavioral activation treatment for depression: revised treatment manual. *Behav Modif*. (2011) 35:111–61. doi: 10.1177/0145445510390929

49. McCauley E, Schloredt K, Gudmundsen G, Martell C, Dimidjian S. *Behavioral activation with adolescents: a Clinician's Guide*. New York, NY: Google Docs (2016). https://drive.google.com/file/d/1eBv-0U-fDRmPM9YIMZE44kSjhbrm-UwQ/view?usp=embed_facebook [Accessed March 11, 2022].

50. Dimidjian S, Davis KJ. Newer variations of cognitive-behavioral therapy: behavioral activation and mindfulness-based cognitive therapy. *Curr Psychiatry Rep*. (2009) 11:453–8. doi: 10.1007/s11920-009-0069-y

51. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. (2019) 20:273–86. doi: 10.1093/biostatistics/kxx069

52. Wu J, Sun Y, Zhang G, Zhou Z, Ren Z. Virtual reality-assisted cognitive behavioral therapy for anxiety disorders: a systematic review and meta-analysis. *Front Psychiatry*. (2021) 12:575094. doi: 10.3389/fpsyt.2021.575094

53. Feasibility, acceptability, and preliminary evidence of efficacy of a digital intervention for adolescent depression. *JMIR Preprints* (2023). doi: 10.2196/preprints.43260 https://preprints.jmir.org/preprint/43260 (Accessed February 9, 2023)

54. Cassar J, Ross J, Dahne J, Ewer P, Teesson M, Hopko D, et al. Therapist tips for the brief behavioural activation therapy for depression—revised (BATD-R) treatment manual practical wisdom and clinical nuance. *Clin Psychol (Aust Psychol Soc)*. (2016) 20:46–53. doi: 10.1111/cp.12085

55. Teen Depression: More Than Just Moodiness. *National Institute of Mental Health (NIMH)* https://www.nimh.nih.gov/health/publications/teen-depression (Accessed September 8, 2022)

56. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care*. (2004) 42:1194–201. doi: 10.1097/00005650-200412000-00006

57. ASQ Screening Tool. *National Institute of Mental Health (NIMH)* https://www.nimh.nih.gov/research/research-conducted-at-nimh/asq-toolkit-materials/asq-tool/asq-screening-tool (Accessed September 19, 2022)

58. CFR—Code of Federal Regulations Title 21. US Department of Food & Drug Administration https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=312.32 (Accessed September 21, 2022)

59. Office of the Commissioner. *What is a serious adverse event?* Silver Spring, MD: US Food and Drug Administration (2016). https://www.fda.gov/safety/reporting-serious-problems-fda/what-serious-adverse-event (Accessed September 21, 2022)

60. Clinical investigation of medical devices for human subjects — Good clinical practice. https://www.iso.org/obp/ui/#iso:std:iso:14155:ed-3:v1:en (Accessed September 21, 2022)

61. Shin C, Lee S-H, Han K-M, Yoon H-K, Han C. Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for Major depressive disorder: analysis of psychiatric outpatient data. *Psychiatry Investig*. (2019) 16:300–5. doi: 10.30773/pi.2019.02.01

62. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. (2001) 16:606–13. doi: 10.1046/j.1525-1497.2001.016009606.x

63. Brooke J. SUS: a "quick and dirty" usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. *Usability evaluation in industry*. London, UK: CRC Press (1996). p. 189–94

64. Sauro J. *A practical guide to the system usability scale: background, benchmarks & best practices*. Denver, CO: Measuring Usability LLC (2011). 162.

65. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. *Int J Human–Computer Inter*. (2008) 24:574–94. doi: 10.1080/10447310802205776

66. O'Brien HL, Cairns P, Hall M. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *Int J Hum Comput Stud*. (2018) 112:28–39. doi: 10.1016/j.ijhcs.2018.01.004

67. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. *Med Care*. (2008) 46:266–74. doi: 10.1097/MLR.0b013e318160d093

68. Forrest CB, Bevans KB, Pratiwadi R, Moon J, Teneralli RE, Minton JM, et al. Development of the PROMIS® pediatric global health (PGH-7) measure. *Qual Life Res*. (2014) 23:1221–31. doi: 10.1007/s11136-013-0581-8

69. Forrest CB, Tucker CA, Ravens-Sieberer U, Pratiwadi R, Moon J, Teneralli RE, et al. Concurrent validity of the PROMIS® pediatric global health measure. *Qual Life Res*. (2016) 25:739–51. doi: 10.1007/s11136-015-1111-7

70. Rhew IC, Simpson K, Tracy M, Lymp J, McCauley E, Tsuang D, et al. Criterion validity of the short mood and feelings questionnaire and one- and two-item depression screens in young adolescents. *Child Adolesc Psychiatry Ment Health*. (2010) 4:8. doi: 10.1186/1753-2000-4-8

71. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The brief resilience scale: assessing the ability to bounce back. *Int J Behav Med*. (2008) 15:194–200. doi: 10.1080/10705500802222972

72. Little RJA. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*. (1988) 83:1198–202. doi: 10.1080/01621459.1988.10478722

73. Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. *J Comput Graph Stat*. (2002) 11:437–57. doi: 10.1198/106186002760180608

74. Cyranowski JM, Frank E, Young E, Shear MK. Adolescent onset of the gender difference in lifetime rates of major depression: a theoretical model. *Arch Gen Psychiatry*. (2000) 57:21–7. doi: 10.1001/archpsyc.57.1.21

75. Piccinelli M, Wilkinson G. Gender differences in depression. Critical review. *Br J Psychiatry*. (2000) 177:486–92. doi: 10.1192/bjp.177.6.486

76. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. (1989) 10:407–15. doi: 10.1016/0197-2456(89)90005-6

77. Kroenke K, Wu J, Yu Z, Bair MJ, Kean J, Stump T, et al. Patient health questionnaire anxiety and depression scale: initial validation in three clinical trials. *Psychosom Med*. (2016) 78:716–27. doi: 10.1097/PSY.0000000000000322

78. Essau CA, Lewinsohn PM, Seeley JR, Sasagawa S. Gender differences in the developmental course of depression. *J Affect Disord*. (2010) 127:185–90. doi: 10.1016/j.jad.2010.05.016

79. Polo AJ, Makol BA, Castro AS, Colón-Quintana N, Wagstaff AE, Guo S. Diversity in randomized clinical trials of depression: a 36-year review. *Clin Psychol Rev*. (2019) 67:22–35. doi: 10.1016/j.cpr.2018.09.004

80. United States Census Bureau. QuickFacts: United States. https://www.census.gov/quickfacts/fact/table/US/AGE295221 (Accessed September 22, 2022)

81. Jacobs RH, Klein JB, Reinecke MA, Silva SG, Tonev S, Breland-Noble A, et al. Ethnic differences in attributions and treatment expectancies for adolescent depression. *Int J Cogn Ther*. (2008) 1:163–78. doi: 10.1521/ijct.2008.1.2.163

82. Rushton JL, Forcier M, Schectman RM. Epidemiology of depressive symptoms in the national longitudinal study of adolescent health. *J Am Acad Child Adolesc Psychiatry*. (2002) 41:199–205. doi: 10.1097/00004583-200202000-00014

83. Morgan AJ, Jorm AF, Mackinnon AJ. Internet-based recruitment to a depression prevention intervention: lessons from the mood memos study. *J Med Internet Res*. (2013) 15:e31. doi: 10.2196/jmir.2262

84. Kutok ER, Doria N, Dunsiger S, Patena JV, Nugent NR, Riese A, et al. Feasibility and cost of using Instagram to recruit adolescents to a remote intervention. *J Adolesc Health*. (2021) 69:838–46. doi: 10.1016/j.jadohealth.2021.04.021

85. McAlindon T, Formica M, Kabbara K, LaValley M, Lehmer M. Conducting clinical trials over the internet: feasibility study. *Br Med J*. (2003) 327:484–7. doi: 10.1136/bmj.327.7413.484

86. Menezes P, Quayle J, Garcia Claro H, da Silva S, Brandt LR, Diez-Canseco F, et al. Use of a Mobile phone app to treat depression comorbid with hypertension or diabetes: a pilot study in Brazil and Peru. *JMIR Ment Health*. (2019) 6:e11698. doi: 10.2196/11698

87. Burns MN, Begale M, Duffecy J, Gergle D, Karr CJ, Giangrande E, et al. Harnessing context sensing to develop a mobile intervention for depression. *J Med Internet Res*. (2011) 13:e55. doi: 10.2196/jmir.1838

88. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. *J Med Internet Res*. (2019) 21:e14567. doi: 10.2196/14567

89. Baumel A, Edan S, Kane JM. Is there a trial bias impacting user engagement with unguided e-mental health interventions? A systematic comparison of published reports and real-world usage of the same programs. *Transl Behav Med*. (2019) 9:1020–33. doi: 10.1093/tbm/ibz147

90. Strauss G, Flannery JE, Vierra E, Koepsell X, Berglund E, Miller I, et al. Meaningful engagement: a crossfunctional framework for digital therapeutics. *Front Digit Health*. (2022) 4, 890081. doi: 10.3389/fdgth.2022.890081

91. Clarke G, Kelleher C, Hornbrook M, Debar L, Dickerson J, Gullion C. Randomized effectiveness trial of an internet, pure self-help, cognitive behavioral intervention for depressive symptoms in young adults. *Cogn Behav Ther*. (2009) 38:222–34. doi: 10.1080/16506070802675353

92. Gladstone T, Marko-Holguin M, Henry J, Fogel J, Diehl A, Van Voorhees BW. Understanding adolescent response to a technology-based depression prevention program. *J Clin Child Adolesc Psychol*. (2014) 43:102–14. doi: 10.1080/15374416.2013.850697

93. Strohmaier S. The relationship between doses of mindfulness-based programs and depression, anxiety, stress, and mindfulness: a dose-response meta-regression of randomized controlled trials. *Mindfulness (N Y)*. (2020) 11:1315–35. doi: 10.1007/s12671-020-01319-4

94. White-House-Report-on-Mental-Health-Research-Priorities.pdf. https://www.whitehouse.gov/wp-content/uploads/2023/02/White-House-Report-on-Mental-Health-Research-Priorities.pdf

95. Geraedts AS, Kleiboer AM, Twisk J, Wiezer NM, van Mechelen W, Cuijpers P. Long-term results of a web-based guided self-help intervention for employees with depressive symptoms: randomized controlled trial. *J Med Internet Res*. (2014) 16:e168. doi: 10.2196/jmir.3539

96. Mantani A, Kato T, Furukawa TA, Horikoshi M, Imai H, Hiroe T, et al. Smartphone cognitive behavioral therapy as an adjunct to pharmacotherapy for refractory depression: randomized controlled trial. *J Med Internet Res*. (2017) 19:e373. doi: 10.2196/jmir.8602

97. Graham AK, Greene CJ, Kwasny MJ, Kaiser SM, Lieponis P, Powell T, et al. Coached Mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. *JAMA Psychiatry*. (2020) 77:906–14. doi: 10.1001/jamapsychiatry.2020.1011

98. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, Wenzel J, Weiske J, Jessen F, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. *NPJ Digit Med*. (2022) 5:144. doi: 10.1038/s41746-022-00677-8

99. Wannachaiyakul S, Thapinta D, Sethabouppha H, Thungjaroenkul P, Likhitsathian S. Randomized Controlled Trial of Computerized Cognitive Behavioral Therapy Program for Adolescent Offenders with Depression. (2017). https://www.semanticscholar.org/paper/36b0e39cad6c0834642b53e4bafd1ff27eb8d4a4 (Accessed October 4, 2022).

100. Van Voorhees BW, Fogel J, Reinecke MA, Gladstone T, Stuart S, Gollan J, et al. Randomized clinical trial of an internet-based depression prevention program for adolescents (project CATCH-IT) in primary care: 12-week outcomes. *J Dev Behav Pediatr*. (2009) 30:23–37. doi: 10.1097/DBP.0b013e3181966c2a

101. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR Ment Health*. (2017) 4:e19. doi: 10.2196/mental.7785

102. Bevan Jones R, Thapar A, Stone Z, Thapar A, Jones I, Smith D, et al. Psychoeducational interventions in adolescent depression: a systematic review. *Patient Educ Couns*. (2018) 101:804–16. doi: 10.1016/j.pec.2017.10.015

103. Girgus JS, Yang K. Gender and depression. *Curr Opin Psychol*. (2015) 4:53–60. doi: 10.1016/j.copsyc.2015.01.019

104. Cyranowski JM, Frank E, Young E, Katherine Shear M. Adolescent onset of the gender difference in lifetime rates of Major depression: a theoretical model. *Ann Prog in Child Psychiatry and Child Develop 2000-2001*. (2002) 57(1):383–98. doi: 10.4324/9780203449523-19

frontiers | Frontiers in Digital Health

TYPE Brief Research Report
PUBLISHED 02 June 2023
DOI 10.3389/fdgth.2023.1063165

OPEN ACCESS

EDITED BY
Kirsten Smayda,
MedRhythms, United States

REVIEWED BY
Angel Enrique Roig,
Silvercloud Health, Ireland
Anis Davoudi,
University of Florida, United States

*CORRESPONDENCE
Andrew C. Heusser
aheusser@akiliinteractive.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 06 October 2022
ACCEPTED 10 May 2023
PUBLISHED 02 June 2023

CITATION
Heusser AC, DeLoss DJ, Cañadas E and Alailima T (2023) Leveraging machine learning to examine engagement with a digital therapeutic.
Front. Digit. Health 5:1063165.
doi: 10.3389/fdgth.2023.1063165

Leveraging machine learning to examine engagement with a digital therapeutic

Andrew C. Heusser*†, Denton J. DeLoss†, Elena Cañadas and Titiimaea Alailima

Akili Interactive, Boston, MA, United States

Digital Therapeutics (DTx) are evidence-based software-driven interventions for the prevention, management, and treatment of medical disorders or diseases. DTx offer the unique ability to capture rich objective data about when and how a patient engages with a treatment. Not only can one measure the quantity of patient interactions with a digital treatment with high temporal precision, but one can also assess the quality of these interactions. This is particularly useful for treatments such as cognitive interventions, where the specific manner in which a patient engages may impact likelihood of treatment success. Here, we present a technique for measuring the quality of user interactions with a digital treatment in near-real time. This approach produces evaluations at the level of a roughly four-minute gameplay session (mission). Each mission required users to engage in adaptive and personalized multitasking training. The training included simultaneous presentation of a sensory-motor navigation task and a perceptual discrimination task. We trained a machine learning model to classify user interactions with the digital treatment to determine if they were "using it as intended" or "not using it as intended" based on labeled data created by subject matter experts (SME). On a held-out test set, the classifier was able to reliably predict the SME-derived labels (Accuracy = .94; F1 Score = .94). We discuss the value of this approach and highlight exciting future directions for shared decision-making and communication between caregivers, patients and healthcare providers. Additionally, the output of this technique can be useful for clinical trials and personalized intervention.

KEYWORDS

digital therapeutics, engagement, machine learning, cognition, brain health

Introduction

Digital mental health interventions target the prevention or treatment of mental health disorders and associated impairments (i.e., functional, affective, cognitive) delivered via a digital platform (e.g., web browser, mobile apps, text messaging, or virtual reality) (1). They offer the potential to overcome availability and accessibility limitations, including geographical location and time (2–4).

While there are thousands of digital interventions claiming to improve various aspects of mental health, many of them have never gone through clinical trials or regulatory scrutiny. Also, due to a number of factors, including fast growth of the industry and an absence of well-accepted standards, there are widely varying definitions of what constitutes a "good" DTx (5). Contrary to wellness apps, DTx products are typically validated in rigorous clinical trials measuring safety and efficacy as well as evidence from real-world outcomes, whereas there is no such standard for wellness products (5).

Similar to traditional behavioral interventions (e.g., Cognitive Behavioral Therapy), the success of a DTx depends largely on a user's engagement (6). Broadly speaking, engagement can be described as "(a) the extent (e.g., amount, frequency, duration, depth) of usage and (b) a subjective experience characterized by attention, interest, and affect" (7). Engagement is considered to be a dynamic process that is expected to vary both within and across individuals over time (7). While data for traditional behavioral interventions is typically limited to attendance/adherence, a DTx affords the opportunity to collect rich data on when and importantly how a user interacts with the intervention.

Stakeholders across academic and industry settings acknowledge that the current measures of engagement (e.g., extent of usage) may not be sufficient (1) especially if they are not strong mediators of outcomes (8). For example, users might come back to the app every day for months (strong retention), but their symptoms do not improve. This could be interpreted as the intervention not being effective, but adherence/retention alone does not ensure that the DTx is being used as intended. The user may not have followed the instructions for use correctly, or may not have put forth significant cognitive effort and/or were distracted during use of the DTx. Another example of why standard adherence/retention methods may not be sufficient is that a user may abandon a treatment once they have achieved the desired benefits. Standard methods would predict attenuated efficacy, whereas methods focused on the quality of engagement could tell a different story. Due to their ability to collect rich data not just when but *how* a user interacts with a treatment, DTx products afford the unique opportunity to identify when a DTx product is being used as intended. Thus, an approach that provides a clinically-informed and data-driven way to measure the quality of DTx engagements may shed some light on the effectiveness of a DTx (8).

While measuring adherence and retention can be achieved by simply tracking the number of user interactions over time, assessing the quality of interactions is much more nuanced and time-consuming, and requires the expertise of trained clinicians or individuals deeply familiar with the intervention. Machine learning enables us to capture the wisdom of such experts into a classification algorithm, making this task efficient and scalable. In other words, once the classifier is trained it can be applied to large quantities of new data without the need for additional human labeling.

This manuscript introduces a machine learning-based approach to examine engagement with a pair of related DTx targeting attentional control function. These devices use proprietary algorithms designed to improve cognitive interference management in an adaptive manner and thereby personalized to the patient. Interference is instantiated through a video game-like interface presenting two tasks that are performed simultaneously (multitasking): a perceptual discrimination task (selecting the correct target from a number of distractor stimuli) and a sensory motor navigation task (continuously adjusting their position to steer towards some objects and away from others). Performance in each task is assessed during single and multitasking conditions. The interference training is adapted in real time based on the individual's performance. Thus the training is tailored specifically to each individual's performance level to achieve a consistent and optimal challenge at a predefined level of difficulty, continually challenging them to improve while providing rewards and positive feedback when they succeed.

We propose to evaluate engagement not only by examining simple adherence metrics of (e.g., sessions or total time played) but also the *quality* of the interactions with the DTx. In other words, is the user engaging with the DTx as intended (i.e., following the instructions provided, putting in an appropriate level of effort)?

Methods & results

Visualizing and labeling gameplay data

As described in the introduction, users engage in a perceptual discrimination and a sensory-motor task simultaneously for approximately 4 min per "mission". The perceptual discrimination task is performed by tapping on the screen of the device while the sensory-motor task is performed by tilting the device left and right. For a video example of gameplay, see here. Missions are the basic unit of interaction with our DTx. To develop and assess the approach we used gameplay data from 1,308 missions sampled from 427 users, including users from four studies from which data has been previously published (9–12) and users of the commercial product. We pseudo-randomly sampled missions with the goal of balancing across data source (clinical or commercial), the 4 sequentially played worlds in the game, balancing types of missions (training or assessment) proportional to how frequently they occurred in the game, and selecting a variety of performance levels. The particular software build varied across studies with some differences in game content between builds, but the tasks were substantially identical and all task difficulties were governed by our proprietary cognitive training technology, the Selective-Stimulus Management Engine (SSME^{TM}). Details on the particular instantiation of SSME for a given study can be found in the papers referenced above. The data was sourced from a number of studies across a number of indications (ADHD, Multiple Sclerosis, Major Depressive Disorder) so that our classifier could learn patterns that are not indication-specific. For the clinical trials, consent for health research and publication was provided by caregivers in the form of IRB consent (please see individual studies for details). For the commercial data, retrospective IRB-exempt status was granted under 45 CFR 46 116(f)[2018 Requirements] 45 CFR 46.116(d) [Pre-2018 Requirements] for the analysis of de-identified data by the WCG Institutional Review Board on April 21, 2023 (Study Number: 1353416).

To facilitate label generation, we created a set of plots that depict how a user is interacting with the treatment and how the treatment is dynamically responding to the user's input. The plots are generated from telemetry data that is captured as a participant engages in a mission. **Figure 1** is a schematic representing a mission "played as intended" (left panel) and "not played as intended" (right panel). For each panel, the top two plots represent game difficulty levels (solid lines) that varied

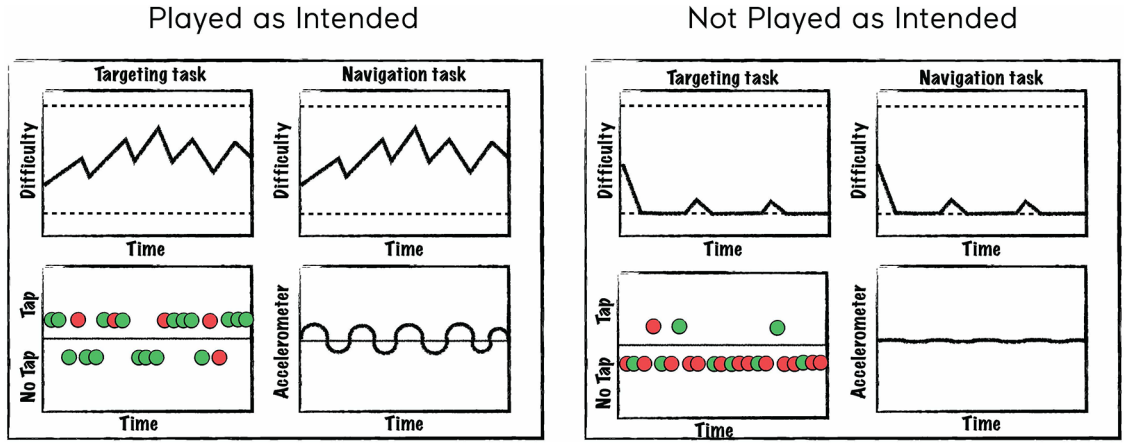


FIGURE 1
Schematic of plots created to assist in labeling. A schematic of a mission "played as intended" is represented in the left panel and "not played as intended" in the right panel. The top two plots depict difficulty levels for each of two tasks changing over time. The bottom left plot represents tapping behavior where colors indicate whether the trial was correct (green) or incorrect (red) and points above the line represent taps and points below the line represent no taps. The bottom right plot represents accelerometer input where deflections from the 0 axis indicate the degree to which the device was tilted (which controls steering in the sensory-motor navigation task).

dynamically between the top and bottom difficulty limits (dashed lines) for each of two tasks played simultaneously. The bottom left plot represents screen tapping in response to the targeting task (green = correct tap, red = incorrect tap) and the bottom right plot represents accelerometer measurements. As seen in the left panel, "playing as intended" is characterized by dynamic changes in task difficulty as the user engages with the tasks, tapping during a reasonable percentage of targeting trials with a reasonable correct rate, and continuously varying accelerometer input representing movement of the device to perform the navigation task. In contrast, the right panel depicts a mission "not played as intended", which is often characterized by task difficulty levels at the lower difficulty limit, infrequent taps and excessive errors, and little to no accelerometer input.

Labeling the plots of mission data

We trained human labelers to analyze the data presented in these plots (Figure 1) and label the missions using an agreed upon strategy. The labelers included Akili employees from various departments such as Cognitive Science, Clinical Operations, and Data Science. Before they labeled the data, they were trained by reviewing a number of plots representing various examples of gameplay (e.g., playing effortfully with the correct rules for the entire mission or playing one or both tasks with the rules systematically wrong). A labeling application was created to allow labelers to indicate the proportion of time (e.g., 0, 25, 50, 75, or 100%) during each mission where each task was "played as intended", as well as check a series of boxes if certain conditions were met (for example, if it appears they did not understand the targeting rules). These labels were only used in cases where there was reasonable certainty and typically result in accuracy levels that are far below what are in the typical range for missions. For each mission, labels were collected from 3 human labelers to increase accuracy/reduce human error. Numeric labels were transformed to binary ones for the purposes of training a binary classifier using the following operational definition for "playing as intended": multitasking for greater than 75% of a mission while playing with the correct rules. Above this threshold is considered "playing as intended" and below the threshold is considered "not playing as intended". We considered a full consensus from all raters of requiring 100% to be too stringent (and would lead to many false negatives) and that >50% was too liberal, which left us with >75% as the best option. A final label was determined based on the majority of the labels for each mission. For example, if 2 out of 3 labelers coded the mission as "playing as intended", the final label was "playing as intended".

Model features

Features were created based on aspects of the mission data that were informative in making a decision on whether the mission was "played as intended". To create features, we extracted the raw gameplay telemetry data that is captured for each mission played and transformed the data into a set of summary statistics. The feature set included statistics such as task accuracy, tapping frequency and accelerometer variance. These feature vectors were paired with the labels described above and were used to train a machine learning model to predict the most likely label ("played as intended" or "not played as intended").

Model fitting

The labeled dataset was split into a training (80%) and test (20%) set using a stratified random sampling approach (stratified

by data source and label (0 or 1). A grid search was performed on the training data over hyperparameters of a Random Forest Classifier (implemented with scikit-learn version 0.23.2) using leave-one-user-out cross validation. Using the hyperparameter combination with the highest cross-validated F1-score, the random forest was retrained on the full training set. To ensure that the classifier accuracy was not inflated due to overlap in users in the training and test set, we separately analyzed the accuracy for users' data who were only in the test set (F1-score = .96) and found them to be comparable to users in both the training and test sets (F1-score = .94).

Model results

We validated the model by assessing performance on the held out test set. Overall, test set accuracy was 94%. An ROC curve analysis representing the model's true/false positive rate at different thresholds is shown in **Figure 2**. Precision, Recall and F1-score for both "playing as intended" (1) and "not playing as intended" (0) were all exactly 94% (see **Table 1** for positive label metrics).

In addition to the model validation outlined above, we ran additional validation on new data ($n = 600$ missions) from a different set of users ($n = 220$). The labeling procedure was identical to what is described above except that there were only 2 labelers. This data was used for the purposes of model drift monitoring (i.e., the model was not retrained with this data).

Any disagreement between the labelers ("playing as intended" vs. "not playing as intended") were reviewed together live until a consensus was reached. The F1-score for this additional validation step was similar (.92), providing additional support that the model performance was high and that model drift was unlikely to be of concern.

Discussion

In this manuscript, we introduce a machine learning-based approach to examining engagement with a DTx targeting attentional control function. Our results suggest that it is feasible to label missions (the "units" of interaction with our DTx) based on whether or not they are "played as intended" with high accuracy. Importantly, this labeling can be done in an automated and scalable way (without continual expert assessment), which opens the door for many potential use cases centered around measuring the quality of engagement with a DTx.

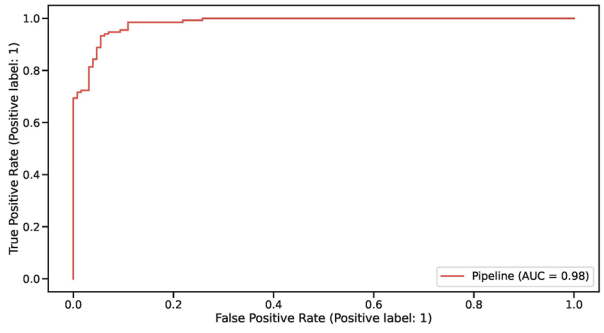


FIGURE 2
ROC curve representing model performance on the held out test set.

TABLE 1 Model performance statistics: accuracy, precision, recall, F1-score and support.

Accuracy	Precision	Recall	F1-score	# of Samples (positive/negative)
.94	.94	.94	.94	134/128

A recent opinion piece (1) calls for better measures of attrition and engagement. The approach described herein fills that unmet need in the DTx space, opening a new dimension for assessing engagement. It can also help to tease apart whether attrition is due to lack of use, or due to the manner of use, and improve the product experience accordingly. The proposed approach helps the DTx be more accurate and proactive in determining whether the patient is engaging with the product as intended and can help address many of the issues brought up in the opinion piece. The authors mention gamification of the DTx as means to increase engagement, which is core to the DTx under examination. The interactivity of the game experience produces a rich data stream that enables an approach like this to be developed. But the output of this approach can enable further gamification, such as points for completing your daily tasks in the intended manner, or simple rewards for periods of significant effort when the DTx is used as intended. The approach we describe also enables more direct feedback to the patient and/or their caregivers in a near real-time manner as to whether they were using the DTx as intended, potentially paired with further messaging to encourage proper engagement to maximize benefit. These messages should be tailored to each app and given in a positive/motivational manner, to avoid inducing frustration or dissatisfaction with the DTx for the patient.

This kind of approach has several other potential uses. It could also be used to discover cohorts of patients for any number of analyses. These cohorts could be used to identify responders, examine dosage at a much finer level, or even to predict whether a user is likely to cease using the product altogether. Different cohorts may require different types of messaging to the patient depending on their usage patterns. The approach could also provide patients or their caregivers additional insights on (a) the time course of engagement over the course of a treatment, (b) ways to get more out of the DTx by using the product as intended, and (c) any number of other communication strategies to give the patient a behavioral cue to move them into a pattern of use that is more likely to lead to greater benefit from the DTx.

Future development of this approach could include moving from a binary classification of whether the patient was using the product as intended to a continuous outcome, for example indicating the total amount of time in each daily task where they were using the product in the intended manner. This could be useful for a number of reasons. First, missions which are currently near the classification boundary and would register as

"not played as intended" could instead result in more granular feedback as to the level of engagement as intended (e.g., 73% of a given mission). A continuous output could also serve better as a feature in other models or analyses to examine patterns of usage, effectiveness of different messaging campaigns around proper use of the product, or differences in effectiveness of the DTx for different patients.

A limitation of the current approach is that the machine learning model (a random forest classifier) is moderately complex, and so explaining how the model arrived at a particular decision is not straightforward. Explainable Boosting Machines can be used to create a model that can be as accurate as a random forest while simultaneously providing output that can be easily interpreted (13). We have experimented with these models and found that they produce similar results.

While the specific methods and tooling used for our DTx will not likely transfer directly to another DTx the overall approach could be replicated with similar labeling, feature engineering, and model training efforts. It will require sufficient telemetry recorded, such that an expert observer might discern with high confidence whether or not the data stream represents use as intended. This sort of tool could become a standard feature of DTx, ensuring that products that have undergone such rigorous clinical validation can consistently prove out their benefit in the real world.

Data availability statement

The authors agree to share de-identified labels ("played as intended" or "not played as intended") and associated classifier predicted labels for each mission included in the training and test datasets following the completion of a Data Use Agreement. Proposals should be directed to medinfo@akiliinteractive.com.

Ethics statement

The studies involving human participants were reviewed and approved by WIRB-Copernicus Group [14 sites], Duke University Health System, Cincinnati Children's Hospital Medical Center, University of California Davis, University of California San Francisco, Johns Hopkins Medical Center, Western Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

All authors made a substantial, intellectual contribution to this work and approved the submitted version.

Acknowledgments

Data analysis support, under the direction of the authors, was provided by Xinyu D. Song, a former employee of Akili Interactive, in accordance with Good Publications Practice guidelines.

Conflict of interest

Akili Interactive Labs sponsored all of the studies included in this manuscript. AH, DD, EC, and TA work for or have worked for Akili Interactive and may have stock options.

Publisher's note

References

1. Nwosu A, Boardman S, Husain MM, Doraiswamy PM. Digital therapeutics for mental health: is attrition the achilles heel? *Front Psychiatry*. (2022) 13:900615. doi: 10.3389/fpsyt.2022.900615

2. Rudd BN, Beidas RS. Digital mental health: the answer to the global mental health crisis? *JMIR Ment Health*. (2020) 7(6):e18472. doi: 10.2196/18472

3. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. *JMIR Ment Health*. (2021) 8(12):e30000. doi: 10.2196/30000

4. Smith E, Cummings J, Bellgrove M, Robertson I, Wolfe J, Kirk H, et al. Addressing the cognition crisis in our COVID-19 world. *Psychiatric Times*. (2021). Available at: https://www.psychiatrictimes.com/view/addressing-the-cognition-crisis-in-our-covid-19-world (Accessed September 7, 2022).

5. Healthcare decision maker considerations. *DTx value assessment & integration guide. digital therapeutics alliance*. Digital Therapeutics Alliance (2022). Available at: https://dtxalliance.org/wp-content/uploads/2022/05/DTx-Value-Assessment-Guide_May-2022.pdf

6. Glenn D, Golinelli D, Rose RD, Roy-Byrne P, Stein MB, Sullivan G, et al. Who gets the most out of cognitive behavioral therapy for anxiety disorders? The role of treatment dose and patient engagement. *J Consult Clin Psychol*. (2013) 81 (4):639–49. doi: 10.1037/a0033403

7. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Transl Behav Med*. (2017) 7(2):254–67. doi: 10.1007/s13142-016-0453-1

8. Strauss G, Flannery JE, Vierra E, Koepsell X, Berglund E, Miller I, et al. Meaningful engagement: a crossfunctional framework for digital therapeutics. *Front Digit Health*. (2022) 4:890081. doi: 10.3389/fdgth.2022.890081

9. Bove R, Rowles W, Zhao C, Anderson A, Friedman S, Langdon D, et al. A novel in-home digital treatment to improve processing speed in people with multiple sclerosis: a pilot study. *Mult Scler J*. (2020) 27(5):778–89. doi: 10.1177/1352458520930371

10. Keefe RSE, Cañadas E, Farlow D, Etkin A. Digital intervention for cognitive deficits in major depression: a randomized controlled trial to assess efficacy and safety in adults. *Am J Psychiatry*. (2022) 179(7):482–9. doi: 10.1176/appi.ajp.21020125

11. 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 Digit Health*. (2020) 2(4):e168–78. doi: 10.1016/S2589-7500(20)30017-0

12. Kollins SH, Childress A, Heusser AC, Lutz J. Effectiveness of a digital therapeutic as adjunct to treatment with medication in pediatric ADHD. *Npj Digit Med*. (2021) 4 (1):58. doi: 10.1038/s41746-021-00429-0

13. Lou Y, Caruana R, Gehrke J, Hooker G. *Accurate intelligible models with pairwise interactions. In: proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining [internet]*. New York, NY, USA: Association for Computing Machinery; 2013. p. 623–31. (KDD '13). doi: 10.1145/2487575.2487579

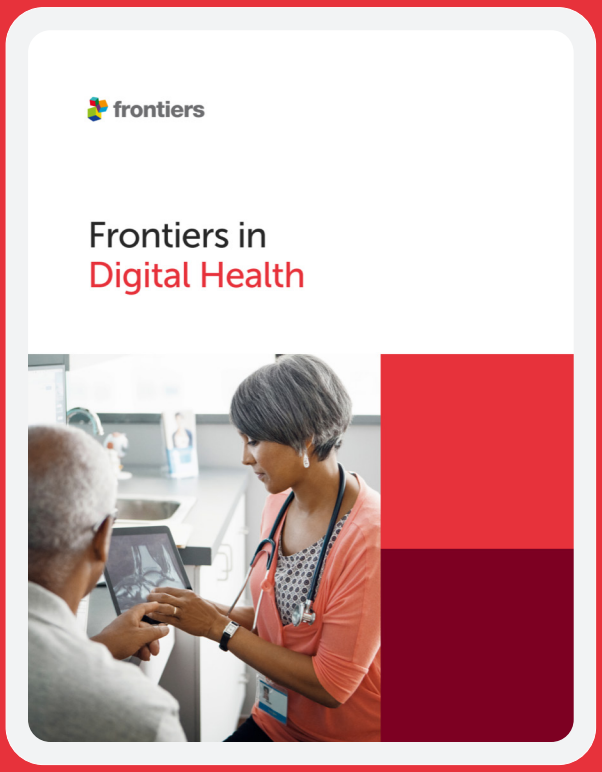
Frontiers in
Digital Health

Explores digital innovation to transform modern healthcare

A multidisciplinary journal that focuses on how we can transform healthcare with innovative digital tools. It provides a forum for an era of health service marked by increased prediction and prevention.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact