frontiers | Research Topics

# Mental health, epidemiology and machine learning

**Edited by**
Marcos Del Pozo Banos, Robert Stewart and Ann John

**Published in**
Frontiers in Psychiatry
Frontiers in Digital Health

ISSN 1664-8714
ISBN 978-2-8325-5927-7
DOI 10.3389/978-2-8325-5927-7

## 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

# Mental health, epidemiology and machine learning

**Topic editors**

Marcos Del Pozo Banos — Swansea University Medical School, United Kingdom
Robert Stewart — King's College London, United Kingdom
Ann John — Swansea University Medical School, United Kingdom

**Citation**

Del Pozo Banos, M., Stewart, R., John, A., eds. (2025). *Mental health, epidemiology and machine learning*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-5927-7

# Table of contents

frontiers | Frontiers in Psychiatry

TYPE Editorial
PUBLISHED 09 January 2025
DOI 10.3389/fpsyt.2024.1536129

OPEN ACCESS

EDITED AND REVIEWED BY
Stefan Borgwardt,
University of Lübeck, Germany

*CORRESPONDENCE
Marcos DelPozo-Banos
✉ m.delpozobanos@swansea.ac.uk

RECEIVED 28 November 2024
ACCEPTED 24 December 2024
PUBLISHED 09 January 2025

CITATION
DelPozo-Banos M, Stewart R and John A (2025) Editorial: Mental health, epidemiology and machine learning. *Front. Psychiatry* 15:1536129. doi: 10.3389/fpsyt.2024.1536129

# Editorial: Mental health, epidemiology and machine learning

Marcos DelPozo-Banos[1]*, Robert Stewart[2,3] and Ann John[1]

[1]Swansea University Medical School, Swansea, United Kingdom, [2]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [3]South London and Maudsley NHS Foundation Trust, London, United Kingdom

KEYWORDS

**mental health, epidemiology, machine learning (ML), artificial intelligence (AI), research methods, challenges and opportunities**

Editorial on the Research Topic
Mental health, epidemiology and machine learning

Globally, one in eight people live with a mental health condition, contributing to approximately 16% of disability-adjusted life years (1, 2). The significant impact of mental disorders on quality of life and life expectancy is well established and highlights significant health inequalities.[2] However, despite this, progress in mental health has lagged behind other medical fields, hindered by social stigma, cultural barriers, resource constraints, and the intrinsic complexity of mental health conditions (2).

Accessing data for mental health research is inherently challenging, due to the relevance of social and environmental factors beyond traditional health systems. Advances in data collection and linkage—including the integration of electronic health records with data from education, employment, and criminal justice—has enabled more comprehensive studies on these determinants (3, 4). However, this new data landscape presents unique analytical challenges. The DATAMIND initiative (https://datamind.org.uk/) aims to optimise the use of UK's rich mental health data, coordinating research efforts and fostering multidisciplinary collaboration.

Machine learning (ML) has emerged as a promising tool to address these new challenges, offering the power to work with large-scale data resources and produce new insights. However, ML applications in mental health must be rooted in sound epidemiological practices to ensure clinical relevance and to gain the trust of both healthcare community and the public. Our opinion piece (DelPozo-Banos et al.) discussed some of these challenges, particularly: (i) the risk of losing sight of mental health objectives in favour of technical performance; (ii) underlying biases and heightened privacy requirements; and (iii) the difficulties of building, validating and approving ML-enabled clinical devices for mental health disorders with insufficiently clear underlying mechanisms. These ideas, and the setting up of the DATAMIND hub provided impetus for the current Research Topic, titled "*Mental Health, Epidemiology, and Machine Learning*." With it, we aimed to highlight ML's potential role in mental health research and to illustrate clinically and epidemiologically sound ML applications in mental health, making the most of novel data sources and linkages.

One of the most evident applications of ML in mental health is in diagnosing complex conditions, enhancing early detection and decision support. Wright-Berryman et al. developed NLP models to identify depression, anxiety, and suicide risk in clinical records; these understandably performed better in cases where symptoms were severe or well-documented. Oh et al. also proposed NLP for depression diagnoses, but their model analyzed the emotional content in patient-psychiatrist interviews. They found that the expression of "disgust" prominently helped to distinguish patients with depression, highlighting the utility of linguistic analysis for capturing emotional markers in mental health diagnostics. Chen et al. presented a decision support tool for ADHD diagnosis, integrating ML with clinical knowledge and processing not only related symptoms, but also comorbid conditions. Their approach pointed to specific features in the Diagnostic Interview for ADHD in Adults that help distinguish ADHD from other conditions, and crucially, their model also identified and flagged complex ADHD cases for expert review. Finally, Merhbene et al. conducted a systematic review on ML for eating disorder detection, revealing challenges such as insufficient data quantity and quality, alongside a lack of representation of minority groups, reduced clinical involvement in development, and culturally driven heterogeneities. Overall, the number and heterogeneity of symptom presentations makes clinical diagnoses a highly complex task in mental healthcare (5), and these papers highlight how ML might be of value to professionals in this regard.

ML can also help to personalise mental health services and treatments to better meet patients' individual needs. Bernard et al. applied ML clustering to identify usage patterns among young users of a digital mental health platform, with a battery of sensitivity analyses across clustering methods. Their results, validated through hypothesis testing, indicated that user engagement profiles change over time, highlighting the importance of adaptive digital services tailored to changing user behaviors. Garriga et al. developed an ML model that tailors monitoring duration for psychiatric patients with a depression crisis. For over 20% of patients, their model prescribed monitoring beyond the standard one-week period, suggesting that a "one-size-fits-all" approach may overlook important individual needs. Additionally, Yao et al. analysed the satisfaction levels of Chinese psychotherapy patients, identifying cultural factors as critical determinants. While the use of ML for personalised psychiatry is not new (6), it is still under-explored. For example, in their systematic review, Rollmann et al. found only four papers investigating ML applications in psychodynamic psychotherapy, but these foundational models suggest that ML could support tailored treatments, predict treatment responses, and match therapists to patients more effectively. The need for additional research is clear, especially as personalised approaches are critical to improving therapeutic outcomes.

Suicide risk assessment and crisis prediction are areas where ML-driven personalized psychiatry can make a difference in both clinical practice and research. Chou et al. evaluated multiple ML models in a suicide risk identification task based on data from a Japanese population. They found trauma-related emotional distress and functional impairment to be important factors, demonstrating the importance of culturally contextualized risk profiles. Dutta et al. and Wright-Berryman et al. assessed suicide risk using NLP, the former on routinely collected electronic patient records from a mental health service, and the latter on 5-to-10-minute semi-structured interview data. Overall, although ML models may enhance our risk assessment capabilities, they should only be used as complements and not replacements for comprehensive clinical evaluations of patient needs.

Finally, ML can also drive the discovery of new insights on the social and environmental influences on mental health, helping to inform policies and practices. Mason et al. first used NLP to extract indicators of violence from routinely collected clinical notes of a mental healthcare provider. They fed these indicators to an ANN to identify actual experiences of violence. They found that violence-related records were more common among women, mid-life adults, ethnic minorities, and those with PTSD or schizophrenia, highlighting the intersection between demographic and clinical factors. Qasrawi et al. showed that children in violent environments exhibit cognitive and mental health patterns that align with general findings on trauma's developmental impacts. Castillo-Toledo et al. used NLP to study public perceptions of cocaine use on a large sample of social media posts, providing insights into the way some healthcare professionals openly discussed cocaine's perceived benefits. These studies demonstrate ML's capacity to identify and analyze social factors critical to mental health, contributing insights that can shape public health strategies.

In summary, the studies in this Research Topic demonstrate manifold ways in which ML might be of benefit to the field of psychiatry. They maintained a clinical focus and helpfully went beyond simple reporting and comparison of ML performance metrics. They studied the behaviour of such algorithms across varied sub-populations (e.g., by disorder severity) and tried to extract novel clinical insights, aided by additional classical statistical methods. They also openly acknowledged and discussed the limitations of their ML models and sought to validate their findings through traditional epidemiological methods.

Putting all of the above into perspective is Speechley and McTernan's central work, an opinion piece authored by people with mental health lived experience. In it, they reflected on how ML might help make sense of their lives. They highlighted the need for researchers to foster public trust, cautioning against language that could exacerbate health inequalities and stigma, and emphasizing the need to inform the public that "[their] data saves lives" and how.

Our hope is that this Research Topic serves as a catalyst for deeper conversations on ML's appropriate role in mental health research and clinical care. Most importantly, researchers must ensure that ML's transformative potential remains a positive force, advancing mental health research and clinical practice in ways that are ethical, inclusive, and grounded in real-world needs.

## Author contributions

MDPB: Writing – original draft. RS: Writing – review & editing. AJ: Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. All authors were funded by the Medical Research Council and Health Data Research UK (Grant DATAMIND: Data Hub for Mental Health INformatics research Development, with Ref.: MR/W014386/1). RS is additionally part-funded by: i) the NIHR Maudsley Biomedical Research Centre at the South London and Maudsley NHS Foundation Trust and King's College London; ii) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust; iii) the UK Prevention Research Partnership (Violence, Health and Society; MR-VO49879/1), an initiative funded by UK Research and Innovation Councils, the Department of Health and Social Care (England) and the UK devolved administrations, and leading health research charities.

## Conflict of interest

AJ chairs the National Advisory Group on Suicide and Self-harm Prevention to Welsh Government. RS declares research funding/support from GSK and Takeda in the last 3 years.

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

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. *EClinicalMedicine*. (2022) 54. doi: 10.1016/j.eclinm.2022.101675

2. World mental health report: transforming mental health for all. Geneva: World Health Organization (2022). Licence: CC BY-NC-SA 3.0 IGO.

3. John A, Friedmann Y, DelPozo-Banos M, Frizzati A, Ford T, Thapar A. Association of school absence and exclusion with recorded neurodevelopmental disorders, mental disorders, or self-harm: a nationwide, retrospective, electronic cohort study of children and young people in Wales, UK. *Lancet Psychiatry*. (2022) 9:23–34. doi: 10.1016/S2215-0366(21)00367-9

4. John A, Rouquette OY, Lee SC, Smith J, del Pozo Baños M. Trends in incidence of self-harm, neurodevelopmental and mental health conditions among university students compared with the general population: nationwide electronic data linkage study in Wales. *Br J Psychiatry*. (2024) 225:389–400. doi: 10.1192/bjp.2024.90

5. Newson JJ, Hunter D, Thiagarajan TC. The heterogeneity of mental health assessment. *Front Psychiatry*. (2020) 11:76. doi: 10.3389/fpsyt.2020.00076

6. Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, Irving J, Catalan A, Oliver D, et al. Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. *Schizophr Bull*. (2021) 47:284–97. doi: 10.1093/schbul/sbaa120

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 04 August 2022
DOI 10.3389/fpsyt.2022.918667

# A machine-learning model to predict suicide risk in Japan based on national survey data

Po-Han Chou[1,2*†], Shao-Cheng Wang[3,4,5†], Chi-Shin Wu[6,7*], Masaru Horikoshi[8] and Masaya Ito[8]

[1]Department of Psychiatry, China Medical University Hsinchu Hospital, China Medical University, Hsinchu, Taiwan, [2]Department of Psychiatry, China Medical University Hospital, China Medical University, Taichung, Taiwan, [3]Department of Psychiatry, Taoyuan General Hospital, Ministry of Health and Welfare, Taoyuan, Taiwan, [4]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States, [5]Department of Medical Laboratory Science and Biotechnology, Chung Hwa University of Medical Technology, Tainan, Taiwan, [6]National Center for Geriatrics and Welfare Research, National Health Research Institutes, Zhunan Town, Yunlin County, Taiwan, [7]Department of Psychiatry, National Taiwan University Hospital, Douliu, Taiwan, [8]National Center for Cognitive-Behavior Therapy and Research, National Center of Neurology and Psychiatry, Tokyo, Japan

Check for updates

OPEN ACCESS

EDITED BY
Ashwani Kumar Mishra,
All India Institute of Medical Sciences, India

REVIEWED BY
Shaw-Ji Chen,
Taitung MacKay Memorial Hospital, Taiwan
Chih-Wei Hsu,
Kaohsiung Chang Gung Memorial Hospital, Taiwan

*CORRESPONDENCE
Po-Han Chou
phchou1980@gmail.com
Chi-Shin Wu
chishinwu@nhri.edu.tw

†These authors have contributed equally to this work and share first authorship

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

RECEIVED 12 April 2022
ACCEPTED 11 July 2022
PUBLISHED 04 August 2022

CITATION
Chou P-H, Wang S-C, Wu C-S, Horikoshi M and Ito M (2022) A machine-learning model to predict suicide risk in Japan based on national survey data.
*Front. Psychiatry* 13:918667.
doi: 10.3389/fpsyt.2022.918667

**Objective:** Several prognostic models of suicide risk have been published; however, few have been implemented in Japan using longitudinal cohort data. The aim of this study was to identify suicide risk factors for suicidal ideation in the Japanese population and to develop a machine-learning model to predict suicide risk in Japan.

**Materials and methods:** Data was obtained from Wave1 Time 1 (November 2016) and Time 2 (March 2017) of the National Survey for Stress and Health in Japan, were incorporated into a suicide risk prediction machine-learning model, trained using 65 items related to trauma and stress. The study included 3,090 and 2,163 survey respondents >18 years old at Time 1 and Time 2, respectively. The mean (standard deviation, SD) age was 44.9 (10.9) years at Time 1 and 46.0 (10.7) years at Time 2. We analyzed the participants with increased suicide risk at Time 2 survey. Model performance, including the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity, were also analyzed.

**Results:** The model showed a good performance (AUC = 0.830, 95% confidence interval = 0.795–0.866). Overall, the model achieved an accuracy of 78.8%, sensitivity of 75.4%, specificity of 80.4%, positive predictive value of 63.4%, and negative predictive value of 87.9%. The most important risk factor for suicide risk was the participants' Suicidal Ideation Attributes Scale score, followed by the Sheehan Disability Scale score, Patient Health Questionnaire-9 scores, Cross-Cutting Symptom Measure (CCSM-suicidal ideation domain, Dissociation Experience Scale score, history of self-harm, Generalized Anxiety Disorder-7 score, Post-Traumatic Stress Disorder check list-5 score, CCSM-dissociation domain, and Impact of Event Scale-Revised scores at Time 1.

**Conclusions:** This prognostic study suggests the ability to identify patients at a high risk of suicide using an online survey method. In addition to confirming several well-known risk factors of suicide, new risk measures related to trauma and trauma-related experiences were also identified, which may help guide future clinical assessments and early intervention approaches.

KEYWORDS

machine learning, Super Learner, suicide, prediction model, suicide risk, Japan

# Introduction

Despite advances in the diagnosis and treatment of mental illness, suicide remains to be a major public health problem, with annual suicide rates at approximately 10–12 per 100,000 people over the past 60 years (1). Therefore, an increased understanding of the risk factors for suicide is important for early intervention and prevention. For the past 50 years, extensive work has been conducted to improve the prediction of suicide, yet a recent published meta-analysis demonstrated that using known suicide risk factors leads to modest results (weighted area under the receiver operating characteristic curve [AUC], 0.58) (2). Several factors may have led to this prediction failure (3). Firstly, suicidal rate in the population is relatively low, making prospective studies not practical (1). Second, prior studies were often limited to small samples, measured at a single time point, and examined few number of factors. Finally, the traditional method for statistical analysis of the suicide data mainly focus on inference, which resulted in simple prediction models; lastly, they are not designed to incorporate new clinical data to continuously update the existing models (3).

Recently, novel statistical analyses, such as machine learning, and big data sources, such as electronic health records or national survey data, have led to enormous improvements in predicting suicide risk in clinical practice (AUC, 0.63–0.94) (1, 4–7). However, most of the published work mainly focused on high-risk groups who sought for medical treatment (8). As it has been reported that more than one-third of the people attempting suicide do not actively seek medical treatment (9), it is essential to extend suicide prediction models beyond the treatment-seeking populations to the general population. Previous studies using population-based cohort data to build suicide prediction models have also yielded fair performance in general adult (10, 11) and adolescent (12) populations (AUC, 0.62–0.86). However, because of the variable suicide rates among countries with different cultural backgrounds, the suicide prediction model in one country may not be generalizable to another.

In view of the aforementioned limitations, we aimed to identify important risk factors for future suicide risk in a longitudinal cohort from the National Survey for Stress and Health (NSSH) dataset (13) in Japan, using an explanatory machine-learning model. This study aims to extend prior research in two directions. First, we used a large longitudinal sample to identify risk factors for suicidal ideation in the Japanese population. Second, we included an extensive assessment instrument that included detailed psychometric assessments for substance use, psychiatric disorders, personality traits, and clinical symptoms, which are not routinely available in electronic health records or administrative data. Overall, we expected to develop a model predicting suicide risk in a longitudinal cohort in Japan.

# Methods

## Database

The data of the present study were extracted from the National Survey for Stress and Health (NSSH), conducted between 2016 and 2017. Detailed information on NSSH can be found in our previous work (13–15). In brief, two waves of surveys were conducted. Wave 1 ($n = 3{,}090$) consisted of screening (November 2016), Time 1 (November 2016), and Time 2 surveys (March 2017). Wave 2 ($n = 3{,}090$) consisted of screening and the Time 1 survey (both in March 2017) (15). Recruitment emails were sent to 100,077 panelists in November (Wave 1). The target sample size in our study was 6,000 individuals, including 3,000 patients who met the probable diagnostic criteria based on the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) for posttraumatic stress disorder (PTSD) using the PCL-5, 1,000 non-clinical responders denting any past traumatic experience, and 2,000 non-clinical or subclinical responders with traumatic experiences. We terminated the screening when reaching half of the target sample size (i.e., 3,000 participants). The screened participants answered questions measuring their psychiatric symptoms and psychological processes at Times 1 and 2. Only participants at Wave 1 participated the Time 2 survey, which was conducted 4 months after Time 1.

All participants had read a full explanation of the research project and gave informed consent before answering the questionnaires. All survey contents were examined with design, logical flow, validity, and checking for errors by nine experienced psychologists and double-checked by two macromill survey engineers. To improve the data quality, the online survey system automatically excluded responders who answers the questions rapidly. Because the survey was designed to not allow participants to proceed if there are unanswered items, no data were missing except for income. This study was approved by the Institutional Review Board of the National Center of Neurology and Psychiatry (approval number: A2015-086).

## Participants

This study used longitudinal data collected in Wave 1, including Time 1 survey data ($n$ = 3,090) and self-reported suicide ideation at the follow-up 4 months later (Time 2, $n$ = 2,163). The cumulative response rate for Wave 2 was 66.7 %.

## Assessment of risk factors at time 1

### Demographics

Personal information, including sex, age, income, marital status, substance use, history of physical or psychological abuse, or self-harm behavior; diagnosis and treatment for any psychiatric disorder, including major depressive disorder (MDD), bipolar disorder, dysthymic disorder, seasonal affective disorder, obsessive compulsive disorder, panic disorder, PTSD, generalized anxiety disorder, psychotic disorder, and eating disorder were recorded (Table 1).

## Measures

### PCL-5

We used the Japanese version of the PCL-5 to assess PTSD symptoms of the responders. The PCL-5 comprises a 20-item assessment, available from the National Center for PTSD (13). The 20 items are concordant with the DSM-5 diagnostic items for PTSD. Each question were answered with a 5-point Likert scale (0 = not at all, 1 = a little bit, 2 = moderately, 3 = quite a bit, 4 = extremely).

### Trauma-related guilt inventory (TRGI)

The TRGI was developed by Kubany to assess the emotional and cognitive aspects of guilt associated with a specific traumatic event (16). The final version consists of 32 items on six scales: the Guilt Cognition Scale (which comprises three empirically derived subscales: Hindsight-Bias/Responsibility (seven items), Wrongdoing (five items), and Insufficient Justification (four items), along with an additional six general cognition items), the Distress Scale (six items), and the Global Guilt Scale (four items). The answers for all 32 items were recorded on a 5-point scale, with poles from "extremely true/always true" to "not at all true/never true" (eight items were reverse-scored).

TABLE 1 Characteristics of study participants.

| Characteristic | Participants of wave 1 | Wave 1 participants Follow-ups |
|---|---|---|
| | $N = 3090$ | $N = 2163$ |
| **Female** | 1,509 (48.8%) | 990 (45.8%) |
| **Age (mean ± SD)** | 44.9 ± 10.9 | 46 ± 10.7 |
| **Marital status** | | |
| Married | 1,466 (47.4%) | 1,038 (48.0%) |
| **Personal yearly income (Japanese yen)**[a] | | |
| 0–1,999,999 | 1,460 (54.3%) | 1,019 (49.7%) |
| 2,000,000–3,999,999 | 572 (21.3%) | 420 (20.5%) |
| 4,000,000–5,999,999 | 320 (11.9%) | 237 (11.6%) |
| 6,000,000–7,999,999 | 197 (7.3%) | 140 (6.8%) |
| 8,000,000–9,999,999 | 80 (3.0%) | 60 (2.9%) |
| Over 10,000,000 | 40 (1.5%) | 36 (1.7%) |
| **Hx of physical abuse** | 1,222 (39.6%) | 818 (37.8%) |
| **Hx of emotional abuse** | 1,875 (60.7%) | 1,284 (59.4%) |
| **Hx of self-harm behavior** | 778 (25.2%) | 539 (24.9%) |
| **Psychiatric comorbidities diagnosed and treated in medical settings** | | |
| MDD | 1,630 (52.8%) | 1,159 (53.6%) |
| Bipolar disorder | 335 (10.8%) | 232 (10.7%) |
| Dysthymic disorder | 341 (11.0%) | 234 (10.8%) |
| SAD | 489 (15.8%) | 350 (16.2%) |
| GAD | 487 (15.8%) | 350 (16.2%) |
| Panic disorder | 630 (20.4%) | 434 (20.0%) |
| OCD | 471 (15.2%) | 332 (15.4%) |
| PTSD | 422 (13.7%) | 289 (13.4%) |
| Psychosis | 321 (10.4%) | 210 (9.7%) |
| Eating disorder | 293 (9.5%) | 202 (9.3%) |
| **SIDAS score at Time 1** | 20.0 +/− 9.7 | 19.8 +/− 9.5 |
| **SIDAS score at Time 2** | | 19.6 +/− 9.1 |

[a] 1 Japanese yen is approximately equal to 0.0074 US dollar.

SD, standard deviation; MDD, major depressive disorder; SAD, seasonal affective disorder; GAD, generalized anxiety disorder; OCD, obsessive compulsive disorder; PTSD, post-traumatic stress disorder; SIDAS, suicidal ideation attributes scale.

### Impact of event scale-revised (IES-R)

The Impact of Event Scale-revised (IES-R) is a widely used self-report measure in the field of traumatic stress (17). It contains 22 questions used to assess the core psychological phenomena of traumatic stress: intrusion (eight questions), avoidance (eight questions), and hyperarousal (six questions). A scoring scheme with intervals of 0, 1, 2, 3, and 5 was adopted for responders regarding the degree to which they were distressed or bothered by the listed conditions in the past 7 days from "not at all," "a little bit," "moderately," "quite a bit," to "extremely."

### Patient health questionnaire-9 (PHQ-9)

The PHQ-9 is a nine-item assessment for depressive symptoms experienced for the past 2 weeks (18). Responses were rated on a 4-point Likert scale (0 = not at all, 3 = nearly every day). The reliability and validity of PHQ-9 have been established in previous studies.

### Generalized anxiety disorder 7-item scale (GAD-7)

The GAD-7 assesses symptoms of generalized anxiety experienced over the past 2 weeks (19). A seven-item questionnaire was developed that asked participants how often they were bothered by the listed anxiety symptoms during the past 2 weeks. The response options were "not at all," "several days," "more than half the days," and "nearly every day," scored as 0, 1, 2, and 3, accordingly. Scores of 5, 10, and 15 were used as the cutoff points for mild, moderate, and severe anxiety, respectively.

### Sheehan disability scale (SDS)

The SDS is a three-item assessment for functional impairment in three domains: work/school, social life, and family life/home responsibility (20); higher scores imply more severe functional impairment.

### Cut-annoyed-guilty-eye (CAGE) questionnaire

The CAGE is used for brief assessment of alcoholism (21). This questionnaire comprises four items: desire to reduce drinking, annoyance at being criticized for drinking, feeling guilty about drinking, and drinking in the morning to wake up. Participants responded with yes/no answers.

### Tobacco dependence screener (TDS)

The TDS is a ten-item questionnaire for screening tobacco dependence, as defined by the Tenth revision of the International Statistical Classification of Diseases and Related Health Problems, DSM-III-Revision, and DSM-IV (22). The participants provided yes/no answers on each item.

### Patient-reported version of the level 1 cross-cutting symptom measure for the DSM-5 (CCSM)

The Level 1 CCSM is a 23-item assessment of 13 domains of symptoms common to psychiatric disorders (23). Test-retest reliability for each domain was fair in a DSM-5 field trial.

### Eysenck personality questionnaire revised-short form (EPQR-S)

The EPQR-S is a self-report questionnaire consisting of 48 items, 12 for each trait of neuroticism, extraversion, and psychoticism, and 12 on the lie scale. Each question has a binary "yes" or "no" response. The dichotomous item is scored as 1 or 0, and each scale has a maximum score of 12 and a minimum of zero (24).

### Posttraumatic maladaptive beliefs scale (PMBS)

The PMBS is a 15-item scale developed to measure maladaptive beliefs about current life circumstances following trauma exposure (25). This scale assesses maladaptive beliefs in three domains: (a) threat of harm, (b) self-worth and judgment, and (c) reliability and trustworthiness of others. Each item included in the PMBS was rated using a 7-point Likert-type response format, ranging from one (not at all true) to seven (completely true). A list of subscale items and reverse code directions are indicated on the measure. The possible scores range from 15 to 105, and the subscale scores range from 5 to 35.

### Emotion regulation questionnaire (ERQ)

The ERQ is a 10-item self-report scale to assess habitual use of two commonly used strategies for emotional regulation: cognitive reappraisal and expressive suppression (26). Responders answered each item with a 7-point Likert scale ranging from one (strongly disagree) to seven (strongly agree). Cognitive reappraisal involves thinking about a situation in a different perspective to change its meaning to alter one's emotional experience. Expressive suppression means a decrease in the outward expression of emotions. There are six items contributing to the subscale for cognitive reappraisal (e.g., "When I'm faced with a stressful situation, I make myself think about it in a way that helps me stay calm") and four items contributing to the subscale for expressive suppression (e.g., "When I am feeling negative emotions, I make sure not to express them").

### Satisfaction with life scale (SWLS)

The SWLS is a 5-item scale designed to measure global cognitive judgments of one's life satisfaction (not a measure of either positive or negative affect) (27). Participants indicated their agreement/disagreement with each of the five items using a 7-point scale that ranges from seven (strongly agree) to one (strongly disagree).

### Dissociative experiences scale (DES)

The DES is a 28-item self-report measure of dissociative experience. In the newer DES format, respondents circle a percentage, ranging from 0 to 100% at 10% intervals, indicating their agreement with the question. The DES score is the average of all questions; therefore, the minimum score is 0 and the maximum score is 100. All the questions are scored by dropping the zero on the percentage of each answer, e.g., 30% = 3; 80% = 8, these numbers are then added up. Scores of 30 or higher indicate high levels of dissociation.

### The anxiety sensitivity index-3 (ASI-3)

The ASI-3 is an 18-item self-report measure developed to assess anxiety sensitivity (28). Each item is rated on a 5-point Likert scale ranging from zero ("not at all") to four ("very much"); the higher the score, the more severe the anxiety sensitivity.

### Positive emotion in distress scale (PEIDS)

The PEIDS is a 10-item Japanese self-report scale that assesses positive emotions during negative affective states, including broaden-and-build theory (29). Participants were asked to read each item and indicate the extent of their agreement or disagreement with 10 statements. Items were scored on a five-point Likert scale (1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, and 5 = strongly agree).

### Affective style questionnaire (ASQ)

The ASQ is a 20-item scale used to assess emotion regulation in terms of three affective styles: concealing, adjusting, and tolerating (30). The items were measured using a five-point Likert scale.

## Outcome at time 2: Suicidal risk

### Suicidal ideation attributes scale (SIDAS)

The SIDAS was used assesses the severity of suicidal ideation over the preceding month. There are five items asking the frequency, controllability, closeness to suicide attempt, level of distress associated with suicidal thoughts, and impact on daily function (31). Answers are responded with an 11-point Likert scale. The SIDAS was assessed both at Time 1 and Time 2. Responders with SIDAS scores of 21 or higher were regarded as having a risk of suicide (31). In the present study, we used SIDAS score at time 1 as a covariate and SIDAS score at time 2 as the outcome in the prediction model.

## Statistical analysis

### Sensitivity analysis

We used Student's *t*-tests and chi-square tests to compare the characteristics between those who were lost to follow-up and participants at Time 2.

### Model building and validation

We used Super Learner to develop the suicidality risk prediction model. Super Learner is an ensemble algorithm that uses a stacking process to determine the optimal weighted combination of a set of candidate algorithms using cross-validation to minimize the value of loss function (32). The values of weighted and loss function are considered the coefficient and risk. Super Learner can include many diverse algorithms and perform equally or better than the best-performing candidate algorithms. In process, we divided the data randomly into two sets: 70% into a training set and 30% into a test set. We estimated the risk of each algorithm using a 10-fold cross-validation. Super Learner combined all the candidate algorithms to generate a new algorithm with the best performance. All analyses were conducted using R version 4.2 (https://cran.r-project.org) and Super Learner 2.0–28 to develop the prediction models. In this study, we used 20 candidate algorithms to generate SuperLearners, including generalized linear mode, Bayesian generalized linear models, general additive model, five elastic-net regularized generalized linear models with alpha from zero to one with an increment of 0.25, kernel k nearest neighbors, support vector machine, linear discriminant analysis, neural networks, multivariate adaptive polynomial spline regression, random forests, and six extreme gradient boosting models by a grid of shrinkage parameter (0.1 and 0.01) with the number of terminal nodes (1, 2, and 4) (32). We found the performance of Super Learner is better than that of any specific algorithm (Supplementary Material I). The risk and coefficients are shown in Table 2. We then evaluated the performance of the model using a test dataset. The indicators of model performance included the AUC, sensitivity, specificity, and accuracy.

### Identifying the top 10 risk factors

Given that Super Learner is a black box model, we used the random forest algorithm to train a model for the predicted value from Super Learner and to identify the variable importance measures for each predicator by calculating the increase in mean-squared errors, which indicated a decrease in accuracy after permutation of a predictor. The top 10 important risk factors were identified in this study. Furthermore, to address the problem of collinearity of the included variables, we measured the co-linearity using variable inflation factors (VIF). If the VIF ≥10, it indicated there is serious collinearity requiring correction. The results showed that the maximum of the VIFs

TABLE 2 Relative importance of the 10 top factors based on the suicide prediction model using measurements collected from the Time 1 responses of National Survey of Stress and Health.

| | %IncMSE |
|---|---|
| SIDAS | 5.85E−03 |
| SDS | 3.26E−03 |
| PHQ-9 | 3.19E−03 |
| CCSM-suicidal | 2.69E−03 |
| DES score | 2.67E−03 |
| Past history of self-harm | 2.22E−03 |
| GAD-7 | 1.97E−03 |
| PCL-5 | 1.78E−03 |
| CCSM-dissociation | 1.76E-03 |
| IES-R | 1.66E−03 |

SIDAS, suicidal ideation attributes scale; SDS, Sheehan disability scale; PHQ-9, Patient health questionnaire-9; CCSM, cross-cutting symptom measure for DSM-5; DES, Dissociative Experiences Scale; GAD-7, Generalized anxiety disorder 7-item scale; PCL-5, The PTSD Checklist for DSM-5; IES-R, Impact of event scale-revised.

of the variables is 6.038. Therefore, the possibility of collinearity of the included variables is less likely.

# Results

## Clinical characteristics of the study population

The baseline characteristics of study participants are shown in Table 1. Data from a total 3,090 respondents were analyzed (mean age, 44.9 ± 10.9 years; 48.8% female) at Time 1 of Wave 1, and 2,163 participants completed the survey at Time 2. There were no significant differences in the demographic characteristics between those who were lost to follow-up and those who remained in the study at Time 2. Among the responders, the most common traumatic experience was emotional abuse (60.7%, $n = 1{,}875$), followed by physical violence (39.6%, $n = 1\ 222$). The most common psychiatric comorbidity was MDD (52.8%), followed by panic disorder (20.4%), seasonal affective disorder (15.8%), and generalized anxiety disorder (15.8%).

## Performance of the suicide prediction model

A total of 65 factors were included as features to build the model. The model trained with 65 features showed a good performance (AUC = 0.830, 95% confidence interval [CI] = 0.795–0.866) in predicting future suicide risk (Figure 1). Overall, the model achieved an accuracy of 78.8%, sensitivity of 75.4%, specificity of 80.4%, positive predictive value (PPV) of 63.4%, and negative predictive value of 87.9%.

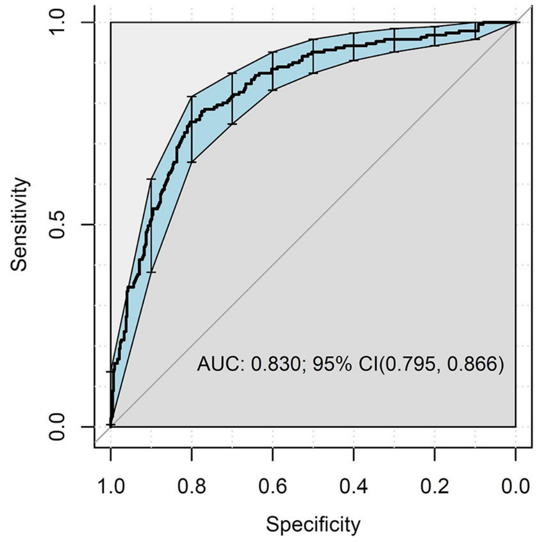


FIGURE 1
Area under the receiver operating curve of the predictive models of increased suicide risk.

## Variable importance

The mean square error (%IncMSE) was used to evaluate variable importance in the model. Table 2 shows the 10 most important variables from the super-learner model. The most important risk factor was SIDAS score at Time 1. Other risk factors included the SDS score, PHQ-9 scores, CCSM-suicidal ideation domain, DES score, history of self-harm, GAD-7 score, PCL-5 score, CCSM-dissociation domain, and IES-R scores at baseline.

# Discussion

This is the first study to apply a machine-learning algorithm to online survey data to develop a model for predicting suicide risk in the general Japanese population. To our knowledge, few studies have integrated population-based datasets with machine-learning methods to predict suicide risk (10–12). The performance of our prediction model (AUC = 0.83, sensitivity = 75.4%, specificity = 80.4%) was similar to those previous studies using machine-learning approach in the general adult population in the United States (10) (AUC = 0.86, sensitivity = 85.3%, specificity = 73.3%) and South Korea (11) (AUC = 0.85, sensitivity = 83.6%, specificity = 80.7%), and much better than those using traditional methodology (AUC = 0.58) (2).

Models for predicting suicidal outcomes that were developed in prior studies have been criticized for having low PPVs (≤50% in most models), which precluded their readiness for clinical applications in health care systems (33). Our model achieved a clinically actionable PPV (63.4%). These results are encouraging, given the recent emphasis on models in the general adult population using big data and their usefulness in developing precision treatment protocols for individuals at risk for suicide (8, 18).

One noteworthy finding in our study was that the most important risk factor in our prediction model was the baseline SIDAS. The SIDAS has proven to be a valid web-based measure for the severity of suicidal ideation. A previous study reported that scores ≥21 had a 95.8% specificity for the presence of a suicide plan in the past year and a 94.9% specificity for the presence of suicidal preparation/attempt in the past year (31). Our results indicated that SIDAS could be a good predictor of suicide risk.

Moreover, our results extend prior work by revealing the predictive value of variables related to functional impairment in three major life domains: work, social life/leisure activities, and family life/home responsibilities, as assessed by SDS, which are not covered in commonly used screening tools for suicide risk assessment. These findings may offer a new direction for improving suicidal behavior prediction through functional assessments.

Other important novel risk factors were related to emotional responses to traumatic experiences. The PCL-5, DES, and IES-R scores were moderate risk factors. The IES-R (34), PTSD symptoms (14, 15) and dissociative symptoms (35) are known risk factors for suicide in patients with traumatic experiences. Therefore, future assessment tools for suicide should include responders' past traumatic experiences and their related psychological consequences.

## Limitations

This study had several limitations. First, we employed an online self-report survey methodology to assess suicide risk and clinical and functional correlates. Although participation in the study was anonymous, the use of online surveys may increase the endorsement of sensitive responses due to increased anonymity (36). Furthermore, our results cannot be generalized to face-to-face interview assessments; however, our psychometric information may be useful for online epidemiological surveys or telemedicine. Second, we only included data from participants aged 18 years and older, and the risk factors identified might not be generalizable to children and adolescents. Third, we lacked information about suicide among participants lost to follow-up (i.e., Time 2 non-responders). Yet, the results of the sensitivity analysis showed that there was no significant difference in baseline demographic characteristics between those who were lost to follow-up and those under follow-up at Time 2. Fourth, the participants were limited to those who had Internet access and were registered as panelists for the survey company. To be specific, our study sample was relatively young and had lower personal income than the general Japanese adult population. The generalizability of these findings to other population remains unclear.

## Conclusions

Our study demonstrated the usefulness of machine learning methods to generate powerful suicide prediction models in a longitudinal cohort. We confirmed several well-known risk factors for suicide, such as the SIDAS and PHQ-9, while identifying new important risks. Specifically, functional impairment and emotional distress related to traumatic experiences emerged as novel, important factors in suicidal behavior. We hope that these results deepen our understanding of the etiology of suicide in adults and improve suicide prediction by identifying new risk variables to guide the future development of suicide risk assessment tools.

## 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 National Center of Neurology and Psychiatry. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

C-SW analyzed the data. P-HC and S-CW drafted the manuscript. MI conceived and designed the study, managed study administration, and including the ethical review process. MH and MI provided critical comments on the manuscript related to intellectual content. All authors contributed to the article and approved the submitted version.

## Funding

This study was supported by a Grant-in-Aid for Scientific Research (A) (15H01979), awarded to MH, from the Japan Society for the Promotion of Science, Tokyo, Japan. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

## 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/fpsyt.2022.918667/full#supplementary-material

## References

1. Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky ZA, et al. Machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ Digit Med.* (2020) 3:78. doi: 10.1038/s41746-020-0287-6

2. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull.* (2017) 143:187–232. doi: 10.1037/bul0000084

3. Boudreaux ED, Rundensteiner E, Liu F, Wang B, Larkin C, Agu E, et al. Applying machine learning approaches to suicide prediction using healthcare data: overview and future directions. *Front Psychiatry.* (2021) 12:707916. doi: 10.3389/fpsyt.2021.707916

4. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *j Child Psychol Psychiatry.* (2018) 59:1261–70. doi: 10.1111/jcpp.12916

5. choi sb, lee w, yoon jh, won ju, kim dw. ten-year prediction of suicide death using cox regression and machine learning in a nationwide retrospective cohort study in South Korea. *J Affect Disord.* (2018) 231:8–14. doi: 10.1016/j.jad.2018.01.019

6. Walsh CG, Johnson KB, Ripperger M, Sperry S, Harris J, Clark N, et al. Prospective validation of an electronic health record-based, real-time suicide risk model. *JAMA network open.* (2021) 4:e211428. doi: 10.1001/jamanetworkopen.2021.1428

7. Su C, Aseltine R, Doshi R, Chen K, Rogers SC, Wang F. Machine learning for suicide risk prediction in children and adolescents with electronic health records. *Transl Psychiatry.* (2020) 10:413. doi: 10.1038/s41398-020-01100-0

8. Kessler RC, Bossarte RM, Luedtke A, Zaslavsky AM, Zubizarreta JR. Suicide prediction models: a critical review of recent research with recommendations for the way forward. *Mol Psychiatry.* (2020) 25:168–79. doi: 10.1038/s41380-019-0531-0

9. Luoma JB, Martin CE, Pearson JL. Contact with mental health and primary care providers before suicide: a review of the evidence. *Am J Psychiatry.* (2002) 159:909–16. doi: 10.1176/appi.ajp.159.6.909

10. Garcia de. la Garza A, Blanco C, Olfson M, Wall MM. Identification of suicide attempt risk factors in a National US Survey Using Machine Learning. *JAMA psychiatry.* (2021) 78:398–406. doi: 10.1001/jamapsychiatry.2020.4165

11. Ryu S, Lee H, Lee DK, Park K. Use of a machine learning algorithm to predict individuals with suicide ideation in the general population. *Psychiatry Investig.* (2018) 15:1030–6. doi: 10.30773/pi.2018.08.27

12. Navarro MC, Ouellet-Morin I, Geoffroy MC, Boivin M, Tremblay RE, Cote SM, et al. Machine learning assessment of early life factors predicting suicide attempt in adolescence or young adulthood. *JAMA network open.* (2021) 4:e211450. doi: 10.1001/jamanetworkopen.2021.1450

13. Ito M, Takebayashi Y, Suzuki Y, Horikoshi M. posttraumatic stress disorder checklist for DSM-5: psychometric properties in a Japanese population. *J Affect Disord.* (2019) 247:11–9. doi: 10.1016/j.jad.2018.12.086

14. Chou PH, Ito M, Horikoshi M. Associations between PTSD symptoms and suicide risk: a comparison of 4-factor and 7-factor models. *J Psychiatr Res.* (2020) 129:47–52. doi: 10.1016/j.jpsychires.2020.06.004

15. Chu CS, Chou PH, Wang SC, Horikoshi M, Ito M. Associations between PTSD symptom custers and longitudinal changes in suicidal ideation: comparison between 4-factor and 7-factor models of DSM-5 PTSD Symptoms. *Front Psychiatry.* (2021) 12:680434. doi: 10.3389/fpsyt.2021.680434

16. Kubany ES, Haynes SN, Abueg FR, Manke FP, Brennan JM. Stahura C. Development and validation of the Trauma-Related Guilt Inventory (TRGI). *Psychol Assess.* (1996) 8:428. doi: 10.1037/1040-3590.8.4.428

17. Weiss DS, Marmar CR. The Impact of Event Scale—Revised. In: Wilson JP, Keane TM, editors. *Assessing Psychological Trauma and PTSD: A Handbook for Practitioners.* New York: Guilford Press (1997). pp. 399–411.

18. 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

19. Spitzer RL, Kroenke K, Williams JB, Lowe B, A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med.* (2006) 166:1092–7. doi: 10.1001/archinte.166.10.1092

20. Sheehan DV, Harnett-Sheehan K, Raj BA. The measurement of disability. *Int Clin Psychopharmacol.* (1996) 11 (Suppl 3):89–95. doi: 10.1097/00004850-199606003-00015

21. Ewing JA. Detecting alcoholism. The CAGE questionnaire. *JAMA.* (1984) 252:1905–7. doi: 10.1001/jama.252.14.1905

22. Kawakami N, Takatsuka N, Inaba S, Shimizu H. Development of a screening questionnaire for tobacco/nicotine dependence according to ICD-10, DSM-III-R, and DSM-IV. *Addict Behav.* (1999) 24:155–66. doi: 10.1016/S0306-4603(98)00127-0

23. Narrow WE, Clarke DE, Kuramoto SJ, Kraemer HC, Kupfer DJ, Greiner L, et al. DSM-5 field trials in the United States and Canada, Part III: development and reliability testing of a cross-cutting symptom assessment for DSM-5. *Am J Psychiatry.* (2013) 170:71–82. doi: 10.1176/appi.ajp.2012.12071000

24. Eysenck SBG, Eysenck H, J. Barrett P. A revised version of the psychoticism scale. *Pers Individ Differ.* (1985) 6:21–9. doi: 10.1016/0191-8869(85)90026-1

25. Vogt DS, Shipherd JC, Resick PA. Posttraumatic maladaptive beliefs scale: evolution of the personal beliefs and reactions scale. *Assessment.* (2012) 19:308–17. doi: 10.1177/1073191110376161

26. Gross JJ, John OP. Individual differences in two emotion regulation processes: implications for affect, relationships, and wellbeing. *J Pers Soc Psychol.* (2003) 85:348–62. doi: 10.1037/0022-3514.85.2.348

27. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. *J Pers Assess.* (1985) 49:71–5. doi: 10.1207/s15327752jpa4901_13

28. Taylor S, Zvolensky MJ, Cox BJ, Deacon B, Heimberg RG, Ledley DR, et al. Robust dimensions of anxiety sensitivity: development and initial

validation of the anxiety sensitivity index-3. *Psychol Assess.* (2007) 19:176–88. doi: 10.1037/1040-3590.19.2.176

29. Yamaguchi K, Ito M, Takebayashi Y. Positive emotion in distress as a potentially effective emotion regulation strategy for depression: a preliminary investigation. *Psychol Psychother.* (2018) 91:509–25. doi: 10.1111/papt.12176

30. Hofmann SG, Kashdan TB. The affective style questionnaire: development and psychometric properties. *J Psychopathol Behav Assess.* (2010) 32:255–63. doi: 10.1007/s10862-009-9142-4

31. van Spijker BA, Batterham PJ, Calear AL, Farrer L, Christensen H, Reynolds J, et al. The Suicidal Ideation Attributes Scale (SIDAS): community-based validation study of a new scale for the measurement of suicidal ideation. *Suicide Life Threat Behav.* (2014) 44:408–19. doi: 10.1111/sltb.12084

32. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol.* (2007) 6:1–21. doi: 10.2202/1544-6115.1309

33. Belsher BE, Smolenski DJ, Pruitt LD, Bush NE, Beech EH, Workman DE, et al. Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA psychiatry.* (2019) 76:642–51. doi: 10.1001/jamapsychiatry.2019.0174

34. Sharif Nia H, Kaur H, Fomani FK, Rahmatpour P, Kaveh O, Pahlevan Sharif S, et al. Psychometric Properties of the Impact of Events Scale-Revised (Ies-R) among general iranian population during the COVID-19 pandemic. *Front Psychiatry.* (2021) 12:692498. doi: 10.3389/fpsyt.2021.692498

35. Calati R, Bensassi I, Courtet P. the link between dissociation and both suicide attempts and non-suicidal self-injury: meta-analyses. *Psychiatry Res.* (2017) 251:103–14. doi: 10.1016/j.psychres.2017.01.035

36. Brown LA, Contractor A, Benhamou K. Posttraumatic stress disorder clusters and suicidal ideation. *Psychiatry Res.* (2018) 270:238–45. doi: 10.1016/j.psychres.2018.09.030

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 19 January 2023
DOI 10.3389/fpsyt.2023.947081

# Prediction of Chinese clients' satisfaction with psychotherapy by machine learning

Lijun Yao[1], Ziyi Wang[2], Hong Gu[1], Xudong Zhao[1], Yang Chen[2] and Liang Liu[1]*

[1]Clinical Research Center for Mental Disorders, Shanghai Pudong New Area Mental Health Center, School of Medicine, Tongji University, Shanghai, China, [2]Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China

Check for updates

OPEN ACCESS

EDITED BY
Jinsong Tang,
Zhejiang University, China

REVIEWED BY
Hong Huang,
Huazhong University of Science and Technology, China
Felix Beierle,
National Institute of Informatics, Japan
Yanli Luo,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Liang Liu
✉ goldwalker@163.com

SPECIALTY SECTION
This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry

RECEIVED 18 May 2022
ACCEPTED 02 January 2023
PUBLISHED 19 January 2023

CITATION
Yao L, Wang Z, Gu H, Zhao X, Chen Y and Liu L (2023) Prediction of Chinese clients' satisfaction with psychotherapy by machine learning.
*Front. Psychiatry* 14:947081.
doi: 10.3389/fpsyt.2023.947081

**Background:** Effective psychotherapy should satisfy the client, but that satisfaction depends on many factors. We do not fully understand the factors that affect client satisfaction with psychotherapy and how these factors synergistically affect a client's psychotherapy experience.

**Aims:** This study aims to use machine learning to predict Chinese clients' satisfaction with psychotherapy and analyze potential outcome contributors.

**Methods:** In this cross-sectional investigation, a self-compiled online questionnaire was delivered through the WeChat app. The information of 791 participants who had received psychotherapy was used in the study. A series of features, for example, the participants' demographic features and psychotherapy-related features, were chosen to distinguish between participants satisfied and dissatisfied with the psychotherapy they received. With our dataset, we trained seven supervised machine-learning-based algorithms to implement prediction models.

**Results:** Among the 791 participants, 619 (78.3%) reported being satisfied with the psychotherapy sessions that they received. The occupation of the clients, the location of psychotherapy, and the form of access to psychotherapy are the three most recognizable features that determined whether clients are satisfied with psychotherapy. The machine-learning model based on the CatBoost achieved the highest prediction performance in classifying satisfied and psychotherapy clients with an F1 score of 0.758.

**Conclusion:** This study clarified the factors related to clients' satisfaction with psychotherapy, and the machine-learning-based classifier accurately distinguished clients who were satisfied or unsatisfied with psychotherapy. These results will help provide better psychotherapy strategies for specific clients, so they may achieve better therapeutic outcomes.

KEYWORDS

**psychotherapy, therapy satisfaction, online survey, machine learning, prediction model**

## Introduction

Psychotherapy is regarded as an approach in which professionally trained clinicians inspire and facilitate changes in the perspectives, emotions, and behaviors of clients using guided conversations and special techniques (1). To date, psychotherapy has proven effective for clients or patients with various clinical complaints, such as depression, anxiety, obsessive-compulsive disorder, alcohol abuse, personality disorder, and children's mental health complaints (2, 3). However, previous studies have indicated that not all clients were satisfied with psychotherapy, and many factors may influence clients' responses to psychotherapy (4, 5).

Ignoring the factors affecting client's satisfaction with psychotherapy may generate many problems. For example, some clients with special complaints, certain ages or occupations, specific education levels, or economic conditions may not be suitable for certain types of psychotherapy (6). The lack of target clients, clinical complaints and theoretically therapy-oriented practices may lead to excessive energy consumption for clients and practitioners (7). Therefore, research that clarifies factors predicting client response to psychotherapy may contribute significantly to the design and improvement in clinicians' daily interventions. It may potentially reduce the time and economic costs related to psychotherapy for both clinicians and clients (5, 7).

Previously, efforts to predict clients' responses to psychotherapy have focused on theoretically motivated variables that may influence therapeutic outcomes. In general, the variables contributing to good therapy response can be grouped into five categories. First, previous research has suggested a significant link of client's satisfaction with strong, supportive, trustworthy, and collaborative *therapist-client alliance* (5, 6, 8, 9). Second, previous studies implied that *settings of psychotherapy*, including more efficient registration procedures, appropriate appointment times and venues, involvement of family members, moderate frequency of interviews, and appropriate session durations, were related to better results in psychotherapy (8, 10, 11). Third, associations between client satisfaction and *personality traits and professional competence of therapists* have also been demonstrated in previous studies. Therapist factors that positively affect the clients' psychotherapy experience include the therapist's patience, affinity, enthusiasm, humor, meticulousness, authenticity, professional sensitivity, ability to sort out complex information, empathy, theoretical interpretation, training background of psychiatry, psychoanalytic investigation ability, and facilitative interpersonal skills (5, 8, 12, 13). Fourth, previous research implied that some pretreatment patient characteristics, such as client preferences, severe mental symptoms, depression with less comorbid anxiety, middle age, and unwillingness to accept psychotherapy and medication, were correlated with poorer therapy outcomes (5, 6, 14–16). Fifth, therapy theoretical orientations, strategies, and skills have also been found to be important factors affecting therapeutic outcomes, although conclusions have varied across different studies. Regarding the theoretical orientation of psychotherapy, previous research implied that clients accepting psychodynamic therapy reported more experience with side effects than other treatments, such as family (systemic) therapy, humanistic psychotherapy, and cognitive behavioral therapy (CBT) (17, 18). Concerning therapy skills, in family and psychodynamic therapy, therapeutic techniques and strategies that have shown to be helpful include circular questioning, genograms, homework, visualization techniques, reformulating, metaphor, reflecting team, reframing, promoting individual development, expressing acknowledgment, facilitating emotional flow, self-exploration, and coping with daily practical issues (5, 8, 19–22).

Another issue requiring clarification is the relationship between psychotherapy satisfaction and objective clinical outcomes such as symptom reduction, social function improvement, and increased wellbeing. Psychotherapy satisfaction refers to the client's positive appraisal of the outcomes and process attributes of a therapy. It is a prominent indicator of the quality of therapy and belongs to the subjective experience of clients (23). Although psychotherapy satisfaction does not necessarily demonstrate a one-to-one relationship with objectively assessed therapy outcomes (24), previous research suggested that they were closely correlated and contributed to each other (25–27). Meanwhile, prior studies have taken both as important indicators for psychotherapy effectiveness (28). However, the mechanism of how these two variables influence each other remains unclear. In the current study, we took psychotherapy satisfaction as an indicator of the client's therapy effectiveness.

Previous literature suggests that most previous analyses on the predictors of client satisfaction with psychotherapy have been conducted using *a priori* programming of fixed solutions with a specific theoretical hypotheses or through qualitative approaches (29). Only recently have machine-learning approaches been used to predict outcomes of psychotherapy. Machine learning is an emerging area of artificial intelligence that implements a classification or prediction model in a data-driven and no-hypotheses way. To date, studies using machine learning to predict client response to psychological talk therapies can be categorized into two groups. The first cluster of studies includes those predicting psychotherapy outcomes from certain *pretreatment characteristics* of the clients. These characteristics included the client's demographic, psycho-social and clinical characteristics (e.g., age, ethnicity, gender, economic status, social support, life events, personality trait, the severity of symptoms, and comorbidity), electronic medical records, structured interview data, and brain function (29–34). For example, regarding demographic, psycho-social, and clinical characteristics, Green et al. (35) built a machine-learning model with five pretreatment factors to predict depressed patient's response to psychotherapy. Those variables included the client's ethnicity, gender, deprivation, and initial depression and anxiety severity. Their model predicted a reduction of depression symptoms with an accuracy of 74.9% (35). Similarly, Buckman et al. (36) used clinical data such as anxiety and depression symptoms, alcohol use, life events, and social support to predict depressive patients' remission after 3–4 months of therapy in primary care settings. The prediction power of the nine machine learning models they built was acceptable. Additionally, Gori et al. (37) applied artificial neural network (ANN) technology to analyze the predictive effect of clients' personality data on their psychotherapy outcome. Their model showed a mean rate of correct classification of 81% in forecasting successful and unsuccessful treatment cases (37). As for brain function, a machine learning analysis on the CBT outcomes of 38 schizophrenic patients implied that psychotic and affective symptom improvement was related to participants' neural responses to facial affect across frontal-limbic, sensorimotor, and frontal regions (38). A longitudinal study on 49 panic patients who received CBT found that patients' pretreatment whole brain signals were good predictors of their response to therapy (39).

The second group refers to the process-outcome studies that predicted psychotherapy outcomes based on data *during or in between sessions*, such as theoretical orientations of therapy (e.g., CBT, interpersonal therapy), therapist's interventions (e.g., therapist's specific conversation strategies, psychodynamic assessment, and intensity of therapy), therapist-client interactions (e.g., psychotherapy conversation text, smartphone messages, session notes and transcripts, session audio acoustics, and video), client's real-time response to therapy (e.g., completion of the homework assignment, ambient smartphone data, and biomarkers during a session) (32, 40–44). As for theoretical orientations of psychotherapy, Chekroud et al. (33) suggested that some multivariable modeling methods, such as "personalized advantage index" (PAI), could be used to identify which evidence-based therapy approach (e.g., CBT, interpersonal therapy, and psychodynamic therapy) might be effective in patients with complaints including major depression and post-traumatic stress disorder (PTSD). Regarding therapist's interventions, in a large-scale study on the discourse of text-message-based psychotherapy conversations, Althoff et al. (45) found that actionable conversation strategies were linked with clients' higher therapy satisfaction. Similarly, an investigation of 14,899 patients suggested that certain therapist utterances of CBT, such as change methods, were associated with more patient engagement and improvement in symptoms (46). As for therapist-client interactions, Nasir et al. (41) found that couple therapy outcomes were closely related to the behavioral interaction and acoustics of the spoken interactions, such as vocal intonation and intensity, during the therapy sessions. Regarding the client's real-time response, Wallert et al.'s (40) study applied machine-learning technologies to estimate patients' adherence to internet CBT for depression and anxiety after myocardial infarction. The strongest predictors included self-assessed cardiac-related fear, sex and the number of words the patient used to finish the homework assignment (40). Meanwhile, Chekroud et al. (33) suggested that a valuable future research direction is to track a patient's real-time response during treatment (e.g., self-reported outcome/symptom measures) and enter them into a machine learning computer system. Then the computerized system might predict the patient's improvement trajectories by comparing it to an established clinical database (47).

Generally, the current research using machine learning approach to predict psychotherapy outcome and satisfaction is still preliminary. As described above, client satisfaction with psychotherapy is affected by many factors, including clients' factors, psychotherapists' factors, and specific strategic factors. We hope to use a variety of machine-learning methods to study this issue from multiple angles. Meanwhile, to date, the majority of studies predicting clients' response to psychotherapy using machine learning were from Western countries. Although several studies conducted external validation to test the generalizability of certain machine learning (ML) models (43, 48), there is a lack of studies that build an artificial model to predict Chinese clients' satisfaction with psychotherapy. Comparatively, the Chinese tradition emphasizes more on individuals' emotional bonding with their families than Western culture. Meanwhile, Chinese families are more influenced by Confucianism and place more emphasis on an individual's obedience to authority (5, 49). This implies that some special psychotherapy theories, methods or settings, such as systemic family therapy, psychoeducation and being treated in medical institutes, may act as important contributors to client satisfaction with psychotherapy (5, 6). However, the potential factors that contribute to Chinese clients' therapy satisfaction and the mechanism by which these underlying factors interact with each other remain unclear. This restriction in research may hinder the development and tailoring of more effective psychotherapy strategies.

Thus, in this study, we collected information from both clients and psychotherapists and applied seven types of machine learning algorithms, as well as biostatistics, aiming to (1) identify the most important factors that affect client satisfaction with psychotherapy and (2) design and implement a classifier based on supervised machine learning to predict whether clients are satisfied with psychotherapy. Based on the results of our study, the classifier can provide a predictive outcome of a specific client's satisfaction with psychotherapy. Meanwhile, we can improve our treatment strategy to provide clients with more personalized therapy services.

# Materials and methods

## Participants

From 5 July to 28 August, 2021, individuals who had received psychotherapy were recruited *via* the WeChat platform. Each of them was asked to complete an electronic questionnaire using their WeChat account. WeChat is a representative mobile social networking platform in China, with more than one billion users. The inclusion criteria for participants were as follows: (1) received or were receiving psychotherapy; (2) had at least one therapy session in the past 4 months; (3) aged 12–60; and (4) agreed to join the investigation and signed the informed consent. Participants were excluded for the following reasons: (1) being diagnosed with severe physical diseases; (2) unable to understand the questions in the investigation; and (3) severe mental disorders with a risk of self-harm.

## Questionnaire

Based on the purpose of the research and a review of literatures, a questionnaire containing the relevant demographic information of the participants and their therapists, as well as certain characteristics of psychotherapy, was compiled. The questionnaire collected the information from participant's demographic data (age, gender, ethnic group, marriage status, occupation, education, and family economic status), status of psychotherapy (finished or ongoing), psychotherapist's gender and age, time of the last session, the form of therapy (individual, group, family/couple, and integrative form), the form of access to psychotherapy (face to face, audio, and video), the location where they received the therapy (welfare organization, medical institutes, commercial counseling agency, school, and other), cost per session, qualification of the therapist, the theoretical orientation of therapy (humanistic therapy, systemic therapy, psychodynamic or psychoanalysis therapy, CBT, integrated therapy, or unclear), number of psychotherapists (how many therapists the participant had seen by the time of survey), order of therapy being reflected on, number of sessions (how many sessions had taken place during the therapy being reflected on), diagnosis by psychiatrists, and whether the participant received medication. The client's satisfaction with psychotherapy was judged based on the answer to the last question: "In general, are you satisfied with the psychotherapy you received?" "Yes" was classified as being satisfied with the psychotherapy; otherwise, the participant was dissatisfied. Detailed information on each feature is listed in **Table 1**. A total of 15 participants were invited to complete the initial questionnaire and provided their feedback on the content. The final version of the questionnaire was achieved through revisions based on the feedback.

## Procedure

The questionnaire was shared and distributed *via* the WeChat platform. When clicking on the online questionnaire, participants first read a brief introduction about the investigation, such as the aims of the study and the inclusion and exclusion criteria. Then, they decided whether to participate in the survey. The participants were asked to agree and click "yes, I agree to join this investigation" to indicate their informed consent before starting the questionnaire. It was an anonymous investigation. WeChat users or participants who joined the investigation were encouraged to share the investigation in their WeChat moments. They were also asked to forward the investigation to other WeChat groups that they belonged to and to share the questionnaire with their WeChat friends including clients, psychotherapists, psychiatrists, social workers, and schoolteachers. The investigation was accomplished by the participants using either the mobile app or the PC-based interface of WeChat. The completion time for the whole survey was about 5 min. Every participant could complete the survey only once. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Shanghai Pudong New Area Mental Health Center, Tongji University School of Medicine. Informed consent was obtained from each participant.

## Modeling using machine learning

In this study, we leveraged machine-learning technologies to predict participants' satisfaction with psychotherapy and evaluated the predictive performance of training models. Our machine-learning-based modeling process has four key steps, namely, pre-processing of raw data, selection of features, selection of algorithms, and tuning of the parameters. Finally, we compare the predictive performance of all the models and choose the best classifier. **Figure 1** shows the detailed workflow. The method was also described in our previous study (50). Scikit-learn 1.1.3, [1]a well-known machine-learning library based on the Python language, was applied to train prediction models (51).

## Pre-processing of raw data and selection of features

In our dataset, 619 participants were satisfied- and 172 participants were unsatisfied with psychotherapy (**Table 1**). We select 30 features according to the mutual information, and use L1 normalization to pre-process the features. We randomly split the whole dataset into one training/validation subset and one test subset (**Figure 1**). We used 70% of the participants' data for training and validation, and used the remaining 30% of the participants' data for testing (52–54). The training/validation subset was used to both train and validate the prediction model, and the test subset was used to evaluate the performance of the model. To address the problem of unbalanced samples, for the training/validation subset, we used the synthetic minority oversampling technique (SMOTE) approach (55) to oversample the minority type of participants. We further applied the five-fold cross-validation approach to prepare the training/validation subset, where the training/validation subset was randomly divided into five groups of equal size. Of the five groups, one group was retained as the validation data to evaluate the model, and the remaining four groups were used for training. We repeated the cross-validation procedure five times, and each of the five groups was used once for validation.

The selected features were designed to reflect the different aspects of participants who underwent psychotherapy. In this study, our features included the demographic information of the participants and information related to psychotherapy. These features are either numerical or categorical. **Table 1** describes the selected features in detail.

## Selection of the algorithm and parameter tuning

To obtain the best prediction model, we selected classical algorithms to implement supervised machine learning, such as logistic regression, decision tree, random forest, and support vector machine (SVM), as well as some emerging approaches, including LightGBM, XGBoost, and CatBoost. For each of these selected algorithms, we aimed to find a "best" parameter set. Based on the training/validation subset, we swept through the parameter space using the grid search approach. We selected a set of possible values of each parameter to form the parameter space. The grid search iterated through each combination of parameters. For each parameter combination, we calculated the prediction performance. To avoid bias, we apply a nested approach, i.e., we repeat the random split for training/validation subset and the test subset for 10 times, and record the average prediction performance. In the end, the parameters leading to the best average prediction performance will be recorded. Our model can now be used to judge a new client's satisfaction with psychotherapy based on the input information.

## Evaluation of the model performance

To quantify the predictive performance of the trained models, we adopted the following three classic metrics: precision, recall, and F1 score (56). Precision indicates the fraction of the model comprising participants who actually satisfied with their psychotherapy. Recall indicates the fraction of participants with psychotherapy satisfaction who have been correctly uncovered by the model. The F1 score represents the harmonic mean of the precision and the recall metric. The best F1 score is 1, and the worst is zero. A higher F1 score indicates better predictive performance of a model.

[1] https://scikit-learn.org/stable/

TABLE 1 Features of participants involved in the study.

| Feature | Satisfied (*n* = 619) | Unsatisfied (*n* = 172) | Overall (%) (*n* = 791) | *P*-value |
|---|---|---|---|---|
| Gender | | | | |
| *Female* | 291 (47.0%) | 97 (56.4%) | 388 (49.1%) | 0.031* |
| *Male* | 328 (53.0%) | 74 (43.0%) | 402 (50.8%) | |
| *Transgender* | 0 | 1 (0.6%) | 1 (0.1%) | |
| Age (mean ± SD) | 29.8 ± 8.3 | 31.3 ± 9.7 | 31.1 ± 8.6 | 0.071 |
| Ethnic group | | | | 0.638 |
| *Han* | 594 (96.0%) | 163 (94.8%) | 757 (95.7%) | |
| *Minority* | 25 (4.0%) | 9 (5.2%) | 34 (4.3%) | |
| Marriage status | | | | 0.806 |
| *Single* | 197 (31.8%) | 54 (31.4%) | 251 (31.7%) | |
| *Single with partner* | 97 (15.7%) | 26 (15.1%) | 123 (15.5%) | |
| *Married without children* | 46 (7.4%) | 13 (7.6%) | 59 (7.5%) | |
| *Married with children* | 262 (42.3%) | 71 (41.3%) | 333 (42.1%) | |
| *Divorced or widowed* | 17 (2.8%) | 8 (4.6%) | 25 (3.2%) | |
| Occupation | | | | 0.0004*** |
| *Enterprise and institution staff* | 255 (41.2%) | 57 (33.1%) | 312 (39.4%) | |
| *Student* | 101 (16.3%) | 28 (16.3%) | 129 (16.3%) | |
| *Civil servant* | 41 (6.6%) | 6 (3.5%) | 47 (5.9%) | |
| *Medical personnel* | 14 (2.3%) | 10 (5.8%) | 24 (3.0%) | |
| *Teacher* | 41 (6.6%) | 12 (7.0%) | 53 (6.7%) | |
| *Self-employed* | 101 (16.3%) | 21 (12.2%) | 122 (15.4%) | |
| *Others* | 66 (10.7%) | 38 (22.1%) | 104 (13.3%) | |
| Education | | | | 0.011** |
| *Junior high school and below* | 29 (4.7%) | 12 (7.0%) | 40 (5.2%) | |
| *High school* | 126 (20.4%) | 30 (17.4%) | 156 (19.7%) | |
| *Undergraduate* | 409 (66.1%) | 101 (58.7%) | 510 (64.5%) | |
| *Master and Ph.D. degree* | 55 (8.8%) | 29 (16.9%) | 84 (10.6%) | |
| Family economic status | | | | 0.160 |
| *Poor* | 19 (3.1%) | 10 (5.8%) | 29 (3.7%) | |
| *Ordinary* | 419 (67.7%) | 119 (69.2%) | 538 (68.0%) | |
| *Good* | 181 (29.2%) | 43 (25.0%) | 224 (28.3%) | |
| Status of psychotherapy | | | | 0.179 |
| *Ongoing* | 425 (68.7%) | 101 (58.7%) | 526 (66.5%) | |
| *Finished* | 194 (31.3%) | 71 (41.3%) | 265 (33.5%) | |
| Gender of psychotherapist | | | | 0.153 |
| *Female* | 370 (59.8%) | 92 (53.5%) | 462 (58.4%) | |
| *Male* | 247 (39.9%) | 78 (45.3%) | 325 (41.1%) | |
| *Transgender* | 2 (0.3%) | 2 (1.2%) | 4 (0.5%) | |
| Age of psychotherapist | | | | 0.039* |
| *< 30* | 75 (12.1%) | 19 (11.0%) | 94 (11.9%) | |
| *30–40* | 364 (58.8%) | 90 (52.3%) | 454 (57.4%) | |
| *40–50* | 159 (25.7%) | 49 (28.5%) | 208 (26.3%) | |
| *≥ 50* | 21 (3.4%) | 14 (8.1%) | 35 (4.4%) | |

*(Continued)*

TABLE 1 (Continued)

| Feature | Satisfied ($n$ = 619) | Unsatisfied ($n$ = 172) | Overall (%) ($n$ = 791) | $P$-value |
|---|---|---|---|---|
| Time of the last session | | | | 0.128 |
| *Last week* | 169 (27.3%) | 37 (21.5%) | 206 (26.0%) | |
| *1 week∼1 month* | 252 (40.7%) | 67 (39.0%) | 319 (40.3%) | |
| *>1 month* | 198 (32.0%) | 68 (39.5%) | 266 (33.7%) | |
| Form of psychotherapy | | | | 0.811 |
| *Individual therapy* | 411 (66.4%) | 116 (67.4%) | 537 (67.9%) | |
| *Group therapy* | 19 (3.1%) | 7 (4.1%) | 26 (3.3%) | |
| *Family/couple therapy* | 139 (22.5%) | 35 (20.3%) | 174 (22.0%) | |
| *Integrative therapy* | 40 (6.5%) | 14 (1.8%) | 54 (6.8%) | |
| Form of access to psychotherapy | | | | 0.003** |
| *Video* | 43 (7.0%) | 27 (15.7%) | 70 (8.8%) | |
| *Audio* | 94 (15.2%) | 27 (15.7%) | 121 (15.3%) | |
| *Face-to-face* | 358 (57.8%) | 82 (47.7%) | 440 (55.6%) | |
| *Mixed* | 122 (19.7%) | 34 (19.8%) | 156 (19.7%) | |
| *Others* | 2 (0.3%) | 2 (1.1%) | 4 (0.5%) | |
| Location of psychotherapy | | | | 0.002** |
| *Welfare organization* | 31 (5.0%) | 16 (9.3%) | 47 (6.0%) | |
| *Hospital* | 201 (32.5%) | 44 (25.6%) | 245 (31.0%) | |
| *School* | 61 (9.8%) | 19 (11.0%) | 80 (10.1%) | |
| *Commercial counseling agency* | 309 (50.0%) | 79 (45.9%) | 388 (49.0%) | |
| *Others* | 17 (2.7%) | 14 (8.2%) | 31 (3.9%) | |
| Cost per session (mean ± SD) | 460.5 ± 810.5 | 643.9 ± 1,142 | 500.4 ± 895.5 | 0.050* |
| Qualification of the psychotherapist | | | | 0.735 |
| *School teacher* | 47 (7.6%) | 12 (7.0%) | 59 (7.5%) | |
| *Psychologist* | 308 (49.8%) | 92 (53.5%) | 400 (50.6%) | |
| *Psychotherapist* | 195 (31.5%) | 45 (26.2%) | 240 (30.3%) | |
| *Social worker* | 14 (2.2%) | 4 (2.3%) | 18 (2.3%) | |
| *Psychiatrist* | 54 (8.7%) | 19 (11.0%) | 73 (9.2%) | |
| *No qualification* | 1 (0.2%) | 0 | 1 (0.1%) | |
| Theoretical orientation of psychotherapy | | | | 0.042* |
| *Humanistic therapy* | 51 (8.2%) | 14 (8.1%) | 65 (8.2%) | |
| *Systemic therapy* | 111 (17.9%) | 19 (11.1%) | 130 (16.4%) | |
| *Integrated therapy* | 66 (10.7%) | 18 (10.5%) | 84 (10.6%) | |
| *Psychodynamic or psychoanalysis* | 247 (39.9%) | 90 (52.3%) | 337 (42.6%) | |
| *Cognitive behavioral therapy* | 134 (21.7%) | 27 (15.7%) | 161 (20.4%) | |
| *Unclear* | 10 (1.6%) | 4 (2.3%) | 14 (1.8%) | |
| Number of psychotherapists (mean ± SD) | 2.4 ± 1.6 | 2.4 ± 1.9 | 2.4 ± 1.7 | 0.961 |
| Order of the therapy (mean ± SD) | 2.2 ± 1.9 | 2.1 ± 1.1 | 2.2 ± 1.8 | 0.439 |
| Number of sessions (mean ± SD) | 9.4 ± 23.1 | 14.9 ± 29.8 | 10.6 ± 24.8 | 0.027* |
| Diagnosis by psychiatrist | | | | 0.140 |
| *Yes* | 308 (49.8%) | 74 (43.0%) | 382 (48.3%) | |
| *No* | 311 (50.2%) | 98 (57.0%) | 409 (51.7%) | |
| Receiving medicine | | | | 0.110 |
| *Yes* | 177 (28.6%) | 38 (22.1%) | 215 (27.2%) | |
| *No* | 442 (71.4%) | 134 (77.9%) | 576 (72.8%) | |

Satisfied, participants satisfied with psychotherapy; unsatisfied, participants unsatisfied with psychotherapy. *$P < 0.05$ was considered statistically significant; **$P < 0.01$; ***$P < 0.001$.

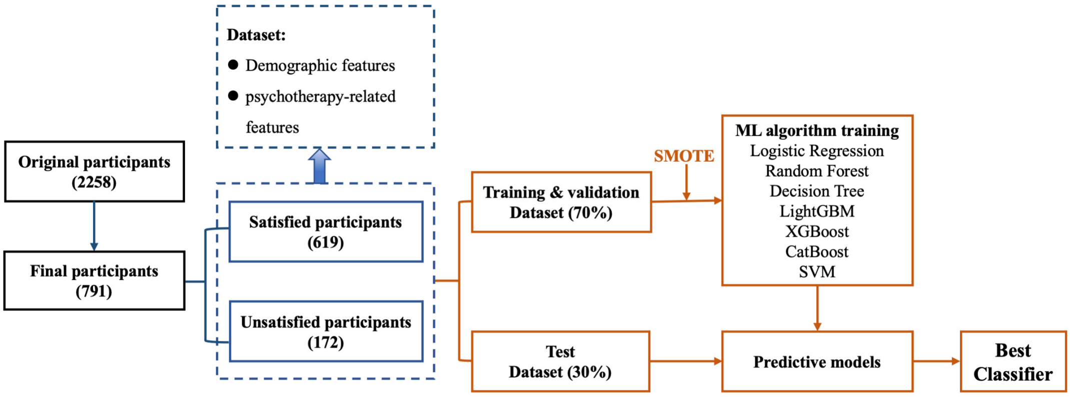


FIGURE 1
The flowchart of data processing and machine learning-based modeling. The raw dataset was processed by removing non-compliant data entries to form the dataset used in the study. The dataset consisted of participants' demographic features, and psychotherapy-related features were split into a training and validation dataset and a test dataset. Different machine learning algorithms were selected for training based on the training and validation dataset. Predictive models were obtained after parameter tuning. The final classifier was determined according to the comparison of each trained model's prediction performance using the test dataset.

## Statistics

We implemented statistical analysis using the Python programming language. The numerical variables were represented in the form of the mean ± standard deviation (SD) (Table 1); categorical variables were shown as numbers and percentages. *P*-values in Table 1 were obtained by using the chi-square test (57). A *p*-value less than 0.05 was considered statistically significant. Chi-square ($\chi^2$) statistics were used to quantify the dependence of each selected feature and the groups of participants (satisfied or dissatisfied with psychotherapy) (57). A larger $\chi^2$ value indicates that a feature has higher discriminative power.

# Results

## The demographics of the participants

In total, 925 participants completed the original questionnaire. By removing non-compliant data entry, the information of 791 (85.5%) participants were finally analyzed in our study. In the dataset, 619 participants reported that they were satisfied with the psychotherapy they received, while 172 participants expressed dissatisfaction with the therapy (Table 1). The incidence of participants who were satisfied with psychotherapy was 78.3%. Each participant's data contained 22 main features, which can be numerical (such as age, cost per session, number of therapies received, order of therapy being reflected on, and number of sessions) or categorical (other features). Detailed information on the participant number, percentage, and *p*-value of each feature is shown in Table 1.

A total of 328 (53.0%) male participants were satisfied with the therapy they received, while fewer female participants (291, 47.0%) were satisfied ($p = 0.031$, see also Figure 2F). The average cost per session of the unsatisfied group was much higher than that of the satisfied group (643.9 ± 1,142 vs. 460.5 ± 810.5, $p = 0.05$). Regarding the number of sessions underwent by the participants, the average number of sessions in the unsatisfied group was much higher than that in the satisfied group (14.9 ± 29.8 vs. 9.4 ± 23.1, $p = 0.027$). Besides, features, such as occupation, education, psychotherapist's age, method of psychotherapy, therapy location, and psychotherapy theoretical orientations, were significantly different between the two groups of participants (Table 1).

## Important features distinguishing clients who were satisfied or unsatisfied with psychotherapy

Next, chi-square analysis was used to evaluate each feature's discriminative power for the categories of clients who were satisfied or unsatisfied with psychotherapy. The top 10 features that most contributed to distinguishing clients' psychotherapy satisfaction include occupation, therapy location, form of access to psychotherapy, theoretical orientation of psychotherapy, education, gender, psychotherapist's age, psychotherapy status, time of the last session, and psychotherapist's gender, with chi-square values of 24.913, 16.856, 16.046, 11.490, 11.134, 8.645, 8.370, 5.530, 4.116, and 3.759, respectively (Table 2).

To visualize the difference between clients satisfied or unsatisfied with psychotherapy, we compared the distribution of the top six features in the feature importance ranking of the two types of clients in Figure 2. The client's occupation is the feature that most strongly distinguishes between those participants who were satisfied and those who were unsatisfied with psychotherapy. Enterprise and institution staff, civil servants, and self-employed individuals had a higher percentage of psychotherapy satisfaction, while medical personnel and others had a higher percentage of dissatisfaction (Figure 2A). Concerning therapy location, clients of medical institutes, and counseling agencies were more satisfied with therapy, while clients of public welfare organizations and other consulting agencies were relatively less satisfied (Figure 2B). Regarding the form of access to psychotherapy, the patients engaged in face-to-face therapy showed higher satisfaction than the patients receiving psychotherapy by other methods (Figure 2C). In terms of the theoretical orientation of therapy, individuals undergoing systemic family therapy and CBT showed a higher percentage of satisfaction than those in psychodynamic therapy (Figure 2D). Interestingly, clients with an

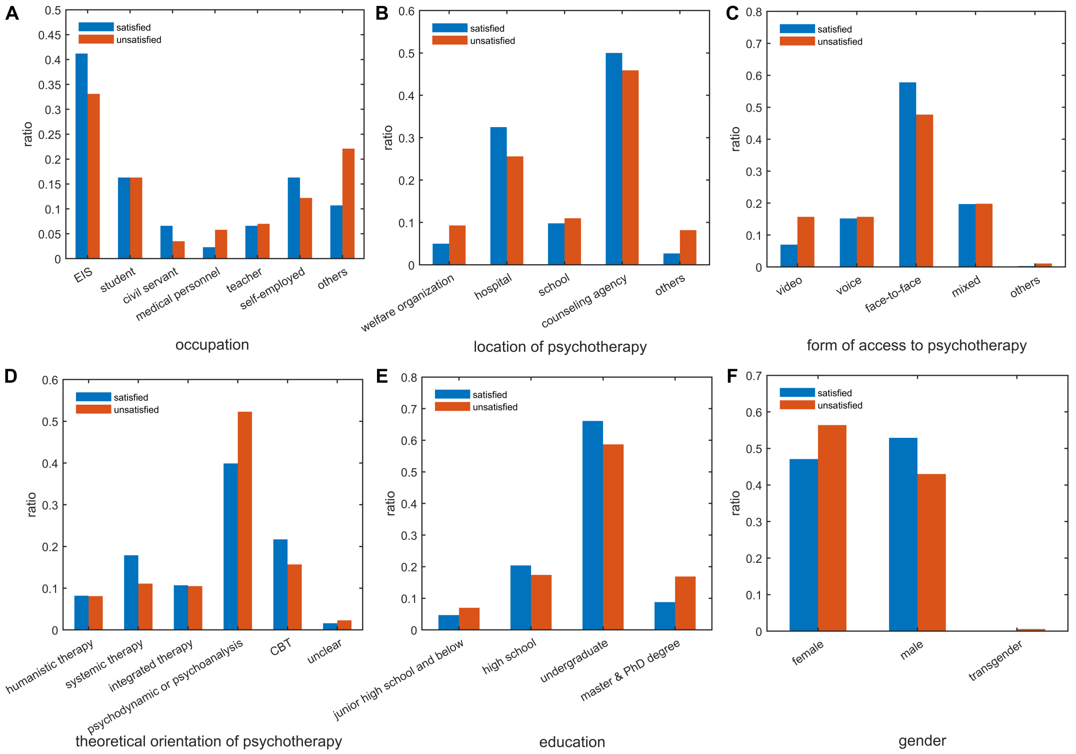


FIGURE 2
Comparison of clients' satisfaction with psychotherapy based on graph metrics. **(A)** client's occupation; **(B)** location of psychotherapy; **(C)** way of psychotherapy; **(D)** theoretical orientation of psychotherapy; **(E)** client's education; **(F)** client's gender. satisfied: participant satisfied with psychotherapy; unsatisfied: participant unsatisfied with psychotherapy. EIS, enterprise and institution staff; CBT, cognitive behavioral therapy.

TABLE 2 The ranking of feature importance.

| Rank | Feature | Chi-square value |
|---|---|---|
| 1 | Occupation | 24.913 |
| 2 | Location of psychotherapy | 16.856 |
| 3 | Form of access to psychotherapy | 16.046 |
| 4 | Theoretical orientation of psychotherapy | 11.490 |
| 5 | Education | 11.134 |
| 6 | Gender | 8.645 |
| 7 | Age of psychotherapist | 8.370 |
| 8 | Status of psychotherapy | 5.530 |
| 9 | Time of the last session | 4.116 |
| 10 | Gender of psychotherapist | 3.759 |

education level of junior high school and below or master's and doctoral degrees had a lower rate of psychotherapy satisfaction (Figure 2E). Moreover, compared with female clients, male clients were relatively more satisfied with psychotherapy (Figure 2F).

## Machine learning algorithms applied for the prediction of client psychotherapy satisfaction

Next, we used a series of supervised machine-learning algorithms to predict clients' psychotherapy satisfaction. First, we chose seven representative machine-learning algorithms, i.e., CatBoost, LightGBM, XGBoost, random forest, decision tree, SVM, and logistic regression, to build prediction models. Then, based on the test subset, we compared all the models' predictive performances to find the best prediction model. The F1 scores of the seven models, CatBoost, XGBoost, random forest, LightGBM, SVM, decision tree, and logistic regression, were 0.758, 0.735, 0.734, 0.725, 0.716, 0.701, and 0.612, respectively (Table 3). The precision value and recall value of each model were also listed in Table 3, and the precision-recall curves of the seven models were presented in Figure 3. By comparing these models' prediction performances, the model based on the CatBoost algorithm achieved the largest F1 score of 0.758, leading to the best performance in predicting client psychotherapy satisfaction.

## Discussion

To the best of our knowledge, our work is the first study applying machine-learning algorithms to predict clients' satisfaction with psychotherapy in China. There were two main findings in the current study: (1) the most relevant six features in distinguishing clients with or without satisfaction with psychotherapy were the client's occupation, location of therapy, the form of access to psychotherapy, theoretical orientation of therapy, client's education, and gender; (2) the CatBoost algorithm-based model performed the best in distinguishing between satisfied- and unsatisfied participants with psychotherapy, with an F1 score of 0.758. Meanwhile, our study demonstrated the value and feasibility of using machine-learning

TABLE 3 Compare the performance of different ML algorithms to predict clients' satisfaction with psychotherapy.

| Classifier | Precision | Recall | F1-Score |
|---|---|---|---|
| CatBoost | 0.757 | 0.789 | 0.758 |
| LightGBM | 0.726 | 0.768 | 0.735 |
| XGBoost | 0.731 | 0.776 | 0.734 |
| Random forest | 0.725 | 0.776 | 0.725 |
| Decision tree | 0.715 | 0.717 | 0.716 |
| SVM | 0.684 | 0.730 | 0.701 |
| Logistic regression | 0.680 | 0.578 | 0.612 |

approaches to predict clients' psychotherapy satisfaction based on the features of the participants and their therapists.

Among various features determining participants' psychotherapy satisfaction, occupation was identified as the best feature for distinguishing between those satisfied and those unsatisfied participants with psychotherapy. Our analysis showed that enterprise and institution staff, civil servants, and self-employed individuals had a higher percentage of psychotherapy satisfaction than medical personnel and others (Figure 2A). The reasons for this difference might be related to the different mindsets and educational experiences of different groups. Medical personnel may be more trained in the mindset of biological medicine (1, 58). It may be somewhat at odds with interpretative, speculative, and circular philosophies of psychotherapy (1, 58). This may hinder their engagement with the therapist, and psychotherapy process triggers their dissatisfaction with psychotherapy (6). Hence, for medical personnel, more practical and linear interventions (e.g., linear questioning and action suggestions) might be more applicable (6). However, future research to explore improving medical personnel's satisfaction with psychotherapy is still needed.

Another finding is that participants with a graduate degree or less than a high school educational level reported more dissatisfaction than those with high school and undergraduate education levels (Figure 2E). Previous research did not strongly support an association between the client's education level and psychotherapy outcomes (59–61). A possible explanation for this finding might be that for participants with less than a high school education, the process and pragmatic system of psychotherapy may be confusing for them. Therefore, they may receive fewer gains from psychotherapy. By comparison, for participants with high school and undergraduate education levels, their cognitive levels and expectations may be more suitable for therapists in the current study. This might be partially consistent with previous studies, which reported that a higher cognitive level similarity between the therapist and the client could improve the clients' therapy experience (5). Our finding implies that psychotherapy strategies should be tailored to clients' cognitive and education levels. However, why participants with a graduate degree had lower therapy satisfaction still needs to be explored in the future.

Concerning the location of psychotherapy, we found that clients who received psychotherapy in mental health institutes, comprehensive hospitals, and commercial agencies were more satisfied with psychotherapy than those whose sessions were conducted at public welfare organizations. The potential factors contributing to this difference may include: (1) Chinese culture advocates worship and trust in authority (5, 49). The participants may think clinicians in "official and professional" medical institutes are more trustworthy. The stereotype of "authorized professionals in official hospitals" may increase participants' trust in their therapists and adherence to the therapy. This coincides with the findings of previous research that Chinese clients expect

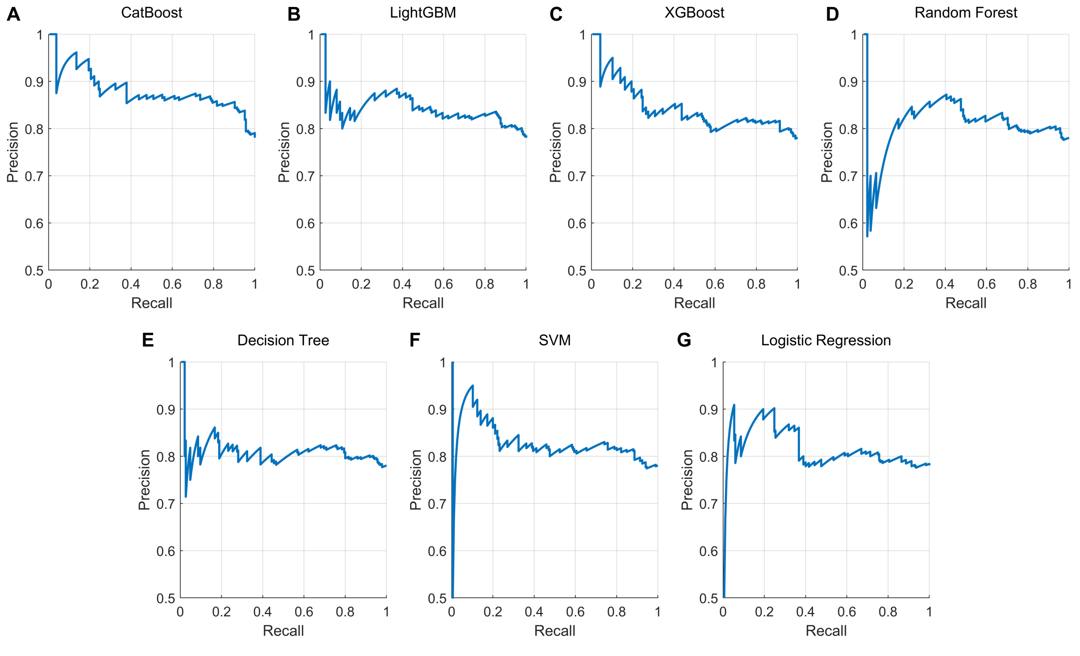


FIGURE 3
Precision-recall curve for each trained model in predicting client psychotherapy satisfaction. The precision-recall curve for each prediction model predicts whether the participant is satisfied with psychotherapy. Seven machine learning algorithms were selected for training, namely, **(A)** CatBoost, **(B)** LightGBM, **(C)** XGBoost, **(D)** random forest, **(E)** decision tree, **(F)** support vector machine (SVM), and **(G)** logistic regression.

their psychotherapists to provide guidance and suggestions from a professional authoritative perspective (5). (2) In China, the admission criteria for psychotherapists in official hospitals and commercial agencies may be stricter than those in public welfare organizations. Clinicians may be more experienced and competent in prescribing various therapy strategies and interventions. In contrast, the percentage of novice therapists and interns in public welfare organizations may be higher (62). Previous studies have also suggested that clients' therapy outcomes are positively related to the professional competence of therapists (5, 8, 12). (3) In hospitals and commercial agencies, participants need to pay for psychotherapy. Therefore, they may take the therapy more seriously and be more engaged and attentive (63). Meanwhile, they may be more convinced by the therapist's feedback and suggestions than those participants who receive free treatment in public welfare institutions. Interestingly, our analysis also showed that for participants who were satisfied with their paid sessions, the average cost per session was lower than those who were unsatisfied (**Table 1**). This finding was in accordance with the study by Stanley et al. (64) that financial incentives that reward therapy attendance with discounted fees were associated with clinical improvement in the clients. It implies that psychotherapy with modest charges may be more helpful. However, the mechanism of this phenomenon and the strategies to solve this problem remain unclear. Future research exploring the strategies (e.g., recruiting more experienced psychotherapists with a medical training background or giving flexible charging policies for specific clients) to improve clients' satisfaction in welfare organizations is strongly suggested.

In our study, participants who received systemic family therapy and CBT reported a higher percentage of satisfaction than those in psychodynamic therapy (**Figure 2D**). This finding was partially consistent with previous studies that psychodynamic therapy showed a higher risk of side effects in psychotherapy (6, 17, 18). Psychoanalysis therapy emphasizes the exploration of past traumatic experiences, clients' defects, and their self-reflection on internal conflicts and pain. It may trigger participants to blame themselves or others for their problems, thus taking the role of an isolated victim and presenting a defect-orientation thinking model (6). Even if participants experience sudden gains from the therapy (65), it may put much pressure on the clients. Comparatively, systemic family therapy focused on the resources and flexibility of the participants' families. Meanwhile, there is a greater focus on improving and reframing dysfunctional family interactions (5, 6, 66). Chinese culture attaches greater importance to the influence of the family environment on an individual's mental health, and family interpersonal conflicts are correlated with various clinical complaints of Chinese clients (5, 67–69). Hence, participants who received systemic family therapy may have more positive perceptions of themselves and their families while also experiencing more significant adjustments in their family relationships. This may contribute to the relief of their symptoms and increase their satisfaction with psychotherapy. Previous research has also implied that the involvement of family members was associated with better outcomes in psychotherapy (10). CBT emphasizes finding more effective coping strategies and cognitive schemes to solve clients' difficulties (30). This therapeutic philosophy may coincide with Chinese culture, which advocates obedience to professional authority, useful knowledge, and effective coping strategies (5, 49). As a result, participants experienced higher therapy satisfaction.

In terms of the association of therapy satisfaction with participants' gender, the results of previous studies have varied, and no consistent conclusions have been reached (59). For example, both Vitinius et al. (59) and Schneider and Heuft's (70) studies found that client gender did not have a significant impact on psychotherapy success. Although our analysis implied that male participants were more satisfied with psychotherapy than female participants, more research is strongly suggested to clarify the potential factors and mechanisms contributing to this gender difference. Regarding the form of access to psychotherapy, face-to-face therapy rated higher in satisfaction than other psychotherapy methods. The main reason may be because the flow of emotions, exchange of ideas, and behavioral interactions between clients and therapists are more fluent during offline psychotherapy (9). Thus, a high-quality therapeutic alliance might be more easily fostered. As suggested by previous research, client satisfaction with therapy was positively correlated with supportive and trustworthy therapist-client alliance (5, 9).

Regarding therapists' age and gender, our analysis implied that they worked as two of the top 10 features that most contributed to distinguishing client psychotherapy satisfaction (**Table 2**). However, their contributions were relatively lower (with chi-square values of 8.370 and 3.759, respectively) than the other six features mentioned above. As suggested by previous research, clients whose preferences for the therapist's gender and age were met reported better therapy outcomes (5, 9). However, no linear correlation between client satisfaction and the therapist's gender and age have been identified by prior studies. Although our results suggested that clients treated by younger therapists were more satisfied with their treatment (**Table 1**), future research exploring how therapists' age affects the participants' satisfaction is still suggested.

Another issue concerns the impact of diagnosis and symptom severity on participant satisfaction. These two features were not involved in our investigation, but previous research implied that their influence on clients' responses to psychotherapy could be complicated. Some studies suggested that symptom severity and diagnosis would work as predictors for psychotherapy outcomes (71). Meanwhile, previous research implied that symptom severity might also moderate the associations between other predictors and psychotherapy outcomes (72). However, the mechanisms they interact with other potential features influencing Chinese clients' psychotherapy satisfaction remain unclear. Therefore, future studies exploring the impact of Chinese clients' diagnosis and symptom severity on their therapy satisfaction are strongly suggested.

Clients seeking psychotherapy always expect satisfactory outcomes. However, due to differences in clients, therapists, and other relevant factors, not all clients will have satisfactory outcomes. Although the psychotherapeutic process is relatively subjective, our attempts show that its outcomes can still be predicted by models. By applying seven different types of supervised machine-learning-based algorithms, our research showed that, for clients undergoing psychotherapy, the model can accurately distinguish between those who are satisfied or unsatisfied with therapy based on the features of participants and therapists. Among the seven models, the CatBoost algorithm-based model showed the best performance in predicting clients' psychotherapy satisfaction, with an F1 score of 0.758. CatBoost is based on gradient-boosted decision trees developed by Yandex researchers and engineers (73). It has been employed in a wide variety of fields due to its great performance for classification and regression tasks (74). To date, there are few studies using machine-learning methods to predict satisfaction with psychotherapy. Different studies have used different features to predict satisfaction with psychotherapy from different aspects (29). We are the first study to incorporate both client and therapist factors in a model to predict treatment satisfaction. In addition, previous studies often used a single algorithm or a few (37, 40, 41, 75), making the prediction effect relatively limited. Our research used a variety of machine-learning algorithms, including traditional algorithms such as logistic regression, decision tree, random forest, and SVM, also some emerging approaches such as CatBoost, XGBoost, and LightGBM. Different algorithms have different predictive effects due to different principles and computing abilities. By comparing the performance of multiple algorithms and using automatic parameter tuning, the model we trained can achieve the best prediction performance to the greatest extent possible.

The CatBoost-based machine-learning model achieved in the current study is sufficiently accurate and could provide meaningful implications for psychotherapy practice. The CatBoost-based prediction can be implemented as software or app on the mobile device without special equipment or materials. Input the relevant information of the future client and the counselor, and the model can easily and quickly give whether the client is satisfied with the psychotherapy. The results provided by the model can be used as part of an auxiliary diagnosis and treatment. Still, the precise treatment for a specific client requires the therapist to consider both the given by the model and their own experience, then formulate an appropriate plan to improve the effectiveness of psychotherapy.

## Limitations

This study still has some limitations. First, in the current study, the psychotherapy satisfaction of clients was not assessed in an independent survey. This may make our finding less valid. Meanwhile, the objectively measured outcomes (e.g., symptom reduction and improvement in social function) of psychotherapy were not included in the investigation. The relationships between client satisfaction and treatment outcomes were not explored. Hence, randomized control trial and longitudinal research using objective and experimental data will be introduced in the future. Second, the self-assessed questionnaires were disseminated and completed online *via* social media applications according to the inclusion and exclusion criteria. Although our sample size was relatively large, the validity and accuracy of the survey on some variables (e.g., whether participants actually received the therapy as indicated, the age and qualification of the therapist, and the theoretical orientation of therapy) might not be sufficiently guaranteed. Third, some other potential features associated with client satisfaction, such as therapists' education level, career stage, professional experience and mental activity, competence, participant's detailed diagnosis, symptom severity, therapy contents, frequency, and treatment quality were not investigated. Future studies using qualitative approaches or quantitative frameworks to explore the underlying mechanisms of how these factors influence client satisfaction with psychotherapy are strongly suggested. Forth, we conducted a binary assessment of psychotherapy satisfaction in the current study. In future studies, we will increase the sample size, conduct a more nuanced classification of psychotherapy satisfaction, and refine the current predictive model.

## Conclusion

The current study clarified several major factors influencing client satisfaction with psychotherapy, including the client's occupation, gender, education, location of psychotherapy, the form of access to psychotherapy, and theoretical orientation of therapy. It suggests that good therapy strategies should be designed in accordance with the certain demographic characteristics of the clients and their specific preferences for therapy settings and approaches. Meanwhile, we built a supervised machine-learning-based model which could distinguish between satisfied or unsatisfied participants with psychotherapy. The model based on the CatBoost algorithm achieved an F1 score of 0.758. These results provide meaningful implications for designing and tailoring better psychotherapy strategies for specific clients to achieve better therapeutic outcomes.

## Data availability statement

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

## Ethics statement

The study involving human participants was reviewed and approved by the Ethics Committee of Tongji University and the Shanghai Pudong New Area Mental Health Center (No. PWRd2020-01). The patients/participants provided their written informed consent to participate in this study.

## Author contributions

LY and LL substantially contributed to the design, participant recruitment, data analysis, and draft the manuscript. HG was responsible for recruiting data. XZ contributed to the study conception and critical review of the manuscript for content. ZW and YC implemented machine learning algorithms and statistical analysis. All authors contributed to the article and approved the submitted version.

## Funding

This study was funded by the Training Plan of Health System Academic Leader of the Shanghai Pudong Municipality Health Commission (Grant Number: PWRd2020-01), Open Research Program of Shanghai Key Laboratory of Intelligent Information Processing (Grant Number: IIPL201911), and Medical Discipline Construction Project of Pudong Health Committee of Shanghai (Grant Number: PWYgy2021-02).

## Acknowledgments

We would like to thank Congcong Ge, Hongping Zhuang, and Yunhan Zhao for their assistance in recruiting participants and data collection. Meanwhile, we greatly appreciate the contributions of all the participants.

## 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.

## References

1. Rybarczyk B. Psychotherapy. In: Kreutzer J, DeLuca J, Caplan B editors. *Encyclopedia of Clinical Neuropsychology*. Cham: Springer (2018).

2. Li J, Wang X, Meng H, Zeng K, Quan F, Liu F. Systemic family therapy of comorbidity of anxiety and depression with epilepsy in adolescents. *Psychiatry Investig.* (2016) 13:305–10. doi: 10.4306/pi.2016.13.3.305

3. Dragioti E, Karathanos V, Gerdle B, Evangelou E. Does psychotherapy work? An umbrella review of meta-analyses of randomized controlled trials. *Acta Psychiatr Scand.* (2017) 136:236–46.

4. Levitt HM, Pomerville A, Surace FI. A qualitative meta-analysis examining clients' experiences of psychotherapy: a new agenda (vol 142, pg 801, 2016). *Psychol Bull.* (2016) 142:1067–1067. doi: 10.1037/bul0000057

5. Liu L, Wu J, Wang J, Wang Y, Tong Y, Ge C, et al. What do chinese families with depressed adolescents find helpful in family therapy? A qualitative study. *Front Psychol.* (2020) 11:1318. doi: 10.3389/fpsyg.2020.01318

6. Yao L, Zhao X, Xu Z, Chen Y, Liu L, Feng Q, et al. Influencing factors and machine learning-based prediction of side effects in psychotherapy. *Front Psychiatry.* (2020) 11:537442. doi: 10.3389/fpsyt.2020.537442

7. Liu L, Miller JK, Zhao X, Ma X, Wang J, Li W. Systemic family psychotherapy in China: a qualitative analysis of therapy process. *Psychol Psychother.* (2013) 86:447–65. doi: 10.1111/j.2044-8341.2012.02075.x

8. Løvgren A, Røssberg JI, Nilsen L, Engebretsen E, Ulberg R. How do adolescents with depression experience improvement in psychodynamic psychotherapy? A qualitative study. *BMC Psychiatry.* (2019) 19:95. doi: 10.1186/s12888-019-2080-0

9. Flückiger C, Del Re AC, Wampold Bruce E, Horvath Adam O. The alliance in adult psychotherapy: a meta-analytic synthesis. *Psychotherapy.* (2018) 55:316–40. doi: 10.1037/pst0000172

10. Dardas LA, van de Water B, Simmons LA. Parental involvement in adolescent depression interventions: a systematic review of randomized clinical trials. *Int J Ment Health Nurs.* (2018) 27:555–70. doi: 10.1111/inm.12429

11. Williams R, Farquharson L, Palmer L, Bassett P, Clarke J, Clark DM, et al. Patient preference in psychological treatment and associations with self-reported outcome: national cross-sectional survey in England and Wales. *BMC Psychiatry.* (2016) 16:4. doi: 10.1186/s12888-015-0702-8

12. Schöttke H, Flückiger C, Goldberg SB, Eversmann J, Lange J. Predicting psychotherapy outcome based on therapist interpersonal skills: A five-year longitudinal study of a therapist assessment protocol. *Psychother Res.* (2017) 27:642–52. doi: 10.1080/10503307.2015.1125546

13. Anderson T, McClintock AS, Himawan L, Song X, Patterson CL. A prospective study of therapist facilitative interpersonal skills as a predictor of treatment outcome. *J Consult Clin Psychol.* (2016) 84:57–66. doi: 10.1037/ccp0000060

14. Spinelli MG, Endicott J, Goetz RR, Segre LS. Reanalysis of efficacy of interpersonal psychotherapy for antepartum depression versus parenting education program: initial severity of depression as a predictor of treatment outcome. *J Clin Psychiatry.* (2016) 77:535–40. doi: 10.4088/JCP.15m09787

15. Lopes RT, Gonçalves MM, Sinai D, Machado PP. Predictors of dropout in a controlled clinical trial of psychotherapy for moderate depression. *Int J Clin Health Psychol.* (2015) 15:76–80. doi: 10.1016/j.ijchp.2014.11.001

16. Lindhiem O, Bennett CB, Trentacosta CJ, McLear C. Client preferences affect treatment satisfaction, completion, and clinical outcome: a meta-analysis. *Clin Psychol Rev.* (2014) 34:506–17.

17. Leitner A, Märtens M, Koschier A, Gerlich K, Liegl G, Hinterwallner H, et al. Patients' perceptions of risky developments during psychotherapy. *J Contemp Psychother.* (2013) 43:95–105.

18. Crawford MJ, Thana L, Farquharson L, Palmer L, Hancock E, Bassett P, et al. Patient experience of negative effects of psychological treatment: results of a national survey. *Br J Psychiatry.* (2016) 208:260–5.

19. de Leon CGP. "What was i thinking?!' rhetorical questions as a technique to identify and explore impasses in therapy. *Aust NZ J Fam Ther.* (2018) 39:21–37.

20. Earl RM. Video game use as a tool for assessing and intervening with identity formation and social development in family therapy. *Aust NZ J Fam Ther.* (2018) 39:5–20.

21. Larner G. Utilising gaming, rhetorical questions, deception genograms, and other useful techniques in family therapy. *Aust NZ J Fam Ther.* (2018) 39:3–4.

22. North J, Shadid C, Hertlein KM. Deception in family therapy: recognition, implications, and intervention. *Aust NZ J Fam Ther.* (2018) 39:38–53.

23. Sidani S, Epstein DR, Fox M, Collins L. The contribution of participant, treatment, and outcome factors to treatment satisfaction. *Res Nurs Health.* (2018) 41:572–82.

24. De Smet M, Below C, Acke E, Werbart A, Meganck R, Desmet M, et al. When 'good outcome' does not correspond to 'good therapy': reflections on discrepancies between outcome scores and patients' therapy satisfaction. *Eur J Psychother Counsel.* (2021) 23:156–76.

25. Ring M, Gysin-Maillart A. Patients' satisfaction with the therapeutic relationship and therapeutic outcome is related to suicidal ideation in the attempted suicide short intervention program (ASSIP). *Crisis.* (2020) 41:337–43. doi: 10.1027/0227-5910/a000644

26. Viefhaus P, Döpfner M, Dachs L, Goletz H, Görtz-Dorten A, Kinnen C, et al. Parent- and therapist-rated treatment satisfaction following routine child cognitive-behavioral therapy. *Eur Child Adolesc Psychiatry.* (2021) 30:427–39. doi: 10.1007/s00787-020-01528-1

27. Harper-Jaques S, Foucault D. Walk-in single-session therapy: client satisfaction and clinical outcomes. *J System Ther.* (2014) 33:29–49.

28. Keum BT, Wang L. Supervision and psychotherapy process and outcome: a meta-analytic review. *Transl Issues Psychol Sci.* (2021) 7:89–108.

29. van Doorn KA, Kamsteeg C, Bate J, Aafjes M. A scoping review of machine learning in psychotherapy research. *Psychother Res.* (2021) 31:92–116.

30. Lutz W, Rubel JA, Schwartz B, Schilling V, Deisenhofer AK. Towards integrating personalized feedback research into clinical practice: development of the trier treatment navigator (TTN). *Behav Res Ther.* (2019) 120:103438. doi: 10.1016/j.brat.2019.103438

31. Reggente N, Moody TD, Morfini F, Sheen C, Rissman J, O'Neill J, et al. Multivariate resting-state functional connectivity predicts response to cognitive behavioral therapy in obsessive-compulsive disorder. *Proc Natl Acad Sci USA.* (2018) 115:2222–7.

32. Rubel JA, Zilcha-Mano S, Giesemann J, Prinz J, Lutz W. Predicting personalized process-outcome associations in psychotherapy using machine learning approaches-A demonstration. *Psychother Res.* (2020) 30:300–9. doi: 10.1080/10503307.2019.1597994

33. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry.* (2021) 20:154–70.

34. Lutz W, Leach C, Barkham M, Lucock M, Stiles WB, Evans C, et al. Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. *J Consult Clin Psychol.* (2005) 73:904–13.

35. Green SA, Honeybourne E, Chalkley SR, Poots AJ, Woodcock T, Price G, et al. A retrospective observational analysis to identify patient and treatment-related predictors of outcomes in a community mental health programme. *BMJ Open.* (2015) 5:e006103. doi: 10.1136/bmjopen-2014-006103

36. Buckman JEJ, Cohen ZD, O'Driscoll C, Fried EI, Saunders R, Ambler G, et al. Predicting prognosis for adults with depression using individual symptom data: a comparison of modelling approaches. *Psychol Med.* (2021). [Epub ahead of print]. doi: 10.1017/S0033291721001616

37. Gori A, Lauro-Grotto R, Giannini M, Schuldberg D. Predicting treatment outcome by combining different assessment tools: Twoward an integrative model of decision support in psychotherapy. *J Psychother Integr.* (2010) 20:251–69.

38. Tolmeijer E, Kumari V, Peters E, Williams SCR, Mason L. Using fMRI and machine learning to predict symptom improvement following cognitive behavioural therapy for psychosis. *Neuroimage Clin.* (2018) 20:1053–61. doi: 10.1016/j.nicl.2018.10.011

39. Hahn T, Kircher T, Straube B, Wittchen H, Konrad C, Stroehle A, et al. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. *JAMA Psychiatry.* (2015) 72:68–74. doi: 10.1001/jamapsychiatry.2014.1741

40. Wallert J, Gustafson E, Held C, Madison G, Norlund F, von Essen L, et al. Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the U-CARE heart randomized controlled trial. *J Med Internet Res.* (2018) 20:e10754. doi: 10.2196/10754

41. Nasir M, Baucom BR, Georgiou P, Narayanan S. Predicting couple therapy outcomes based on speech acoustic features. *Plos One.* (2017) 12:e0185123. doi: 10.1371/journal.pone.0185123

42. Wahle F, et al. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR Mhealth Uhealth.* (2016) 4:e111. doi: 10.2196/mhealth.5960

43. Symons M, Feeney GFX, Gallagher MR, Young RMD, Connor JP. Machine learning vs addiction therapists: a pilot study predicting alcohol dependence treatment outcome from patient data in behavior therapy with adjunctive medication. *J Subst Abuse Treat.* (2019) 99:156–62. doi: 10.1016/j.jsat.2019.01.020

44. Villmann T, Liebers C, Bergmann B, Gumz A, Geyer M. Investigation of psycho-physiological interactions between patient and therapist during a psychodynamic therapy and their relation to speech using in terms of entropy analysis using a neural network approach. *New Ideas Psychol.* (2008) 26:309–25.

45. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. *Trans Assoc Comput Linguist.* (2016) 4:463–76.

46. Ewbank MP, Cummins R, Tablan V, Bateup S, Catarino A, Martin AJ, et al. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry.* (2020) 77:35–43. doi: 10.1001/jamapsychiatry.2019.2664

47. de Jong K, Conijn JM, Gallagher RAV, Reshetnikova AS, Heij M, Lutz MC, et al. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. *Clin Psychol Rev.* (2021) 85:102002. doi: 10.1016/j.cpr.2021.102002

48. Carcone AI, Hasan M, Alexander GL, Dong M, Eggly S, Hartlieb KB, et al. Developing machine learning models for behavioral coding. *J Pediatr Psychol.* (2019) 44:289–99.

49. Hou J, Chen Z. The trajectories of adolescent depressive symptoms: Identifying latent subgroups and risk factors. *Acta Psychol Sinica.* (2016) 48: 957–68.

50. Yao L, Cai M, Chen Y, Shen C, Shi L, Guo Y, et al. Prediction of antiepileptic drug treatment outcomes of patients with newly diagnosed epilepsy by machine learning. *Epilepsy Behav.* (2019) 96:92–7.

51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Machine Learn Res.* (2011) 12:2825–30. doi: 10.5555/1953048.2078195

52. Mcdonald A, Sasangohar F, Jatav A, Rao A. Continuous monitoring and detection of post-traumatic stress disorder (PTSD) triggers among veterans: a supervised machine learning approach. *IISE Trans Healthc Syst Eng.* (2019) 9: 201–11.

53. Gonzalez SDP, Delgadillo J, Lutz W. Predicting early dropout in online versus face-to-face guided self-help: A machine learning approach. *Behav Res Ther.* (2022) 159:104200. doi: 10.1016/j.brat.2022.104200

54. Gunther MP, Kirchebner J, Lau S. Identifying Direct Coercion in a High Risk Subgroup of Offender Patients With Schizophrenia via Machine Learning Algorithms. *Front Psychiatry.* (2020) 11:415. doi: 10.3389/fpsyt.2020.00415

55. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* (2002) 16:321–57.

56. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett.* (2006) 27: 861–74.

57. Yang Y, Pedersen JOA. Comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning.* San Francisco, CA: Morgan Kaufmann Publishers Inc (1997). p. 412–20. doi: 10.5555/645526.657137

58. Feinstein R, Heiman N, Yager J. Common factors affecting psychotherapy outcomes: some implications for teaching psychotherapy. *J Psychiatr Pract.* (2015) 21:180–9.

59. Vitinius F, Tieden S, Hellmich M, Pfaff H, Albus C, Ommen O, et al. Perceived psychotherapist's empathy and therapy motivation as determinants of long-term therapy success-results of a cohort study of short term psychodynamic inpatient psychotherapy. *Front Psychiatry.* (2018) 9:660. doi: 10.3389/fpsyt.2018.00660

60. Hiller W, Fichter MM, Rief W. A controlled treatment study of somatoform disorders including analysis of healthcare utilization and cost-effectiveness. *J Psychosom Res.* (2003) 54:369–80.

61. Swift JK, Callahan JL, Cooper M, Parkin SR. The impact of accommodating client preference in psychotherapy: a meta-analysis. *J Clin Psychol.* (2018) 74:1924–37. doi: 10.1002/jclp.22680

62. Wang M, Jiang G-R, Yan Y-P, Zhou Z-Y. The way for certifying and psychotherapists in China. *Chin Ment Health J.* (2015) 29:503–9.

63. Clark P, Sims PL. The practice of fee setting and collection: implications for clinical training programs. *Am J Fam Ther.* (2014) 42:386–97.

64. Stanley IH, Chu C, Brown TA, Sawyer KA, Thomas E. Improved clinical functioning for patients receiving fee discounts that reward treatment engagement. *J Clin Psychol.* (2016) 72:15–21. doi: 10.1002/jclp.22236

65. Aderka IM, Kauffmann A, Shalom JG, Beard C, Björgvinsson T. Using machine-learning to predict sudden gains in treatment for major depressive disorder. *Behav Res Ther.* (2021) 144:103929. doi: 10.1016/j.brat.2021.103929

66. Sexton TL. Functional family therapy: an evidence-based, familyfocused, and systemic approach for working with adolescents and their families. In: Fiese BH, et al. editors. *APA handbook of contemporary family psychology: family therapy and training.* Washington DC: American Psychological Association (2019).

67. Wang JK, Zhao XD. Family functioning and social support for older patients with depression in an urban area of Shanghai, China. *Arch Gerontol Geriatr.* (2012) 55:574–9. doi: 10.1016/j.archger.2012.06.011

68. Chen W, Huang Y, Riad A. Family environment and depression: a population-based analysis of gender differences in rural China. *J Fam Issues.* (2014) 35: 481–500.

69. Lui M, Lau G, Tam V, Chiu H, Li S, Sin K. Parents' impact on children's school performance: marital satisfaction, parental involvement, and mental health. *J Child Fam Stud.* (2020) 29:1548–60. doi: 10.1186/s12913-016-1423-5

70. Schneider G, Heuft G. Operationalized psychodynamic diagnosis system and outcome of psychodynamic inpatient psychotherapy in male and female patients. *Zeitschrift Fur Psychosomatische Medizin Und Psychotherapie.* (2018) 64:281–97.

71. Simon NM, Otto MW, Worthington JJ, Hoge EA, Thompson EH, Lebeau RT, et al. Outcome prediction of cognitive behaviour therapy for panic disorder: initial symptom severity is predictive for treatment outcome, comorbid anxiety or depressive disorder, cluster c personality disorders and initial motivation are not. *Behav Cogn Psychother.* (2008) 36:99–112.

72. Seow LLY, Page AC, Hooke GR. Severity of borderline personality disorder symptoms as a moderator of the association between the use of dialectical behaviour therapy skills and treatment outcomes. *Psychother Res.* (2020) 30:920–33.

73. Prokhorenkova L, Gusev G, Vorobev A, Dorogush A, Gulin A. CatBoost: unbiased boosting with categorical features. *Adv. Neural Inf. Process.Syst.* (2018) 31:6639–49. doi: 10.5555/3327757.3327770

74. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J Big Data.* (2020) 7:94. doi: 10.1186/s40537-020-00369-8

75. Månsson KN, Frick A, Boraxbekk CJ, Marquand AF, Williams SC, Carlbring P, et al. Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. *Transl Psychiatry.* (2015) 5:e530. doi: 10.1038/tp.2015.22

frontiers | Frontiers in Psychiatry

TYPE Mini Review
PUBLISHED 09 May 2023
DOI 10.3389/fpsyt.2023.1055868

Check for updates

OPEN ACCESS

EDITED BY
Andreas Stengel,
University Hospital Tübingen, Germany

REVIEWED BY
David Benrimoh,
McGill University, Canada

*CORRESPONDENCE
Ivo Rollmann
✉ ivo.rollmann@med.uni-heidelberg.de

RECEIVED 28 September 2022
ACCEPTED 17 April 2023
PUBLISHED 09 May 2023

CITATION
Rollmann I, Gebhardt N, Stahl-Toyota S, Simon J, Sutcliffe M, Friederich H-C and Nikendei C (2023) Systematic review of machine learning utilization within outpatient psychodynamic psychotherapy research. *Front. Psychiatry* 14:1055868. doi: 10.3389/fpsyt.2023.1055868

# Systematic review of machine learning utilization within outpatient psychodynamic psychotherapy research

Ivo Rollmann*, Nadja Gebhardt, Sophia Stahl-Toyota, Joe Simon, Molly Sutcliffe, Hans-Christoph Friederich and Christoph Nikendei

Department for General Internal Medicine and Psychosomatics, University Hospital Heidelberg, Heidelberg, Germany

**Introduction:** Although outpatient psychodynamic psychotherapy is effective, there has been no improvement in treatment success in recent years. One way to improve psychodynamic treatment could be the use of machine learning to design treatments tailored to the individual patient's needs. In the context of psychotherapy, machine learning refers mainly to various statistical methods, which aim to predict outcomes (e.g., drop-out) of future patients as accurately as possible. We therefore searched various literature for all studies using machine learning in outpatient psychodynamic psychotherapy research to identify current trends and objectives.

**Methods:** For this systematic review, we applied the Preferred Reporting Items for systematic Reviews and Meta-Analyses Guidelines.

**Results:** In total, we found four studies that used machine learning in outpatient psychodynamic psychotherapy research. Three of these studies were published between 2019 and 2021.

**Discussion:** We conclude that machine learning has only recently made its way into outpatient psychodynamic psychotherapy research and researchers might not yet be aware of its possible uses. Therefore, we have listed a variety of perspectives on how machine learning could be used to increase treatment success of psychodynamic psychotherapies. In doing so, we hope to give new impetus to outpatient psychodynamic psychotherapy research on how to use machine learning to address previously unsolved problems.

KEYWORDS

**machine learning (ML), psychodynamic psychotherapy, outpatient therapy, review—systematic, perspectives**

## Introduction

Outpatient psychodynamic psychotherapy is effective in treating various psychological disorders (1–3). Further positive effects include a reduced number of sick leaves, a reduction of health care utilization, less psychiatric hospitalizations after therapy, and a reduced relapse rates for depression (4–6). A number of factors that predict successful therapy are also known, such as improving the working alliance (7, 8), therapeutic agency (7, 9, 10) or the patient's ability to perceive emotions (11, 12) which lead to a reduction in symptom burden. However, as Leichsenring et al. (13) point out, recent substantial improvements in treatment success have been scarce. The authors (13) recommend that future studies focus primarily on non-responders and drop-outs to improve available treatments. Identifying characteristics and features of non-responders and drop-outs would allow treatment to be tailored more specifically to the patients (13). However, Leichsenring et al. (13) neglect the fact that theoretical or statistical models are needed to accurately predict whether the patient's treatment will be successful or unsuccessful. We argue that such models can be developed with machine learning.

Machine learning is a field of computer science, in which the "computer" is supposed to "learn" models from data (14). "Learning" in this context means that statistical models are adapted to the data until they optimally perform a previously defined task, for example predicting drop-out rates of psychotherapy (14, 15). These statistical models can be the same methods that are used in classical statistical approaches, such as regression analyses. There is therefore no clear boundary between machine learning and classical statistical approaches (14). However, machine learning differs from classical statistical approaches in the primary way the developed models are evaluated. In classical statistical approaches, the developed models are assessed primarily with the help of statistical significance, explained variance, and many other characteristic values (16). In contrast, models in the machine learning approach are assessed primarily by how well they can perform the task they have "learned" on new data that is unknown to the often iterative model fitting process (14, 15, 17). In practice, this means that a model is "taught" by means of a first data set and then evaluated on a second data set. Machine learning approaches can further be divided into unsupervised and supervised learning (18). The primary goal of unsupervised learning is to discover relationships and structures in the data (14, 15). Commonly used statistical models for unsupervised machine learning include explanatory factor analysis, k-means clustering, and hierarchical clustering (19). While supervised learning also discovers correlations and structures in the data, the goal is to determine the value of a dependent variable as accurately as possible (14, 15). A prerequisite for this is that the dependent variable is known, both in the data set in which the model is being "taught", and in which it is being evaluated. Commonly used statistical models for supervised machine learning include regression analysis, support vector machines, random forest, and latent discriminant analysis (19). For a more detailed description of machine learning and its own terminology, the interested reader is referred here to Dwyer et al. and Bi et al. (14, 15).

Because models developed using the machine learning approach are primarily evaluated for their ability to perform a previously defined task for new unknown data, they often perform better (assuming access to an appropriate dataset) in tasks such as predicting whether a patient's treatment will be successful, as opposed to models developed using the classical statistical approach. Machine learning thus has the potential to develop models that could lead to psychodynamic psychotherapeutic treatments being more successful. However, it is unclear whether machine learning is currently used in psychodynamic psychotherapy research. In 2019, Aafjes-van Doorn et al. (19) found 51 studies which utilized machine learning to analyse psychotherapy. Most of those studies were initial proof-of-concept studies, which either predicted the outcome of therapy, or automatically rated patient behavior for further analyses. Most of these 51 studies used transcripts of psychotherapy sessions as data and utilized supervised machine learning to answer their research questions. However, Aafjes-van Doorn et al. (19) did not differentiate between specific treatment approaches. Machine learning may have other applications in psychodynamic psychotherapy research because of the focus on the patient's unconscious. Yet, to the best of our knowledge, there seem to be no current systemic reviews about machine learning within outpatient psychodynamic psychotherapy research. Therefore, we focused our literature review on the use of machine learning in outpatient psychodynamic psychotherapy research.

# Methods

For this review, we applied the Preferred Reporting Items for systematic Reviews and Meta-Analyses (PRISMA) Guidelines.

## Search strategy and eligibility criteria

We searched the two most comprehensive databases regarding psychotherapy research, "PsycInfo" and "PubMed". In preparation for this mini-review, we searched several databases (PsycInfo, PubMed, Heidi, Google Scholar and IEEE Xplore) for relevant literature and did not obtain any additional results beyond those from PubMed and PsycInfo. Therefore, we estimated that there would be little loss of knowledge if we omitted further databases. All 36 combinations of the terms (psychothera* OR thera* OR clinical assessment) AND (machine learning OR artificial intelligence OR neural network OR deep learning) AND (patient* OR client* OR mental health) with no limitation on publication year were searched. As there is a corpus of theoretical work comparing neural processes to artificial intelligence that considers how to use the conclusions for psychotherapy, we added the term (patient* OR client* OR mental health) to the search. We sought to omit such work. The two searches were conducted on 17th September 2021 and 2nd January 2023, respectively.

## Eligibility criteria

To be eligible, studies had to be original works, treat their patients with outpatient psychotherapy and use machine learning as a statistical method. Results were limited to publications in English. There were no further eligibility criteria.

## Selection process

The first and second author read all abstracts of the articles and selected the studies which appeared to meet the eligibility criteria. In a second step, the manuscripts were read and discussed among the first and second author. During this stage, studies whose psychotherapeutic treatment was not psychodynamic, or whose treatment consisted only of diagnostics, were not considered for this review. Furthermore, all studies that did not use machine learning as a statistical method were removed. Lastly, studies which treated patients with several treatment approaches, yet did not differentiate between them, were excluded, as it was impossible to attribute the results to one specific treatment approach.

## Data items

The first and second author independently retrieved the research question and the respective use of machine learning from the included studies. They also retrieved the sample size and data used for machine learning.

## Bias assessment

To assess outcome bias, we checked which characteristics the studies reported. According to Lantz (17), a study should at least report accuracy, specificity and sensitivity for an unbiased report of a machine learning model. Therefore, these aspects were taken into consideration during the selection process. Since it could be assumed that machine learning is a new field in outpatient psychodynamic psychotherapy research (19), our primary goal was to gain an overview of research conducted in this area. Therefore, a full assessment of report quality using the TRIPOD (20) criteria would have been beyond the scope of this work.

# Results

## Study selection

The initial search of both databases yielded 1,358 results, the second 6,206. In total, 3,216 were duplicate records, which were removed before screening. Of the 4,348 records screened, 4,289 were excluded, as they did not meet the eligibility criteria. Most excluded records were related to brain research and prediction of recovery processes after surgery. The second largest group of excluded records were associated with treating patients with other approaches than outpatient psychodynamic psychotherapy. In total, 59 records appeared to have met our eligibility criteria. Two records were inaccessible *via* any platform. Another 27 reports were excluded because outpatient psychodynamic psychotherapy was not a form of treatment. Nine of the 27 studies were conducted by a research group under Professor Atkins, who successfully created an automatic transcription and evaluation tool for short term psychotherapy, namely motivational interviewing (21–29). Ten of the 27 excluded studies either tried to predict the outcome of cognitive behavior therapy by using natural language processing, or tried to predict optimal therapeutic interventions with sociodemographic data (30–40). Another 16 studies excluded were reviews about machine learning and its utilization within psychotherapy, psychopharmacotherapy, and diagnostics (19, 27, 38, 41–51). Another 8 Studies were excluded because they treated patients with psychotherapy, yet did not specify which kind of treatment approach they used (22, 29, 52–57). Lastly, two studies retrieved were theoretical studies (58, 59). To summarize, only 4 out of 4,348 records screened were eligible for our review.

## Study characteristics and results

An overview of the four studies that were deemed eligible, as well as the utilized machine learning methods and their bias assessment, can be seen in Tables 1, 2. Two of the four studies were single-dyad studies. Villmann et al. (63) examined the possible use of artificial neural networks to investigate psycho-physiological parameters derived during the therapy sessions. The authors measured five physiological parameters across 37 therapy sessions for both the patient and therapist and transcribed all sessions. The machine learning model applied was a growing-self-organizing map to combine the psychophysiological data into emotional entropy. Emotional entropy can be understood as emotional variability or emotional energy. The session transcripts were processed with the Mergenthaler Cycle Model (64), which groups the words in the transcripts into four topics: relaxing, experiencing, reflecting and connecting. The authors then compared the emotional entropy with the Mergenthaler Cycle topics. In doing so, they found a cyclic process (64). The patient experiences an interpersonal conflict which increases emotional entropy. The conflict is then reflected upon, and the patient connects the interpersonal conflict with an inner conflict. This connection unleashes emotional energy, which enables a structural change within the patient. Afterwards, a period of relaxation and stabilization follows. Villmann et al. (63) described their proof-of-concept study as an initial first step, which should be verified in future studies.

The second single-dyad study was done by Laskoski et al. (62). They used a random-forest model to predict patient distress based on coded interventions from a videotaped psychoanalysis, consisting of 120 sessions. Trained judges rated the psychotherapist's interventions with the Psychotherapy Process Q-Set (65). The patient answered the Outcome Questionnaire after each session (66). The random-forest model had an AUC of 0.725, sensitivity of 79%, specificity of 79%, and accuracy of 70.5 % in predicting patient stress after therapy sessions. Additionally, the authors calculated the variable importance of the predictors and found standard techniques of psychodynamic therapy, e.g., drawing the patient's attention to unconscious content or linking the patient's feelings to past situations, to be the most important factors in reducing patient distress.

The third study was conducted by Atzil-Slonim et al. (60). They used Latent Dirichlet allocation to extract various topics from session transcripts. Then, a sparse multinomial logistic regression was used to predict the social functioning and symptom distress after each therapy session based on the topics discussed. Social functioning and symptom distress of the patient were measured with the Outcome Rating Scale and Symptom-Checklist (67, 68). In total, they analyzed 873 therapy sessions deriving from 58 patients and 52 therapists. Results showed that an increase in positive topics was positively correlated with high social functioning and associated with a decrease in symptoms distress. Conversely, an increase of negatively connotated topics correlated with an increase of symptom distress. Accuracy of the final model was at 75.6 % with regard to predicting social functioning.

Halfon et al. (61) tried to predict four basic emotions (joy, anger, sadness and anxiety) of children within a psychodynamic play therapy. Their sample consisted of at least two randomly

TABLE 1 Study characteristics.

| First Author | Study Design | Sample | Research question | Data used |
|---|---|---|---|---|
| Atzil-Slonim et al. (60) | Longitudinal study,<br>no RCT,<br>exploratoty study | 873 therapy sessions from 58 patients and 52 therapists | Prediction of symptom reduction and social functioning based on topics spoken about in therapy | Transcript of therapy sessions,<br>Symptom Checklist,<br>Outcome Rating Scale |
| Halfon et al. (61) | Longitudinal study,<br>no RCT,<br>exploratory study | 148 therapy sessions from 53 children and 24 therapists | Prediction of a child's affect expressions based on video or transcription of therapy sessions | Video and Transcript of sessions,<br>Affect expression scale of the Children's Play Therapy Instrument |
| Laskoski et al. (62) | Longitudinal study,<br>proof-of-concept | 120 therapy sessions from 1 patient and 1 therapist | Prediction of Patient distress based on therapist behavior and interventions | Psychotherapy Process Q-set (requires video of sessions),<br>Outcome questionnaire |
| Villmann et al. (63) | Longitudinal study,<br>proof-of-concept | 37 therapy sessions from 1 patient and 1 therapist | Studying behavior of psychophysiological variables within therapy | Heart rate, respiratory frequency,<br>muscular tension,<br>skin conductance response,<br>skin conductance level, transcript of sessions |

TABLE 2 Utilized machine learning models.

| First author | Machine learning models | Type of machine learning | Application in paper | Bias assessment |
|---|---|---|---|---|
| Atzil-Slonim et al. (60) | Latent dirichlet allocation | Unsupervised | Extracting topics discussed in psychotherapy session | Potentially biased. Authors only report accuracy. |
| | Sparse multinomial logistic regression | Supervised | Predicting Patient Outcome based on topics discussed in therapy session | |
| Halfon et al. (61) | Dictionary approach | Unsupervised | Generating affect scores from session transcripts, based on existing corpora | Not applicable. Authors correlate results with human raters. Instead of assessing accuracy, specificity and sensitivity of model. |
| | Deep neural network | Both | To generate valence and arousal scores from therapy videos (pre-trained on affect net database) | |
| | Support vector machine | Supervised | Predicting the affect of children with on the generated affect scores | |
| | Extreme learning machine | Supervised | Predicting the affect of children with the generated affect scores | |
| Laskoski et al. (62) | Random forest algorithm | Supervised | Predicting patient distress with coded therapist behavior and interventions | Potentially biased. Authors only report values from best model. |
| Villmann et al. (63) | Dictionary approach | Unsupervised | Grouping Words of session transcripts into Mergenthaler Cycle topics | Potentially biased. Authors report neither accuracy, sensitivity nor specificity. |
| | Growing Self-Organizing Map | Unsupervised | Creating a lower dimensional description of psychophysiological data (comparable to a non-linear Principal component analysis) | |

drawn videotaped sessions per therapy. In total, 148 videotaped sessions of 53 children and 24 psychotherapists were selected. Emotional expressions of children were coded by trained judges using the affect expression items of the Children's Play Therapy Instrument (69). The videos were transcribed separately. Afterwards, the authors trained several supervised machine learning models to predict the affect expressions of children based on the transcript or the video. Overall, a fusion strategy, which combined text analysis and facial recognition to predict affect expressions, achieved the best results. Still, affect expression predictions of the final model correlated on average $r = 0.30$ with the ratings of trained judges. Halfon et al. (61) concluded that the "automatic affect analysis is promising, however, needs further development."

## Synthesis of results

Our review identified four studies that utilized machine learning within outpatient psychodynamic psychotherapy research. All four studies are proof-of-concept studies. Furthermore, none of the four studies reported their results without bias. All four studies differed in the type of data they used for their machine learning models and their study aims. All studies trained their machine learning models with data from completed outpatient psychodynamic psychotherapies only. Lastly, three of the four studies were done within the last 5 years.

## Discussion

Some authors call for psychodynamic psychotherapies to be tailored to patients as much as possible in order to be more successful (13). However, this requires models that allow an accurate prediction of whether the therapy will be successful or unsuccessful. We argue that models developed using the machine learning approach are particularly well suited for this purpose. Within machine learning, models are evaluated primarily for their ability to perform a predetermined task on new data (14, 15). However, since it was unknown how widely machine learning is represented in outpatient psychodynamic psychotherapy research, we conducted a review. Our systematic review identified four proof-of-concept studies that utilized machine learning. Three studies were published between 2019 and 2021 and two studies had a single-dyad sample. All four studies utilized machine learning to evaluate completed outpatient psychodynamic psychotherapies. It seems that machine learning has only recently entered outpatient psychodynamic psychotherapy research. However, this could also mean that researchers are not yet aware of what machine learning can be used for in psychodynamic psychotherapy research. We therefore want to present perspectives and ideas that can be used for future psychodynamic psychotherapy studies. We would then like to highlight a possible risk that might occur when using machine learning in psychotherapy.

To make psychotherapies more successful, Leichsenring et al. (13) suggest that therapy should be tailored to the needs of patients with high non-response and drop-out probability. A possible implementation of this idea would be to predict therapy success or drop-out at the beginning of the therapy and to include these predictions in the therapy planning. Some studies within cognitive behavioral therapy research attempted to implement this (42, 53, 70, 71). A commonality of these studies is that socio-demographic data was collected before the start of psychotherapy and used to predict psychotherapy success with the help of machine learning models. Psychotherapy success was operationalised differently, either as symptom improvement, drop-out or improvement in quality of life.

Another implementation of the idea of tailoring therapy to patients' needs could include feedback to the therapists. De Jong et al. (72) were able to show that feedback to the therapists reduces the drop-out probability of patients by 20% and leads to stronger symptom improvement. Machine learning can be used in this context to develop models that automatically evaluate audio transcripts of psychotherapy sessions and provide feedback to the therapist. The therapist would then be able to get timely feedback about possible pathological developments and could intervene accordingly. A research group led by Professor Atkins is currently attempting to implement this idea (21, 22, 25, 26, 28). Currently, they are only successful in doing this for the very standardized Motivational Interviewing (21). However, it seems possible to build on the work of this group and provide therapists with feedback on variables that are relevant to psychodynamic therapy, such as the agency, working alliance, and the patient's structural integration of personality (7, 8, 11).

A third way to tailor therapy to patients' needs is to predict the fit between therapist and patient. Delgadillo et al. (33) found that there are differences between therapists in the effectiveness with which they treat individual patient groups. Their final machine learning model identified 17 classes of patient-to-therapist matches, which vary greatly in their effectiveness. Building on this idea of Delgadillo et al. (33), it would be conceivable to develop a machine learning models that can predict which therapist has the highest probability of achieving a successful therapy with a patient.

On the other hand, the use of machine learning in psychotherapy research should be carefully considered. As the previous ideas illustrate, models developed with machine learning have the ability to automate many processes, such as feedback to and allocation of patients to therapists. This poses the risk that such models could become an unreflective and potentially discriminatory standard (26, 73). In other words, minorities and vulnerable groups are disadvantaged, for instance, by being denied psychotherapeutic treatment because the model predicts that treatment will be unsuccessful. In this context, Hirsch et al. (26) examined how well their model, which gave feedback on Motivational Interviewing, was accepted by therapists. They found that novices in particular tended to accept the feedback without reflection. Therefore, Besse et al. (73) warned that this can also systematically create discrimination, especially if the model was not developed on the basis of theoretical considerations and representative data (73, 74). The use of machine learning in psychotherapy research should therefore be embedded in existing theories.

## Limitations

Several limitations of the presented work must be mentioned. Some studies which utilized machine learning as a method within outpatient psychodynamic psychotherapy research may not have been considered in our review, as we had rather strict criteria for inclusion. Although Zilcha-Mano et al. (56) treated some patients with psychodynamic psychotherapy, it was excluded in our review, as they did not differentiate their results between treatment approaches. Furthermore, we only included studies which explicitly mentioned, in their abstracts, that they used machine learning, deep learning or a form of artificial intelligence. It is conceivable that articles referring to their methodology with the name of statistical model, instead of machine learning, were not included. Furthermore, articles that did not mention their methodology within the abstract, although they used machine learning, may also have been disregarded. As we only found four studies within outpatient psychodynamic psychotherapy research that used machine learning, our review summarizes the first attempts at adopting machine learning into this field of research. Therefore, various limitations mentioned by Aafjes-van Doorn et al. (19) also apply to this review. As these are among the first studies in this area, they should be interpreted cautiously, namely as proof-of-concepts studies, so that the significance of their results is not overestimated. Therefore, we also refrained from assessing the quality of the studies using TRIPOD criteria (20). Thus, we recommend that future reviews in this field use the TRIPOD criteria (20) to assess the quality of studies.

## Conclusion

Although much research has been done on psychodynamic psychotherapy, the treatment success of this therapy method has not improved. We argue that machine learning is a way to develop models that detect non-responders and patients with high drop-out probability early and enable intervention. However, since it was unknown how widespread machine learning is in outpatient psychodynamic psychotherapy research, we felt it necessary to conduct a review of current research. We found four studies, three of which were carried out between 2019 and 2021. Thus, machine learning seems to have entered this field of research only recently and researchers might not yet be aware of its possible uses. We have therefore outlined some possibilities, ideas, and perspectives on how machine learning can be used to improve the success of psychodynamic psychotherapies. Thus, we hope to give new impetus to outpatient psychodynamic psychotherapy research on how to use machine learning to address previously unsolved problems.

## Author contributions

IR planned the study, conducted the literature search, and wrote the manuscript. NG was responsible with IR for selecting the studies to include into the review. SS-T used her expertise to correct the parts about machine learning in the manuscript. JS was an advisor about machine learning. MS was substantially aided in improving the accuracy of the results and discussion. H-CF and CN were the supervisors for this study. All authors contributed to the article and approved the submitted version.

## 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.

## References

1. Steinert C, Munder T, Rabung S, Hoyer J, Leichsenring F. Psychodynamic therapy: as efficacious as other empirically supported treatments? A meta-analysis testing equivalence of outcomes. *Am J Psychiatry.* (2017) 174:943–53. doi: 10.1176/appi.ajp.2017.17010057

2. Ehrenthal JC, Dinger U, Nikendei C. Aktuelle entwicklungen der psychodynamischen psychotherapieforschung. *Psychotherapeut.* (2014) 59:212–8. doi: 10.1007/s00278-014-1045-5

3. Leichsenring F, Leweke F, Klein S, Steinert C. The empirical status of psychodynamic psychotherapy—an update: Bambi's alive and kicking. *Psychother Psychosom.* (2015) 84:129–48. doi: 10.1159/000376584

4. Maljanen T, Knekt P, Lindfors O, Virtala E, Tillman P, Harkanen T, et al. The cost-effectiveness of short-term and long-term psychotherapy in the treatment of depressive and anxiety disorders during a 5-year follow-up. *J Affect Disord.* (2016) 190:254–63. doi: 10.1016/j.jad.2015.09.065

5. Yonatan-Leus R, Strauss AY, Cooper-Kazaz R. Psychodynamic psychotherapy is associated with sustained reduction in health care utilization and cost. *Clin Psychol Psychother.* (2021) 28:642–55. doi: 10.1002/cpp.2527

6. Rosso G, Aragno E, Cuomo A, Fagiolini A, Di Salvo G, Maina G. Five-year follow-up of first-episode depression treated with psychodynamic psychotherapy or antidepressants. *Psychiatry Res.* (2019) 275:27–30. doi: 10.1016/j.psychres.2019.02.073

7. Huber J, Jennissen S, Nikendei C, Schauenburg H, Dinger U. Agency and alliance as change factors in psychotherapy. *J Consult Clin Psychol.* (2021) 89:214–26. doi: 10.1037/ccp0000628

8. Volz M, Jennissen S, Schauenburg H, Nikendei C, Ehrenthal JC, Dinger U. Intraindividual dynamics between alliance and symptom severity in long-term psychotherapy: why time matters. *J Couns Psychol.* (2021) 68:446. doi: 10.1037/cou0000545

9. Jennissen S, Connolly Gibbons MB, Crits-Christoph P, Schauenburg H, Dinger U. Insight as a mechanism of change in dynamic therapy for major depressive disorder. *J Couns Psychol.* (2021) 68:435. doi: 10.1037/cou0000554

10. Jennissen S, Huber J, Ehrenthal JC, Schauenburg H, Dinger U. Association between insight and outcome of psychotherapy: systematic review and meta-analysis. *Am J Psychiatry.* (2018) 175:961–9. doi: 10.1176/appi.ajp.2018.17080847

11. Cierpka M. *Operationalisierte Psychodynamische Diagnostik OPD-2: das Manual für Diagnostik und Therapieplanung.* Berlin: Huber (2006).

12. Halstensen K, Gjestad R, Luyten P, Wampold B, Granqvist P, Stalsett G, et al. Depression and mentalizing: a psychodynamic therapy process study. *J Couns Psychol.* (2021) 68:705–18. doi: 10.1037/cou0000544

13. Leichsenring F, Steinert C, Ioannidis JPA. Toward a paradigm shift in treatment and research of mental disorders. *Psychol Med.* (2019) 49:2111–7. doi: 10.1017/S0033291719002265

14. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol.* (2019) 188:2222–39. doi: 10.1093/aje/kwz189

15. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol.* (2018) 14:91–118. doi: 10.1146/annurev-clinpsy-032816-045037

16. Döring N, Bortz J. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.* Berlin, Heidelberg: Springer (2016). doi: 10.1007/978-3-642-41089-5

17. Lantz B. *Machine Learning with R: Expert Techniques for Predictive Modeling.* Birmingham: Packt publishing ltd. (2019).

18. Barber D. *Bayesian Reasoning and Machine Learning.* Cambridge: Cambridge Univ Press. (2015).

19. Aafjes-van Doorn K, Kamsteeg C, Bate J, Aafjes M. A scoping review of machine learning in psychotherapy research. *Psychother Res.* (2021) 31:92–116. doi: 10.1080/10503307.2020.1808729

20. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Bmj.* (2015) 350:g7594. doi: 10.1136/bmj.g7594

21. Flemotomos N, Martinez VR, Chen Z, Singla K, Ardulov V, Peri R, et al. Automated evaluation of psychotherapy skills using speech and language technologies. *Behav Res Methods.* (2021) 54:690–711. doi: 10.3758/s13428-021-01623-4

22. Gaut G, Steyvers M, Imel ZE, Atkins DC, Smyth P. Content coding of psychotherapy transcripts using labeled topic models. *IEEE J Biomed Health Inform.* (2017) 21:476–87. doi: 10.1109/JBHI.2015.2503985

23. Goldberg SB, Flemotomos N, Martinez VR, Tanana MJ, Kuo PB, Pace BT, et al. Machine learning and natural language processing in psychotherapy research: alliance as example use case. *J Couns Psychol.* (2020) 67:438–48. doi: 10.1037/cou0000382

24. Goldberg SB, Tanana M, Imel ZE, Atkins DC, Hill CE, Anderson T. Can a computer detect interpersonal skills? Using machine learning to scale up the facilitative interpersonal skills task. *Psychotherapy Res.* (2020) 31:281–88. doi: 10.1080/10503307.2020.1741047

25. Hirsch T, Merced K, Narayanan S, Imel ZE, Atkins DC. Designing contestability: interaction design, machine learning, and mental health. *DIS.* (2017) 2017:95–9. doi: 10.1145/3064663.3064703

26. Hirsch T, Soma C, Merced K, Kuo P, Dembe A, Caperton DD, et al. "It's hard to argue with a computer:" investigating psychotherapists' attitudes towards automated evaluation. *DIS.* (2018) 2018:559–71. doi: 10.1145/3196709.3196776

27. Imel ZE, Caperton DD, Tanana M, Atkins DC. Technology-enhanced human interaction in psychotherapy. *J Couns Psychol.* (2017) 64:385–93. doi: 10.1037/cou0000213

28. Imel ZE, Pace BT, Soma CS, Tanana M, Hirsch T, Gibson J, et al. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy.* (2019) 56:318–28. doi: 10.1037/pst0000221

29. Tanana MJ, Soma CS, Kuo PB, Bertagnolli NM, Dembe A, Pace BT, et al. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behav Res Methods*. (2021) 1–14. doi: 10.3758/s13428-020-01531-z

30. Aalbers G, Engels T, Haslbeck JMB, Borsboom D, Arntz A. The network structure of schema modes. *Clin Psychol Psychother.* (2021) 28:1065–78. doi: 10.1002/cpp.2577

31. Bohannon J. The synthetic therapist: Some people prefer to bare their souls to computers rather than to fellow humans. *Science.* (2015) 349:250–1. doi: 10.1126/science.349.6245.250

32. Delgadillo J, Gonzalez Salas Duhne P. Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach. *J Consult Clin Psychol.* (2020) 88:14–24. doi: 10.1037/ccp0000476

33. Delgadillo J, Rubel J, Barkham M. Towards personalized allocation of patients to therapists. *J Consult Clin Psychol.* (2020) 88:799–808. doi: 10.1037/ccp0000507

34. Ewbank MP, Cummins R, Tablan V, Bateup S, Catarino A, Martin AJ, et al. Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. *JAMA Psychiatry.* (2020) 77:35–43. doi: 10.1001/jamapsychiatry.2019.2664

35. Ewbank MP, Cummins R, Tablan V, Catarino A, Buchholz S, Blackwell AD, et al. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychother Res.* (2020) 31:300–12. doi: 10.1080/10503307.2020.1788740

36. Gómez Penedo JM, Schwartz B, Giesemann J, Rubel JA, Deisenhofer A-K, Lutz W, et al. For whom should psychotherapy focus on problem coping? A machine learning algorithm for treatment personalization. *Psychotherapy Res.* (2021) 32:151–64. doi: 10.1080/10503307.2021.1930242

37. Hilbert K, Jacobi T, Kunas SL, Elsner B, Reuter B, Lueken U, et al. Identifying cbt non-response among ocd outpatients: a machine-learning approach. *Psychotherapy Res.* (2020) 31:52–62. doi: 10.1080/10503307.2020.1839140

38. Hilbert K, Lueken U. Prädiktive analytik aus der perspektive der klinischen psychologie und psychotherapie = Predictive analytics from a mental health perspective. *Verhaltenstherapie.* (2020) 30:8–17. doi: 10.1159/000505302

39. Probst T, Kleinstäuber M, Lambert MJ, Tritt K, Pieh C, Loew TH, et al. Why are some cases not on track? An item analysis of the assessment for signal cases during inpatient psychotherapy. *Clin Psychol Psychotherapy*. (2020) 27:559–66. doi: 10.1002/cpp.2441

40. Yao L, Zhao X, Xu Z, Chen Y, Liu L, Feng Q, et al. Influencing factors and machine learning-based prediction of side effects in psychotherapy. *Front Psychiatry.* (2020) 11:537442. doi: 10.3389/fpsyt.2020.537442

41. Andersson G, Titov N, Dear BF, Rozental A, Carlbring P. Internet-delivered psychological treatments: from innovation to implementation. *World Psychiatry.* (2019) 18:20–8. doi: 10.1002/wps.20610

42. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry.* (2021) 20:154–70. doi: 10.1002/wps.20882

43. Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. *Mol Psychiatry.* (2019) 24:1583–98. doi: 10.1038/s41380-019-0365-9

44. Geoghegan L, Scarborough A, Wormald JCR, Harrison CJ, Collins D, Gardiner M, et al. Automated conversational agents for post-intervention follow-up: a systematic review. *BJS Open.* (2021) 5:zrab070. doi: 10.1093/bjsopen/zrab070

45. Horn RL. Weisz JR. Can artificial intelligence improve psychotherapy research and practice? *Administrat Policy Mental Health Mental Health Services Res.* (2020) 47:852–5. doi: 10.1007/s10488-020-01056-9

46. Huys QJM, Browning M, Paulus MP, Frank MJ. Advances in the computational understanding of mental illness. *Neuropsychopharmacology.* (2021) 46:3–19. doi: 10.1038/s41386-020-0746-4

47. Lee Y, Ragguett R-M, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *J Affect Disord.* (2018) 241:519–32. doi: 10.1016/.jad.2018.08.073

48. Lee Y, Ragguett R-M, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. 'Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review': corrigendum. *J Affect Disord.* (2020) 274:1211–5. doi: 10.1016/j.jad.2020.02.037

49. Tahan M. Artificial Intelligence applications and psychology: an overview. *Neuropsychopharmacol Hung.* (2019) 21:119–26. doi: 10.32598/ajnpp.4.3.210

50. Tracey TJG. The scientific future of counseling psychology: five specific areas of predictions. *J Couns Psychol.* (2017) 64:347–8. doi: 10.1037/cou0000234

51. Zale A, Lasecke M, Baeza-Hernandez K, Testerman A, Aghakhani S, Munoz RF, et al. Technology and psychotherapeutic interventions: bibliometric analysis of the past four decades. *Internet Interv.* (2021) 25:100425. doi: 10.1016/j.invent.2021.100425

52. Bavaresco R, Barbosa J, Vianna H, Buttenbender P, Dias L. Design and evaluation of a context-aware model based on psychophysiology. *Comput Methods Programs Biomed.* (2020) 189:105299. doi: 10.1016/j.cmpb.2019.105299

53. Bone C, Simmonds-Buckley M, Thwaites R, Sandford D, Merzhvynska M, Rubel J, et al. Dynamic prediction of psychological treatment outcomes: development and validation of a prediction model using routinely collected symptom data. *Lancet Digit Health.* (2021) 3:e231–e40. doi: 10.1016/S2589-7500(21)00018-2

54. Bruijniks SJE, DeRubeis RJ, Lemmens LHJM, Peeters FPML, Cuijpers P, Huibers MJH. The relation between therapy quality, therapy processes and outcomes and identifying for whom therapy quality matters in CBT and IPT for depression. *Behav Res Ther.* (2021) 139:103815. doi: 10.1016/j.brat.2021.103815

55. de Mello FL, de Souza SA. Psychotherapy and artificial intelligence: a proposal for alignment. *Front Psychol.* (2019) 10:263. doi: 10.3389/fpsyg.2019.00263

56. Zilcha-Mano S, Errazuriz P, Yaffe-Herbst L, German RE, DeRubeis RJ. Are there any robust predictors of "sudden gainers," and how is sustained improvement in treatment outcome achieved following a gain? *J Consult Clin Psychol.* (2019) 87:491–500. doi: 10.1037/ccp0000401

57. Ziobrowski HN, Cui R, Ross EL, Liu H, Puac-Polanco V, Turner B, et al. Development of a model to predict psychotherapy response for depression among veterans. *Psychol Med.* (2022) 1–10. doi: 10.1017/S0033291722000228

58. Kinley JL, Reyno SM. Project for a scientific psychiatry: A neurobiologically informed, phasic, brain-based model of integrated psychotherapy. *J Psychother Integr.* (2016) 26:61–73. doi: 10.1037/a0039636

59. Caspar F, Rothenfluh T, Segal Z. The appeal of connectionism for clinical psychology. *Clin Psychol Rev.* (1992) 12:719–62. doi: 10.1016/0272-7358(92)90022-Z

60. Atzil-Slonim D, Juravski D, Bar-Kalifa E, Gilboa-Schechtman E, Tuval-Mashiach R, Shapira N, et al. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. *Psychotherapy.* (2021) 58:324–39. doi: 10.1037/pst0000362

61. Halfon S, Doyran M, Türkmen B, Oktay EA, Salah AA. Multimodal affect analysis of psychodynamic play therapy. *Psychother Res.* (2020) 31:313–28. doi: 10.1080/10503307.2020.1839141

62. Laskoski PB, Serralta FB, Passos IC, Hauck S. Machine-learning approaches in psychotherapy: a promising tool for advancing the understanding of the psychotherapeutic process. *Braz J Psychiatry.* (2019) 41:568–9. doi: 10.1590/1516-4446-2018-0295

63. Villmann T, Liebers C, Bergmann B, Gumz A, Geyer M. Investigation of psycho-physiological interactions between patient and therapist during a psychodynamic therapy and their relation to speech using in terms of entropy analysis using a neural network approach. *New Ideas Psychol.* (2008) 26:309–25. doi: 10.1016/j.newideapsych.2007.07.010

64. Mergenthaler E. Emotion-abstraction patterns in verbatim protocols: a new way of describing psychotherapeutic processes. *J Consult Clin Psychol.* (1996) 64:1306–15. doi: 10.1037/0022-006X.64.6.1306

65. Jones E. *Manual for the Psychotherapy Process Q-set. Unpublished Manuscript.* Berkeley: University of California. (1985).

66. Lambert MJ, Burlingame GM, Umphress V, Hansen NB, Vermeersch DA, Clouse GC, et al. The reliability and validity of the outcome questionnaire. *Clin Psychol Psychother.* (1996) 3:249–58.

67. Franke GH. Symptom-Checklist-90®-Standard: SCL-90®-S. *Göttingen [ua]: Hogrefe.* (2014).

68. Miller SD, Duncan B, Brown J, Sparks J, Claud D. The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *J Brief Therapy.* (2003) 2:91–100.

69. Kernberg PF, Chazan SE, Normandin L. The children's play therapy instrument (CPTI). Description, development, and reliability studies. *J Psychother Pract Res.* (1998) 7:196–207. doi: 10.1037/t72860-000

70. Koppe G, Meyer-Lindenberg A, Durstewitz D. Deep learning for small and big data in psychiatry. *Neuropsychopharmacology.* (2021) 46:176–90. doi: 10.1038/s41386-020-0767-z

71. Poster K, Bennemann B, Hofmann SG, Lutz W. Therapist interventions and skills as predictors of dropout in outpatient psychotherapy. *Behavior Therapy*. (2021) 52:1489–1501. doi: 10.1016/j.beth.2021.05.001

72. de Jong K, Conijn JM, Gallagher RAV, Reshetnikova AS, Heij M, Lutz MC. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. *Clin Psychol Rev.* (2021) 85:102002. doi: 10.1016/j.cpr.2021.102002

73. Besse P, Castets-Renard C, Garivier A, Loubes J-M. Can everyday ai be ethical? Machine learning algorithm fairness (english version). *Statistiques et Société.* ( 2018) 6:8.

74. Elhai JD, Montag C. The compatibility of theoretical frameworks with machine learning analyses in psychological research. *Current Opinion Psychol.* (2020) 36:83–8. doi: 10.1016/j.copsyc.2020.05.002

TYPE Original Research
PUBLISHED 26 May 2023
DOI 10.3389/fpsyt.2023.1071622

# Machine learning techniques for identifying mental health risk factor associated with schoolchildren cognitive ability living in politically violent environments

Radwan Qasrawi[1,2]*, Stephanny Vicuna Polo[3], Rami Abu Khader[3], Diala Abu Al-Halawa[4], Sameh Hallaq[5], Nael Abu Halaweh[1] and Ziad Abdeen[4]

[1]Department of Computer Sciences, Al-Quds University, Jerusalem, Palestine, [2]Department of Computer Engineering, Istinye University, Istanbul, Türkiye, [3]Al-Quds Center for Business Innovation and Entrepreneurship, Al-Quds University, Jerusalem, Palestine, [4]Faculty of Medicine, Al-Quds University, Jerusalem, Palestine, [5]Al-Quds Bard College for Arts and Sciences, Al-Quds University, Jerusalem, Palestine

OPEN ACCESS

EDITED BY
Marcos Del Pozo Banos,
Swansea University Medical School,
United Kingdom

REVIEWED BY
Olivier Y. Rouquette,
Swansea University Medical School,
United Kingdom
Deepak Gupta,
National Institute of Technology,
Arunachal Pradesh, India

*CORRESPONDENCE
Radwan Qasrawi
✉ radwan@staff.alquds.edu

RECEIVED 16 October 2022
ACCEPTED 09 May 2023
PUBLISHED 26 May 2023

CITATION
Qasrawi R, Vicuna Polo S, Abu Khader R, Abu Al-Halawa D, Hallaq S, Abu Halaweh N and Abdeen Z (2023) Machine learning techniques for identifying mental health risk factor associated with schoolchildren cognitive ability living in politically violent environments. *Front. Psychiatry* 14:1071622. doi: 10.3389/fpsyt.2023.1071622

**Introduction:** Mental health and cognitive development are critical aspects of a child's overall well-being; they can be particularly challenging for children living in politically violent environments. Children in conflict areas face a range of stressors, including exposure to violence, insecurity, and displacement, which can have a profound impact on their mental health and cognitive development.

**Methods:** This study examines the impact of living in politically violent environments on the mental health and cognitive development of children. The analysis was conducted using machine learning techniques on the 2014 health behavior school children dataset, consisting of 6373 schoolchildren aged 10–15 from public and United Nations Relief and Works Agency schools in Palestine. The dataset included 31 features related to socioeconomic characteristics, lifestyle, mental health, exposure to political violence, social support, and cognitive ability. The data was balanced and weighted by gender and age.

**Results:** This study examines the impact of living in politically violent environments on the mental health and cognitive development of children. The analysis was conducted using machine learning techniques on the 2014 health behavior school children dataset, consisting of 6373 schoolchildren aged 10-15 from public and United Nations Relief and Works Agency schools in Palestine. The dataset included 31 features related to socioeconomic characteristics, lifestyle, mental health, exposure to political violence, social support, and cognitive ability. The data was balanced and weighted by gender and age.

**Discussion:** The findings can inform evidence-based strategies for preventing and mitigating the detrimental effects of political violence on individuals and communities, highlighting the importance of addressing the needs of children in conflict-affected areas and the potential of using technology to improve their well-being.

KEYWORDS

**mental health, cognitive abilities, machine learning, prediction, health, social support, nutrition**

# 1. Introduction

The normal cognitive development of children living in conflict areas is crucial, given the adverse situation they face, which can have a long-lasting impact on their well-being. Minimizing the exposure to and impact of other risk factors is important, as research has shown that there is a relationship between cognitive development and future ill mental health. Several studies have highlighted the negative effects of exposure to violence and mental health difficulties on children's cognitive development (1–4). Mental health disorders, such as depression, anxiety, and stress, have been linked with decreased cognitive functioning, whereas political violence can have a significant impact on the mental health and cognitive ability of children. Exposure to violence and traumatic events, such as war and civil conflict, can result in mental health disorders such as post-traumatic stress disorder (PTSD), depression, anxiety, and stress (1, 2). These disorders can have negative impacts on children's cognitive development, including their memory, attention, and executive function (4). Moreover, mental health and cognitive ability are two important aspects of human functioning that are closely related and can influence each other. Mental health problems can affect cognitive abilities, while cognitive abilities can also play a role in the development and maintenance of mental health (5, 6).

Several studies have investigated the association between mental health symptoms and cognitive ability and have demonstrated a complex interplay between the two (4–7). Depression has been found to impact attention, memory, and executive function (8). Individuals with depression have been found to perform poorly on tasks requiring sustained attention and memory recall. Moreover, depression has also been found to impair cognitive functions, and impact working memory, which is a crucial component of executive function (9, 10). Additionally, anxiety has also been linked with decreased cognitive functioning, particularly in the areas of attention and memory. Several studies assessed the relationship between anxiety and cognitive ability (1, 11, 12). Findings show that children and youth with mental health symptoms were associated with low cognitive functions, such as children with social fears, who showed specific types of memory deficit, and children with social problems may have neurodevelopmental delays compared to other children (11). Further studies evidenced that children with greater anxiety symptoms are more likely to have difficulties in cognitive activities, such as problem-solving (11, 13). Furthermore, stress has also been found to have negative effects on cognitive functioning. Chronic stress has been linked with decreased performance in tasks involving memory and executive function (14–16). Particularly, stress has been found to impact working memory, which is a crucial component of executive function (17). Stress has also been found to impact attention and memory recall, particularly when the information being remembered is emotionally charged (15).

Children living in political violence and conflict environments were found to be subject to mental health problems that consequently affect their cognitive skills. Children who are exposed to political violence often experience high levels of stress and trauma, which can affect their ability to form healthy attachments and relationships (1, 2, 4). This, in turn, can negatively impact their cognitive and social development. Additionally, children who are exposed to political violence are more likely to experience disrupted sleep patterns, which can have negative impacts on their cognitive abilities, including attention and memory (18).

In recent years, machine learning (ML) has been increasingly used as a tool for identifying and understanding the associated factors affecting cognitive ability, such as mental health, sociodemographic characteristics, lifestyle, and political violence (19–23). Several studies have utilized ML algorithms, such as decision trees, random forests, and support vector machine algorithms, to identify the impact of political violence on children's mental health outcomes, such as PTSD and depression (24, 25). These studies have found that ML models can accurately identify children's mental health outcomes based on factors such as exposure to violence, trauma history, and demographic characteristics (26). Moreover, ML has been used in estimating mental health issues, and to explore the potential benefits and challenges associated with this approach, including the ability to analyze large amounts of data from multiple sources, improving the accuracy of identifications, and its cost-effectiveness (21, 26–28).

To the best of our knowledge, this is the first study that used different ML techniques to identify the associated risk factors with children's cognitive ability living in a conflict area. The study compared ML techniques and identified the most important factors that affect cognitive ability. Furthermore, the study produced a ML model that could be used in clinical and educational applications for improving cognitive ability and mental health identification among schoolchildren.

# 2. Materials and methods

## 2.1. Dataset

The dataset used in this study consists of a sample of 6,374 schoolchildren, and 31 associated features. The dataset consists of primary data extracted from the national Health Behavior in School-Aged Children (HBSC) study [derived from the international HBSC study (29)] conducted in the Palestinian territories by Al-Quds University and the Ministry of Education in the academic year 2013–2014. The data set was weighted and adjusted by gender (50% Boys, and 50% Girls) and grade, including students in grades 5th, 6th,7th, 8th, and 9th (19.9%, 20.2%, 21.1%, 19.3%, and 19.5%, respectively), within the ages of 10–15 years.

The data type is a mix of numerical and categorical variables. Numerical variables include the children's age, cognitive score, and academic performance, while categorical variables include sociodemographic, mental health, political violence, physical health, and lifestyle variables. To prepare the data for analysis, the interquartile range (IQR) method was used to identify the outliers, and manual inspection and data cleansing methods were used to identify the incorrect data entry values, while the missing values were input using the median imputation method, and irrelevant features were removed through a feature selection process, in which we used the correlation analysis, mutual information techniques to identify the most relevant and informative features.

The labels of the target variables indicated whether children had low, average, or above-average cognitive scores. The random under sampling technique was used to balance the class distribution with 3,187 samples for each class. The classes were balanced to have unbiased, more accurate and to ensure fairness in models' identification, The data was split into training sets for learning (70%), testing (20%), and validation (10%). Evaluation metrics used in this study include accuracy, F1 score, and receiver operating characteristic (ROC) curve analysis. These metrics were chosen to provide a comprehensive evaluation of the model's ability to correctly classify schoolchildren's cognitive abilities. The dataset is published on DANS EASY open access database: https://doi.org/10.17026/dans-zzt-guh7.

## 2.2. Study variables

The ML model features were listed in Table 1, including variable names, values, and levels of classification. The balanced cognitive score was used as the study target variable.

### 2.2.1. Sociodemographic variables

The sociodemographic variables describe the social and demographic characteristics of the study participants, including age, parents' education, family income, place of residence, school type, and Body Mass Index (BMI) [BMI = weight in kg/(height in $m^2$)].

### 2.2.2. Lifestyle

The lifestyle variables describe the children's behaviors and habits that might impact their overall health and well-being. In this study, physical activity, leisure time activity, smoking, sleeping, and food consumption were included. The physical activity and leisure time activities were measured by collecting data on the rate of physical exercises over 60 min and categorized according to the WHO definition [low (>3 days per week), Moderate (3–5 days per week), and High (6–7 days per week)], while the leisure time (screen time) activity

TABLE 1 List of machine learning model variables.

| Variable | Description |
|---|---|
| Cognitive_Ability | Below average, average, and above |
| Gender | Boys, Girls |
| Age (years) | 10–11, 12–13, 14–15 |
| Living_Place | Urban, Rural, Camp |
| Body Mass Index | Underweight, normal, overweight, and obese |
| Father_Education | ≤Secondary, >Secondary |
| Mother_Education | ≤Secondary, >Secondary |
| School_type | Public, UNRWA |
| Family_Income | Low, Moderate, High |
| Physical_Activity | Low, Moderate, High |
| Leisure_Time_Activity | Low, Moderate, High |
| Smoking_Tobacco | Yes, No |
| Healthy_Food_Consumption | Yes, No |
| Depression_Symptoms | Normal: 0–11, Depressed: ≥12 |
| Anxiety_Symptoms | (Low: 0–9, Moderate: 10–14, and High:15–21) |
| Mental_Health_Difficulties | Low: 0–14, Moderate: 15–17, and High:18–40 |
| Post-traumatic_stress disorder | 0 = No PTSD, 1 = Moderate PTSD, and 3 = Severe |
| PsychosomaticSymptom | Yes, No |
| Exposure_Political Violence | No Exposure, Moderate, and Severe |
| Child_Maltreatment | Never, Some and Severe |
| Sleeping_Hours | ≥8 h per day, <8 h per day |
| Family_Support | Low, Moderate, High |
| Peer_Support | Low, Moderate, High |
| School_Support | Low, Moderate, High |
| Positive_Health_Perception | Positive, Negative |
| Life_Satisfaction | Satisfied, Unsatisfied |
| Facing_School_Violence | Low, Moderate, High |
| Child_Abuse | Low, Moderate, High |
| Bullying | Never, Mild, High |
| Academic_Performance | Low, Moderate, High |
| Suicide_Attempt | Yes, No |

was categorized into [low (>2 h per day), moderate (2–3 h per day), and high (≥4 h per day)] (30). The smoking variable was categorized into "smoker and non-smoker," the sleeping variable was categorized into whether subject sleeps on average less than or equal to 8 h/day, or more than 8 h/day (≥8 h/day, or <8 h/day), and the food consumption was categorized into "healthy" based on the consumption rate of vegetables, fruits, milk or yogurt, and dairy products, or "unhealthy" based on the rate of consumption of soft drinks, energy drink, sweets, and sugar, both on a weekly basis.

### 2.2.3. Mental health symptoms

The mental health variables include depression, anxiety, stress, psychosomatic, and posttraumatic stress disorder (PTSD). The depression symptoms were measured by the 18-item Birleson Depression Self-Rating Scale for Children (DSRS), which was calculated by summing the items' answers and categorizing them into two categories (Normal: 0–11, Depressed: ≥12) (31). The anxiety levels were measured using the General Anxiety Disorder-7 (GAD-7) scale, which was calculated by summing the items' answers and categorizing them into three categories (Low: 0–9, Moderate: 10–14, and High:15–21) (32). Overall emotional and behavioral problems among children were measured using the Strengths, and Difficulties Questionnaire (SDQ) scale, which was categorized into three groups (Low: 0–14, Moderate: 15–17, and High:18–40) (33).

The psychosomatic symptoms were assessed using an 8-item psychosomatic symptoms scale (Cronbach's alpha = 0.85), in which children were asked if they experience the following symptoms at least once a week: headache, stomachache, backache, or dizziness. They were further asked if they experienced the following symptoms at least once a day: feeling depressed, irritability or bad temper, feeling nervous, difficulties in getting to sleep, and/or feeling dizzy. Participants answered on a scale from 1 (every day) to 5 (rarely or never). The answers were grouped into a continuous variable, which was categorized into a dichotomous variable [1: occurrence of the symptom (every day or more than once a week), and 0: no symptoms (Once a week, once a month, or never)] (34).

Posttraumatic Stress Disorder (PTSD) was measured by the 20 items index scale used by the HBSC survey, which determines the level of posttraumatic stress severity among children (34). The scale is composed of a 5-point scale from 0- "not at all" to "very much." The PTSD level was measured by categorizing the total score into three groups: 0 = No PTSD, 1 = Moderate PTSD, and 3 = Severe PTSD (34).

### 2.2.4. Political violence

Children's exposure to political violence was measured using the political violence inventory scale regarding children's exposure to military violence designed by Haj-Yahia et al. (35). The scale is composed of 40 statements that measure three levels of exposure: (1) very severe exposure (Personal or family member injured or hurt by military incursion), (2) moderate exposure (present at military incursion or seeing someone hurt or injured by military attack), and (3) no exposure (no direct contact with military incursion) (36).

### 2.2.5. Maltreatment

Child maltreatment measures any act or series of acts of bad treatment by parents or family members that results in harming the children. The scale is composed of 8 items (Cronbach's alpha = 0.87) that measure physical abuse, sexual abuse, neglect, and exposure to domestic violence. Participants responded on a scale of (Very True, True, or Not True), the scale was categorized into Never: (21–24 that includes participants who responded to not true to most or all of the items), Moderate: (16–20 that includes participants who responded true to all or most of the items) and Severe score: (8–15 that includes participants who responded to very true to all or most of the items) (18).

### 2.2.6. Social support

Social support measures the relationship and the help that children receive from their parents, friends, and school. Social support is measured by three subscales, each subscale is composed of a list of items that include (1) "family help," "emotional help," "ability to talk," and "help in making decisions," (2) "Friends try to help," "can count on friends," "having friends to share the joy with," and "can talk to friends about problems," and (3) "Teacher accepts me," "teacher cares about me," and "feel trust in teacher." These items were summed and categorized into three groups: Low (0–12), Moderate (13–16), and High levels of support (17–24) (34).

### 2.2.7. Positive health perceptions

The Positive Health Perception Scale (PHPS) to assess an individual's perception of their own health. It was designed to measure positive health perceptions, including attitudes, beliefs, and values related to health. The students were asked to rate their agreement with each statement on a Likert scale, that ranged from strongly disagree to strongly agree. The scale total score was classified into two groups: The scores of 35 or above was used to indicate a positive perception of health, while a score below 35 was considered a negative perception of health (34).

### 2.2.8. Life satisfaction

The Cantril Ladder satisfaction scale was used to measure children's life satisfaction. The scale ranged from 0 to 10, where 10 is the best possible life and 0 is the worst possible life (37). The scale was classified according to Mazur et al. in which the scale was classified into: Low (0–6), average (7–8), and high (9–10). While in this study, the ML model was designed to focus on the low level of satisfactions, so we regrouped the responses into: Unsatisfied (0–6) and Satisfied (7–10) (38).

### 2.2.9. School violence

Violence was measured by asking children if they were involved in physical fights; carrying weapons such as solid objects, knives, or other objects; how many times they were injured and treated by physical fights; or if they were involved in bullying other students. The scale is divided into two groups: Low (0–2); and High level (3–4). Higher scores indicated higher levels of violence.

### 2.2.10. Academic performance

The student's academic performance was measured based on the students' average grades score; the grades were collected from the school grading system. The Grade Point Average (GPA) score was used for classifying the total grades into three groups: Low: ≤59; Moderate 60–79; and High ≥80.

### 2.2.11. Cognitive abilities

Students' cognitive ability scores were assessed through the Cognitive Abilities Test (CogAT) is a standardized test used to measure cognitive abilities in students from kindergarten through grade 12 (or grade 13 in some regions). The test assesses students' abilities in three areas: verbal, quantitative, and nonverbal reasoning. The verbal reasoning section of the CogAT assesses a student's ability to use and understand language, including the ability to detect relationships between words, to recognize synonyms and antonyms, and to understand figurative language. The quantitative reasoning section assesses a student's ability to reason with numbers and to solve mathematical problems, including arithmetic, algebra, and geometry. The nonverbal reasoning section assesses a student's ability to reason with shapes and images, including the ability to recognize patterns, to complete sequences, and to understand spatial relationships. Each feature had a detailed content item, such as vocabulary, series, analogies, and inference. Overall, 181 content items were used for assessing the students' intelligence abilities. The total score was estimated from the detailed scores with an average of 60.7 ± 16.7 points (39). The cognitive scores were further classified into two categories: (1) below average, and (2) average and above average, for enhancing the performance of the ML algorithms.

### 2.2.12. The suicidal ideation and behavior

The HBSC survey designed a scale of 4 items for measuring suicide ideation among school children. The scale measures the severity level of suicidal ideation and behavior. In this study, only the question related to serious thoughts of attempting suicide was considered. The variable is composed of two categories: Yes or No (34).

## 2.3. Machine learning models

The ML models include Gradient Boosting (GB), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), k-nearest neighbors (k-NN), and Decision Tree (DT) algorithms, these models were built and compared based on their performance measures. The performance of the models was evaluated using a variety of metrics, including accuracy, precision, recall, and F1-score. The models' features were structured based on the target variable (cognitive ability), and associated factors that include the list of variables as indicated in Table 1. The selected model was trained on a 70% random sample of the data, and the remaining 30% was used for model testing and validation. The parameter optimization was performed using the grid search method and 10-fold cross-validation approach for the used models. The optimal parameters for each model were selected as follows:

1. The ANN model had a hidden layer with 1,000 neurons, a regularization parameter of 0.0001, and a maximum of 600 iterations using the logistic activation function.
2. The Random Forest model had 1,000 trees with a maximum depth of 5, a minimum number of samples at each leaf node set to 1, and a maximum number of samples to split internal nodes set to 2.
3. The SVM model had a regularization parameter of 20, a Radial Basis Function (RBF) kernel with a value of 0.001, and a bias error control factor set to 1.
4. The Gradient Boosting model had 1,000 trees with a learning rate of 0.1 and a maximum depth of 3 for individual trees.
5. The KNN model used 10 nearest neighbors, a uniform weighting function, and the Euclidean distance metric.
6. The Decision Tree model had a maximum tree depth of 100, the number of instances in leaves set to 2, and the smallest subsets set to 5.

Based on the optimized parameters, the algorithms were used to identify cognitive abilities.

## 2.4. Data analysis

Three approaches of data analysis were used to identify the association between cognitive ability and the associated risk factors. Statistical analysis, machine learning analysis and Gini importance analysis.

### 2.4.1. Statistical analysis

To summarize the demographic characteristics of the study population, we conducted descriptive statistics, which involve the use of summary statistics to describe the central tendency, variability, and distribution of the data. This analysis provided an overview of the characteristics of the study population, including age, gender, and socioeconomic status.

To test the relationship between the study variables and cognitive ability, we used inferential statistics, which involve the use of statistical tests to determine the significance of the relationships between variables. Specifically, we used the binary regression analysis to explore the relationship between variables, including the odd ratio, which is a measure of the strength of the association between the cognitive ability and the independent variables. Additionally, we used analysis of variance (ANOVA) to compare the means of different groups, including the calculation of the *F*-value, which is a measure of the overall significance of the model. Furthermore, the Univariate analysis was also conducted to assess the distribution and relationship of cognitive ability with other variables.

### 2.4.2. Machine learning analysis

Data preprocessing techniques were conducted prior to the implementation of the ML models, including cleaning, transformation, and normalization processes. The final dataset consisted of 6,374 participants. The six ML models were built and performed using the Python orange data mining software (40), which was used for testing and validating the ML models. The study employed a 10-fold cross-validation approach to evaluate the performance of the machine learning models.

The evaluation of ML models for identifying cognitive ability levels in students and associated risk factors involves assessing the effectiveness and reliability of the models using various performance measures. Some commonly used performance measures include balanced accuracy, specificity, precision, recall, and F-measure (the harmonic mean of precision and recall), in addition to the area under the receiver operating characteristic curve (AUC-ROC), was used to evaluate the performance of binary classifiers. The Wilcoxon signed-rank test is a non-parametric statistical test used to compare the performance of two models on a given data set. In this study, we used the Wilcoxon signed-rank test to determine whether there was a

TABLE 2 The statistical analysis of Childrens' cognitive ability scores by sociodemographic variables.

| Variable | Feature | Cognitive ability score | | *F* (*p*-value) | OR (95% CI) |
|---|---|---|---|---|---|
| | | Low | Average and above | | |
| | | *n* (%) | | | |
| Gender | Boys | 1,238 (38.8) | 1,949 (61.2) | 317.3 (0.001) | 2.03 (1.76–2.33) |
| | Girls | 591 (18.5) | 2,596 (81.5) | | |
| | Total | 1,829 (28.7) | 4,545 (71.3) | | |
| Children age | 10–11 | 642 (29) | 1,571 (71) | 2.22 (0.108) | 0.89 (0.82–0.97) |
| | 12–13 | 677 (28) | 1,744 (72) | | |
| | 14–15 | 510 (29.3) | 1,230 (70.7) | | |
| Place of residence | Urban | 909 (31.7) | 1,959 (68.3) | 20.7 (0.001) | 0.87 (0.8–0.95) |
| | Rural | 494 (24.6) | 1,511 (75.4) | | |
| | Camp | 426 (28.4) | 1,075 (71.6) | | |
| Father education | ≤Secondary | 573 (23.9) | 1,826 (76.1) | 5.9 (0.015) | 0.85 (0.73–0.98) |
| | >Secondary | 1,256 (31.6) | 2,719 (68.4) | | |
| Mother education | ≤Secondary | 488 (21.4) | 1,795 (78.6) | 25.7 (0.001) | 0.81 (0.69–0.94) |
| | >Secondary | 1,341 (32.8) | 2,750 (67.2) | | |
| Physical activity | Low | 382 (26.5) | 1,058 (73.5) | 34.1 (0.001) | 1.13 (1.05–1.23) |
| | Moderate | 643 (35.9) | 1,147 (64.1) | | |
| | High | 804 (25.6) | 2,340 (74.4) | | |
| Leisure time activity | Low | 478 (30) | 1,116 (70) | 8.2 (0.001) | 1.17 (1.08–1.26) |
| | Moderate | 564 (32.2) | 1,186 (67.8) | | |
| | High | 787 (26) | 2,243 (74) | | |
| Family income | Low | 802 (29) | 1,960 (71) | 1.4 (0.241) | 1 (0.92–1.09) |
| | Moderate | 688 (29.1) | 1,673 (70.9) | | |
| | High | 339 (27.1) | 912 (72.9) | | |
| School type | GOV[1] | 1,200 (33.1) | 2,425 (66.9) | 145.6 (0.001) | 1.88 (1.62–2.18) |
| | UNRWA[2] | 629 (22.9) | 2,120 (77.1) | | |
| Body mass index (BMI) | Underweight | 84 (25.6) | 244 (74.4) | 1.1 (0.348) | 1.2 (0.83–1.7) |
| | Normal | 1,528 (29.6) | 3,641 (70.4) | | |
| | Overweight | 138 (23.6) | 446 (76.4) | | |
| | Obese | 79 (27) | 214 (73) | | |

[1]GOV, government.

[2]UNRWA, United Nations Relief and Works Agency; OR, Odd ratio; F, Fisher's exact test; (95% CI), 95% Confidence Interval.

significant difference in performance between two machine learning models and helped us to identify the model that performed better on the given dataset.

### 2.4.3. Gini importance analysis

We utilized Gini importance analysis to identify the most important risk factors that contribute to low cognitive ability scores among the study population. Gini importance analysis involved calculating the Gini importance coefficient for each potential risk factor, which allowed us to determine the relative importance of each factor in explaining the variation in cognitive ability scores. We used Python Anaconda software to conduct the analysis and generate the results.

### 2.4.4. Classification and regression trees

In this study, the Classification and Regression Trees (CRT) technique was utilized as a ML approach to identify the patterns of associations between cognitive ability and study variables. The CRT method is a decision tree-based technique that enables the identification of complex nonlinear relationships between predictor variables and an outcome variable. It uses a recursive partitioning algorithm to split the data set into increasingly homogeneous subsets based on the predictor variables' values. The CRT technique produced a decision tree that provided a visual representation of the complex relationships between cognitive ability and other predictor variables in the study and helped us to better understand the factors that contribute to cognitive ability levels.

# 3. Results

## 3.1. Descriptive analysis

The results in Table 2 showed the descriptive univariant analysis of children's cognitive ability with sociodemographic variables. The data set was balanced and weighted by gender and age, with equal representation of both boys and girls across age groups. Of the participants, 28.7% reported low cognitive ability scores, with a higher percentage reported by boys (38.8%) compared to girls (18.5%). The prevalence of low cognitive scores varied by age, with an average of 29% across the three age groups (10–11, 12–13, 14–15). In terms of place of residence, urban and camp residents had a higher percentage

TABLE 3 The statistical analysis of Childrens' cognitive ability scores by mental health and political violence factors.

| Variable | Feature | Cognitive ability scores | | *F* (*p*-value) | OR (95% CI) |
|---|---|---|---|---|---|
| | | Low | Average and above | | |
| | | *n* (%) | | | |
| Exposure to political violence | No | 860 (46.7) | 980 (53.3) | 19.1 (0.001) | 1.39 (1.2–1.63) |
| | Moderate | 516 (21.4) | 1,899 (78.6) | | |
| | Severe | 453 (21.4) | 1,666 (78.6) | | |
| PTSD | Low | 316 (16.3) | 1,618 (83.7) | 7.7 (0.001) | 0.83 (0.75–0.92) |
| | Moderate | 619 (25.8) | 1,776 (74.2) | | |
| | High | 894 (43.7) | 1,151 (56.3) | | |
| Depression scale | Low | 668 (28.2) | 1,705 (71.8) | 4.6 (0.010) | 1.19 (1.06–1.34) |
| | Moderate | 1,048 (28.8) | 2,588 (71.2) | | |
| | Severe | 113 (31) | 252 (69) | | |
| Behaviours strengths and difficulties | Normal | 797 (33) | 1,620 (67) | 7.6 (0.001) | 1.14 (1.06–1.23) |
| | Mild | 538 (25.1) | 1,608 (74.9) | | |
| | Abnormal | 494 (27.3) | 1,317 (72.7) | | |
| Child maltreatment | No | 692 (19.2) | 2,910 (80.8) | 40.1 (0.001) | 0.77 (0.7–0.84) |
| | Moderate | 297 (25.7) | 859 (74.3) | | |
| | High | 840 (52) | 776 (48) | | |
| Positive health | Negative | 436 (18.3) | 1,946 (81.7) | 19.5 (0.001) | 0.69 (0.6–0.81) |
| | Positive | 1,393 (34.9) | 2,599 (65.1) | | |
| Life satisfaction | Satisfied | 1,186 (30.5) | 2,703 (69.5) | 15.9 (0.001) | 2.08 (1.91–2.26) |
| | Unsatisfied | 643 (25.9) | 1,842 (74.1) | | |
| Academic performance | Below average | 1,338 (42) | 1,847 (58) | 169.6 (0.001) | 2.1 (1.9–2.3) |
| | Average | 295 (19.4) | 1,224 (80.6) | | |
| | Above average | 187 (11.6) | 1,431 (88.4) | | |
| Healthy food consumption | No | 573 (22) | 2,032 (78) | 18.9 (0.001) | 0.65 (0.56–0.76) |
| | Yes | 1,256 (33.3) | 2,513 (66.7) | | |
| Tobacco risk | No | 1,226 (23.9) | 3,895 (76.1) | 28.9 (0.001) | 1.24 (1.13–1.35) |
| | Yes | 603 (48.1) | 650 (51.9) | | |
| Parental support | Low | 641 (32.3) | 1,346 (67.7) | 12.3 (0.001) | 1.03 (0.94–1.13) |
| | Moderate | 855 (31.5) | 1,861 (68.5) | | |
| | High | 333 (19.9) | 1,338 (80.1) | | |
| School violence | Never | 512 (23) | 1,710 (77) | 14.2 (0.001) | 0.91 (0.94–1 0.13) |
| | 1–3 times | 684 (25) | 2,048 (75) | | |
| | 4+ times | 633 (44.6) | 787 (55.4) | | |
| Seriously thought of attempting suicide | Yes | 572 (44) | 729 (56) | 191.9 (0.001) | 0.92 (0.78–1.09) |
| | No | 1,257 (24.8) | 3,816 (75.2) | | |

OR, Odd ratio; F, Fisher's exact test; (95% CI), 95% Confidence Interval.

of low cognitive scores (31.7 and 28.4%, respectively). Additionally, participants with parents who had lower than a secondary school education had a higher percentage of low cognitive scores (31.6% for father's education and 32.8% for mother's education). Moreover, public school students had a higher percentage of low cognitive scores compared to UNRWA schools (33.1 and 22.9%, respectively). These results provide important insights into the sociodemographic factors associated with low cognitive ability scores among schoolchildren living in politically violent environments.

Univariate analysis showed significant associations between cognitive ability and gender, place of residence, father's education, mother's education, physical activity, leisure time activity, school type, exposure to political violence, PTSD, depression, SDQ, maltreatment, positive health perception, academic performance, healthy food consumption, tobacco risk, parental support, school violence, and suicidal ideation (*F* values ranging from 4.6 to 317.3, all $ps < 0.05$ except for anxiety, friend support, school support, family income, physical activity, and life satisfaction).

The results in Table 2 showed that there were no significant associations with age, family income, and BMI. The analysis by academic performance showed that participants with below-average academic scores had reported a lower cognitive ability (42%). Moreover, participants who did not consume healthy food on a regular basis, smoke, have low levels of parental support, were exposed to school violence ≥4 times, and thought of attempting suicide reported a higher percentage of low cognitive scores (33.3, 48.1, 32.3, 44.6, and 44%, respectively).

Results in Table 3 showed that 71.1% of participants had moderate or severe exposure to political violence, of which 42.8% had low cognitive ability scores. Participants with moderate or severe PTSD

TABLE 4 10-Folds cross validation performance measures analysis of the different ML models.

| Model | AUC[1] | CA[2] | F1[3] | Precision | Recall | Execution time (s) |
|---|---|---|---|---|---|---|
| RF | 0.91 | 0.87 | 0.86 | 0.88 | 0.87 | 0.62 |
| ANN | 0.85 | 0.84 | 0.84 | 0.84 | 0.84 | 20.0 |
| SVM | 0.84 | 0.84 | 0.83 | 0.84 | 0.84 | 10.2 |
| GB | 0.85 | 0.83 | 0.82 | 0.82 | 0.83 | 8.1 |
| Decision tree | 0.77 | 0.80 | 0.81 | 0.81 | 0.80 | 0.03 |
| KNN | 0.78 | 0.77 | 0.76 | 0.76 | 0.77 | 0.001 |

[1]AUC, area under the curve.
[2]CA, correspondence analysis.
[3]F1-score, a harmonic mean between precision and recall.

TABLE 5 Model evaluation comparison through Wilcoxon signed rank test.

| Pair-wise comparison | *Z*-value | *p*-value |
|---|---|---|
| RF-GB | −2.5 | 0.027 |
| RF-SVM | −2.8 | 0.002 |
| RF-KNN | −2.8 | 0.002 |
| RF-DT | −2.8 | 0.002 |
| RF-ANN | −2.14 | 0.16 |
| GB-ANN | −2.75 | 0.004 |
| GB-KNN | −2.8 | 0.002 |
| GB-SVM | −2.8 | 0.002 |
| GB-DT | −2.75 | 0.004 |
| SVM-KNN | −2.8 | 0.002 |
| SVM-DT | −2.8 | 0.002 |
| SVM-ANN | −1.43 | 0.87 |
| KNN-DT | −2.8 | 0.002 |
| KNN-ANN | −2.8 | 0.002 |
| DT-ANN | −2.8 | 0.002 |

had a higher rate of low cognitive ability (43.7%) than other participants (25.8%). Participants with severe depression reported a higher percentage of low cognitive scores compared to moderate and low depression scores (31, 28.8, and 28.2%, respectively). The measurement of emotional and behavioral problems among children indicated that the participants with no emotional or behavioral problems reported a higher rate of low cognitive scores than other groups (33%). The study further assessed the effect of family maltreatment on children's cognitive ability, and findings indicated that a high level of maltreatment was associated with a high level of low cognitive scores (52%). Furthermore, the results of the logistic regression analysis showed that academic performance, gender, satisfaction, school type, and exposure to political violence were the most five important factors affecting cognitive abilities among school children.

### 3.1.1. Machine learning performance analysis

Several ML models were used to assess the performance of ML techniques in identifying cognitive ability from the associated factors. Results in Table 4 showed the ML models' performance analysis AUC, Balanced accuracy, F1, recall, and precision. The results indicated that the RF had the highest performance of balanced accuracy (87%), followed by NN and SVM. While the lowest balanced accuracy rate was found in the KNN algorithm. In terms of execution time, KNN, Decision Tree and RF reported the lowest execution time (0.001 s, 0.03 s, and 0.62 s). However, all ML models used in our analysis showed an accuracy rate (F1-score) above 75% in identifying cognitive ability. It was determined that the RF algorithm's predictive power differed significantly from that of the other models (Table 5).

The CRT analysis results are illustrated in Figure 1 showed that cognitive ability was highly affected by children's maltreatment, in which the high level of maltreatment had a high percentage of low cognitive scores (52%). Participants who reported experiencing high levels of maltreatment exhibited lower academic performance, which subsequently decreased their cognitive ability. This was evidenced by the fact that 60.1% of participants who had below-average academic scores reported low cognitive ability. Furthermore, the below-average academic performance group was associated with school type and exposure to political violence, in which the public schools had lower cognitive scores, and were more likely to be affected by exposure to political violence. The exposed students in governmental schools were further classified by level of PTSD.

On the other hand, the UNRWA schools were classified by age and positive health perception. Interestingly, the 12–13 and 14–15 age groups were merged into one cluster and reported a high level of low cognitive score (52.6%). Moreover, the negative health perception group reported a higher rate of low cognitive ability (59.4% compared to 26.9% of positive perception).

The other side of the classification tree identified different patterns of association, the maltreatment (never, and moderate) groups were affected by academic performance, whereby the students with below-average academic performance had a higher rate of low cognitive ability (32.5%). Furthermore, the same group was affected by gender, whereby boys reported a higher rate of low cognitive scores than girls. Interestingly, boys were affected by the exposure to political violence.

Gini Importance analysis was conducted to identify the factors that have the most impact on the likelihood of developing low cognitive ability among schoolchildren. The results of the Gini Importance analysis are illustrated in Figure 2. The findings indicate the relative importance of each risk factor, with the highest-scoring factors being the most significant in identifying cognitive ability. Maltreatment, exposure to political violence, PTSD, academic performance, smoking, suicide attempts, school type, and gender were the most important factors affecting cognitive development among children.

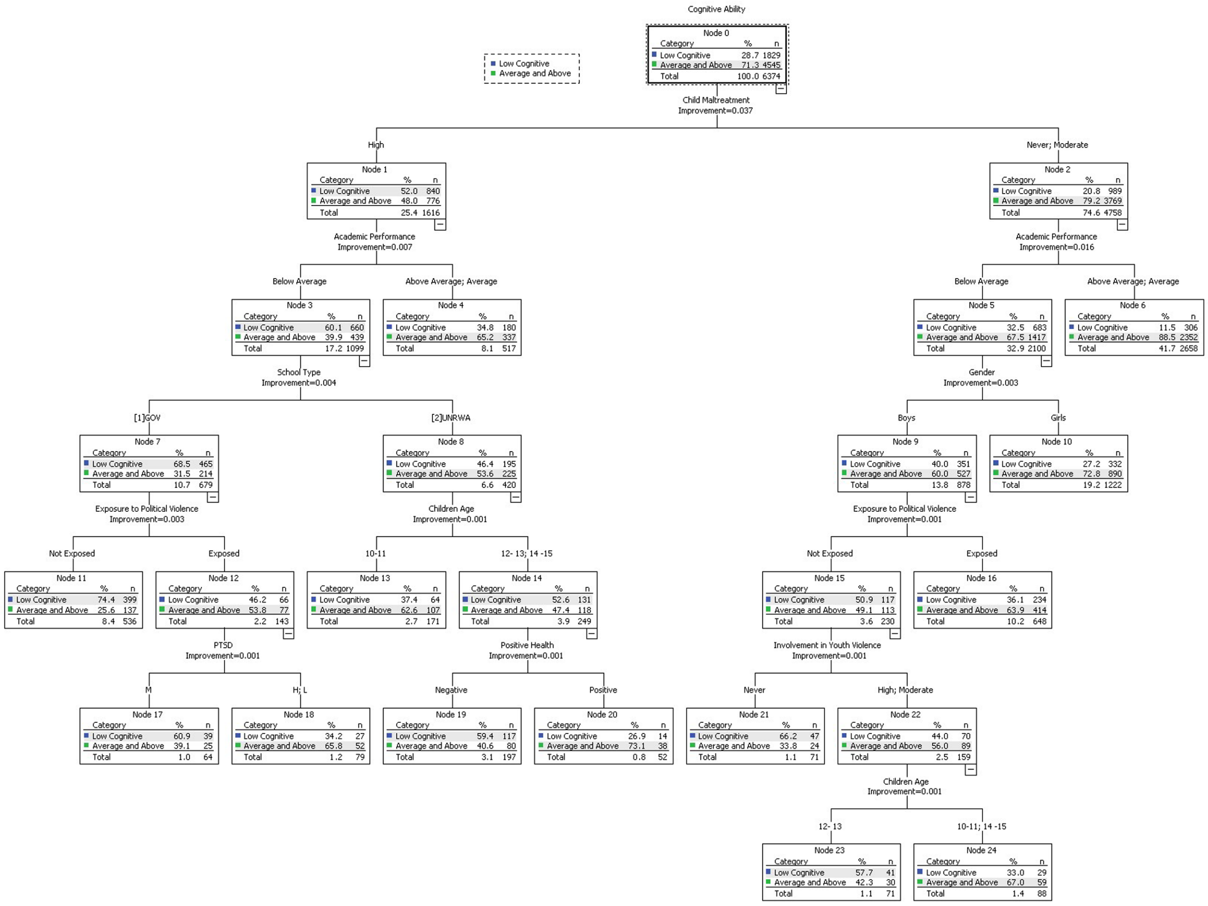


FIGURE 1
Cognitive ability and the associated factors classification tree.

## 4. Discussion

This study utilized a novel approach using machine learning to gain a deeper understanding of the complex relationships between important determinants of cognitive functions (1, 2, 5, 35) that can significantly impact children's emotional and psychological well-being, as well as their cognitive abilities such as memory, attention, and problem-solving skills (34). The study also examined the impact of severe stress caused by living in a violent environment, which can lead to physiological changes in the brain and body that affect cognitive functioning even after the violence has subsided (3, 41). By using machine learning, the study was able to identify new patterns of associations that traditional statistical models may have missed.

The study evidenced a significant association between cognitive ability scores and gender, place of residence, father's education, mother's education, physical activity, leisure time activity, parents' support, and school type. The findings indicated that the prevalence of low cognitive ability among boys is higher than among girls. Our result is consistent with other research studies that showed girls had better cognitive skills and academic achievement (42, 43). Children living in urban places where more exposed political violence are at higher risk of developing ill mental health than children living in rural areas, which might negatively affect their cognitive development. These findings are consistent with other studies that evidenced the significance of place of residence with cognitive development (13, 14, 34).

Furthermore, the findings indicate that exposure to political violence, PTSD, depression, Behaviours Strengths, and Difficulties scale (emotional symptoms, conduct problems, hyperactivity, peer relationship problems, and prosocial behavior), positive health perception, and serious thoughts of attempting suicide were highly significant with cognitive ability scores. The findings indicated that this effect might be higher among children with low socio-economic status. Our results are consistent with other studies that investigated the effect of ongoing political violence on children's mental health and cognitive development, in which a strong association between political violence and low academic achievement were found (1, 2, 41). In terms of cognitive ability, research has shown that exposure to political violence can lead to a decline in IQ scores, memory loss, and decreased attention span. This is likely due to the high levels of stress and traumatic experiences associated with such violence, which can disrupt normal brain functioning and lead to long-lasting changes in the brain structure and function (4, 36).

In our study, the CRT classification model revealed different patterns of associations among participants. The study showed that maltreatment was the most important factor related to cognitive development. The CRT model classified maltreatment into two clusters: (1) High levels of maltreatment, and (2) Never or moderate maltreatment. The first cluster (high levels of maltreatment) showed associations with academic performance, school type, and exposure to political violence. This is evidenced by the fact that children with

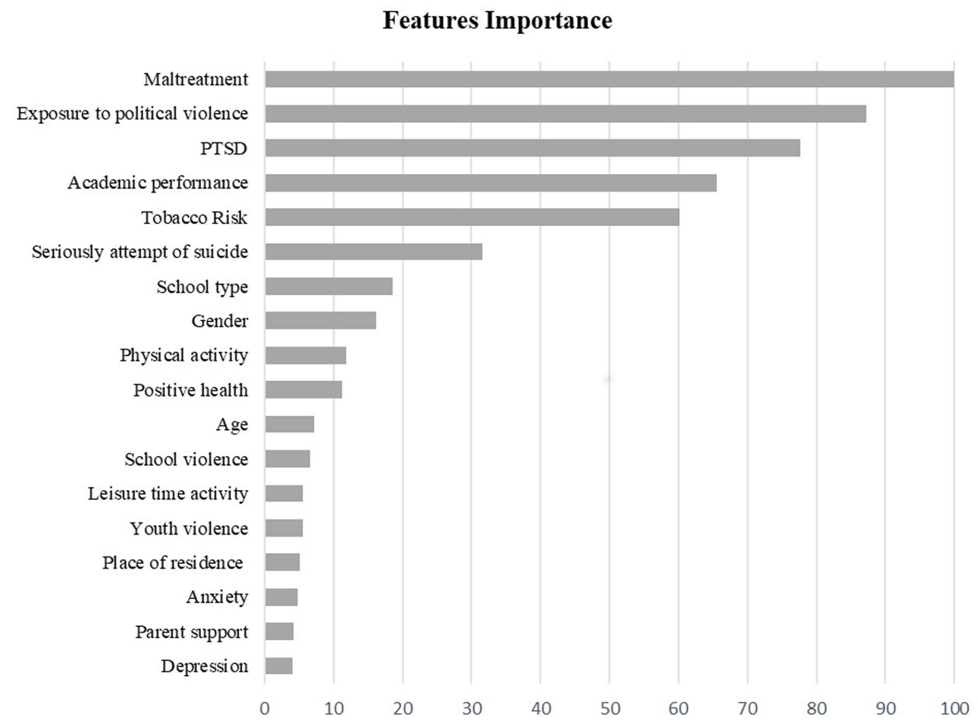


FIGURE 2
Cognitive ability associated risk factors importance ranking.

lower academic performance, as well as public-school students had higher levels of association to political violence. In turn, children with higher levels of exposure to political violence showed higher levels of PTSD, which further negatively affected cognitive ability.

The second cluster (moderate or no levels of maltreatment) evidenced a direct association with higher academic performance and gender. Boys' academic performance is conversely associated to their extent of exposure to political violence, and youth violence, whereby lower exposure to violence among boys is associated with higher academic scores. This can be due to the direct effects of traumatic experiences, exposure to loss and harm, and disruptions to social and personal relationships, especially among boys. These findings are consistent with other studies that indicated a strong association between child abuse, exposure to political violence, and posttraumatic events (2, 5, 35, 36).

The study evidenced that children who experience mental health problems, such as depression and anxiety, may have difficulties with attention and memory, which can affect their academic performance. Furthermore, the study emphasized the fact that the negative impacts of political violence on children's mental health and cognitive ability can have serious effects, including difficulties with learning, emotional regulation, and overall well-being.

Upon comparing the results of the ML model with those of the logistic regression model, it was found that the ML was able to detect novel patterns of associations between cognitive ability and risk factors. Specifically, the ML model identified maltreatment, PTSD, smoking, and suicide attempts as important factors that might affect the cognitive ability of school children. On the other hand, these factors were not identified as significant in the logistic regression model. The advantage of using the ML model is that it can identify non-linear relationships and interactions between predictor variables that may be missed by traditional statistical models such as logistic regression.

In particular, the ML model identified maltreatment as a significant risk factor for decreased cognitive ability, which is consistent with previous research in this area. However, the model also identified PTSD, smoking, and suicide attempts as additional risk factors that have not been widely studied in the context of cognitive ability in school children. The ability of the ML model to identify novel risk factors underscores the potential usefulness of this approach in identifying previously unknown or overlooked factors that impact cognitive ability. These findings suggest that the ML model may be a valuable tool for future research in this area, as well as for identifying interventions and treatments that may help to mitigate the negative impact of these risk factors on cognitive ability in school children.

The study showed the power of ML tools and algorithms in understanding and addressing cognitive development in a holistic and comprehensive approach that considers the complex interplay between mental health, cognitive ability, and exposure to political violence. Moreover, the utilization of advanced algorithms and tools, researchers can more accurately identify patterns and trends in complex datasets, which can help them to better understand the relationship between mental health and cognitive ability in this vulnerable population. It is important to note that the use of ML in identifying mental health and cognitive ability is still in its early stages, and there is a need for further research to validate and improve the balanced accuracy of these models.

## 4.1. Strengths and limitations

The relationship between cognitive development and mental health has been investigated by several research studies evidencing a strong interrelation between the two factors. To the best of our knowledge, the current study is considered the first of its kind that deploys ML techniques in assessing the relationship between socio-economic, mental health social factors, lifestyle and exposure to political violence, and cognitive ability among children living within an ongoing politically violent environment. The use of ML provides an in-depth understanding of the nature of these associations and identified new patterns of associations. The studied ML models are less dependent on the linear relationship between risk factors, which could provide a more precise and accurate association.

The ML model was developed to detect novel patterns of associations and identified maltreatment, PTSD, smoking, and suicide attempts as important risk factors affecting the cognitive ability of school children. On the other hand, these factors were not identified as significant in the logistic regression model. Our results highlight the potential usefulness of ML in identifying non-linear relationships and interactions between predictor variables that may be missed by traditional statistical models such as logistic regression. This, in turn, will enhance the development of precise and efficient intervention programs that improve children's growth and cognitive skills development in Palestine and other politically violent environments. Thus, this research study not only introduces the methodology of ML techniques in identifying cognitive abilities but also provides decision-makers with the power of ML in the early diagnosis of schoolchildren's cognitive skills.

Nonetheless, the study is limited by several factors, such as the target region was selected from the Palestinian community only, the levels of political violence, and the participants' age group that does not include children under 12 years and adults >18 years. However, future research will benefit from this study by adding other risk factors. Additional studies can be conducted related to other external factors, such as forced displacement, house demolitions, poverty, family social problems, and mobility limitations. The integration of the above variables would provide an in-depth understanding of cognitive abilities development among schoolchildren in the Palestinian context. The presence of these variables would further enhance the accuracy of the ML models' identification for cognitive abilities.

## 5. Conclusion

The findings of this study offer important insights into the complex interplay between various risk factors and cognitive development in children in conflict zones. The use of ML techniques allowed for the identification of factors associated with cognitive ability and mental health risks in a novel and data-driven approach, highlighting significant risk factors such as exposure to political violence, maltreatment, severe mental health disorders, and below-average academic performance.

This study contributes to the growing literature on the effects of adverse childhood experiences and underscores the need for policymakers, practitioners, and researchers to address the negative impact of political violence on children's well-being. Additionally, the study's identification of gender differences in the relationship between academic performance and cognitive ability emphasizes the need for targeted interventions to support boys exposed to political violence.

Practically, this study's findings can inform evidence-based strategies for preventing and mitigating the detrimental effects of political violence on individuals and communities. Policymakers, practitioners, and researchers can utilize the insights gained from this study to design interventions aimed at supporting and treating children impacted by violence and promoting safer environments for their growth and development. Moreover, the performance of the developed ML model (precision and recall >85%) is encouraging and can guide future research in the development of a tool to support the identification of children at risk of having low cognitive abilities.

## Data availability statement

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

## Ethics statement

The studies involving human participants were reviewed and approved by Al-Quds University IRB. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

## Author contributions

RQ, ZA, SH, and DA: conceptualization and methodology. RQ: formal analysis, validation, and writing original draft preparation. DA and SV: editing and review. RQ, NA, and RA: data curation and data pre-processing. All authors contributed to the article and approved the submitted version.

## Funding

The work included in this study is part of the wider project Determinants of Cognitive Development in Deprived Environments: evidence from the West Bank funded by the German Research Foundation (DFG) under Grant Number JU 2769/2.

## Acknowledgments

The authors would like to express their sincere gratitude to all those who have contributed to the successful completion of this research. Our heartfelt thanks go out to our colleagues in the field who have offered their valuable time and expertise in providing insightful suggestions and feedback. The authors would also like to extend our appreciation to the participants of the study, who have generously given their time and effort towards making this research a success.

## 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.

## References

1. Moussa S, El KM, Enaba D, Salem K, Ali A, Nasreldin M, et al. Impact of political violence on the mental health of school children in Egypt. *J Ment Health*. (2015) 24:289–93. doi: 10.3109/09638237.2015.1019047

2. Agbaria N, Petzold S, Deckert A, Henschke N, Veronese G, Dambach P, et al. Prevalence of post-traumatic stress disorder among Palestinian children and adolescents exposed to political violence: a systematic review and meta-analysis. *PLoS One*. (2021) 16:e0256426–17. doi: 10.1371/journal.pone.0256426

3. Islam A, Raschky P, Smyth R. The long-term health effects of mass political violence: evidence from China's cultural revolution. *Soc Indic Res*. (2017) 132:257–72. doi: 10.1007/s11205-015-1030-6

4. Bosqui TJ, Marshoud B. Mechanisms of change for interventions aimed at improving the wellbeing, mental health and resilience of children and adolescents affected by war and armed conflict: a systematic review of reviews. *Confl Heal*. (2018) 12:15–7. doi: 10.1186/s13031-018-0153-1

5. Nyarko F, Peltonen K, Kangaslampi S, Punamäki RL. Emotional intelligence and cognitive skills protecting mental health from stress and violence among Ghanaian youth. *Heliyon*. (2020) 6:e03878. doi: 10.1016/j.heliyon.2020.e03878

6. Davis SK, Humphrey N. Emotional intelligence predicts adolescent mental health beyond personality and cognitive ability. *Pers Individ Dif*. (2012) 52:144–9. doi: 10.1016/j.paid.2011.09.016

7. Biddle SJH, Asare M. Physical activity and mental health in children and adolescents: a review of reviews. *Br J Sports Med*. (2011) 45:886–95. doi: 10.1136/bjsports-2011-090185

8. Scult MA, Paulli AR, Mazure ES, Moffitt TE, Hariri AR, Strauman TJ. The association between cognitive function and subsequent depression: a systematic review and meta-analysis. *Psychol Med*. (2017) 47:1–17. doi: 10.1017/S0033291716002075

9. Bora E, Harrison BJ, Yücel M, Pantelis C. Cognitive impairment in euthymic major depressive disorder: a meta-analysis. *Psychol Med*. (2013) 43:2017–26. doi: 10.1017/S0033291712002085

10. Rock PL, Roiser JP, Riedel WJ, Blackwell AD. Cognitive impairment in depression: a systematic review and meta-analysis. *Psychol Med*. (2014) 44:2029–40. doi: 10.1017/S0033291713002535

11. Kristensen H, Torgersen S. Is social anxiety disorder in childhood associated with developmental deficit/delay? *Eur Child Adolesc Psychiatry*. (2008) 17:99–107. doi: 10.1007/s00787-007-0642-z

12. Qasrawi R, Vicuna Polo SP, Abu Al-Halawa D, Hallaq S, Abdeen Z. Assessment and prediction of depression and anxiety risk factors in schoolchildren: machine learning techniques performance analysis. *JMIR Format Res*. (2022) 6:e32736. doi: 10.2196/32736

13. Broeren S, Muris P. The relation between cognitive development and anxiety phenomena in children. *J Child Fam Stud*. (2009) 18:702–9. doi: 10.1007/s10826-009-9276-8

14. Pesonen AK, Eriksson JG, Heinonen K, Kajantie E, Tuovinen S, Alastalo H, et al. Cognitive ability and decline after early life stress exposure. *Neurobiol Aging*. (2013) 34:1674–9. doi: 10.1016/j.neurobiolaging.2012.12.012

15. McEwen BS, Gianaros PJ. Stress-and allostasis-induced brain plasticity. *Annu Rev Med*. (2011) 62:431–45. doi: 10.1146/annurev-med-052209-100430

16. Zeidner M, Matthews G, Shemesh DO. Cognitive-social sources of wellbeing: differentiating the roles of coping style, social support and emotional intelligence. *J Happiness Stud*. (2016) 17:2481–501. doi: 10.1007/s10902-015-9703-z

17. Vogel S, Schwabe L. Learning and memory under stress: implications for the classroom. *NPJ Sci Learn*. (2016) 1:16011–0. doi: 10.1038/npjscilearn.2016.11

18. Coates DR, Chin JM. Chung STL. 基因的改变NIH public access. *Bone*. (2011) 23:1–7. doi: 10.1007/s10567-009-0041-8.Children

19. Wang S, Wang W, Li X, Liu Y, Wei J, Zheng J, et al. Using machine learning algorithms for predicting cognitive impairment and identifying modifiable factors among Chinese elderly people. *Front Aging Neurosci*. (2022) 14:1–12. doi: 10.3389/fnagi.2022.977034

20. Ansart M, Epelbaum S, Bassignana G, Bône A, Bottani S, Cattai T, et al. Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review. *Med Image Anal*. (2021) 67:101848. doi: 10.1016/j.media.2020.101848

21. Qasrawi R, Vicuna Polo S, Al-Halawa DA, Hallaq S, Abdeen Z. Predicting school children academic performance using machine learning techniques. *Adv Sci Technol Eng Syst J*. (2021) 6:8–15. doi: 10.25046/aj060502

22. Bowe AK, Lightbody G, Staines A, Murray DM. Big data, machine learning, and population health: predicting cognitive outcomes in childhood. *Pediatr Res*. (2022) 93:1–8. doi: 10.1038/s41390-022-02137-1

23. Park JH, Cho HE, Kim JH, Wall MM, Stern Y, Lim H, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *NPJ Digit Med*. (2020) 3:46. doi: 10.1038/s41746-020-0256-0

24. Wshah S, Skalka C, Price M. Predicting posttraumatic stress disorder risk: a machine learning approach. JMIR Ment. *Health*. (2019) 6:e13946. doi: 10.2196/13946

25. Ramos-Lima LF, Waikamp V, Antonelli-Salgado T, Passos IC, Freitas LHM. The use of machine learning techniques in trauma-related disorders: a systematic review. *J Psychiatr Res*. (2020) 121:159–72. doi: 10.1016/j.jpsychires.2019.12.001

26. Chung J, Teo J. Mental health prediction using machine learning: taxonomy, applications, and challenges. *Appl Comput Intell Soft Comput*. (2022) 2022:1–19. doi: 10.1155/2022/9970363

27. Poudel GR, Barnett A, Akram M, Martino E, Knibbs LD, Anstey KJ, et al. Machine learning for prediction of cognitive health in adults using sociodemographic, Neighbourhood environmental, and lifestyle factors. *Int J Environ Res Public Health*. (2022) 19:10977. doi: 10.3390/ijerph191710977

28. Tate AE, McCabe RC, Larsson H, Lundström S, Lichtenstein P, Kuja-Halkola R. Predicting mental health problems in adolescence using machine learning techniques. *PLoS One*. (2020) 15:e0230389–13. doi: 10.1371/journal.pone.0230389

29. Currie C, Inchley J, Molcho M, Lenzi M, Veselska Z, Wild F. *Health behaviour in school-aged children (HBSC) study protocol: Background, methodology and mandatory items for the 2013/14 survey*. (2014).

30. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *Br J Sports Med*. (2020) 54:1451–62. doi: 10.1136/bjsports-2020-102955

31. Birleson P. The validity of depressive disorder in childhood and the development of a self-rating scale: a research report. *J Child Psychol Psychiatry*. (1981) 22:73–88. doi: 10.1111/j.1469-7610.1981.tb00533.x

32. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*. (2006) 166:1092–7. doi: 10.1001/archinte.166.10.1092

33. Goodman R. The strengths and difficulties questionnaire: a research note. *J Child Psychol Psychiatry*. (1997) 38:581–6. doi: 10.1111/j.1469-7610.1997.tb01545.x

34. Harel-Fisch Y, Radwan Q, Walsh SD, Laufer A, Amitai G, Fogel-Grinvald H, et al. Psychosocial outcomes related to subjective threat from armed conflict events (STACE): findings from the Israeli-Palestinian cross-cultural HBSC study. *Child Abuse Negl*. (2010) 34:623–38. doi: 10.1016/j.chiabu.2009.12.007

35. Haj-Yahia MM. Political violence in retrospect: Its effect on the mental health of Palestinian adolescents. *International Journal of Behavioral Development* (2008) 32, 283–289.

36. Pat-Horenczyk R, Qasrawi R, Lesack R, Haj-Yahia M, Peled O, Shaheen M, et al. Posttraumatic symptoms, functional impairment, and coping among adolescents on both sides of the israeli-palestinian conflict: a cross-cultural approach. *Appl Psychol*. (2009) 58:688–708. doi: 10.1111/j.1464-0597.2008.00372.x

37. Levin KA, Currie C. Reliability and validity of an adapted version of the Cantril ladder for use with adolescent samples. *Soc Indic Res*. (2014) 119:1047–63. doi: 10.1007/s11205-013-0507-4

38. Szkultecka-Dębek M, Dzielska A, Drozd M, Małkowska-Szkutnik A, Mazur J. What does the Cantril ladder measure in adolescence? *Arch Med Sci*. (2018) 1:182–9. doi: 10.5114/aoms.2016.60718

39. Jones K, Christopher PM, John HP, Scott T. Screening cognitive performance with the resident assessment instrument for mental health cognitive performance scale. *La Revue Canadienne de Psychiatrie*. (2010) 55:736–40. doi: 10.1177/070674371005501108

40. Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, et al. Orange: data mining toolbox in Python. *J Mach Learn Res*. (2013) 14:2349–53.

41. Cummings EM, Merrilees CE, Taylor LK, Mondi CF. Developmental and social–ecological perspectives on children, political violence, and armed conflict. *Dev Psychopathol*. (2017) 29:1–10. doi: 10.1017/S0954579416001061

42. Hamid Jan JM, Mitra AK, Hasmiza H, Pim CD, Ng LO, Wan Manan WM. Effect of gender and nutritional status on academic achievement and cognitive function among primary school children in a rural district in Malaysia. *Malays J Nutr*. (2011) 17:189–200.

43. Qasrawi R. Links between nutrition, life style habits and academic acheivment in Palestinian schoolchildren: a cross-sectional study. *Al-Quds J Acad Res*. (2021) 01:90–102. doi: 10.47874/2021p6

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 09 June 2023
DOI 10.3389/fpsyt.2023.1164433

OPEN ACCESS

EDITED BY
Robert Stewart,
King's College London, United Kingdom

REVIEWED BY
Radwan Qasrawi,
Al-Quds University, Palestine
Rabia Saleem,
University of Leicester, United Kingdom

*CORRESPONDENCE
Tianhua Chen
✉ t.chen@hud.ac.uk

RECEIVED 12 February 2023
ACCEPTED 12 May 2023
PUBLISHED 09 June 2023

CITATION
Chen T, Tachmazidis I, Batsakis S, Adamou M, Papadakis E and Antoniou G (2023) Diagnosing attention-deficit hyperactivity disorder (ADHD) using artificial intelligence: a clinical study in the UK. *Front. Psychiatry* 14:1164433. doi: 10.3389/fpsyt.2023.1164433

# Diagnosing attention-deficit hyperactivity disorder (ADHD) using artificial intelligence: a clinical study in the UK

Tianhua Chen[1]*, Ilias Tachmazidis[1], Sotiris Batsakis[1,2], Marios Adamou[3], Emmanuel Papadakis[1] and Grigoris Antoniou[1,4]

[1]Department of Computer Science, University of Huddersfield, Huddersfield, United Kingdom, [2]School of Production Engineering and Management, Technical University of Crete, Chania, Greece, [3]South West Yorkshire Partnership National Health Service (NHS) Foundation Trust, Wakefield, United Kingdom, [4]L3S Research Center, Leibniz University Hannover, Hannover, Germany

Attention-deficit hyperactivity disorder (ADHD) is a neurodevelopmental disorder affecting a large percentage of the adult population. A series of ongoing efforts has led to the development of a hybrid AI algorithm (a combination of a machine learning model and a knowledge-based model) for assisting adult ADHD diagnosis, and its clinical trial currently operating in the largest National Health Service (NHS) for adults with ADHD in the UK. Most recently, more data was made available that has lead to a total collection of 501 anonymized records as of 2022 July. This prompted the ongoing research to carefully examine the model by retraining and optimizing the machine learning algorithm in order to update the model with better generalization capability. Based on the large data collection so far, this paper also pilots a study to examine the effectiveness of variables other than the Diagnostic Interview for ADHD in adults (DIVA) assessment, which adds considerable cost in the screenining process as it relies on specially trained senior clinicians. Results reported in this paper demonstrate that the newly trained machine learning model reaches an accuracy of 75.03% when all features are used; the hybrid model obtains an accuracy of 93.61%. Exceeding what clinical experts expected in the absence of DIVA, achieving an accuracy of 65.27% using a rule-based machine learning model alone encourages the development of a cost effective model in the future.

KEYWORDS

**attention-deficit hyperactivity disorder (ADHD), diagnostic system, artificial intelligence, machine learning, explainable AI, mental health**

## 1. Introduction

Attention-deficit hyperactivity disorder (ADHD) is a neurodevelopmental disorder characterized by symptoms of inattention, hyperactivity, and/or impulsivity that causes significant impairment across domains. People with ADHD also exhibit deficits in executive functions, behavior and emotion regulation and motivation (1). Global demand for ADHD diagnostic assessment is rapidly growing due to increased awareness of the condition and other possible factors like impact of the pandemic (2). Within the UK where the conducted research is trialed in clinical practice, ADHD affects about 3–5% of children and 2% of adults (3).

In case of ADHD diagnosis the modes of intervention according to the National Collaborating Centre for Mental Health, UK, are both pharmacological and psychological

(4). The first line treatment for adult ADHD is psychostimulants (5) and medication is safe and effective, with 70% of patients reported improvement compared to 7% of controls (5, 6).

Delayed diagnosis and treatment for ADHD can be harmful to people and may cause broader mental health conditions, relationship and employment problems, criminal activities, and substance misuse. Specifically the adverse effects of untreated ADHD are well-documented with negative effects on academic outcomes (7), social functioning (8), employment (9) but also life itself leading to increased mortality (10).

For the UK, the National Institute for Health and Clinical Excellence (NICE) suggested in 2008 that the standard benchmark rate for referral to a Service in adults is 25 per 100,000 per year. The largest challenge at the moment for the adult population, bearing in mind the relative recency of acceptance amongst the professional community that ADHD can persist into adulthood (11), is the dearth of clinicians appropriately trained and confidence to place the diagnosis. Such bottleneck prevents patients receiving appropriate treatments and hence contributes to the morbidity of the adult ADHD.

The increased demand for assessments combined with the shortage of adequate healthcare capacity led to excessively long waiting lists, with an average waiting time up to 3 years. This puts a significant economic burden on the NHS, social services and the state overall. The total yearly costs to the individual and state combined were recently estimated to be €17,769 per person, per year (12) thus suggesting there is strong impetus for action.

In order to handle these challenges and coupled with the fact that Artificial Intelligence (AI) is enjoying an increasing number of successes in medical applications (13–16), an AI system, called NeuroIntel, was developed. For this work, clinical information collected from an NHS adult ADHD Service, which delivers a clinical pathway compliant with NICE recommendations (i.e., the gold standard), was used for creating a decision support tool that can first automate the process of making a diagnosis and second prioritize the ADHD cases based on levels of complexity. This prioritization serves to select the patients which would require a more in-depth clinical assessment. The clinical data collected were in the form of screening questionnaires and validated clinical diagnostic interviews, which are routinely collected as part of a clinical diagnostic assessment.

Applying machine learning for ADHD diagnosis (17) is a recent approach for dealing with this issue. Being commonly used in medical settings where the demand of interpretability is generally considered high, knowledge-based systems aim to represent knowledge explicitly via tools such as production or if-then rules, which allow such a system to reason about how it reaches a conclusion and to provide explanation of its reasoning to the user (18). In order to combine the strengths of machine learning-based approaches with the interpetability of knowledge based systems these approaches were combined in a hybrid setting (19), such that patterns extracted by machine learning and expertise directly given by clinicians can be unified in a single framework that best maximizes both approaches.

A series of efforts (17, 19) have been invested that has lead to the deployment of existing hybrid systems in the Adult ADHD Service of South-West Yorkshire Partnership NHS Foundation Trust (SWYPFT). The initial exploration (17) made use of data sources including both structured patient information as well as unstructured textual medical notes, but only on the basis of available electronic records from 69 patients only, with decision tree learning algorithm identified as an optimum choice to construct the diagnostic model, owing to its superior performance and interpretability. Another outcome of the underlying study also suggested the inclusion of features extracted from medical notes did not necessarily enhance the predictive capability, but might ran the risk of overfitting the models. A hybrid model (19) was subsequently proposed, which aims to not only utilize patterns learned by machine learning, but also incorporate expertise from senior clinicians, with results showing great promise of the technology, as it can accurately identify clear-cut cases where a decision can be safely made and can be verified by a less senior clinician, while referring the more complex cases for further assessment by a senior clinical specialist. With an ongoing trial operating in the largest NHS Service for adults with ADHD in the UK, the ongoing data collection has lead to the accumulation of 285 total patient records, with the evaluations and results as a retrospective study currently under review.

Most recently, more data was made available that has lead to a total collection of 501 anonymized records by 2022 July. This prompted the ongoing research to carefully examine the model with the large data collection so far, as reported in this paper, with the following major contributions:

1. In this paper, we aggregated all cases collected so far into one data set, followed by retraining and optimizing the machine learning algorithm in order to update the model with better generalization capability.
2. Efforts so far has made full use of all available variables in all the models, including the Diagnostic Interview for ADHD in adults (DIVA) assessment. While relevant to the assessment process, DIVA adds considerable burden as it relies on specially trained clinicians. Given the huge demand for diagnosis, both primary and secondary healthcare providers have been seeking for screenings without the possible use of DIVA. Such clinical demand warrants a test on the effectiveness of predictors other than DIVA that is also reported in this research.

The results reported in this paper demonstrate that the newly trained machine learning model reaches an accuracy of 75.03% when all features (including DIVA) are used; the hybrid model combining the new ML model with the knowledge model from (19) obtains an accuracy of 93.61%. When DIVA attributes are disregarded, the best performing machine learning model reaches an accuracy of 65.27%.

The remainder of this paper is organized as follows. Section 2 describes the data used as well as the data analysis framework. Section 3 presents the results of applying machine learning to all available data, as well as to partial data omitting attributes originating from the DIVA. Section 4 analyses and discusses the results, and Section 5 concludes the paper with a summary of the contributions and an outlook on future research.

# 2. Materials and methods

## 2.1. Data collection

For this project, the need for ethics approval was waived by South West Yorkshire Partnership Foundation Trust (SWYPFT) Research and Development Department as data were gathered retrospectively. Data was gathered as part of the clinical operations of the Service and was classed as a service improvement activity. The Caldicott Guardian at SWYPFT endorsed access to data following Caldicott Principles presented at: https://www.highspeedtraining.co.uk/hub/7-caldicott-principles/. Data was gathered from electronic records and patients accessing the Service are routinely informed that their data can be used for research purposes and can opt out if they wish.

The patient's data are provided by an NHS specialist mental health provider (South West Yorkshire Partnership NHS Foundation Trust-SWYPFT). In this study, we have included all cases from the time period it covers, without excluding any patients. For each case, we considered all clinical data routinely collected by the NHS service, following NICE guidelines, ahead of an appointment with a specialist clinician. The approach is to identify not only symptoms of ADHD but also consider comorbid conditions which could also present as ADHD before a diagnosis is made. This is consistent with what the DSM-5 criteria requires which in criterion E requires the clinician to make a judgement that the comorbid conditions do not better explain the presentation.

The dataset consists of 501 anonymized assessments for ADHD patients in the period between 2019 and 2022 July. The dataset contains demographic information about these patients in addition to self-reported screening questionnaires and clinical interview results. A total of 66 independent attributes are included into the dataset for each case with the last column of the dataset being the diagnostic outcome to predict. With 236 positive cases and 265 negative cases, the distribution of class labels is relatively balanced; whereas male subjects (322) are nearly twice that of female (179), in Figure 1, where it also shows the gender distribution for each of the diagnoses. In terms of the age distribution, Figure 2 suggests the age group between 20 and 30 has the most patients, with the youngest patient being 17 while the most senior being 72. The swarm plot shown in Figure 3, further suggests that, in general, ages of both positive and negative cases span from just below 20 to around 55, except for a couple of positive cases where age is around 70.

The data for each patient includes a patient identifier and the patient's gender and age. This is followed by the results of the Mood Disorder Questionnaire (MDQ) (20), the HELPS brain injury screening tool (21), the Drug Abuse Screening Test (DAST-10) (22), the GAD-7 test results measuring Generalized Anxiety (23), the Patient Health Questionnaire (PHQ-9) which measures the severity of depression (24), the Iowa Personality Disorder Screen (IOWA) (25), the Alcohol Use Disorders Identification Test (AUDIT) (26), the Conner's ADHD Rating Scales (27) and the Diagnostic Interview for ADHD in adults (DIVA) (28) results.

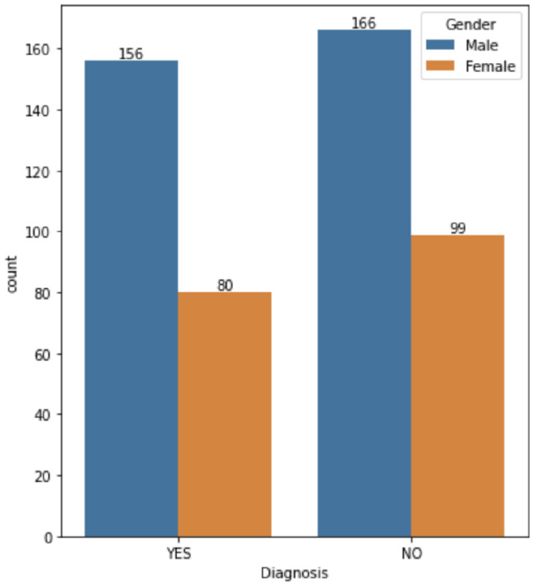


FIGURE 1
ADHD case vs. gender.

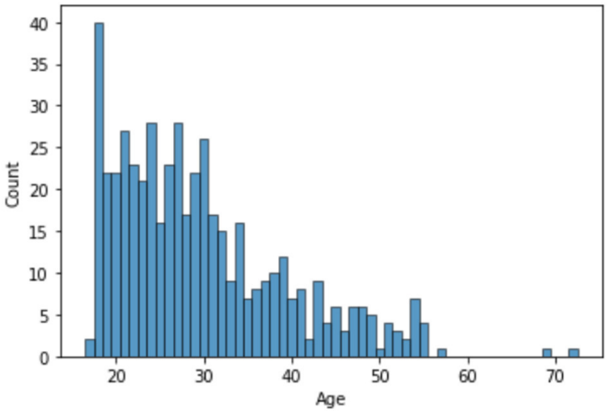


FIGURE 2
Age distribution.

## 2.2. Diagnostic process

The diagnostic process follows best practice approach recommended by the Royal College of Psychiatrists in the UK ADHD in adults: Good practice guidance (CR235) (29). The approach recommends a list of validated screening and diagnostic tools as well as a formal exploration of comorbidity. The inputs we used to construct the diagnostic tool capture these recommendations by capturing all components of the diagnostic process by using screening tools for ADHD and mental health, validated diagnostic tools for ADHD and a process for considering comorbidity with other conditions. As such the clinical diagnostic process has three steps: first, collection of information using screening tools (which also include input by carer); second, administration of a validated diagnostic tool (DIVA); third, full

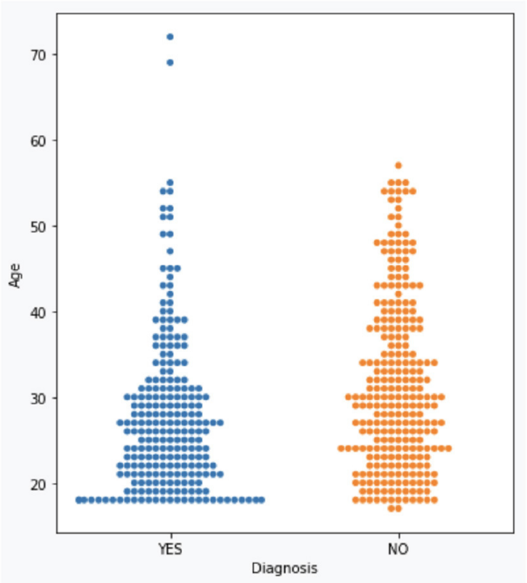


FIGURE 3
ADHD case vs. age.

psychiatric history to validated findings and identify comorbidities which could explain the presentation.

A schematic description of the current approach is shown in Figure 4A, while the use of the AI tool is illustrated in Figure 4B.

## 2.3. Framework architecture

Based on the analysis of the dataset, a diagnostic system for adult ADHD diagnosis has been developed consisting of two parts: (a) the machine learning (ML) model and (b) the knowledge representation-based (KR) model. The diagnostic outcomes of these two models are also combined producing the hybrid model. The system consisting of the ML, KR, and hybrid models has been deployed and used for assisting clinicians for ADHD diagnosis, offering an intuitive Web-based interface. In the following, the components of the system are presented.

### 2.3.1. Machine learning model

In order to develop a prediction model using machine learning, a number of mainstream algorithms were evaluated, with the evaluation results presented in detail in Section 3. The fact that a decision tree model is adopted, is due partially to the robust performance it offers in comparison with alternative popular machine learning models, but also to the interpretability it offers to represent the learning model through a set of IF-THEN rules (18). Such rules are highly recommended by healthcare professionals who not only are enabled to interrogate inference made by a machine learning model, but also makes it possible to integrate human knowledge for the ultimate generation of a hybrid system that incorporates both patterns extracted by a learning system and expertise from clinicians, as reported in our recent work (19).

The input to the model is a set of numerical values aggregating the full set of features of the initial dataset. Apart from the Age attribute, a number of psychological measures are used, i.e., PHQ9: Severity of self reported depression (numerical, having values between 0 and 27); GAD: Severity of self reported anxiety (values 0–21); MDQ: Self reported symptoms of bipolar disorder (Boolean value); AUDIT: Harmful alcohol consumption scale (values 0–40); DAST10: Drugs use score in the last 12 months (values 0–10); HELPS: Exposure to brain injury during lifetime (Boolean value); IOWA: Personality disorders evaluation (values 0 to 11); CAARS: CAARS ADHD TT1 score (values 1–100); DIVA Child IA: Attention deficit during childhood score (values 0–9); DIVA Child HI: Hyperactivity/impulsivity during childhood score (values 0–9); DIVA Adult IA: Attention deficit during adulthood score (values 0–9); DIVA Adult HI: Hyperactivity/impulsivity during adulthood score (values 0–9). The machine learning model receives the above mentioned input and produces as output of the corresponding rules a Boolean diagnostic outcome (“Yes” or “No”).

### 2.3.2. Knowledge based model

The knowledge model (19) for ADHD diagnosis encodes the empirical knowledge of an international expert in adult ADHD. This expert knowledge was extracted through interviews in order to encode the deep understanding of various tests and questionnaires that are routinely conducted by SWYPFT Research and Development Department (see Section 2.1). The meaning of each source of data was explored during the interviews and encoded in a machine-readable format in the form of if-then rules. Once rules were defined, their priority needed to be further specified in order to emulate the rationale of a clinical expert.

The knowledge model relies on DIVA scores, with low DIVA scores indicating that ADHD should not be inferred, while high DIVA scores indicate that ADHD diagnosis is more probable. A holistic approach is required, were patients affected by substance abuse, personality disorder, alcohol use, bipolar disorder, anxiety, depression and brain injury, might exhibit overlapping symptoms with ADHD. Thresholds are set to quantify abstract notions such as low or high DIVA scores, with rules prioritized in order to recreate the decision making process of a clinical expert. The knowledge model allows three possible outcomes, namely “Yes” (positive diagnosis), “No” (negative diagnosis), or “Expert” (the case should be referred to a clinician). For further details about the knowledge model, readers are referred to (19).

### 2.3.3. Hybrid model

The hybrid model (19) combines the results of the knowledge model and the machine learning model. Notice that the hybrid model requires the use of all available data (as opposed to the alternative machine learning model without DIVAs). A key difference between the two models above is that the machine learning model provides yes/no answers, while the knowledge model provides yes/no/expert answers. The hybrid model provides yes/no answers when both machine learning and knowledge model are in agreement. When the two models are in disagreement, patients are referred to a medical expert.

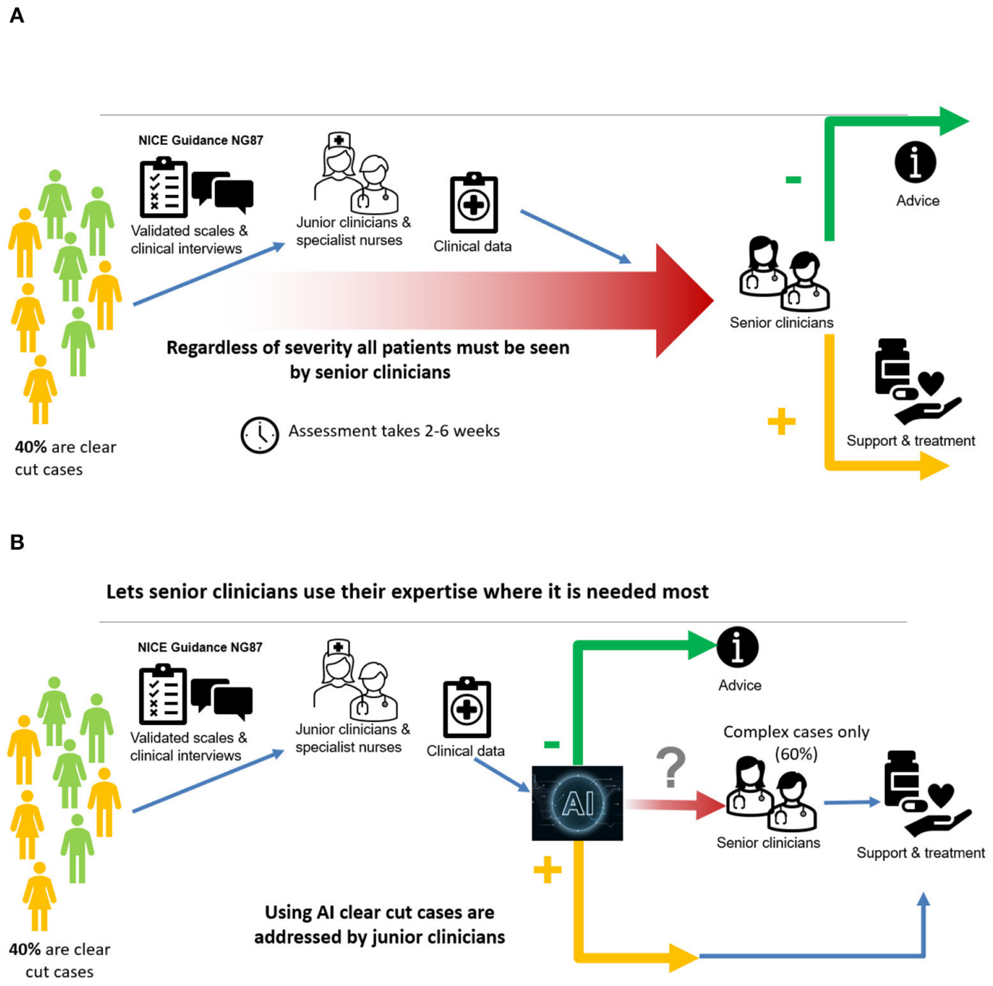

A
NICE Guidance NG87
Validated scales & clinical interviews
Junior clinicians & specialist nurses
Clinical data
Regardless of severity all patients must be seen by senior clinicians
Assessment takes 2-6 weeks
40% are clear cut cases
Senior clinicians
-
Advice
+
Support & treatment
B
Lets senior clinicians use their expertise where it is needed most
NICE Guidance NG87
Validated scales & clinical interviews
Junior clinicians & specialist nurses
Clinical data
AI
-
Advice
Complex cases only (60%)
?
Senior clinicians
Support & treatment
+
Using AI clear cut cases are addressed by junior clinicians
40% are clear cut cases

FIGURE 4
**(A)** A major bottleneck in current pathway; **(B)** AI overcomes the major bottleneck.

The main advantage of the hybrid model is that its yes/no answers are endorsed by both machine learning and knowledge model. Moreover, the machine learning model provides its recommendation to clinical specialists toward a particular outcome, even for patients that are referred to an expert (by the hybrid model). It is worth noting that referring patients to medical experts is a valid and desirable outcome since AI is aimed at streamlining clear-cut cases as well as identifying complicated cases that require expert's analysis. For further details about the hybrid model, readers are referred to (19).

Figure 5 depicts the overall framework of all the components mentioned above, which build the proposed diagnostic system. The visual is organized in three parts reflecting the corresponding models; each model is detailed in terms of its construction pipeline along with a brief overview of its functionality. To this end, a new assessment is processed by creating two diagnostic outcomes based on the rules set and the decision tree underlying the knowledge-based and machine learning component, respectively. Both outcomes, then, are combined by the hybrid model where the convergence of the diagnostic result is assessed, which eventually results into the final diagnostic recommendation.

## 3. Results and discussion on machine learning models

Correlation analysis is a statistical method used to examine the relationship between two variables, which allows determining whether there is a relationship and the strength between the two variables. To understand better such potential relationships before building the predictive models using machine learning algorithms for diagnosing ADHD, a correlation analysis is therefore conducted for each of the independent variables against the "Diagnosis" dependent variable.

This paper adopts the popular Pearson correlation with the value ranging between +1 and −1, where a value of +1 is a total positive linear correlation; 0 is no linear correlation; and −1 is a total negative linear correlation. In order to measure the strength of the correlation, absolute values of the actual correlation are used. Owing to the space limit, Figure 6 demonstrates the strength of the correlation of the top 20 independent variables. The DIVA attention deficit for both childhood and adulthood presents the strongest correlation with the diagnosis, followed by the DIVA hyperactivity/impulsivity for both childhood and adulthood. This

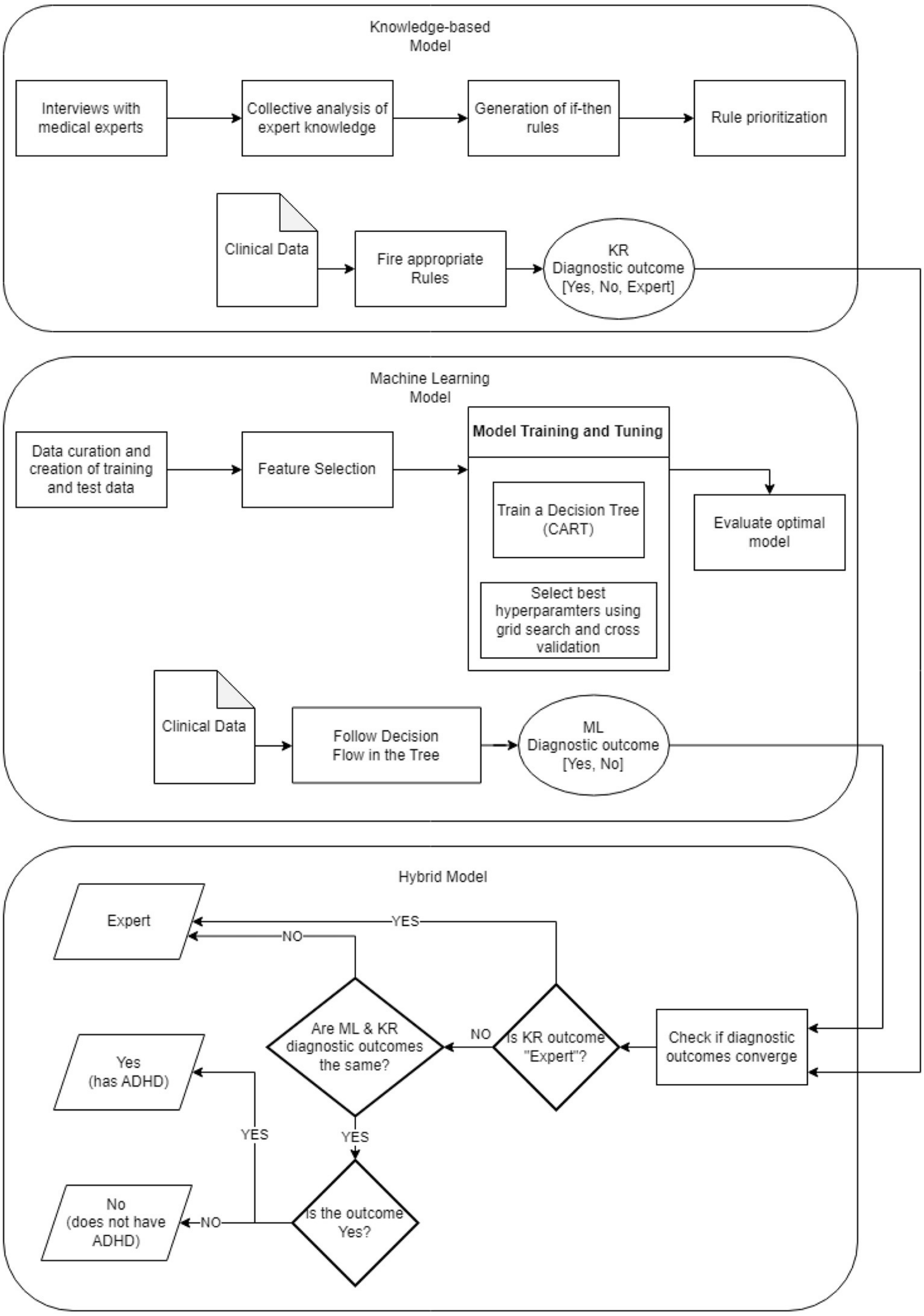

Knowledge-based Model
Interviews with medical experts
Collective analysis of expert knowledge
Generation of if-then rules
Rule prioritization
Clinical Data
Fire appropriate Rules
KR Diagnostic outcome [Yes, No, Expert]
Machine Learning Model
Data curation and creation of training and test data
Feature Selection
Model Training and Tuning
Train a Decision Tree (CART)
Select best hyperparamters using grid search and cross validation
Evaluate optimal model
Clinical Data
Follow Decision Flow in the Tree
ML Diagnostic outcome [Yes, No]
Hybrid Model
Expert
YES
NO
Are ML & KR diagnostic outcomes the same?
NO
Is KR outcome "Expert"?
Check if diagnostic outcomes converge
Yes (has ADHD)
YES
YES
No (does not have ADHD)
NO
Is the outcome Yes?

FIGURE 5
ADHD diagnostic system framework.

is followed by a weaker set of attributes around CAARS ADHD TT1 score, the IOWA personality disorders evaluation and age. It is important to note that correlation analysis does not necessarily imply causation, though they can be useful observations for following analysis.

While the knowledge model, which directly comes from clinical expertise, remains relatively stable; the machine learning model, which is data-driven in nature, is subject to re-train, given the significant recent intake of data from 216 new patients, making the total data entries 501. In particular, the decision tree algorithm, which has been used consistently, for its effectiveness in diagnostic accuracy as well as the inference interpretable by clinical professionals, remains our first choice among alternative machine learning models, which is consolidated by our successes so far as also highly recommended by the clinicians (17, 19).

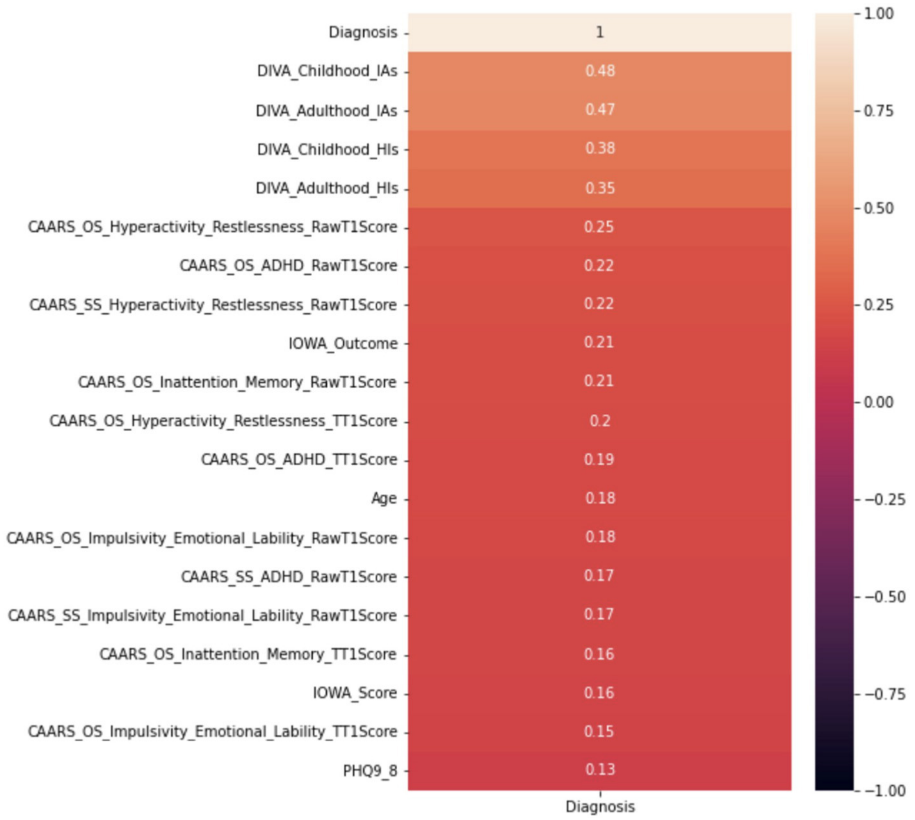


FIGURE 6
Correlation analysis of top 20 attributes including DIVAs.

## 3.1. ML results using all available data

Decision tree learning has been one of the most influential machine learning and data mining algorithms (30), where it recursively selects the most informative attribute that returns best homogeneous sets of the underlying data instances, until all attributes have been considered or the addition of any remaining attribute does not improve its discriminative power. While more details on the induction of a decision tree can be found in (30), the specific Classification and Regression Tree (CART) is utilized for experimenting with the newly collected ADHD data. Using Google Colab, the implementations of the algorithm comes from Scikit-learn (31), which is a free software machine learning library for the Python programming language.

In identifying the optimal decision tree model that best fits the underlying data, a hyper-parameter search is conducted through a grid search to examine a number of hyper-parameter that might affect model construction, so that multiple instances of CART models can be trained and assessed on the same dataset but initialized with different hyperparameters. In particular, "min samples split" was tested with values 2, 4, 6, 8, 10; this parameter specifies the minimum number of data samples required to create an internal decision node, which eventually protects the model from over-fitting. "Max features" was tested on "sqrt" and "log2", which defines the number of features required to make a split decision. "Min samples leaf", similar to "min samples split" was set to [1, 8], this parameter sets a threshold of minimum observations for the creation of final decision nodes (leaf nodes). "Max depth" was tested on range [2,6], which is mainly used for preventing overfitting by controlling the size of the final decision tree. Another test took place in the choice of splitting criteria, which is either "gini" and "entropy"—both quantify the level of impurity and disorder and is used to directly guide the selection of a particular attribute to split the tree.

To ensure that the generated model is not overfitting the data with a more fair estimation of the model's generalization error, the k-fold cross-validation is used whereby a model is given a dataset of known data on which training is run and an independent dataset of unknown data against which the model is tested. Specifically, for k-fold cross-validation, the original dataset is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times, with a different subsample being used as the validation data each time. The performance measure is calculated by averaging the performance across all k iterations. In this research, the value of *k* is set to 10 for conventional purposes (32).

In terms of the specific metric to examine the performance, results are reported using several metrics, including:

- Accuracy (Acc), in the percentage of correct predictions, i.e., the resultant model predicts positive in case the patient to be diagnosed is with ADHD and negative in case the patient is without ADHD. A perfect classification model would always make correct predictions, resulting in 100% accuracy. Given a model trained on training data, the train accuracy reports the performance on the training data; while the test accuracy is the performance when the trained model is validated on test data that model has never seen before.
- Balanced accuracy, is defined as the average of the sensitivity and specificity of the model, where sensitivity is the proportion of positive cases that are correctly identified by the model, while specificity is the proportion of negative cases that are correctly identified by the model.
- Precision measures the proportion of positive predictions that were actually correct, defined as the number of true positive predictions made by the model divided by the total number of positive predictions made by the model.
- Recall measures the proportion of actual positive cases that were correctly identified by the model, defined as the number of true positive predictions made by the model divided by the total number of actual positive cases in the dataset.
- F1-score is used to balance precision and recall as a measure of a model's overall accuracy, defined as the harmonic mean of the model's precision and recall.
- Auc, the Area Under the Receiver Operating Characteristic (ROC) curve, is the curve of sensitivity (a.k. a. true positive rate), plotted against 1-specificity (a.k.a. false positive rate), which is independent of the prior class distribution, i.e., percentages of positive and negative samples. A perfect classification would produce AUC = 1, while random guessing would produce a 0.5 AUC.

Best result of the CART model after the grid search is then reported in Table 1. However, despite the decision tree being the first choice, it's also critical to evaluate learning algorithms of alternative common choices to give a comprehensive view of the general performance landscape. A multitude of mainstream machine learning algorithms (30) was selected including:

- Logistic Regression, a generalized linear model that uses a logistic function to model the probability of a positive/negative diagnosis given the underlying variables.
- Linear Discriminant Analysis, similar to logistic regression, finds the linear combination of features that maximally separates the different classes, but with additional assumptions on data that logistic regression does not make.
- Multiple Layer Perceptron, widely applied in numerous practical applications, is a type of artificial neural network that consists of multiple layers of interconnected nodes, with each layer fully connected to the next.
- K-nearest Neighbor or KNN is the classical instance-based learning approach, where an instance is classified by a majority vote of its neighbors. It works by assigning an instance to the class most common among its *k* nearest neighbors.
- Support Vector Machine is a sequential optimization algorithm for building support vector machines (which form another type of most powerful learning classifiers), with both linear and Radial basis function (RBF) kernel adopted as kernel function.
- Gaussian Naive Bayes, is based on the Bayes Theorem by using the probability distribution of each feature to make predictions about the diagnosis of a new patient, assuming the probability distribution of each feature follows a Gaussian distribution.
- Random Forest is a very powerful ensemble machine learning method, made up of a collection of decision trees, which are trained on different subsets of the data and then combined to make predictions. The final predictions are made by averaging the predictions of all the individual trees in the forest.
- Extreme Gradient Boosting is another mighty ensemble machine learning method that involves training a sequence of weak decision trees models, and then combining their predictions to form a stronger model.

Experiments were conducted using the scikit-learn open source machine learning library that integrates the implementation of all aforementioned ML approaches with default settings unless otherwise explicitly specified. As there also exist missing values in the collected data, all missing values were replaced using simple imputation, i.e., the mean value for numerical data, and the mode for categorical data, though more advanced interpolation technique (33) may be considered in future work.

Table 1 summarizes results of the ten machine learning models across the seven aforementioned performance metrics; the most important observations are highlighted in bold. On the basis of 10-fold cross validation, attention is first drawn to the train and test accuracy, whereby Random Forest and Extreme Gradient Boosting achieves the best possible train accuracy of 100%, which clearly suggest overfitting, with serious gap between train and test accuracies. The remaining model generally achieves 70+% train accuracies with 60+% test accuracies, indicating some slight overfitting. Despite performances of alternative models might be further improved, results based on default parameter settings are generally considered in experimental practice as comparison. In general, testing results all exceed 60%, clearly beating the random guess of 52.9%, which is calculated based on the original distribution of 236 positive and 265 negative cases—these demonstrates the validity of using machine learning models to support decision making of a complex task in clinical practice such as ADHD diagnosis. Among those, the CART decision tree has achieved the test accuracy only very slightly higher train result, suggesting a robustly fitted model when it's trained. Furthermore, its test and balanced accuracy, as highlighted in bold, achieves the best results, clearly beating most competitors by a large margin. In terms of precision and recall, while the precision isn't the best among all, this can be mitigated by the knowledge model and further examination by clinicians, the significantly high recall suggests it only misses a few positive cases that should have been attended to. Whereas the F1, which is an average of the precision and recall, as well as the Auc score, still suggests that CART is among the top accurate models. Overall, with the recent significant

TABLE 1 Results using all available attributes.

| Data set | Train acc | Test acc | Balanced acc | Precision | Recall | F1 | Auc |
|---|---|---|---|---|---|---|---|
| Classification and Regression Tree | 75.05 | **75.03** | **75.74** | 0.69 | **0.88** | **0.77** | 0.78 |
| Logistic Regression | 75.74 | 66.85 | 66.79 | 0.65 | 0.66 | 0.65 | 0.75 |
| Linear Discriminant Analysis | 77.93 | 64.45 | 64.44 | 0.62 | 0.64 | 0.63 | 0.70 |
| Artificial Neural Networks | 76.00 | 66.25 | 68.84 | 0.65 | 0.72 | 0.66 | 0.75 |
| K Nearest Neighbor | 76.36 | 60.27 | 60.04 | 0.58 | 0.57 | 0.57 | 0.62 |
| Support Vector Machine (RBF) | 65.31 | 62.27 | 62.09 | 0.60 | 0.60 | 0.60 | 0.67 |
| Support Vector Machine (Linear) | 79.53 | 64.85 | 64.83 | 0.62 | 0.65 | 0.63 | 0.70 |
| Gaussian Naïve Bayes | 67.31 | 66.07 | 67.06 | 0.60 | 0.84 | 0.70 | 0.71 |
| Random Forest | **100.00** | 74.44 | 73.34 | 0.72 | 0.75 | 0.71 | **0.80** |
| Extreme Gradient Boosting | **100.00** | 72.84 | 72.84 | 0.71 | 0.73 | **0.72** | **0.80** |
| Averaged | 79.80 | 66.48 | 66.70 | 0.64 | 0.68 | 0.65 | 0.72 |

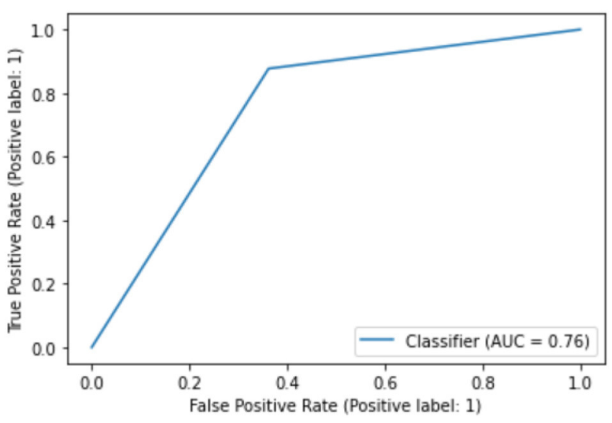


FIGURE 7
AUC on full data.

collection of new patient data, machine learning is able to enhance the diagnostic accuracy for ADHD, and decision tree is still a robust choice from performance perspective.

While the above performance is reported on the basis of cross validation involving both train and test data subsets for model selection and validation, a final decision tree will be trained using all available data so that data can be fully exploited in generating a working model. With an accuracy of 75.04%, this is almost the same as 75.03% as an averaged result of 10-fold cross validation. The associated AUC curve can be found in Figure 7 with more detailed results reported in Table 2. With results closely following that of Table 1, this again demonstrates the reliability of the CART model.

Furthermore, despite that the full decision tree may not be presented due to confidentiality, we are able to show the significance of variables utilized by the CART algorithm, which only includes two variables, i.e., 0.68 for with DIVA attention deficit for adulthood, and 0.32 for DIVA attention deficit for childhood, both are also the top attributes as analyzed by the correlation in Figure 6.

TABLE 2 Performance of CART model on full data.

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| No (does not have ADHD) | 0.85 | 0.64 | 0.73 | 265 |
| Yes (has ADHD) | 0.68 | 0.88 | 0.77 | 236 |
| Accuracy | | | 0.75 (overall acc: 75.04%) | 501 |
| Macro avg | 0.77 | 0.76 | 0.75 | 501 |
| Weighted avg | 0.77 | 0.75 | 0.75 | 501 |

## 3.2. ML results without DIVAs

Whilst it's assuring that the DIVA attributes exhibit strong capabilities in differentiating positive and negative ADHD cases, cost of conducting DIVA tests in practice proves high and they have to executed by senior clinicians—this motivates on-going projects to explore alternative tests that may be able to perform by junior clinicians while also being effective in the diagnosis. As such, the four attributes with DIVA tests are now removed, i.e., the DIVA Child IA and DIVA Adult IA, which is the attention deficit during childhood and adulthood; the DIVA Child HI and DIVA Adult HI. which is the Hyperactivity/impulsivity during childhood and adulthood. Similar to Figure 6, correlation of the top 20 remaining attributes with respect to the diagnosis is now shown in Figure 8, where most of the best correlated results come from the CAARS ADHD TT1 score, but the strength of correlation of these variables are much lower than that of DIVA, generally around 0.2, suggesting that it may not be valid to use each attribute alone to make effective diagnosis.

Following the same grid search of the hyper-parameters to best fit the underlying data in again 10-fold cross validation, the result of CART model is presented in Table 3, in comparison with the same set of mainstream machine learning models as above. From a holistic perspective, the averaged performance (as well as every single result) across all selected machine learning models are significantly worse than that when DIVA attributes are used in all metrics. For instance, the averaged test accuracy has dropped to 58.08% from 66.48% while the Auc is now 0.63 compared to 0.72 before. This clearly suggests how critical DIVA attributes are in establishing an effective learning model and the limited capabilities of remaining attributes. It is worth noting that both random forest and extreme gradient boosting are still able to fit the training data perfectly throughout, but their test accuracy are also very limited, suggesting that it's possible to use complex machine learning techniques like the two ensemble-based methods to fit the data, but predicting unseen ADHD patients can still be challenging, especially in the absence of DIVA attributes. Having said that, most machine learning models are still achieve results clearly better than the random guess of 52.9%.

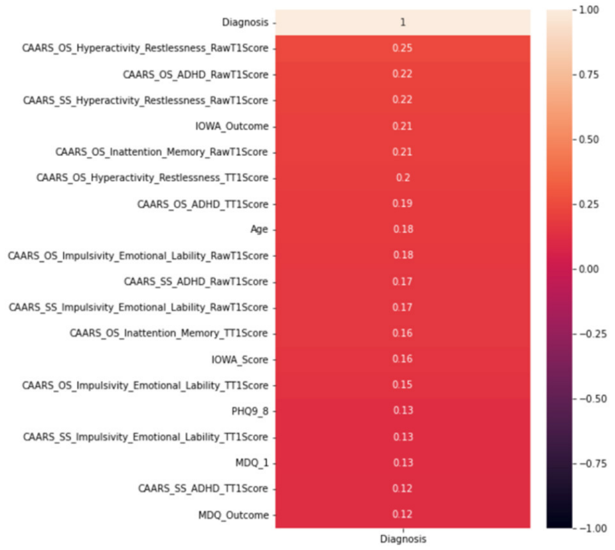


FIGURE 8
Correlation analysis of top 20 attributes excluding DIVAs.

In terms of the CART decision tree model, it still achieves the best test accuracy and balanced accuracy, with relatively small gap between train and test accuracy, indicating it still a robust and effective choice. The final decision tree model is then trained on full data again excluding the use of four DIVA attributes. The final model achieves a slightly higher accuracy of 65.27% than the averaged performance of 61.69% as a result of 10-fold cross validation. The associated AUC curve can be found in Figure 9 with more detailed results reported in Table 4. Overall, these results are slightly better than that of Table 3 obtained through the 10-fold cross validation, which can be expected as the model is trained and tested on the same data instances. As for the specific variables selected by the final CART model, irrespective of already ignored four DIVA attributes, it's observed that "IOWA_Score" is selected with 0.52 significance, followed by "Age" of 0.3 significance, and then "CAARS_OS_Inattention_Memory_TT1Score" of 0.18 importance. In comparison with the top 20 attributes in Figure 8, the three selected attributes are not the top ones as calculated by correlation, which indicates that attributes of lower correlations alone may be significant when combined with others that lead to an effective cohort.

## 4. Results and discussions on KR and hybrid models

The knowledge model is based on if-then rules, encoding the knowledge of medical experts. In addition, the chosen machine learning algorithm, namely the decision tree algorithm, generates a set of if-then rules as well. The results of the two models are combined by the hybrid model as described in Section 2.3.3, leading to an overall prediction of an ADHD diagnosis. We evaluated all three models over the existing dataset for the 501 patients and compared the results to the diagnosis made by the medical experts. Note that the knowledge model, the machine learning model and the hybrid model are referred below as KR, ML and Hybrid models, respectively.

Table 5 shows how patients were classified by the three models. It is evident that in the ML model all patients are classified as

TABLE 3 Results without DIVA.

| Data set | Train acc | Test acc | Balanced acc | Precision | Recall | F1 | Auc |
|---|---|---|---|---|---|---|---|
| Classification and Regression Tree | 65.40 | 61.69 | 61.44 | 0.60 | 0.59 | 0.59 | 0.62 |
| Logistic Regression | 69.86 | 60.47 | 60.37 | 0.58 | 0.59 | 0.58 | 0.63 |
| Linear Discriminant Analysis | 71.55 | 58.66 | 58.60 | 0.56 | 0.57 | 0.56 | 0.60 |
| Artificial Neural Networks | 69.11 | 56.70 | 59.96 | 0.56 | 0.63 | 0.56 | 0.65 |
| K Nearest Neighbor | 75.80 | 59.09 | 59.14 | 0.57 | 0.56 | 0.56 | 0.61 |
| Support Vector Machine (RBF) | 62.74 | 57.71 | 58.84 | 0.60 | 0.49 | 0.51 | 0.66 |
| Support Vector Machine (Linear) | 72.50 | 59.46 | 59.33 | 0.57 | 0.57 | 0.56 | 0.61 |
| Gaussian Naïve Bayes | 57.35 | 54.26 | 55.66 | 0.51 | 0.81 | 0.62 | 0.61 |
| Random Forest | 100.00 | 58.88 | 59.61 | 0.61 | 0.53 | 0.51 | 0.66 |
| Extreme Gradient Boosting | 100.00 | 57.52 | 57.28 | 0.54 | 0.51 | 0.52 | 0.63 |
| Averaged | 75.43 | 58.08 | 58.75 | 0.57 | 0.59 | 0.55 | 0.63 |

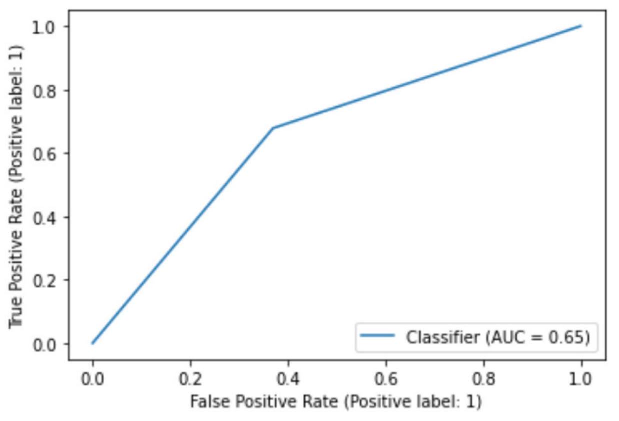


FIGURE 9
AUC on data excluding DIVA attributes.

TABLE 4 Performance of CART model on data excluding DIVA attributes.

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| No (does not have ADHD) | 0.69 | 0.63 | 0.66 | 265 |
| Yes (has ADHD) | 0.62 | 0.68 | 0.65 | 236 |
| Accuracy | | | 0.65 (overall acc: 65.27%) | 501 |
| Macro avg | 0.65 | 0.65 | 0.65 | 501 |
| Weighted avg | 0.65 | 0.65 | 0.65 | 501 |

TABLE 5 Confusion matrix of KR, ML, and hybrid models.

| | | Predicted | | |
|---|---|---|---|---|
| Clinical outcome | Model | Yes | No | Expert |
| Yes (has ADHD) | KR | 57 | 6 | 173 |
| | ML | 207 | 29 | 0 |
| | Hybrid | 52 | 6 | 178 |
| No (does not have ADHD) | KR | 32 | 109 | 124 |
| | ML | 96 | 169 | 0 |
| | Hybrid | 26 | 109 | 130 |

TABLE 6 Accuracy of each model per set of outcomes.

| Model | Yes/no (%) | Yes/no/expert (%) |
|---|---|---|
| KR | 166/204 (81.37%) | 463/501 (92.42%) |
| ML | 376/501 (75.05%) | 376/501 (75.05%) |
| Hybrid | **161/193 (83.42%)** | **469/501 (93.61%)** |

either having ADHD or not having ADHD (Yes/No outcomes only), while in the KR model approximately 40.7% of patients are classified to a Yes/No outcome with 59.3% of patients being referred to a medical expert. It is expected that the Hybrid model will classify the minimum number of patients to a Yes/No outcome (as both KR and ML models must provide the same classification) and the maximum number of patients will be referred to a medical expert (those referred by the KR model as well as all outcome disagreements between KR and ML models). Thus, the results for the Hybrid model classifying 38.5% of patients to a Yes/No outcome and 61.5% of patients referred to a medical expert are in line with model design.

Table 6 presents the accuracy of each model, namely how many patients where correctly classified out of all patients assigned to a specific set of outcomes, where the set of allowed outcomes is either Yes/No or Yes/No/Expert. Note that the highest accuracy is highlighted in bold. Referring complex cases to clinical experts increases the accuracy for both KR (from 81.37% to 92.42%) and Hybrid (from 83.42% to 93.61%) models. Recall that referring patients to medical experts is considered a valid and desirable outcome (see Section 2.3.3). The Hybrid model combines the strengths of both KR and ML models, thus exhibiting better accuracy over both Yes/No and Yes/No/Expect outcomes.

Employing Artificial Intelligence in clinical settings holds great potential to improve healthcare. However, these benefits can be attained only if the underlying ethical implications are addressed (34). Although a comprehensive ethical risk analysis is planned for future research; at this stage of this work we provide a preliminary discussion on major ethical challenges and how they are being currently addressed (when applicable) on the machine learning and knowledge-based component of the proposed framework. We cover the three primary ethical factors, as they are described in (35), namely, data protection, algorithmic fairness and accountability.

The dataset used to train the machine learning component was provided by SWYPFT following Caldicott Principles (Section 2.1) and privacy is ensured via anonymization, where individuals are no longer identifiable. Clinical data are primarily used to train the machine learning component by recognizing data patterns and encoding them into the underlying mathematical formulation of the model. The knowledge-based part, on the other hand, does not rely on data but expert knowledge. Data protection is assured considering that drawing predictions using the trained model or the rules within the knowledge-based model does not provide any access to the initial dataset nor the data of a new case is internally stored.

In terms of algorithmic fairness, the proposed framework incorporates several steps to reduce bias. The training dataset is relatively balanced and includes all the available assessments within the predefined case study period, making the dataset representative of the selected demographic. Several candidate models are trained using cross-validation to mitigate bias by minimizing the odds of over-fitting. This is a crucial step that prevents models from learning particularities of the training dataset and instead enables them to focus on more generic data trends. In the case of hyperparameters, several performance metrics are employed and tuning is achieved via a thorough grid search.

Regarding accountability, the proposed work operates as a recommendation system that aims to assist clinicians instead of independently ruling diagnostic decisions. Even as a decision support tool, explainability plays a pivotal role when applying AI in clinical settings. Consequently, great emphasis is put on transparency, where both knowledge-based model (rule-based format) and the optimal machine learning model (CART—decision tree structure) provide clear reasoning paths that facilitate *in situ* examination of potential outcomes.

Several ethical considerations have been taken into account and the corresponding ethical issues have been mitigated through sophisticated design of the tasks of data processing, training pipeline and knowledge representation. However, the resources that build the proposed framework, clinical data and expert medical knowledge are of high importance and they may raise ethical challenges if the quality is not assured, despite the rigorous design of the methodology. For instance, expert systems that partially capture the available knowledge (e.g., ignoring special cases—outliers—of ADHD) or non-representative clinical data (due to data scarcity) can introduce bias in the end product. In the current work, the models rely on carefully curated information provided by the collaborating healthcare facility that meet quality standards to examine the capabilities of the proposed solution. We plan to advance this work to a wider clinical study and eventually pilot this solution into a fully-fledged AI diagnostic recommendation tool. However, this would require an in-depth analysis that would eliminate any bias, which is one of the first priorities of our future work.

## 5. Conclusion

This paper is part of our long-term effort to introduce automation support to the diagnosis of adult ADHD, using AI technologies. Following clinical trial deployment in an NHS adult ADHD service, this paper reported on results obtained from retraining machine learning models on the richer dataset. The results are encouraging and suggest that the AI algorithm can be used in clinical practice.

Next steps in our efforts will include obtaining a broader evidence base by trialing the AI algorithm in other NHS or private healthcare providers. In addition, performing an in-depth ethical risk analysis and introducing mitigation strategies to eliminate bias is included in our project plan. Furthermore, we will validate and refine the knowledge-based model to confirm it captures all relevant knowledge and introduce flexibility in the contained rules using exceptions of probabilities. We also intend to conduct a study on the acceptance of our approach by clinicians and patients.

We also trained machine learning models *without using DIVA attributes*, so as to potentially lower the burden of the diagnostic process on healthcare services. The accuracy obtained was not sufficiently high to encourage clinical testing. We believe that this outcome is partially due to the absence of a knowledge model that could work in conjunction with the machine learning model—note that the knowledge model of (19) could not be used because it makes uses of DIVA values. In future work, we intend to develop a new knowledge model without use of DIVA attributes, to help increase overall accuracy through a hybrid AI algorithm.

## Data availability statement

The datasets presented in this article are not readily available because in the interests of protecting patients' privacy, the data cannot be shared. Requests to access the datasets should be directed to MA, marios.adamou@swyt.nhs.uk.

## Ethics statement

Ethical approval was waived by South West Yorkshire Partnership Foundation Trust (SWYPFT) Research and Development Department as data were gathered retrospectively. Written informed consent was not required in accordance with institutional requirements and national legislation.

## Author contributions

TC: formal analysis, literature research, methodology, validation, and writing—draft and review. IT: formal analysis, methodology, validation, and writing—draft and review. SB and EP: writing—review and editing. MA: conceptualization and methodology. GA: conceptualization, methodology, and writing—review and editing. All authors contributed to the article and approved the submitted version.

## Funding

This research was partially supported by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor with Grant No. 01DD20003.

## 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.

# References

1. Asherson P, Buitelaar J, Faraone SV, Rohde LA. Adult attention-deficit hyperactivity disorder: key conceptual issues. *Lancet Psychiatry.* (2016) 3:568–78. doi: 10.1016/S2215-0366(16)30032-3

2. Chen T, Lucock M. The mental health of university students during the COVID-19 pandemic: an online survey in the UK. *PLoS ONE.* (2022) 17:e0262562. doi: 10.1371/journal.pone.0262562

3. Riglin L, Leppert B, Langley K, Thapar AK, O'Donovan MC, Davey Smith G, et al. Investigating attention-deficit hyperactivity disorder and autism spectrum disorder traits in the general population: what happens in adult life? *J Child Psychol Psychiatry.* (2021) 62:449–57. doi: 10.1111/jcpp.13297

4. NCCMH. *Attention Deficit Hyperactivity Disorder: Diagnosis and Management of ADHD in Children, Young People and Adults* (2009).

5. Fields SA, Johnson WM, Hassig MB. Adult ADHD: addressing a unique set of challenges. *J Fam Pract.* (2017) 66:68–74.

6. APA. DSM 5 diagnostic and statistical manual of mental disorders. In: DSM 5 Diagnostic and Statistical Manual of Mental Disorders. (2013). p. 947.

7. Arnold LE, Hodgkins P, Kahle J, Madhoo M, Kewley G. Long-term outcomes of ADHD: academic achievement and performance. *J Attent Disord.* (2020) 24:73–85. doi: 10.1177/1087054714566076

8. Cook J, Knight E, Hume I, Qureshi A. The self-esteem of adults diagnosed with attention-deficit/hyperactivity disorder (ADHD): a systematic review of the literature. *Attent Deficit Hyperact Disord.* (2014) 6:249–68. doi: 10.1007/s12402-014-0133-2

9. Adamou M, Arif M, Asherson P, Aw TC, Bolea B, Coghill D, et al. Occupational issues of adults with ADHD. *BMC Psychiatry.* (2013) 13:59. doi: 10.1186/1471-244X-13-59

10. Dalsgaard S, Østergaard SD, Leckman JF, Mortensen PB, Pedersen MG. Mortality in children, adolescents, and adults with attention deficit hyperactivity disorder: a nationwide cohort study. *Lancet.* (2015) 385:2190–6. doi: 10.1016/S0140-6736(14)61684-6

11. Asherson P, Adamou M, Bolea B, Muller U, Morua SD, Pitts M, et al. Is ADHD a valid diagnosis in adults? Yes. *BMJ.* (2010) 340:c549. doi: 10.1136/bmj.c549

12. Vibert S. *Your Attention Please: The Social and Economical Impact of ADHD.* London: Demos (2018).

13. Chen T, Su P, Shen Y, Chen L, Mahmud M, Zhao Y, et al. A dominant set-informed interpretable fuzzy system for automated diagnosis of dementia. *Front Neurosci.* (2022) 16:867664. doi: 10.3389/fnins.2022.867664

14. Bucholc M, Titarenko S, Ding X, Canavan C, Chen T. A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia. *Expert Syst Appl.* (2023) 217:119541. doi: 10.1016/j.eswa.2023.119541

15. Chen T, Carter J, Mahmud M, Khuman A, editors. Artificial intelligence in healthcare: recent applications and developments. in: *Brain Informatics and Health.* Singapore: Springer (2022). p. 179–97.

16. Ahmed S, Nur SB, Hossain F, Kaiser MS, Mahmud M, Chen T, et al. Computational intelligence in detection and support of autism spectrum disorder. In: *Artificial Intelligence in Healthcare*. Springer (2022). p. 179–197.

17. Chen T, Antoniou G, Adamou M, Tachmazidis I, Su P. Automatic diagnosis of attention deficit hyperactivity disorder using machine learning. *Appl Artif Intell.* (2021) 35:657–69. doi: 10.1080/08839514.2021.1933761

18. Chen T, Shang C, Su P, Keravnou-Papailiou E, Zhao Y, Antoniou G, et al. A decision tree-initialised neuro-fuzzy approach for clinical decision support. *Artif Intell Med.* (2021) 111:101986. doi: 10.1016/j.artmed.2020.101986

19. Tachmazidis I, Chen T, Adamou M, Antoniou G. A hybrid AI approach for supporting clinical diagnosis of attention deficit hyperactivity disorder (ADHD) in adults. *Health Inform Sci Syst.* (2021) 9:1–8. doi: 10.1007/s13755-020-00123-7

20. Hirschfeld RM. The mood disorder Questionnaire: a simple, patient-rated screening instrument for bipolar disorder. *Prim Care Companion J Clin Psychiatry.* (2002) 4:9. doi: 10.4088/PCC.v04n0104

21. Picard M, Scarisbrick D, Paluck R. *HELPS: A Brief Screening Device for Traumatic Brain Injury*. New York, NY: Comprehensive Regional Traumatic Brain Injury Rehabiliation Center (1991).

22. Skinner HA. The drug abuse screening test. *Addict Behav.* (1982) 7:363–71.

23. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Internal Med.* (2006) 166:1092–7. doi: 10.1001/archinte.166.10.1092

24. Löwe B, Kroenke K, Herzog W, Gräfe K. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). *J Affect Disord.* (2004) 81:61–6. doi: 10.1016/S0165-0327(03)00198-8

25. Langbehn DR, Pfohl BM, Reynolds S, Clark LA, Battaglia M, Bellodi L, et al. The Iowa Personality Disorder Screen: Development and preliminary validation of a brief screening interview. *J Pers Disord.* (1999) 13:75–89.

26. Saunders JB, Aasland OG, Babor TF, De la Fuente JR, Grant M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction.* (1993) 88:791–804.

27. Conners CK, Erhardt D, Sparrow EP. *Conners' Adult ADHD Rating Scales (CAARS): Technical Manual.* Multi-Health Systems North Tonawanda. New York, NY (1999).

28. Ramos-Quiroga JA, Nasillo V, Richarte V, Corrales M, Palma F, Ibáñez P, et al. Criteria and concurrent validity of DIVA 2.0: a semi-structured diagnostic interview for adult ADHD. *J Attent Disord.* (2019) 23:1126–35. doi: 10.1177/1087054716646451

29. RC PSYCH. *ADHD in Adults: Good Practice Guidance (CR235)* (2019). Available online at: https://www.rcpsych.ac.uk/docs/default-source/improving-care/better-mh-policy/college-reports/cr235-adhd-in-adults---good-practice-guidance.pdf?sfvrsn=7c8cc8e4_12.

30. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. *Knowledge Inform Syst.* (2008) 14:1–37. doi: 10.1007/s10115-007-0114-2

31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* (2011) 12:2825–30.

32. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. vol. 2. Springer (2009).

33. Chen T, Shang C, Yang J, Li F, Shen Q. A new approach for transformation-based fuzzy rule interpolation. *IEEE Transactions on Fuzzy Systems.* (2019) 28:3330–44. doi: 10.1109/TFUZZ.2019.2949767

34. Naik N, Hameed B, Shetty DK, Swain D, Shah M, Paul R, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg.* (2022) 9:862322. doi: 10.3389/fsurg.2022.862322

35. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, editors. *Artificial Intelligence in Healthcare.* Cambridge, MA: Elsevier, Academic Press (2020). p. 295–336. doi: 10.1016/B978-0-12-818438-7.00012-5

TYPE Original Research
PUBLISHED 12 June 2023
DOI 10.3389/fpsyt.2023.1143175

Check for updates

OPEN ACCESS

EDITED BY
Marcos Del Pozo Banos,
Swansea University Medical School,
United Kingdom

REVIEWED BY
Po-Han Chou,
China Medical University, Taiwan
Shaw-Ji Chen,
Taitung MacKay Memorial Hospital, Taiwan
Chun-Hung Chang,
China Medical University Hospital, Taiwan

*CORRESPONDENCE
Jennifer Wright-Berryman
✉ wrigh2jb@ucmail.uc.edu
Joshua Cohen
✉ jcohen@clarigenthealth.com

RECEIVED 12 January 2023
ACCEPTED 22 May 2023
PUBLISHED 12 June 2023

CITATION
Wright-Berryman J, Cohen J, Haq A, Black DP and Pease JL (2023) Virtually screening adults for depression, anxiety, and suicide risk using machine learning and language from an open-ended interview. *Front. Psychiatry* 14:1143175. doi: 10.3389/fpsyt.2023.1143175

# Virtually screening adults for depression, anxiety, and suicide risk using machine learning and language from an open-ended interview

Jennifer Wright-Berryman[1]*, Joshua Cohen[2]*, Allie Haq[2], David P. Black[2] and James L. Pease[1]

[1]Department of Social Work, College of Allied Health Sciences, University of Cincinnati, Cincinnati, OH, United States, [2]Clarigent Health, Mason, OH, United States

**Background:** Current depression, anxiety, and suicide screening techniques rely on retrospective patient reported symptoms to standardized scales. A qualitative approach to screening combined with the innovation of natural language processing (NLP) and machine learning (ML) methods have shown promise to enhance person-centeredness while detecting depression, anxiety, and suicide risk from in-the-moment patient language derived from an open-ended brief interview.

**Objective:** To evaluate the performance of NLP/ML models to identify depression, anxiety, and suicide risk from a single 5–10-min semi-structured interview with a large, national sample.

**Method:** Two thousand four hundred sixteen interviews were conducted with 1,433 participants over a teleconference platform, with 861 (35.6%), 863 (35.7%), and 838 (34.7%) sessions screening positive for depression, anxiety, and suicide risk, respectively. Participants completed an interview over a teleconference platform to collect language about the participants' feelings and emotional state. Logistic regression (LR), support vector machine (SVM), and extreme gradient boosting (XGB) models were trained for each condition using term frequency-inverse document frequency features from the participants' language. Models were primarily evaluated with the area under the receiver operating characteristic curve (AUC).

**Results:** The best discriminative ability was found when identifying depression with an SVM model (AUC=0.77; 95% CI=0.75–0.79), followed by anxiety with an LR model (AUC=0.74; 95% CI=0.72–0.76), and an SVM for suicide risk (AUC=0.70; 95% CI=0.68–0.72). Model performance was generally best with more severe depression, anxiety, or suicide risk. Performance improved when individuals with lifetime but no suicide risk in the past 3 months were considered controls.

**Conclusion:** It is feasible to use a virtual platform to simultaneously screen for depression, anxiety, and suicide risk using a 5-to-10-min interview. The NLP/ML models performed with good discrimination in the identification of depression, anxiety, and suicide risk. Although the utility of suicide risk classification in clinical settings is still undetermined and suicide risk classification had the lowest performance, the result taken together with the qualitative responses from the interview can better inform clinical decision-making by providing additional drivers associated with suicide risk.

KEYWORDS

suicide, depression, anxiety, machine learning, natural language processing, risk assessment, mental health, virtual screening

# 1. Introduction

Each year in the United States (US), more than 47,000 people die by suicide (1). Additionally, based on recent survey data from the US Census Bureau, 28.2% of adults endorsed symptoms of anxiety, 24.4% reported symptoms of depression, and 33.9% suffered from one or both conditions in the past 7 days (2). To address the growing rates of comorbid mental health conditions, there is a need for a singular, patient-centered, accurate, reliable, and objective tool to simultaneously identify patients at risk of suicide and other mental health disorders.

Universal screening tools deployed in a wide spectrum of facilities, including schools, physicians' offices, outpatient, and inpatient facilities could address this problem, but the lack of a person-centered and objective tool along with a shortage of mental health clinicians in these settings is a major barrier to screening for coexisting depression, anxiety, and suicide risk on a public health scale. A common screening procedure involves filling out paper and pencil individual screeners for depression, anxiety, and suicide risk that are selected by the particular healthcare facility. Some common screeners include the nine-item Patient Health Questionnaire (PHQ-9), and the seven-item Generalized Anxiety Disorder (GAD-7), and the Columbia Suicide Severity Scale (C-SSRS), however, multiple others may be used, resulting in a lack of uniformity in scale and administration approach across settings. This method of screening does not allow for engagement between practitioner and patient, nor does it give space for a nuanced conversation about mental health, suicide risk and related patient needs. Also, even when screening instruments are part of a clinic's protocol, they may not be consistently administered due to time constraints as separate instruments are required for each mental health condition (3). Further, these methods can be subject to self-report or clinician rating bias. Employing a brief, qualitative interview that collects the patient's own words could fill a gap in current screening techniques by giving space for patients to discuss their needs ahead of crisis clinical decision-making.

Screening methods for depression, anxiety, and suicide risk have begun to shift in recent years as telehealth and other digital platforms have become increasingly prevalent. The trend towards using virtual methods for screening have gained momentum as the COVID-19 pandemic complicated in-person healthcare visits. Therefore, healthcare service users have become more aware of, and amenable to, accessing screening and treatment options virtually (4).

Natural language processing (NLP) and machine learning (ML) have expanded how mental health conditions can be identified (5, 6). Prior research from the Pestian lab undergirding the methods in this study used a corpus of suicide notes to train ML models (7, 8). From this, investigators developed an interview, called the Ubiquitous Questionnaire (UQ), to obtain language samples to further test the models in two clinical trials (9). The Adolescent Controlled Trial validated the ML model using the C-SSRS with 60 randomly selected emergency department cases (suicide complaints) and controls (orthopedic patients) (10). The model was able to correctly classify 97% of the participants as cases or controls. The second trial, Suicide Thought Markers (STM), randomly selected 379 adolescents and adults from mental health, suicide complaint, and control groups across three study sites. Results from the STM study indicated the model was able to identify the suicide group with 85% accuracy (11). Since the work in Pestian's lab, innovations in NLP have demonstrated its screening efficiency and scalability in clinical and public health settings. Recent studies highlight the feasibility and clinical acceptability of using a digital platform to collect data through a 5-to-10-min interview for NLP analysis to identify suicide risk (12). Additionally, NLP models can perform well despite speakers' varied location and regional dialect (13) making this method geographically portable (6).

Clairity, a depression, anxiety, and suicide risk screening program, uses NLP to identify all three conditions with a single brief interview. The purpose of this study was to (1) evaluate the feasibility of using a virtual platform to collect brief interviews for NLP analysis, and (2) to validate the ML models against the most widely used standardized instruments in a large, national sample. We also highlight the argument that a qualitative approach to screening is necessary to identify patient needs related to risk early in order to form a collaborative relationship and inform next steps in crisis and treatment planning. Given findings by Carter et al. (14, 15), guidance issued by the United Kingdom's National Institute for Health and Care Excellence [NICE; (16)], and the call to action in determining how addressing patient needs is critical to preventing suicide, reducing risk, and improving quality of life, we propose that this method fills this gap by incorporating both a clinically-useful open-ended conversation and objective machine learning risk detection.

# 2. Methods

## 2.1. Study staff and participants

The study staff was composed of 18 clinical research coordinators (CRC). The CRCs completed online training to learn study procedures, principles of human subject protection, and good clinical practice. The CRCs oversaw all study procedures and were supervised by the clinical principal investigator.

Criteria for participant recruitment were: (1) age ≥ 18, (2) able to provide informed consent, and (3) English as a primary language. ResearchMatch (RM) was used to recruit for this study. RM is a national health volunteer registry created by several academic institutions and supported by the US National Institutes of Health as part of the Clinical Translational Science Award program. RM has an extensive pool of volunteers who have consented to be contacted by researchers about health studies for which they may be eligible. Approval for this study and all procedures was granted by a commercial Institutional Review Board. Participants received a $15 gift card for each session they completed (Figure 1).

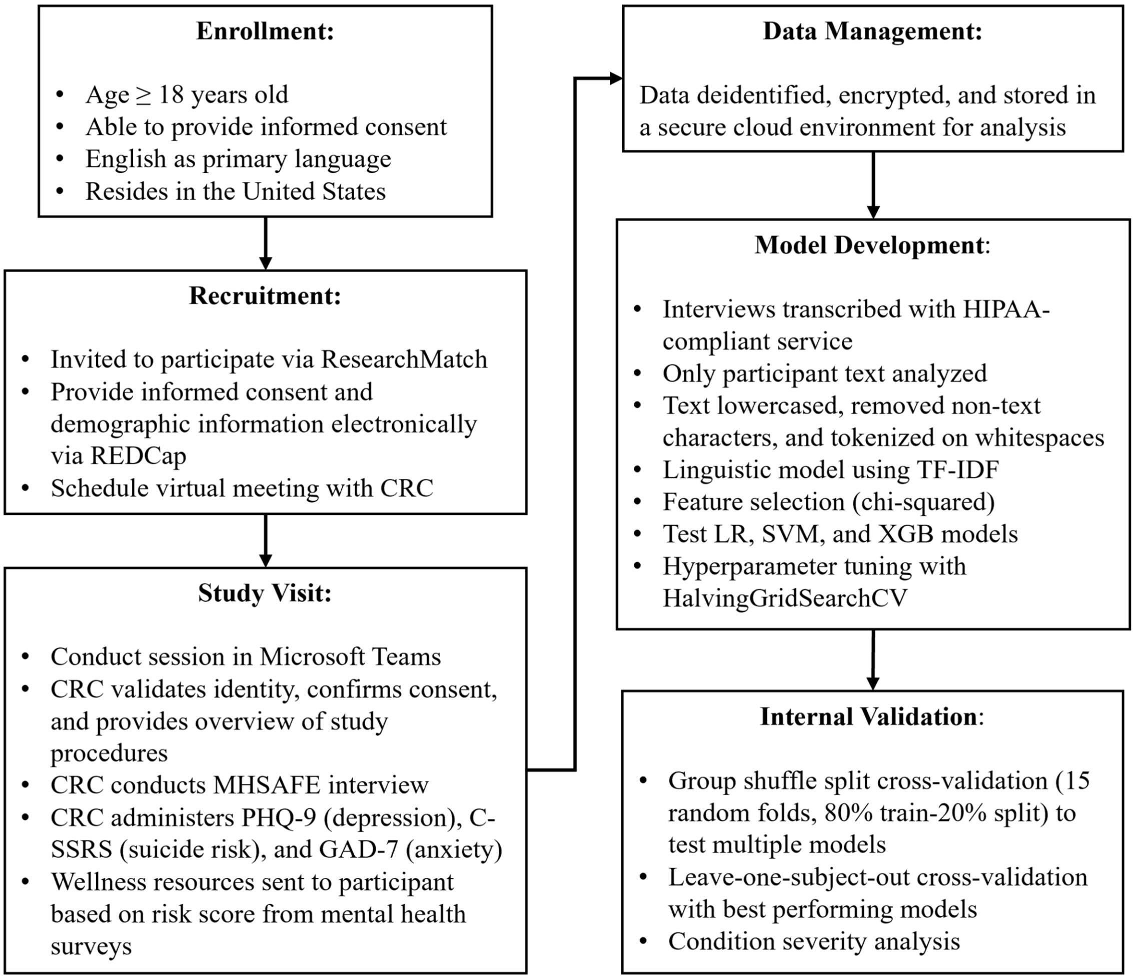

Enrollment:
• Age ≥ 18 years old
• Able to provide informed consent
• English as primary language
• Resides in the United States
Recruitment:
• Invited to participate via ResearchMatch
• Provide informed consent and demographic information electronically via REDCap
• Schedule virtual meeting with CRC
Study Visit:
• Conduct session in Microsoft Teams
• CRC validates identity, confirms consent, and provides overview of study procedures
• CRC conducts MHSAFE interview
• CRC administers PHQ-9 (depression), C-SSRS (suicide risk), and GAD-7 (anxiety)
• Wellness resources sent to participant based on risk score from mental health surveys
Data Management:
Data deidentified, encrypted, and stored in a secure cloud environment for analysis
Model Development:
• Interviews transcribed with HIPAA-compliant service
• Only participant text analyzed
• Text lowercased, removed non-text characters, and tokenized on whitespaces
• Linguistic model using TF-IDF
• Feature selection (chi-squared)
• Test LR, SVM, and XGB models
• Hyperparameter tuning with HalvingGridSearchCV
Internal Validation:
• Group shuffle split cross-validation (15 random folds, 80% train-20% split) to test multiple models
• Leave-one-subject-out cross-validation with best performing models
• Condition severity analysis

FIGURE 1
Schematic of study and modeling procedures.

Once matched, the participant completed the informed consent process, provided demographic information, and selected a session time via an online calendar system. Prior to the session, participants were sent reminders of their session by email and text. Microsoft Teams was used for all interviews.

## 2.2. Study design

Prior to the interview, participants' identities were verified, and the CRC provided a brief overview of the study. Consent was confirmed and the CRC began the recording process. The CRC completed the 5-to-10-minute interview during which the CRC asked about the participant's hopes, secrets, anger, fear, and emotional pain (MHSAFE). The MHSAFE interview is composed of standard prompts based on Pestian's Ubiquitous Questionnaire, developed and tested to elicit emotional language for the screener (9–13, 17).

Survey data collected during the interview included the PHQ-9 (Patient Health Questionnaire-9 item), C-SSRS (Columbia-Suicide Severity Rating Scale) Screener, and GAD-7 (General Anxiety Disorder-7 item) for use in validation of the ML models and to produce a risk score. The resulting risk score prompted the CRC to follow the contingency and safety plan based on the participant's identified risk level. Upon completion of the study, the participant was notified they may participate up to two more times. Participants scoring moderate or high risk were provided with resources including the 988 Suicide & Crisis Lifeline, the Crisis Text Line, and other tools such as the Stanley Brown Safety Plan (18). If the participant scored "high risk" on the mental health surveys, a more comprehensive contingency plan was followed, including asking additional questions about their mental state, access to lethal means, engagement in mental health services, and protective factors. In the event of imminent risk, the contingency safety plan included a warm hand-off to the 988 Suicide & Crisis Lifeline and/or a call to 911. To date, only one call to 911 was required during the study. This participant returned to complete additional interviews and was reported as safe.

Table 1 outlines the thresholds for each mental health condition. The PHQ-9 is a nine-item depression screener and is part of the full-length PHQ. The total score of the nine items ranges from 0 to 27, with a score of 10 used as a depression cut-off score. In a study conducted by Kroenke et al., a score of 10 or higher in the PHQ-9 had high sensitivity and specificity (88%) for detecting depression (19). The findings of this study were externally validated among different patient populations. The GAD-7 is a seven-item anxiety screener. In a

TABLE 1 Assessments and case definitions.

| Condition | Assessment | Case definition |
|---|---|---|
| Depression | PHQ-9 | Total ≥ 10 |
| Anxiety | GAD-7 | Total ≥ 10 |
| Suicide risk | C-SSRS | Risk ≥ Low |

reliability and validity study performed by Spitzer et al., various total cut-off points were analyzed for sensitivity, specificity, and validity. As the cut-off point increased, sensitivity decreases and specificity increases. However, at a total score of 10 or higher, sensitivity and specificity exceed 80% (20). Therefore, a score of 10 indicates a cut-off point for identifying anxiety cases.

The C-SSRS Screener is a structured interview based on the full-length version (21). The first five questions measure suicidal ideation and behaviors in the past month on an ordinal scale. The last question measures suicidal behavior that occurred either in the past 3 months or have ever occurred over the lifetime The C-SSRS Screener designates a participant's suicide risk level as "None" if all answers are negative, "Low" if there are non-specific suicidal ideations, "Moderate" if there is a method along with suicide ideation, or if there was lifetime suicidal behavior, "High" if there is active suicidal ideation with specific plan and intent, or if there was suicidal behavior within the past 3 months. For this study, a case is defined as someone who scores "Low" suicide risk or higher.

## 2.3. Data analysis

All analyses were performed using the Python programming language [version 3.9.12; (22)]. The open-source Python libraries Pandas [version 1.4.2; (23)], NumPy [version 1.22.3; (24, 25)], Scikitlearn [version 1.0.2; (26)], Matplotlib [version 3.5.1; (27)], and SciPy [version 1.8.0; (28)] were also used. Student's *t*-tests were performed with SciPy's *ttest_ind* function.

### 2.3.1. Natural language processing and model development

The NLP/ML pipeline used in this study followed similar techniques used in previous work (10–13), focused on the term frequency-inverse document frequency (TF-IDF) of n-grams (contiguous sequence of n number of words). The text was preprocessed to be all lowercase and to remove any punctuation and non-letter characters. The text was tokenized with a simple whitespace tokenizer. Scikit-learn's *SelectKBest* function was used for feature selection to identify features with the highest chi-square value. The ngram values (e.g., unigrams, bigrams, or trigrams) and the number of features selected were tunable hyperparameters.

We explored performance of three different models including logistic regression (LR), support vector machines (SVM), and extreme gradient boosting (XGB). LR is one of the simplest machine learning models yet still provides acceptable performance. SVMs have demonstrated excellent performance in previous tasks classifying suicidal language from semi-structured interviews, resist overfitting, and perform well in high-dimensional spaces (10–13). XGB has given state-of-the-art results on various problems and displayed promising results in a previous study (12, 29). Models were tuned using Scikit-learn's *HalvingGridSearch* function with a stratified 5-fold cross-validation (CV) technique with non-overlapping subjects. Considered hyperparameters are available in Supplementary Table S1.

### 2.3.2. Internal validation and performance evaluation

Initial model performance estimates were made using a group shuffle split (GSS) CV technique, where the dataset is broken into 15 randomly selected 80% train-20% test groups with non-overlapping subjects. We set the random state of the CV iterator to ensure consistent folds across experiments. This internal validation technique provides a more efficient estimate of model performance over a leave-one-subject-out (LOSO) CV technique. During model training, the only input was the participant's language, labeled as case or control as defined in Table 1. During model testing, participant language was fed into the model and a probability for belonging to the case group was returned. Model performance was then determined by comparing the model predictions to the participant's labeled group. At this stage, models were evaluated by the area under their receiver operating characteristic curves (AUC).

Models with the best GSS performance were then evaluated with a LOSO CV technique, where a model is iteratively trained on all but one subject's sessions, and then makes a prediction on the held-out subject's sessions. Because only one subject's sessions are held out per CV fold, the model is the closest possible approximation to when it is trained on the full corpus. For results of LOSO CV, model performance was primarily evaluated with the AUC and Brier score. AUC values range from 0.5 (random chance) to 1.0 (perfect model). The Brier score is a measure of model calibration and ranges from 0 to 1 where a low score indicate less discrepancy between labels and predicted probabilities. Additional classification metrics calculated include accuracy, sensitivity, specificity, positive predictive value (ppv), and negative predictive value (npv). Thresholds for classification were determined as the maximum of sensitivity and specificity.

Feature weights were extracted for the best performing linear models. Feature weights for SVM models with a radial basis function (RBF), are not easily accessible, therefore the next best performing linear or tree-based model was used to identify important features. For linear models (LR or SVM with a linear kernel), feature weights are either positive or negative, indicating if they contribute to the model predicting a case or control, respectively.

### 2.3.3. Model performance and condition severity

During model training, cases and controls were defined for each condition by accepted thresholds, shown in Table 1. Each instrument also has different severity levels for each condition based on the total scores. For the PHQ-9, severity increases for every 5 points of the total score, ranging from "None" (0–4), "Mild" (5–9), "Moderate" (10–14), "Moderately Severe" (15–19), and "Severe" (≥20). The GAD-7 follows the same severity levels, except there is no Moderately Severe bin, with scores ≥15 classified as "Severe." For the C-SSRS, "None" results from negative answers to all questions; "Low" risk is characterized by passive suicidal ideation (SI); "Medium" risk by SI with methods or suicidal behavior longer than 3 months ago (lifetime); and "High" risk by suicidal intent with or without a plan, or suicidal behavior in the past 3 months.

Using the results from the LOSO CV for each condition, we computed model performance metrics for different severity levels.

In this case the model was not retrained; the performance was computed by only selecting control sessions and the specified severity. For example, for the PHQ-9, "None" and "Mild" severities are considered controls, and the model's performance for discriminating between "None" or "Mild" vs. "Moderate" depression was estimated by only considering sessions from the LOSO results where the severity is "None," "Mild," or "Moderate."

# 3. Results

Between November 2020 and August 2022, 1,433 participants were enrolled. Participants attended 1–3 sessions, which resulted in a total of 2,416 recorded sessions. The PHQ-9, GAD-7, and C-SSRS screeners were collected in all sessions. Participant demographics and the results of the mental health assessments are found in Table 2. Of the 2,416 sessions, 1,361 (56.3%) were classified as at least one case session. Figure 2 shows a Venn diagram of case session overlap. Most case sessions (28%) were positive for all three conditions, while those positive for both depression and anxiety had the next most overlap (16%). Those positive for only suicidal risk (17%) made up the largest number of sessions positive for a single condition.

Demographic information was collected during the informed consent process. Participants were primarily female (79%, *N* = 1,132) and Caucasian (80%, *N* = 1,146). Other races represented in the sample were African American (8.8%, *N* = 126), Asian (5.9%, *N* = 84), Native Hawaiian or Other Pacific Islander (0.2%, *N* = 3) and Other (4%, *N* = 58). Over 7% of the sample reported a Hispanic ethnicity (7.4%, *N* = 106). The mean age of the sample was 39 ± 13.8 years. Of the 2,416 sessions, the average interview length was 8.3 min with a mean of 856.9 words per session. *t*-tests yielded no significant difference between case and control participants for interview length and word count for all mental health conditions.

## 3.1. Internal validation

### 3.1.1. Group shuffle split

Table 3 displays AUCs and standard deviations (SD) across the GSS CV folds for each tuned model and condition. We found the best discrimination for the identification of depression (AUC = 0.76 ± 0.02) with LR and SVM (RBF kernel) models, using 2,048 and 1,024 features, respectively. We found good performance identifying anxiety (AUC = 0.74 ± 0.02) with LR and SVM (RBF kernel), with both models using 2048 features. Suicide had the lowest AUC of the three conditions (AUC = 0.70 ± 0.02), with an SVM (linear kernel) performing the best with 1,024 features. Before rounding, we found the SVM model performed slightly better for depression and suicide risk, while LR performed better for anxiety. Despite promising performance with XGB models in the past (12), XGB consistently had the lowest AUC. The optimal hyperparameters from the HalvingGridSearch for all the classifiers can be found in Supplementary Table S2.

TABLE 2 Participant descriptive statistics and case session summaries.

| | | | Case sessions* | | |
|---|---|---|---|---|---|
| | Participants | Sessions | PHQ-9≥10 | GAD ≥10 | CSSRS ≥ Low |
| Count (%) | 1,433 (100%) | 2,416 (100%) | 861 (35.6%) | 863 (35.7%) | 838 (34.7%) |
| Average word count (SD) | – | 856.9 (589.0) | 830.5 (590.2) | 856.7 (618.0) | 854.1 (604.7) |
| Average interview length (min) (SD) | – | 8.3 (4.6) | 8.2 (4.4) | 8.3 (4.9) | 8.4 (4.7) |
| Average age (SD) | 39.0 (13.8) | 39.3 (14.8) | 38.5 (13.5) | 36.6 (12.7) | 38.3 (14.2) |
| **Sex** | | | | | |
| Female (%) | 1,132 (79.0%) | 1922 (79.5%) | 715 (29.6%) | 724 (30.0%) | 691 (28.6%) |
| Male (%) | 284 (19.8%) | 467 (19.3%) | 131 (5.4%) | 125 (5.2%) | 130 (5.4%) |
| Prefer not to answer (%) | 10 (0.7%) | 17 (0.7%) | 7 (0.3%) | 5 (0.2%) | 10 (0.4%) |
| Other (%) | 5 (0.3%) | 8 (0.3%) | 8 (0.3%) | 8 (0.3%) | 5 (0.2%) |
| **Race** | | | | | |
| American Indian or Alaska Native (%) | 14 (0.9%) | 28 (1.2%) | 12 (0.5%) | 14 (0.6%) | 10 (0.4%) |
| Asian (%) | 84 (5.9%) | 175 (7.2%) | 50 (2.1%) | 51 (2.1%) | 62 (2.6%) |
| Black or African American (%) | 126 (8.8%) | 220 (9.1%) | 80 (3.3%) | 76 (3.1%) | 66 (2.7%) |
| White or Caucasian (%) | 1,146 (80.0%) | 1,891 (78.3%) | 678 (28.1%) | 682 (28.2%) | 658 (27.2%) |
| Native Hawaiian or Other Pacific Islander (%) | 3 (0.2%) | 5 (0.2%) | 2 (0.1%) | 1 (0.04%) | 1 (0.04%) |
| Other (%) | 58 (4.0%) | 95 (3.9%) | 39 (1.6%) | 38 (1.6%) | 40 (1.7%) |
| **Ethnicity** | | | | | |
| Non-Hispanic (%) | 1,325 (92.5%) | 2,245 (92.9%) | 790 (32.7%) | 784 (32.5%) | 773 (32.0%) |
| Hispanic (%) | 106 (7.4%) | 169 (7.0%) | 71 (2.9%) | 78 (3.2%) | 64 (2.6%) |

*Percentages based on total number of sessions.

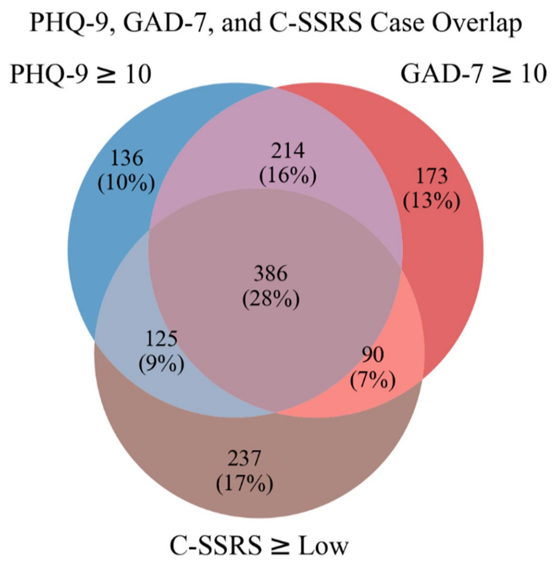


FIGURE 2
Venn diagram of case sessions.

TABLE 3 Model AUC per condition from GSS CV.

| | Model AUC (SD) | | |
|---|---|---|---|
| | LR | SVM | XGB |
| Depression | 0.76 (0.02) | 0.76 (0.02) | 0.71 (0.02) |
| Anxiety | 0.74 (0.02) | 0.74 (0.02) | 0.69 (0.02) |
| Suicide | 0.70 (0.03) | 0.70 (0.02) | 0.66 (0.03) |

### 3.1.2. Leave-one-subject-out

Table 4 shows results of the LOSO CV for the best performing model during GSS CV for each condition. The performance estimates using GSS were consistent with these results, although we saw a 0.01 increase in AUC when identifying depression. The Brier scores ranged from 0.19 to 0.20 and the model score thresholds that optimized the sum of sensitivity and specificity ranged from 0.30 to 0.34.

Table 5 shows the top 10 case and control features by feature weight for the best performing linear models fit to the entire dataset for each condition. Features from the LR model were used for depression because the feature weights from the better performing SVM with an RBF kernel were not readily accessible.

### 3.1.3. Condition severity

Table 6 shows classification performance metrics from the LOSO experiments broken out by severity level for each condition. In general, we saw a decrease in the number of case sessions and an increase in most model performance metrics as a condition's severity increased. The only performance metric that did not improve with increasing severity is PPV, which is correlated with prevalence (30).

We saw a slight decrease in model performance for moderate suicide risk compared to low suicide risk. This may be related to the lookback time of how the C-SSRS classifies moderate risk, where there are two paths to be classified as moderate risk: (1) suicidal ideation (SI) with a method in the past month and (2) suicidal behaviors >3 months ago (lifetime). Of the 447 moderate suicide risk sessions, only 84 (18.8%) answered positively to the SI with method question, while 267 sessions (59.7%) answered negatively to all C-SSRS questions except the lifetime suicidal behavior question. If the performance metrics from the LOSO results for moderate suicide risk are recomputed by considering only those with SI with a method as cases (84 sessions) and the 267 sessions whose suicidal risk was >3 months ago as additional controls, we find a 0.06 increase in AUC (AUC = 0.73; 95% CI = 0.67–0.79), indicating the model tends to score those whose risk was >3 months ago lower than those whose risk is more recent.

## 4. Discussion

The purposes of this study were to demonstrate feasibility of using a virtual platform to collect language data to screen for depression, anxiety, and suicide risk, and to validate machine learning models with a large and diverse national sample.

A recent meta-analysis of studies evaluating the use of technology to address mental health disorders found innovation in screening and mental health treatment has become widespread (31). Additionally, during the COVID-19 pandemic, telehealth and similar web-based methods were increasingly used for both research and treatment of mental disorders (32). However, to date, no single screening technique that simultaneously identifies depression, anxiety, and suicide risk has been developed or tested. Clairity is a program that analyzes linguistic features, collected via language samples, to identify all three disorders from a 5–10-min interview. We found using a virtual platform to conduct this study feasible, as we were able to recruit a large and diverse national sample of participants representing a range of people affected by depression, anxiety, and suicide risk, and comorbidities of the three disorders. This may be of particular importance to the medical and mental health field as the rise in prevalence of mental health disorders (33) warrants screening tools that are both portable and efficient.

Results from this study demonstrate both the efficiency and effectiveness of the Clairity program. First, the average interview time across participants was 8.3 min. In order to screen for these disorders using self-report standardized scales, patients in medical and mental health settings would fill out three separate scales, selecting responses that are the "best fit" for their current mental and suicide risk status. Time to complete these scales is one of the often-cited reasons why they are not completed consistently in settings such as primary care. Additionally, comorbidities (Figure 2) were identified using separate self-report standardized instruments. This type of screening procedure is inefficient and requires clinician scoring and interpretation of separate scales. A brief interview which identifies all three conditions simultaneously with a quick return of results may mitigate time constraints, patient (self)- or clinician-reporting bias, and clinician-nuanced scoring and lack of interpretation expertise (34–36). Lastly, a single interview also eliminates the "one size fits all" standardized screening response options.

The portability of the Clairity program allows for scalability as current mental health resources are stretched thin (37). Clairity is accessed via a web-based platform, allowing service users to complete the interview in various settings. For this study, participants were able to use their phones, desktop or laptop, or tablet, requiring an internet connection. CRCs were able to access results from the scales and

TABLE 4 Model performance summary from LOSO results.

| Condition | Depression | Anxiety | Suicide |
|---|---|---|---|
| Best model | SVM (rbf) | LR (liblinear) | SVM (linear) |
| AUC (95% CI) | 0.77 (0.75–0.79) | 0.74 (0.72–0.76) | 0.70 (0.68–0.72) |
| Number of sessions | 2,416 | 2,416 | 2,416 |
| Number of cases | 861 | 863 | 838 |
| Brier score | 0.19 | 0.20 | 0.20 |
| Threshold | 0.34 | 0.34 | 0.30 |
| Sensitivity (95% CI) | 0.72 (0.69–0.75) | 0.57 (0.53–0.660) | 0.70 (0.67–0.73) |
| Specificity (95% CI) | 0.68 (0.65–0.70) | 0.77 (0.75–0.79) | 0.58 (0.55–0.60) |
| NPV (95% CI) | 0.82 (0.79–0.84) | 0.76 (0.74–0.78) | 0.78 (0.76–0.81) |
| PPV (95% CI) | 0.55 (0.53–0.58) | 0.57 (0.54–0.61) | 0.47 (0.44–0.50) |
| Accuracy (95% CI) | 0.69 (0.68–0.71) | 0.70 (0.68–0.71) | 0.62 (0.60–0.64) |

TABLE 5 Top 10 features for best performing models.

| | Case features | Control features |
|---|---|---|
| Depression | cant, afraid, im, myself, get, just, mental, because, therapy, depression | no, think, good, we, about, well, new, say, so, worry |
| Anxiety | definitely, afraid, im, myself, because, therapy, anxiety, just, me, lot | no, think, not, guess, but, about, new, good, in, share |
| Suicide | its, feels, ive, want, yeah, suicidal, therapy, coping, mental, therapist | vaccinated, think, hold, would, no, hopeful, hopefully, school, you, now |

follow-up with high-risk participants, providing resources including crisis lines and safety plans. In a clinic setting, service providers may elect to use Clairity in office, during a home visit, between visits, or a service user could access the program autonomously at home or in another private setting. Once the interview has been completed, results are then sent to the provider within seconds, with data from the three conditions being displayed in an accessible and intuitive dashboard.

The qualitative and machine learning output data available from the Clairity dashboard provides additional insights for clinicians when making collaborative decisions with patients about next steps in care. The language features (Table 5) related to depression, anxiety, and suicide risk offer clinical information not gleaned via standardized instruments in the form of thought markers. These thought markers (patient's natural language) can aid in understanding the patient's idiosyncratic risk identifiers. With further study of these features in a clinical setting, we can learn more about how these can be used in clinical decision-making and patient risk monitoring. Future studies will include how clinicians can use language features to inform risk levels and changes over time. Additionally, the information gathered through the qualitative interview will provide the clinician with patient-specific details about drivers and needs related to risk. Patient stories, of hope, anger, secrets, fears, and emotional pain can begin to paint the picture of the patient's life experiences leading to their current mental state.

We found the best model performance identifying depression, followed by anxiety, with the poorest performance identifying any suicide risk. This is surprising considering the questions asked in the interview were originally developed with patients admitted to an emergency department for a suicide attempt or severe ideation, with model AUCs ranging from 0.69 to 0.93 depending on the features and CV method used. One possible explanation is that model performance, and consequently the separability of cases and controls, is related to condition severity. Indeed, Table 6 shows model performance tends to improve as the difference in condition severity between cases and controls increases for all conditions; one can imagine the classification task between "None" and "Severe" depression easier than between "Mild" and "Moderate" depression. This is noteworthy as the results from Table 6 did not include retraining the models, therefore the models did not have any explicit information about condition severity yet tended to rate more severe cases higher. In general, the PHQ-9, GAD-7, and C-SSRS Screener follow a linear progression, where more frequent or intense symptoms lead to higher severity classifications, except for the case of "Moderate" suicide risk, where this designation is still possible on the C-SSRS in the absence of any recent SI. Interestingly, this is where we observed the greatest discrepancy between the model and any mental health survey, and model performance improves when the language of those without any recent SI are considered controls. While the relationship between SI, lifetime suicidal behaviors, and risk designations is complex, the model evaluated in this study tends to rate those with lifetime suicidal behavior but no recent SI closer to controls. A person's language with lifetime but no recent SI may accurately be classified as a control (low or no risk) by the model when their mental state reflects no *current or imminent* suicide risk thought or intention although the C-SSRS rates them as "Moderate" suicide risk.

Due to the low prevalence of suicide death, this form of classification of suicide risk has been met with scrutiny and evidence of utility in real-world clinical settings has been called into question by Carter et al. (14, 15). Given the modest predictive values of suicide risk screenings, Carter and colleagues warned that other ways to identify suicide risk are warranted. Although Clairity uses ML methods to provide data and a risk classification for suicide, the qualitative interview is intended to create a more useful path for communication

TABLE 6 Model performance for different condition severities.

| | Depression | | | Anxiety | | Suicide | | |
|---|---|---|---|---|---|---|---|---|
| Severity | Mod. | Mod. Severe | Severe | Mod. | Severe | Low | Moderate | High |
| Sample size | 2,071 | 1,793 | 1,662 | 2,069 | 1,900 | 1,896 | 2,025 | 1,651 |
| Number of cases | 516 | 238 | 107 | 516 | 347 | 318 | 447 | 73 |
| AUC (95% CI) | 0.72 (0.69–0.74) | 0.80 (0.77–0.83) | 0.92 (0.89–0.94) | 0.69 (0.66–0.72) | 0.81 (0.79–0.84) | 0.70 (0.67–0.73) | 0.67 (0.64–0.70) | 0.85 (0.81–0.89) |
| Brier score | 0.18 | 0.14 | 0.12 | 0.21 | 0.19 | 0.16 | 0.18 | 0.13 |
| Threshold | 0.27 | 0.33 | 0.44 | 0.44 | 0.53 | 0.35 | 0.27 | 0.44 |
| Sensitivity (95% CI) | 0.74 (0.70–0.78) | 0.83 (0.77–0.87) | 0.90 (0.83–0.94) | 0.72 (0.67–0.75) | 0.69 (0.64–0.74) | 0.64 (0.59–0.70) | 0.71 (0.67–0.75) | 0.75 (0.64–0.84) |
| Specificity (95% CI) | 0.60 (0.57–0.62) | 0.66 (0.64–0.68) | 0.78 (0.76–0.80) | 0.55 (0.53–0.58) | 0.77 (0.75–0.79) | 0.64 (0.62–0.67) | 0.53 (0.5–0.55) | 0.77 (0.75–0.79) |
| NPV (95% CI) | 0.87 (0.85–0.89) | 0.96 (0.96–0.97) | 0.99 (0.98–0.99) | 0.85 (0.83–0.87) | 0.92 (0.90–0.93) | 0.90 (0.88–0.92) | 0.87 (0.84–0.89) | 0.99 (0.98–0.99) |
| PPV (95% CI) | 0.38 (0.35–0.41) | 0.27 (0.24–0.31) | 0.22 (0.19–0.26) | 0.35 (0.32–0.37) | 0.40 (0.36–0.44) | 0.27 (0.24–0.30) | 0.30 (0.27–0.33) | 0.13 (0.10–0.17) |
| Accuracy (95% CI) | 0.63 (0.61–0.65) | 0.68 (0.66–0.70) | 0.79 (0.77–0.81) | 0.59 (0.57–0.61) | 0.75 (0.73–0.77) | 0.64 (0.62–0.67) | 0.57 (0.55–0.59) | 0.77 (0.75–0.79) |

between the patient and the interviewer. Where most, if not all, commonly used suicide screening techniques employ standardized scales, Clairity offers a new alternative to begin the conversation about risk, providing a patient-centered, non-confrontational exchange in an otherwise potentially volatile interaction. Additionally, Carter et al. suggest that modifiable factors, such as stressors or drivers of suicide, be assessed and addressed in treatment to reduce risk of suicide (14, 15). This is where Clairity's "front door" approach may be particularly useful. By having an open-ended brief risk screening, respondents may disclose these suicide risk factors earlier, allowing for a seamless transition to critical next steps for safety and addressing issues related to suicide risk and treatment. However, additional exploration of how to offer a risk result is warranted, as the current classification system of "low-," "moderate-," and "high-" risk does not allow for the nuance of needs of individual patients (14, 15). The aim of the Clairity program is to disrupt this potentially ineffective system with a screening tool that provides critical information at the time of screening through qualitative responses.

Of the ML models tested during GSS CV, we found the tuned LR and SVM models to have similar performance. Surprisingly, XGB models had the poorest performance for all three conditions despite performing well in previous work and providing excellent results with other classification tasks (12, 29). It is possible additional hyperparameter tuning or feature scaling techniques may improve XGB's classification performance, although it may also be XGB models are not ideal for classification tasks with the high dimensional, sparse matrices used in this study.

During LOSO CV, we found agreement with the GSS CV results, indicating GSS CV was a reasonable method to estimate model performance on the entire dataset. The classification thresholds for the calculation of the additional performance metrics shown in Table 4 were determined as the value that maximized the sum of sensitivity and specificity. Coincidentally, these thresholds are within a few percentage points of the percentage of case sessions for each condition for this slightly imbalanced dataset. There are many methods to determine a classification threshold, and the selection of one will depend on the specific clinical context and an appropriate balance between the cost of false positives and false negatives of the classification task. A recent analysis by Ross et al. examined accuracy requirements for cost-effective suicide risk prediction in US primary care patients and found the two interventions examined – active contact and follow-up (ACF) and cognitive behavioral therapy (CBT) – cost-effective, provided the model performed with a specificity of 95.0% and a sensitivity of 17.0% for ACF and 35.7% for CBT in predicting a suicide *attempt* (38).

Our reference instruments report sensitivity and specificity values exceeding 80%, while our models' performance estimates are lower, as observed in Table 4. However, caution is necessary when drawing comparisons based solely on performance estimates. All performance metrics, whether applied to an ML model or traditional instrument, reflect estimations of the expected performance of the tool in real-world contexts. Thus, it is important to consider the method by which the performance metrics were calculated before making any comparisons. For instance, the initial validation study of the PHQ-9 by Kroenke et al. involved mental health professionals interviewing 580 primary care patients, of whom 41 (7%) were identified as having major depressive disorder (MDD) (19). These cases formed the basis for the sensitivity and specificity estimates of the PHQ-9. Similarly, we found that our models' performance improved as the severity of depression and the likelihood of MDD increased. Furthermore, the classification of depression is complicated by the "gray zone" (a range of depression severity not easily classified as either case or control), which poses a challenge for both the PHQ-9 and our models (19). Thus, the classification task is more challenging when patients fall within this gray zone. Lastly, while our models used a relatively large sample size, accurately representing the complexity and nuance of language remains a challenge. The TF-IDF approach we used in our study provides a simplified representation of language. Thus, more advanced NLP techniques may improve model performance by better accounting for the intricacies of natural language.

Table 5 shows the top 10 case and control features by feature weight for the best performing linear models for each condition. While these features represent a small fraction of the total number of features and a full linguistic analysis is beyond the scope of this paper, some interesting patterns emerge. For each condition analyzed, the condition itself appears as a top feature, and for all models, the word "therapy" was a top case feature. Other work has found increased personal pronoun usage related to depression and suicide risk (39), often interpreted as more inwardly-focused. We found the pronouns "im," "me," "myself," and "ive" as part of the top 10 case features across the three models. We found the top features had the greatest overlap between the depression and anxiety models, especially for the control features, which is congruent with the larger overlap of case sessions for these conditions shown in Figure 2. For the suicide risk model, the word "vaccinated" was the top control feature, likely related to the start of the study coinciding with the national COVID-19 vaccination effort.

## 4.1. Limitations and future directions

While these findings agree with previous studies, some limitations should be noted. First, while the PHQ-9, GAD-7, and C-SSRS screeners have reported acceptable validity, recent literature has highlighted the inadequacy of screeners to accurately identify those at risk at the time of screening (35, 40). These screeners were used as our source of ground truth for model development and validation. Thus, any inaccuracies in our ground truth labels would impact our models' performance estimates. The thresholds used in Table 1 to identify each condition were selected to maximize the instrument's sensitivity and specificity, therefore we anticipate any mislabeling to be roughly equal for cases and controls, and the models' true performance to remain within the reported confidence intervals.

Suicide risk is comprised of numerous factors, including personal, environmental, and time (41–44). Therefore, screening methods should be used as an opportunity to not only identify the presence of risk, but also to begin building a safe space for patients to discuss needs related to risk. The use of brief quantitative screeners cannot be relied upon to engage the patients on their needs or the complicated array of contributors to imminent risk. More comprehensive needs assessments, which are the step after screening is implemented, could help explain discrepancies between our model and the C-SSRS, such as our model's tendency to rate individuals with lifetime but no recent SI closer to controls. Additionally, it should be noted that screening tools, such as Clairity, are not intended to fully assess risk and patient needs for determining the plan of care. In light of the work of Carter et al. (14, 15), it is important that the needs of the individual as assessed during comprehensive and narrative interviewing post-screening should form a collaborative and patient-centered treatment course. If suicide risk, depression, and anxiety screenings are used to decide course of action, patients might be subjected to traumatizing and expensive treatment that does not address their specific needs better assessed by exploring the conditions underlying the risk.

There may be some limitations for users when technology is employed to gather screening data. First, the participant or client must have access to a device and the internet. This may be a challenge for some who have limited access to either of these. However, this challenge can be offset if the provider uses Clairity in office. Second, internet instability, noise in the room, or a participant's style of speaking can alter the language sample. For this study, six sessions (0.2%) were eliminated due to these issues. Some users may find a telehealth approach less patient-friendly and personal, however, many patients report benefit of virtual options in terms of accessibility (45). Last, some clinics' adoption of new technology may be slow, even when clinical users accept new and innovative methods of care (46).

Additional limitations include the use of recruited research volunteers who were incentivized to answer the questions in the MHSAFE interview. Patients in clinical settings in which the provider is conducting the interview may respond differently. This could affect the generalizability of the models when applied in other settings. Additional external validation studies could help to identify the extent of this limitation. In a previous emergency department study, we solicited feedback from clinician users of the MHSAFE interview and gleaned their perception of differences in patient risk disclosure when comparing the C-SSRS to Clairity. Clinicians stated the interview was more patient-centered, and patients were more forthcoming during the open-ended approach. They also reported some patients felt "what they were saying was important" and they felt "seen and heard" (47).

Future work will examine both the use of qualitative data in collaborative clinical decision-making related to patient needs and model performance across different participant demographics and settings (e.g., emergency departments and outpatient therapy), how features can be used to identify patterns in thought markers related to risk, and a repeated measures analysis. We also plan to investigate the use of more advanced NLP techniques that leverage large language models and may account for a more linguistic nuance but may also retain biases (48–51). Lastly, we want to explore how the type of return of results meaningfully informs clinical decision-making whereby the qualitative data from the interview are used in the context of the presence of suicide risk to identify what the individuals needs are related to reducing suicide risk and improving wellbeing.

# 5. Conclusions

The results of this large, national study of the use of a virtual platform to conduct mental health and suicide risk screening suggests it is feasible to simultaneously identify depression, anxiety, and suicide risk from a brief qualitative interview. The methods utilized in this study were modified from those used in outpatient therapy and emergency departments, and they might be easily applied to other settings where early detection may improve outcomes. Although suicide risk classification is still tentative in its utility as suggested in the literature and its lower relative model performance, the MHSAFE interview can offer additional insights about risk factors and related patient needs. The qualitative data along with the risk classification can support clinical decisions and set meaningful next steps in motion. Future work will include a randomized controlled trial to study performance in mental health settings with clinical outcomes.

# Data availability statement

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

## Ethics statement

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

## Author contributions

JW-B, JC, and JP wrote the manuscript. JC, AH, and DB performed statistical analysis on the corpus. JC and JW-B are principal investigators of the study Classification and Assessment of Mental Health Performance Using Semantics—Expanded. All authors contributed to the article and approved the submitted version.

## Funding

The authors declare that this study received funding from Clarigent Health. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

## Acknowledgments

The authors would like to thank Lesley Rohlfs and Marci Campbell for their support with study development and implementation, William Haynes and William Nostheide for their technical support, the clinical research coordinators for conducting the study sessions, the participants of the study for their time, and the community of scientists and researchers that provide open-source machine learning tools and education resources.

## Conflict of interest

JW-B and JP are part-time consultants and are sponsored by a grant from Clarigent Health. JC, DB, and AH are employed full-time by Clarigent Health. The above interests do not alter our adherence to Frontiers Media's policies. Clarigent Health did not influence or restrict the submission of this publication.

## 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/fpsyt.2023.1143175/full#supplementary-material

## References

1. WISQARS (Web-based Injury Statistics Query and Reporting System), *National Center for Injury Prevention and Control, Centers for Disease Control and Prevention*. (2020) Available at: https://www.cdc.gov/injury/wisqars/index.html (accessed December 28, 2021).

2. National Center for Health Statistics. Early release of selected mental health estimates based on data from the January–June 2019 National Health Interview Survey. (2019)

3. English I, Campbell DG. Prevalence and characteristics of universal depression screening in U.S. college health centers. *Fam Syst Health*. (2019) 37:131–49. doi: 10.1037/FSH0000411

4. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull*. (2017) 143:187–232. doi: 10.1037/bul0000084

5. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. *J Med Internet Res*. (2021) 23:e15708. doi: 10.2196/15708

6. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digital Med*. (2022) 5:46. doi: 10.1038/s41746-022-00589-7

7. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide note classification using natural language processing: a content analysis. *Biomed Inform Insights*. (2010) 3:BII.S4706. doi: 10.4137/bii.s4706

8. Pestian JP, Matykiewicz P, Linn-Gust M. What's in a note: construction of a suicide note Corpus. *Biomed Inform Insights*. (2012) 5:BII.S10213:BII.S10213–6. doi: 10.4137/bii.s10213

9. Pestian J. A conversation with Edwin Shneidman. *Suicide Life Threat Behav*. (2010) 40:516–23. doi: 10.1521/suli.2010.40.5.516

10. Pestian JP, Grupp-Phelan J, Bretonnel Cohen K, Meyers G, Richey LA, Matykiewicz P, et al. A controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. *Suicide Life Threat Behav*. (2016) 46:154–9. doi: 10.1111/sltb.12180

11. Pestian JP, Sorter M, Connolly B, Bretonnel Cohen K, McCullumsmith C, Gee JT, et al. A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. *Suicide Life Threat Behav*. (2017) 47:112–21. doi: 10.1111/sltb.12312

12. Cohen J, Wright-Berryman J, Rohlfs L, Wright D, Campbell M, Gingrich D, et al. A feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions. *Int J Environ Res Public Health*. (2020) 17:1–17. doi: 10.3390/ijerph17218187

13. Cohen J, Wright-Berryman J, Rohlfs L, Trocinski D, Daniel L, Klatt TW. Integration and validation of a natural language processing machine learning suicide risk prediction model based on open-ended interview language in the emergency department. *Front Digit Health*. (2022) 4:4. doi: 10.3389/FDGTH.2022.818705

14. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *Br J Psychiatry*. (2017) 210:387–95. doi: 10.1192/BJP.BP.116.182717

15. Carter G, Spittal MJ. Suicide risk assessment: risk stratification is not accurate enough to be clinically useful and alternative approaches are needed. *Crisis*. (2018) 39:229–34. doi: 10.1027/0227-5910/a000558

16. Self-harm: assessment, management and preventing recurrence NICE guideline (2022). Available at: www.nice.org.uk/guidance/ng225 (accessed May 17, 2023).

17. Venek V, Scherer S, Morency LP, Rizzo AS, Pestian J. Adolescent suicidal risk assessment in clinician-patient interaction: a study of verbal and acoustic behaviors In: . *2014 IEEE workshop on spoken language technology, SLT 2014 - proceedings*: Institute of Electrical and Electronics Engineers Inc. New York, NY. (2014). 277–82.

18. Stanley B, Brown GK. Safety planning intervention: a brief intervention to mitigate suicide risk. *Cogn Behav Pract*. (2012) 19:256–64. doi: 10.1016/J.CBPRA.2011.01.001

19. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann*. (2002) 32:509–15. doi: 10.3928/0048-5713-20020901-06

20. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*. (2006) 166:1092–7. doi: 10.1001/ARCHINTE.166.10.1092

21. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, et al. The Columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am J Psychiatr*. (2011) 168:1266–77. doi: 10.1176/appi.ajp.2011.10111704

22. van Rossum G. *Python tutorial*. Amsterdam: Stichting Mathematisch Centrum (1995). 1–65.

23. The pandas development team. pandas-dev/pandas: Pandas 1.1.2 (2020). doi: 10.5281/ZENODO.4019559,

24. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng*. (2011) 13:22–30. doi: 10.1109/MCSE.2011.37

25. Oliphant TE. Python for scientific computing. *Comput Sci Eng*. (2007) 9:10–20. doi: 10.1109/MCSE.2007.58

26. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. *Scikit-learn: Machine learning in Python*. (2011). 2825–2830.

27. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. (2007) 9:90–5. doi: 10.1109/MCSE.2007.55

28. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. (2020) 17:261–72. doi: 10.1038/s41592-019-0686-2

29. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA: Association for Computing Machinery (2016). 785–794.

30. Tenney S, Hoffman MR. Prevalence - StatPearls - NCBI bookshelf. (2022). Available at: https://www.ncbi.nlm.nih.gov/books/NBK430867/ (accessed January 3, 2023).

31. Miralles I, Granell C, Díaz-Sanahuja L, van Woensel W, Bretón-López J, Mira A, et al. Smartphone apps for the treatment of mental disorders: systematic review. *JMIR Mhealth Uhealth*. (2020) 8:e14897. doi: 10.2196/14897

32. Garfan S, Alamoodi AH, Zaidan BB, Al-Zobbi M, Hamid RA, Alwan JK, et al. Telehealth utilization during the Covid-19 pandemic: a systematic review. *Comput Biol Med*. (2021) 138:104878. doi: 10.1016/j.compbiomed.2021.104878

33. Winkler P, Formanek T, Mlada K, Kagstrom A, Mohrova Z, Mohr P, et al. Increase in prevalence of current mental disorders in the context of COVID-19: analysis of repeated nationwide cross-sectional surveys. *Epidemiol Psychiatr Sci*. (2020) 29:e173. doi: 10.1017/S2045796020000888

34. Sales CMD, Neves ITD, Alves PG, Ashworth M. Capturing and missing the patient's story through outcome measures: a thematic comparison of patient-generated items in PSYCHLOPS with CORE-OM and PHQ-9. *Health Expect*. (2018) 21:615–9. doi: 10.1111/HEX.12652

35. Giddens JM, Sheehan KH, Sheehan DV. The Columbia-suicide severity rating scale (C-SSRS): has the 'gold standard' become a liability? *Innov Clin Neurosci*. (2014) 11:66–80.

36. Grazier KL, Smith JE, Song J, Smiley ML. Integration of depression and primary care: barriers to adoption. *J Prim Care Community Health*. (2014) 5:67–73. doi: 10.1177/2150131913491290

37. Ku BS, Li J, Lally CH, Compton MT, Druss BG. Associations between mental health shortage areas and county-level suicide rates among adults aged 25 and older in the USA, 2010 to 2018 HHS public access. *Gen Hosp Psychiatry*. (2021) 70:44–50. doi: 10.1016/j.genhosppsych.2021.02.001

38. Ross EL, Zuromski KL, Ben Reis Y, Nock MK, Kessler RC, Smoller JW. Accuracy requirements for cost-effective suicide risk prediction among primary care patients in the US. *Arch Gen Psychiatry*. (2019) 78:642–50. doi: 10.1001/jamapsychiatry.2021.0089

39. Pennebaker JW. The secret life of pronouns. *New Sci*. (2011) 211:42–5. doi: 10.1016/S0262-4079(11)62167-2

40. Chung TH, Hanley K, Le YC, Merchant A, Nascimento F, De Figueiredo JM, et al. A validation study of PHQ-9 suicide item with the Columbia suicide severity rating scale in outpatients with mood disorders at National Network of depression centers. *J Affect Disord*. (2023) 320:590–4. doi: 10.1016/J.JAD.2022.09.131

41. Bostwick JM, Pabbati C, Geske JR, McKean AJ. Suicide attempt as a risk factor for completed suicide: even More lethal than we knew. *Am J Psychiatry*. (2016) 173:1094–100. doi: 10.1176/APPI.AJP.2016.15070854

42. Chou PH, Wang SC, Wu CS, Horikoshi M, Ito M. A machine-learning model to predict suicide risk in Japan based on national survey data. *Front Psych*. (2022) 13:918667. doi: 10.3389/FPSYT.2022.918667

43. Chu CS, Chou PH, Wang SC, Horikoshi M, Ito M. Associations between PTSD symptom Custers and longitudinal changes in suicidal ideation: comparison between 4-factor and 7-factor models of DSM-5 PTSD symptoms. *Front Psych*. (2021) 12:680434. doi: 10.3389/FPSYT.2021.680434

44. Chou PH, Ito M, Horikoshi M. Associations between PTSD symptoms and suicide risk: a comparison of 4-factor and 7-factor models. *J Psychiatr Res*. (2020) 129:47–52. doi: 10.1016/J.JPSYCHIRES.2020.06.004

45. Jacobs JC, Blonigen DM, Kimerling R, Slightam C, Gregory AJ, Gurmessa T, et al. Increasing mental health care access, continuity, and efficiency for veterans through telehealth with video tablets. *Psychiatr Serv*. (2019) 70:976–82. doi: 10.1176/APPI.PS.201900104

46. Gentry MT, Puspitasari AJ, McKean AJ, Williams MD, Breitinger S, Geske JR, et al. Clinician satisfaction with rapid adoption and implementation of telehealth services during the COVID-19 pandemic. *Telemed J E Health*. (2021) 27:1385–92. doi: 10.1089/TMJ.2020.0575

47. Pease JL, Thompson D, Wright-Berryman J, Campbell M. User feedback on the use of a natural language processing application to screen for suicide risk in the emergency department. *J Behav Health Serv Res*. (2023):1–7. doi: 10.1007/S11414-023-09831-W/METRICS

48. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. *Minds Mach (Dordr)*. (2020) 30:681–94. doi: 10.1007/S11023-020-09548-1/FIGURES/5

49. Bhardwaj R, Majumder N, Poria S. Investigating Gender Bias in BERT. *Cogn Comput*. (2021) 13:1008–18. doi: 10.1007/S12559-021-09881-2

50. Straw I, Callison-Burch C. Artificial intelligence in mental health and the biases of language based models. *PLoS One*. (2020) 15:e0240376. doi: 10.1371/JOURNAL.PONE.0240376

51. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science*. (2017) 356:183–6. doi: 10.1126/SCIENCE.AAL4230

## Glossary

| | |
|---|---|
| AUC | Area under receiver operating characteristic curve |
| C-SSRS | Columbia-suicide severity rating scale |
| CI | Confidence interval |
| CRC | Clinical research coordinator |
| CV | Cross-validation |
| GAD-7 | Generalized anxiety disorder 7-item |
| GSS | Group-shuffle-split |
| LOSO | Leave-one-subject-out |
| LR | Logistic regression |
| MHSAFE | Mental health hopes secrets anger fear and emotional pain |
| ML | Machine learning |
| NLP | Natural language processing |
| NPV | Negative predictive value |
| PHQ-9 | Patient health questionnaire 9-item |
| PPV | Positive predictive value |
| RM | Research match |
| SD | Standard deviation |
| STM | Suicide thought markers (study) |
| SVM | Support vector machine |
| XGB | Extreme gradient boosting |

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 27 July 2023
DOI 10.3389/fpsyt.2023.1143272

# Patterns of engagement in a digital mental health service during COVID-19: a cohort study for children and young people

Aynsley Bernard[1], Santiago de Ossorno Garcia[1], Louisa Salhi[1], Ann John[2] and Marcos DelPozo-Banos[2]*

[1]Kooth Digital Health, London, United Kingdom, [2]Swansea University Medical School, Swansea, United Kingdom

**Introduction:** The COVID-19 pandemic increased public use of digital mental health technologies. However, little is known about changes in user engagement with these platforms during the pandemic. This study aims to assess engagement changes with a digital mental healthcare service during COVID-19.

**Methods:** A cohort study based on routinely collected service usage data from a digital mental health support service in the United Kingdom. Returning users aged 14–25years that interacted for a maximum of two months were included. The study population was divided into pre-COVID and COVID cohorts. Demographic and usage information between cohorts were compared and usage clusters were identified within each cohort. Differences were tested using Chi-squared, two-sample Kolmogorov–Smirnov tests and logit regressions.

**Results:** Of the 624,103 users who joined the service between May 1, 2019, and October 1, 2021, 18,889 (32.81%) met the inclusion criteria: 5,048 in the pre-COVID cohort and 13,841 in the COVID cohort. The COVID cohort wrote more journals; maintained the same focus on messaging practitioners, posting discussions, commenting on posts, and having booked chats; and engaged less in writing journals, setting personal goals, posting articles, and having *ad-hoc* chats. Four usage profiles were identified in both cohorts: one relatively disengaged, one focused on contacting practitioners through chats/messages, and two broadly interested in writing discussions and comments within the digital community. Despite their broad similarities, usage patterns also exhibited differences between cohorts. For example, all four clusters had over 70% of users attempting to have *ad-hoc* chats with practitioners in the pre-COVID cohort, compared to one out of four clusters in the COVID cohort. Overall, engagement change patterns during the COVID-19 pandemic were not equal across clusters. Sensitivity analysis revealed varying strength of these defined clusters.

**Discussion:** Our study identified changes in user activity and engagement behavior within a digital mental healthcare service during the COVID-19 pandemic. These findings suggest that usage patterns within digital mental health services may be susceptible to change in response to external events such as a pandemic. Continuous monitoring of engagement patterns is important for informed design and personalized interventions.

KEYWORDS

**COVID-19, pandemic, engagement, digital mental health, mental health, children and young people, machine learning, clustering**

EDITED BY
Carlos M. M. Travieso-Gonzalez,
University of Las Palmas de Gran Canaria, Spain

REVIEWED BY
Paul Best,
Queen's University Belfast, United Kingdom
Luke Balcombe,
Griffith University, Australia
Lee Valentine,
Orygen Youth Health, Australia

*CORRESPONDENCE
Marcos DelPozo-Banos
✉ m.delpozobanos@swansea.ac.uk

RECEIVED 12 January 2023
ACCEPTED 27 June 2023
PUBLISHED 27 July 2023

CITATION
Bernard A, de Ossorno Garcia S, Salhi L, John A and DelPozo-Banos M (2023) Patterns of engagement in a digital mental health service during COVID-19: a cohort study for children and young people.
*Front. Psychiatry* 14:1143272.
doi: 10.3389/fpsyt.2023.1143272

# Introduction

Research suggests that the COVID-19 pandemic has exacerbated the mental health crisis across many countries, including the United Kingdom (UK) (1). Furthermore, there is a growing body of evidence highlighting effects of the pandemic and consequent lockdowns on children and young people's (CYP) mental health (2–4). At the same time, contacts and interactions with all types of healthcare services reduced dramatically during the pandemic (5). This was also true for mental health-related contacts, particularly face-to-face contacts, with patients having to access to video and over-the-phone contacts (6, 7). Electronic mental healthcare and telemedicine rapidly became the "new normal" (8). By mid-2020, more than 80% of high-income countries shifted to digital mental health technologies to replace or supplement in-person mental health consultations (1).

The use of digital mental health technologies has provided data for several machine learning studies focusing on patterns of engagement within user-led digital support systems. These studies have illustrated several different ways in which mental health support can be personalized in a digital setting: based on engagement type, frequency of access, session duration, timing, and clinical outcomes (9–11). Segmenting users according to behaviors within digital mental health services can be a first step in personalizing support and improving design effectiveness (12).

However, to the best of our knowledge, there are no studies exploring whether user engagement and behaviors within these platforms are subject to change during major events like the COVID-19 pandemic, in which patterns of engagement can be disrupted and present some challenges to machine learning assumptions or solutions based on this type of information. At the same time, such knowledge could inform the use of digital mental health interventions (an important psychological support component during the COVID-19 pandemic) in future disasters (13) and help to resource and prevent overload or saturation of healthcare provision using data-driven technology and decisions.

We hypothesize that stable engagement types exist within Kooth, which could inform the personalization of services. The COVID-19 pandemic provides a unique opportunity to test this hypothesis, as it was an exceptional situation that could potentially affect user engagement patterns with digital mental health services. This cohort study aims to assess changes in engagement within a digital mental health service in the UK during the COVID-19 pandemic, comparing routinely collected usage data between a pre-COVID and a COVID cohort of users.

# Materials and methods

We used data from Kooth Digital Health,[1] the UK's largest provider to the National Health Service of web-based online mental health support (14). This service provides mental health support and interventions through its pseudonymous platform to CYP aged 11–25 years at no cost to the service users. Users can self-refer and find out about Kooth from school, online promotion, primary and secondary health services, social media or word of mouth.

The service allows CYP to self-direct their experience, interacting with their preferred type of support from a range of service features: personal journals, goal setting, discussion boards, articles, asynchronous therapeutic messaging, and live text-based counseling. Comprehensive safeguarding procedures are adhered to by moderators and practitioners following user interactions with the service. Demographic and usage information is stored across databases that can be linked at an individual level under the legal basis of 'legitimate interest' as it informs service improvements (15). In this study, data was used from 1 May 2019 to 31 December 2021.

## Study sample

This study relied on data from 1 May 2019 to 31 December 2021. We included users between the ages of 14–25 years, and who had consented to have their non-identifiable demographic and service usage information used for research purposes. Users who were flagged by practitioners as not having Gillick competence (16) were excluded from the analysis.

To ensure sufficient journey and engagement information per service user, only returning users (i.e., with two or more log-ins) were included in the analysis. Users with a journey longer than 56 days were excluded from the dataset to avoid outliers, in that 99.03% of returning users aged 14–25 had a usage period of 56 days or less. To reduce bias from cut-off or cohort-crossing usage periods, users were excluded if their registration was within 56 days of the end date for each cohort dataset.

We divided users into two cohorts: pre-COVID and COVID. The World Health Organization declared COVID-19 a global pandemic on 11 March 2020 (17). Hence, we defined pre-COVID and COVID cohorts as users who signed up from 1 May 2019 to 11 January 2020 (256 days), and from 11 March 2020 to 1 October 2021 (570 days), respectively.

## Measures

Demographic variables of interest collected routinely included ethnicity ('Asian', 'Black', 'Mixed', 'White' and 'Other'), gender ('Female', 'Male' and 'Non-binary') and age group ('14–17' and '18–25') at the time of registration. We measured interaction with the service through a number of service usage variables: 3 continuous variables ('usage period', 'engagement' and 'activeness') recording the overall level of interaction; and 8 dichotomous variables recording whether users made use of each component of the service (e.g., journals, discussions, *ad-hoc* chats). Table 1 provides the complete list of variables, including activity type, with details.

## Analysis

We performed two-stepped analyses: (1) comparison of pre-COVID and COVID cohorts and (2) identification of usage profiles within pre-COVID and COVID cohorts. All data processing and analyses were done in python v3.9.2 (18). Packages sshtunnel 0.4.0, psycopg2-binary 2.9.5, pymysql 1.0.2, and python-bigquery

[1] Kooth.com

Abbreviation: CYP, children and young people.

TABLE 1 Characteristic, usage and experience variables of interest.

| Variable Type | Variable | Description |
|---|---|---|
| Characteristic | | |
| | Signup Age | Measured in years and split into two groups: 14–18 and 18–25 years. |
| | Ethnicity | One of 'Asian', 'Black', 'Mixed', 'White' and 'Other'. If a user has 'Other' as their ethnicity status, this could be because they selected 'Other' or because they did not state their ethnicity. |
| | Gender | One of 'Agender', 'Female', 'Gender Fluid' and 'Male'. 'Agender' and 'Gender Fluid' are grouped into 'Non-binary' due to low counts. |
| Usage | | |
| Engagement Metrics | Usage Period | Days between first and last login (absolute). |
| | Engagement | Number of active days* divided by Usage Period (defined above). |
| | Activeness | Number of activities divided by active days.* |
| Self Help | Journal Entry | Text journal entry and emoji submitted by a user to signify how the user feels. |
| | Personal Goal Created | Goal set by a user for themselves. |
| Community Engagement | Article Created | Article submitted by a user. |
| | Discussion Created | Discussion thread started by a user. |
| | Comment Created | Comment added to an article or discussion by a user. |
| Asynchronous Practitioner Engagement | Message Sent | Message sent to a practitioner by a user. |
| Synchronous Practitioner Engagement | Drop-in Chat Requested | Impromptu chat requested by joining the chat queue. |
| | Booked Chat Requested | Booked chat with a practitioner scheduled. |
| Experience | | |
| Asynchronous Practitioner Engagement | Administrative message received | Administrative message sent from practitioner to user. |
| | Therapeutic message received | Therapeutic message sent from practitioner to user that includes an assessment, is consistent with a model of intervention and is intended to change behavior. |
| Synchronous Practitioner Engagement | Successful chat | User and practitioner are in an *ad-hoc* chat for >5 min. |
| | Failed chat | User and practitioner do not successfully stay within an *ad-hoc* chat for >5 min. |

*'Active days' is the number of days a user has interacted with the service.

3.3.5 (19–22) were used to query data sources and construct the measures. We used scipy 1.9.3 (23) and statsmodels 0.13.2 packages (24) for statistical modeling and tests. The threshold for statistical significance for all value of *p*s was set at $p < 0.05$.

The main analyses were preceded by an assessment of the generalizability of our results for the service population, comparing separately for pre-COVID and COVID cohorts the study sample (returning, research consenting Kooth users aged 14–25 with a journey of ≤56 days) with the corresponding wider study population (all returning Kooth users aged 14–25). We used the Mann–Whitney U test (25) to measure differences in signup age as a continuous variable, and the Chi-squared test (26) to measure differences in ethnicity and gender.

In the first step of the analysis, we measured proportion and 95% confidence intervals (CI) estimated by Wilson score with continuity correction (27). We measured variations between the pre-COVID and COVID cohorts using Chi-squared tests (26) for demographic variables; two-sample Kolmogorov–Smirnov tests (28) for continuous usage variables; and logit regressions for dichotomous usage variables (as outcomes) using '*cohort*' (i.e., pre-COVID or COVID) as the response variable and controlling for demographic and service change covariates.

In the second step, user groups were identified separately for pre-COVID and COVID cohorts through clustering of the usage variables. Prior to clustering, continuous usage variables were transformed to a logarithmic scale to limit the negative impact of large outliers (29). The usage data was then transformed into a binary indicator of whether the user had interacted with each component of the service, to address the sparsity of the data and improve the efficacy of dimensionality reduction (30). We applied Multiple Correspondence Analysis to reduce dimensionality and allow for the use of Euclidean-based clustering (31).

We ran a sensitivity analysis in which we applied KMeans, Birch, DBSCAN, and Gaussian Mixture Models on the resulting dataset to explore differences across clustering algorithms. We computed the Silhouette Coefficient as $(b_i - a_i)/max(a_i, b_i)$, with $a_i$ the mean intra-cluster distance of sample $i$, $b_i$ the mean nearest-cluster distance of sample $i$, and $N$ the number of samples. This ranges from −1 (worst) to 1 (best) – values near 0 indicate overlapping clusters (32) (Supplementary Tables S3, S4). The cluster number choice was made by inspecting the silhouette analysis plots for the highest scoring algorithms: Birch and KMeans for 2 to 5 clusters (Supplementary Figures S1–S4). After assessing these plots and clusters, we decided to present the output of Birch in the main text, but we ran the full analysis using KMeans for comparison.

We made a deliberate decision not to include service user demographics in the clustering algorithm to minimize the potential biasing effect of demographic factors on the identification of engagement behavior patterns. By doing so, we ensured that the resulting clusters were based solely on the observed engagement behaviors and not influenced by demographic characteristics. Instead, we were interested in observing whether different engagement profiles naturally had demographic differences.

Clusters were then further explored using Chi-squared tests (26) for demographic variables (not used to compute the clusters), Anderson-Darling tests (33) for k-samples for continuous usage variables; and logit regressions for dichotomous usage variables (as outcomes) with '*cluster*' as the response and adjusting for covariates age group, ethnicity and gender (an overall *value of p* for variable '*cluster*' was calculated through a log-likelihood test). For each cluster, we show proportion, 95% CIs and calculated value of *ps* per variable.

# Results

## Study sample

Of the 624,103 individuals who joined Kooth between May 1, 2019, and October 1, 2021, 57,568 (12.82%) were returning users aged 14–25 years. Of these, 18,889 (32.81%) met all the inclusion criteria: 5,048 in the pre-COVID cohort and 13,841 in the COVID cohort. The number of signups per day increased from 19.72 in the pre-COVID cohort to 24.28 in the COVID cohort. The most common place for service users to find out about Kooth is school, which remained consistent across both cohorts but with altered proportions (pre-COVID: 45.38%, COVID: 34.24%). Full details of the cohort selection procedure are in Figure 1.

When comparing demographic distributions between the pre-COVID and COVID cohorts and their corresponding population of interest, we found no significant difference in signup age ($p = 0.220$) and ethnicity ($p = 0.999$) for the pre-COVID cohort, and a significant difference for the COVID cohort ($p < 0.001$) with mean signup age of 0.78 years older and a 0.47 percentage points increase in user proportion that did not select an ethnicity option. For both cohorts, we found no significant differences in gender ($p > 0.888$).

## Comparison between pre-COVID and COVID study cohorts

Full results of the comparison between pre-COVID and COVID cohorts are in Table 2.

Signup age, and the proportion of users who reported gender as 'Non-binary' increased during the pandemic, as did the proportion of users reporting 'Black', 'Mixed' or 'Other' ethnicity against a decrease in those reporting 'White'. More users had relatively longer usage periods during the pandemic, with similar engagement rates but a more active interaction with the service. The COVID cohort wrote more journals; maintained the same focus on messaging practitioners, posting discussions, commenting on posts, and having booked chats; and engaged less in writing journals, setting personal goals, posting articles, and having *ad-hoc* chats. A visual representation of similarities and differences in service usage between pre-COVID and COVID periods can be seen in the right-hand side of Supplementary Figures S1–S4.

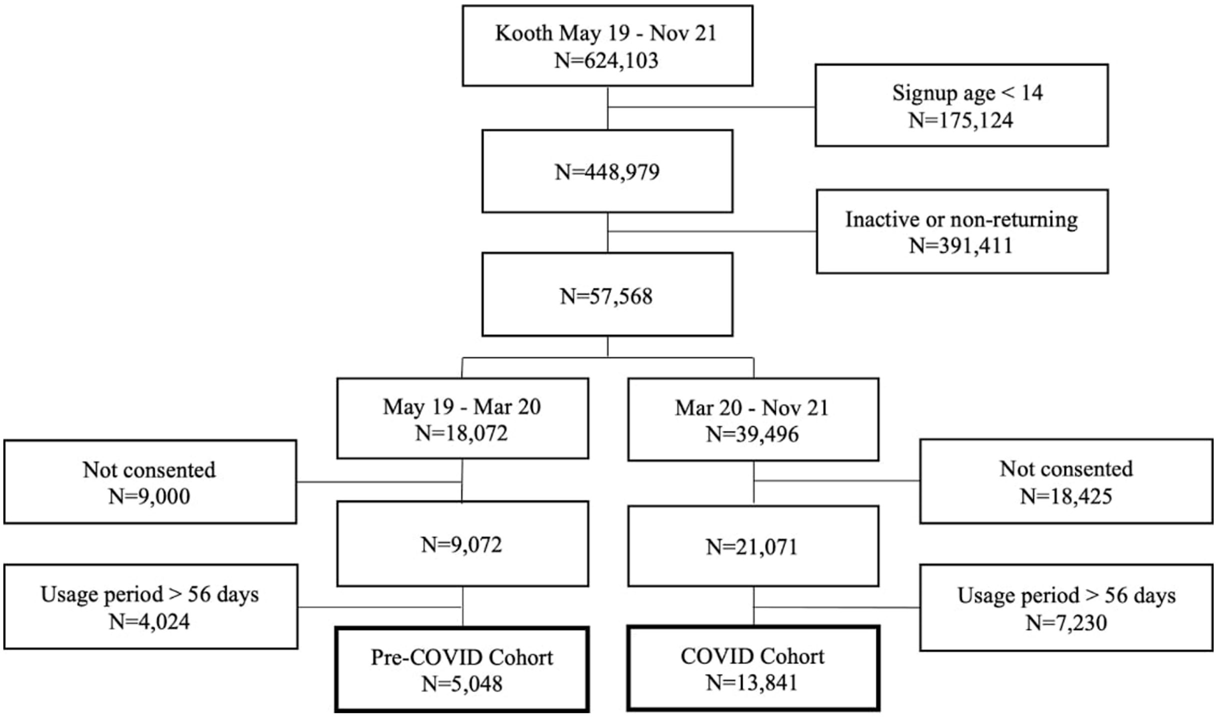


FIGURE 1
Study cohort flow diagram.

TABLE 2 Variable distributions for pre-COVID and COVID cohorts.

| Variable | Pre-COVID | COVID | Adjusted *p* value |
|---|---|---|---|
| Summary statistics | | | |
| Signups | 5,048 | 13,841 | |
| Signups per day | 19.72 | 24.28 | |
| Demographic variables/Control variables not used for clustering | | | |
| Gender | | | <0.001 |
| Female | 3,899 (77.24% [76.06, 78.37]) | 10,406 (75.18% [74.46, 75.89]) | |
| Male | 971 (19.24% [18.17, 20.35]) | 2,629 (18.99% [18.35, 19.66]) | |
| Non-binary | 178 (3.53% [3.05, 4.07]) | 806 (5.82% [5.45, 6.23]) | |
| Age group | | | <0.001 |
| 14–17 | 4,683 (92.77% [92.02, 93.45]) | 12,290 (88.79% [88.26, 89.31]) | |
| 18–25 | 365 (7.23% [6.55, 7.98]) | 1,551 (11.21% [10.69, 11.74]) | |
| Ethnicity group | | | <0.001 |
| White | 4,259 (84.37% [83.34, 85.35]) | 11,280 (81.5% [80.84, 82.14]) | |
| Asian | 331 (6.56% [5.91, 7.27]) | 917 (6.63% [6.22, 7.05]) | |
| Black | 141 (2.79% [2.37, 3.28]) | 471 (3.4% [3.11, 3.72]) | |
| Mixed | 239 (4.73% [4.18, 5.36]) | 727 (5.25% [4.89, 5.64]) | |
| Other | 78 (1.55% [1.24, 1.92]) | 446 (3.22% [2.94, 3.53]) | |
| White | 4,259 (84.37% [83.34, 85.35]) | 11,280 (81.5% [80.84, 82.14]) | |
| Service usage variables/Dependent variables used for clustering | | | |
| Period* | | - | 0.016 |
| Engagement* | | - | 0.074 |
| Activeness* | | - | <0.001 |
| Journal entry | 3,716 (73.61% [72.38, 74.81]) | 11,675 (84.35% [83.74, 84.95]) | <0.001 |
| Personal goal created | 1,007 (19.95% [18.87, 21.07]) | 2,418 (17.47% [16.85, 18.11]) | <0.001 |
| Article created | 296 (5.86% [5.25, 6.55]) | 466 (3.37% [3.08, 3.68]) | <0.001 |
| Discussion created | 1,051 (20.82% [19.72, 21.96]) | 2,741 (19.8% [19.15, 20.48]) | 0.251 |
| Comment created | 1,644 (32.57% [31.29, 33.87]) | 4,772 (34.48% [33.69, 35.27]) | 0.008 |
| Message sent | 771 (15.27% [14.31, 16.29]) | 2,237 (16.16% [15.56, 16.78]) | 0.129 |
| *Ad-hoc* chat | 3,791 (75.1% [73.89, 76.27]) | 7,816 (56.47% [55.64, 57.29]) | <0.001 |
| Booked chat | 196 (3.88% [3.38, 4.45]) | 403 (2.91% [2.64, 3.21]) | 0.001 |
| Service experience variables/Observational variables not used for clustering | | | |
| Administrative message received | 475 (9.41% [8.63, 10.25]) | 7,553 (54.57% [53.74, 55.4]) | <0.0001 |
| Therapeutic message received | 680 (13.47% [12.56, 14.44]) | 7,152 (51.67% [50.84, 52.5]) | <0.0001 |
| Successful chat | 1,567 (31.04% [29.78, 32.33]) | 2,611 (18.86% [18.22, 19.52]) | <0.0001 |
| Failed chat | 1,065 (21.1% [19.99, 22.24]) | 2,675 (19.33% [18.68, 19.99]) | 0.008 |

Reported values are absolute counts, percentages with 95% confidence intervals, and adjusted *p* values assessing whether variables changed during COVID compared to pre-COVID.
*Continuous variable. See Figure 2.

## Identification of usage clusters

During optimization of the clustering algorithms, the silhouette score varied between 0.14 and 0.46 for the pre-COVID cohort and between 0.16 and 0.43 for the COVID cohort across 24 hyperparameter configurations. Full results can be seen in Supplementary Tables S1 and S2 for the pre-COVID and COVID cohorts, respectively. There was no clear best configuration for both pre-COVID and COVID cohorts. Inspecting both the results of the silhouette analysis (Supplementary Figures S1–S4) and the obtained clusters, we decided to continue with the output of 4 usage clusters for each cohort.

## Comparison between pre-COVID and COVID usage clusters

Figure 2 and Tables 3–6 show the size, characteristics, and usage profiles for each cluster. We observed some broad similarities between

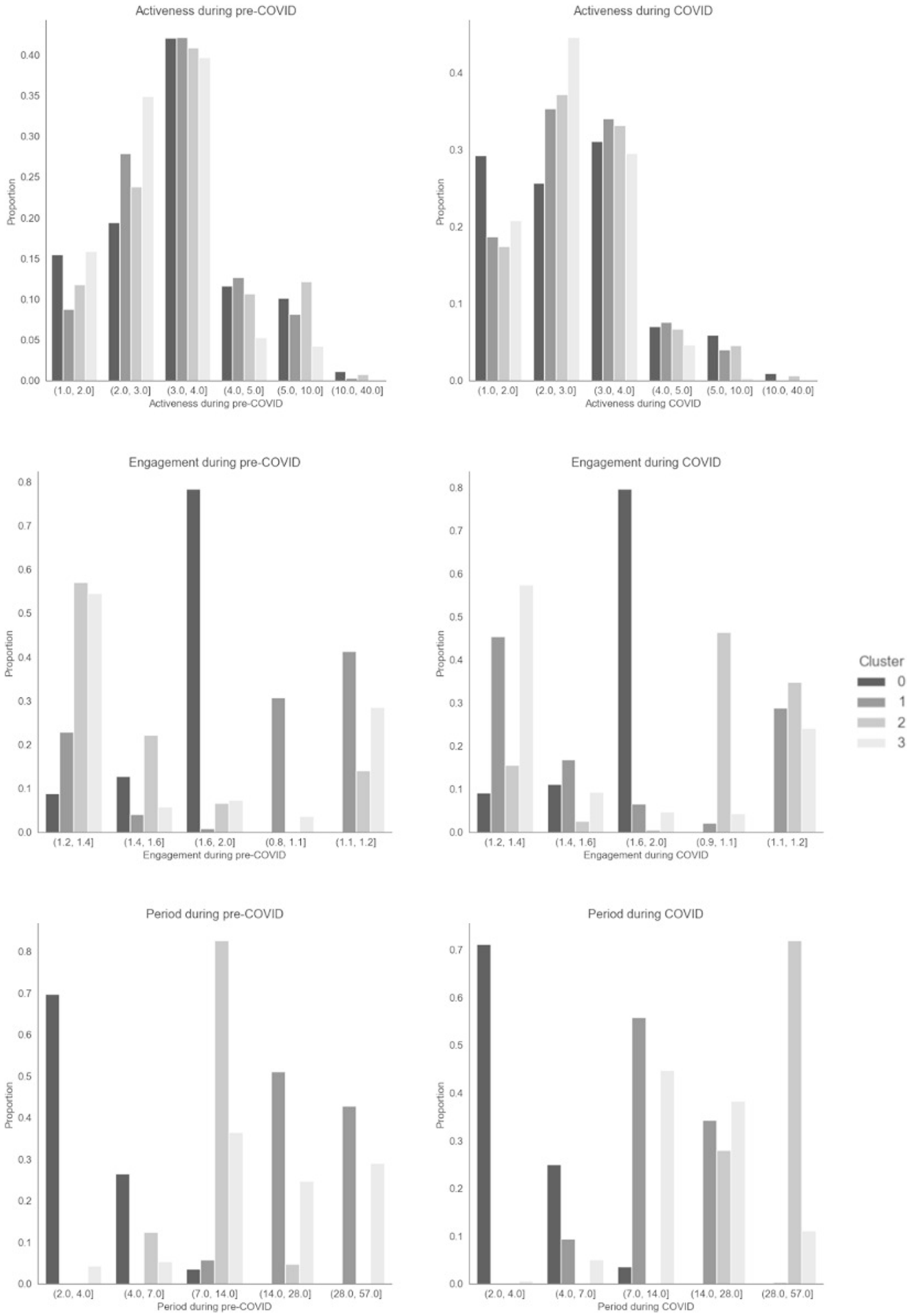


FIGURE 2
Distribution plots for period, activeness and engagement variables across pre-COVID and COVID Birch clusters.

TABLE 3 Control variables across Birch engagement clusters for the pre-COVID cohort.

| Variable | C0 | C1 | C2 | C3 |
|---|---|---|---|---|
| Pre-COVID Birch clusters | | | | |
| Summary statistics | | | | |
| Signups | 2,716 | 1,345 | 798 | 189 |
| Proportion of all pre-COVID signups | 53.80% | 26.64% | 15.81% | 3.74% |
| Signups per day | 10.61 | 5.25 | 3.12 | 0.74 |
| Control variables not used for clustering | | | | |
| Gender ($p$=0.684) | | | | |
| Female | 2091 (76.99% [75.37, 78.53]) | 1,045 (77.7% [75.39, 79.84]) | 612 (76.69% [73.63, 79.49]) | 481 (80.84% [77.48, 83.8]) |
| Male | 535 (19.7% [18.25, 21.24]) | 245 (18.22% [16.24, 20.37]) | 158 (19.8% [17.18, 22.71]) | 91 (15.29% [12.63, 18.41]) |
| Non-binary | 90 (3.31% [2.7, 4.06]) | 55 (4.09% [3.16, 5.28]) | 28 (3.51% [2.44, 5.02]) | 23 (3.87% [2.59, 5.73]) |
| Age group ($p$=0.052) | | | | |
| 14–17 | 2,520 (92.78% [91.75, 93.7]) | 1,233 (91.67% [90.07, 93.03]) | 757 (94.86% [93.1, 96.19]) | 173 (91.53% [86.69, 94.72]) |
| 18–25 | 196 (7.22% [6.3, 8.25]) | 112 (8.33% [6.97, 9.93]) | 41 (5.14% [3.81, 6.9]) | 16 (8.47% [5.28, 13.31]) |
| Ethnicity group ($p$=0.01) | | | | |
| Asian | 183 (6.74% [5.85, 7.74]) | 101 (7.51% [6.22, 9.04]) | 32 (4.01% [2.85, 5.61]) | 15 (7.94% [4.87, 12.68]) |
| Black | 75 (2.76% [2.21, 3.45]) | 37 (2.75% [2.0, 3.77]) | 26 (3.26% [2.23, 4.73]) | 3 (1.59% [0.54, 4.56]) |
| Mixed | 138 (5.08% [4.32, 5.97]) | 54 (4.01% [3.09, 5.2]) | 34 (4.26% [3.06, 5.89]) | 13 (6.88% [4.06, 11.41]) |
| Other | 36 (1.33% [0.96, 1.83]) | 29 (2.16% [1.51, 3.08]) | 10 (1.25% [0.68, 2.29]) | 3 (1.59% [0.54, 4.56]) |
| White | 2,284 (84.09% [82.67, 85.42]) | 1,124 (83.57% [81.49, 85.45]) | 696 (87.22% [84.72, 89.36]) | 155 (82.01% [75.91, 86.83]) |
| Messages received | | | | |
| Message received | 447 (16.46% [15.11, 17.9]) | 344 (25.58% [23.32, 27.98]) | 147 (18.42% [15.88, 21.26]) | 88 (46.56% [39.59, 53.67]) |

Reported values are absolute counts and percentages with 95% confidence intervals. Chi-squared tests across clusters returned $p < 0.001$ for all variables except otherwise specified.

pre-COVID and COVID usage clusters. In both cases, cluster sizes were highly unbalanced, with the largest and smallest clusters containing 53.80 and 3.74% of pre-COVID users, respectively, and 51.29 and 2.15% of COVID users, respectively. The largest cluster in both cohorts was also the one with the shortest enrolment period (1–2 days) and the least engaged in practitioner-based interventions. Similarly, the smallest cluster in both cohorts was also the one with highest proportion of females, older users, and most engaged with practitioner-based interventions. The remaining two clusters where the most interested in community-based interventions.

At the same time, there were substantial differences between the pre-COVID and COVID clusters. For example, gender differences were only significant in the COVID cohort. The unengaged cluster (C0) was less active during the pandemic. Over 70% of users in all pre-COVID clusters requested and *ad-hoc* chat with a practitioner, compared to only one COVID cluster. The two clusters interested in community-based interventions (C1 and C2) showed opposite trends in usage period and engagement. One of these (C2) was also the least engaged on creating articles before COVID, but the second most engaged during COVID. The cluster engaging the most with practitioner-based intervention had the highest proportion of users who selected Asian and Mixed ethnicity pre-COVID, and the lowest during COVID.

Our sensitivity analysis had mixed results (Supplementary Figure S5; Supplementary Tables S3–S5). In the pre-COVID period, the obtained clusters were substantially different to those of Birch, although we still found a cluster of disengaged used (this time even more disengaged). Of the remaining clusters, one was focused on both self-help and community-based interventions, and the other two focused on practitioner-based interventions and had moderate interest in community-based interventions. The size of pre-COVID clusters based on KMeans was also much more valanced, with each accounting for 20–30% of users. Meanwhile, KMeans' result during COVID was similar to Birch's, with a disengaged cluster, a cluster focused on practitioner-based interventions, and a cluster focused on community-based interventions. The fourth cluster was also relatively focused on community-based interventions. Cluster sizes were also similar to Birch's.

# Discussion

## Key findings

We found changes in the usage of Kooth, a UK mental health digital service, by users aged 14–25 years during the COVID-19 pandemic. While the number of signups per day increased, these users were less engaged with the service, most prominently with less activity within each log-in (albeit usage periods were longer on average) and focusing less on creating articles and discussions and requesting *ad-hoc* chats with practitioners. This excess of users during the pandemic may be driven by a lack of capacity on traditional mental health services, a desire to 'protect' these services, and/or fear of COVID-19 infection in physical settings. We also identified changes

TABLE 4 Usage variables across Birch engagement clusters for the pre-COVID cohort.

| Variable | C0 | C1 | C2 | C3 |
|---|---|---|---|---|
| Pre-COVID Birch clusters | | | | |
| Signups | 2,716 | 1,345 | 798 | 189 |
| Dependent variables used for clustering | | | | |
| Journal entry | 2078 (76.51% [74.88, 78.07]) | 992 (73.75% [71.34, 76.04]) | 552 (69.17% [65.88, 72.28]) | 94 (49.74% [42.68, 56.8]) |
| Personal goal created | 520 (19.15% [17.71, 20.67]) | 310 (23.05% [20.88, 25.37]) | 153 (19.17% [16.59, 22.05]) | 24 (12.7% [8.68, 18.2]) |
| Article created | 154 (5.67% [4.86, 6.6]) | 104 (7.73% [6.42, 9.28]) | 32 (4.01% [2.85, 5.61]) | 6 (3.17% [1.46, 6.75]) |
| Discussion created | 512 (18.85% [17.42, 20.37]) | 326 (24.24% [22.02, 26.6]) | 194 (24.31% [21.46, 27.41]) | 19 (10.05% [6.53, 15.17]) |
| Comment created | 810 (29.82% [28.13, 31.57]) | 504 (37.47% [34.92, 40.09]) | 287 (35.96% [32.71, 39.35]) | 43 (22.75% [17.35, 29.24]) |
| Message sent | 344 (12.67% [11.47, 13.97]) | 251 (18.66% [16.67, 20.83]) | 131 (16.42% [14.01, 19.15]) | 45 (23.81% [18.3, 30.37]) |
| *Ad-hoc* chat | 1914 (70.47% [68.73, 72.16]) | 1,065 (79.18% [76.93, 81.27]) | 625 (78.32% [75.33, 81.04]) | 187 (98.94% [96.22, 99.71]) |
| Booked chat (no value of *p*) | 0 (0.0% [0.0, 0.14]) | 6 (0.45% [0.2, 0.97]) | 1 (0.13% [0.02, 0.71]) | 189 (100.0% [98.01, 100.0]) |
| Observational variables not used for clustering | | | | |
| Successful chat | 581 (21.39% [19.89, 22.97]) | 529 (39.33% [36.75, 41.97]) | 271 (33.96% [30.76, 37.32]) | 186 (98.41% [95.44, 99.46]) |
| Failed chat | 425 (15.65% [14.33, 17.06]) | 319 (23.72% [21.52, 26.06]) | 169 (21.18% [18.48, 24.15]) | 152 (80.42% [74.18, 85.45]) |

Reported values are absolute counts and percentages with 95% confidence intervals. Logistic regression models adjusted for demographic variables returned $p < 0.001$ for all variables unless otherwise specified. Some results are omitted due to convergence issues. Cases where 'no value of *p*' is stated indicate where the algorithm did not converge.

in the user experience, with more users being asynchronously contacted and fewer having live chats with practitioners during the pandemic. This is likely the result of service changes implemented to manage the observed increase in the demand, like practitioners actively contacting users.

We conducted cluster analyses individually in each time period (before and during the COVID-19 pandemic) and identified four clusters or usage profiles: one relatively disengaged, one focused on contacting practitioners through chats/messages, and two broadly interested in writing discussions and comments within the digital community. The disengaged profile is likely an extension of our initial observation on the high proportion of users not returning to the system after one visit, as this profile is also the largest of the four (>50% of users), highlighting the importance of this type of interaction and user preference for digital interventions. Users seeking only contact with practitioners returned to the system sporadically. This is a fitting strategy for them, since there are natural idling times between messages and chats. Users more interested on posting articles, discussions and comments seemed to be the most committed overall, with relatively longer usage periods, engagement and activeness metrics. These seemed to be the most valanced users in terms of engagement, showing also high interactions in personal- and practitioner-based interventions. All clusters had over 70% of users requesting *ad-hoc* chats with practitioners, highlighting the importance of this type of interaction for digital interventions.

Pre-COVID and COVID usage profiles, despite being grossly similar, had some stark differences particularly with the two community-focused clusters. These two clusters exhibited opposite changes on some activity in the platform (e.g., practitioner-based interventions), even swapping their ranking as most/least engaged as a result in some instances. They also swapped the length of usage period and engagement. We originally thought that these differences may have been artificially introduced by moving from three to four clusters, but inspection of the Silhouette plots (Supplementary Figures S1, S2) revealed that this step gave way to the practitioner-focused cluster (i.e., the two community-engaged clusters were already present with three clusters). We also observed differences in demographic variables not used for clustering. These may explain part of the usage profile changes between pre-COVID and COVID within the platform, but the demographic differences pertained to a small proportion of the study sample, and therefore unlikely to explain the full range of such changes. Therefore, significant external events such as pandemics may impact how users interact with digital mental health services and may affect how to effectively identify patterns of engagement to form profiles.

Our sensitivity analysis led to a similar conclusion: that usage profiles are susceptible to significant external events. However, it also showed that the resulting usage profiles are not always strongly defined in our data, and thus the selection of clustering algorithm may have a big impact on the results – this may also be weakness of our data, rather than the methodology itself. As such, the utilization of usage profiles to inform the ongoing design of these services and the recommendations of personalized interventions may not be an optimal strategy – at least not during major events and not without the right data, careful sensitivity analyses and a strong methodology leading to robust outcomes.

Our clustering analysis revealed changes in service usage not readily apparent from the analysis using the full pre-COVID and COVID cohorts. Most prominently, despite community engagement variables decreasing (articles created) or not changing (discussions and comments created) during the COVID-19 pandemic, the influx of users focused on community engagement increased. Therefore, even though community engagement decreased during the COVID-19 pandemic, 2 out of 5 users that registered during this period in fact directed their attention to community-based activity. This effect may have been driven by the lockdown, self-isolation, and social distancing measures in place during the pandemic, and thus reflect users longing for social interaction, especially in young people (34). In general terms, different clusters show different patterns of change during COVID. From a methodological point of view, these results suggest that clustering analysis may be a useful tool in the analysis of service

TABLE 5 Control variables across Birch engagement clusters for the COVID cohort.

| Variable | C0 | C1 | C2 | C3 |
|---|---|---|---|---|
| COVID Birch clusters | | | | |
| Summary statistics | | | | |
| Signups | 7,099 | 3,847 | 2,597 | 298 |
| Proportion of all COVID signups | 51.29% | 27.79% | 18.76% | 2.15% |
| Signups per day | 12.45 | 6.75 | 4.56 | 0.52 |
| Control variables not used for clustering | | | | |
| Gender | | | | |
| Female | 5,219 (73.52% [72.48, 74.53]) | 2,925 (76.03% [74.66, 77.36]) | 2026 (78.01% [76.38, 79.56]) | 236 (79.19% [74.23, 83.42]) |
| Male | 1,451 (20.44% [19.52, 21.39]) | 704 (18.3% [17.11, 19.55]) | 420 (16.17% [14.81, 17.64]) | 54 (18.12% [14.16, 22.89]) |
| Non-binary | 429 (6.04% [5.51, 6.62]) | 218 (5.67% [4.98, 6.44]) | 151 (5.81% [4.98, 6.78]) | 8 (2.68% [1.37, 5.21]) |
| Age group (*p*=0.109) | | | | |
| 14–17 | 6,311 (88.9% [88.15, 89.61]) | 3,413 (88.72% [87.68, 89.68]) | 2,316 (89.18% [87.93, 90.32]) | 250 (83.89% [79.29, 87.63]) |
| 18–25 | 788 (11.1% [10.39, 11.85]) | 434 (11.28% [10.32, 12.32]) | 281 (10.82% [9.68, 12.07]) | 48 (16.11% [12.37, 20.71]) |
| Ethnicity group (*p*=0.016) | | | | |
| Asian | 432 (6.09% [5.55, 6.67]) | 282 (7.33% [6.55, 8.2]) | 190 (7.32% [6.38, 8.38]) | 13 (4.36% [2.57, 7.32]) |
| Black | 255 (3.59% [3.18, 4.05]) | 125 (3.25% [2.73, 3.86]) | 81 (3.12% [2.52, 3.86]) | 10 (3.36% [1.83, 6.07]) |
| Mixed | 367 (5.17% [4.68, 5.71]) | 197 (5.12% [4.47, 5.86]) | 150 (5.78% [4.94, 6.74]) | 13 (4.36% [2.57, 7.32]) |
| Other | 246 (3.47% [3.06, 3.92]) | 125 (3.25% [2.73, 3.86]) | 65 (2.5% [1.97, 3.18]) | 10 (3.36% [1.83, 6.07]) |
| White | 5,799 (81.69% [80.77, 82.57]) | 3,118 (81.05% [79.78, 82.26]) | 2,111 (81.29% [79.74, 82.74]) | 252 (84.56% [80.02, 88.22]) |
| Messages received | | | | |
| Message received | 5,217 (73.49% [72.45, 74.5]) | 3,022 (78.55% [77.23, 79.82]) | 2,147 (82.67% [81.17, 84.08]) | 273 (91.61% [87.91, 94.25]) |

Reported values are absolute counts and percentages with 95% confidence intervals. Chi-squared tests across clusters returned $p < 0.001$ for all variables except otherwise specified.

usage and its change over time, as it can provide insight into previously hidden patterns.

## Comparison with prior research

Prior work on mental health service usage profiling incorporates time and typically tries to understand where a user is in their lifetime with a service (9–11). However, since the average user's time with the service studied here is less than two weeks (pre-COVID: 12.35 [12.91], COVID: 12.90 [13.20]), we simplified the analysis by assuming fixed usage profiles throughout the users' journey.

Prior work on mental health service usage profiling incorporates outcome variables and relies on a single time period for examination (9–11). They typically found between 3 and 5 usage profiles, mostly focused on the level of engagement. Since we did not have access to outcome variables in our analysis, direct comparison with other study results is not possible. However, we found a similar number of usage profiles within each cohort, some overall more engaged than others, but we also found differences in the type of engagement, as discussed above.

Previous mental health studies have shown a widespread deterioration of the population's mental health during the COVID-19 pandemic (1), but disproportionately so for young adults and minoritized gender and ethnic groups (35, 36). We found corresponding increases in the number of signups to the service (from 19.72 users/day to 24.28 users/day). The proportion of users increased for adults, users who selected 'Black', 'Mixed' or 'Other' ethnicity and users who selected 'Agender' or 'Gender Fluid' gender, but not in the proportion of females compared to males.

## Strengths and limitations

We have assessed changes in the way users interact with a digital mental health service before and after the COVID-19 pandemic started in the UK, using routinely collected usage data from 18,969 users across 30 months, including the first two waves of the pandemic. We explored whether these differences varied across user types, themselves defined using clustering techniques on usage information. To the authors' knowledge, this is the first work to study how engagement behaviors within a digital mental healthcare service change during a global crisis of this kind.

We approached the use of clustering techniques not as a central part of the research, but as a tool to answer our research question (i.e., whether usage profiles changed during the COVID-19 pandemic). As such, our methodological decisions were not driven by clustering performance, but by domain knowledge (e.g., access routes of users to the different parts of the service, and the way their interactions are recorded) to ensure the relevance of all the included variables. Additionally, we validated and compared the resulting clusters using traditional statistical methods and exploring variable distributions.

There were limitations surrounding the study population, as it included only 32.81% of the total population of users ever using the

TABLE 6 Usage variables across Birch engagement clusters for the COVID cohort.

| Variable | C0 | C1 | C2 | C3 |
|---|---|---|---|---|
| COVID Birch clusters | | | | |
| Signups | 7,099 | 3,847 | 2,597 | 298 |
| Dependent variables used for clustering | | | | |
| Journal entry | 6,147 (86.59% [85.78, 87.36]) | 3,141 (81.65% [80.39, 82.84]) | 2,237 (86.14% [84.76, 87.41]) | 150 (50.34% [44.69, 55.97]) |
| Personal goal created | 1,170 (16.48% [15.64, 17.36]) | 642 (16.69% [15.54, 17.9]) | 585 (22.53% [20.96, 24.17]) | 21 (7.05% [4.66, 10.53]) |
| Article created | 184 (2.59% [2.25, 2.99]) | 177 (4.6% [3.98, 5.31]) | 100 (3.85% [3.18, 4.66]) | 5 (1.68% [0.72, 3.87]) |
| Discussion created | 1,186 (16.71% [15.86, 17.59]) | 916 (23.81% [22.49, 25.18]) | 623 (23.99% [22.39, 25.67]) | 16 (5.37% [3.33, 8.54]) |
| Comment created | 2091 (29.45% [28.41, 30.53]) | 1,626 (42.27% [40.71, 43.83]) | 991 (38.16% [36.31, 40.04]) | 64 (21.48% [17.19, 26.49]) |
| Message sent | 831 (11.71% [10.98, 12.47]) | 692 (17.99% [16.81, 19.23]) | 647 (24.91% [23.29, 26.61]) | 67 (22.48% [18.11, 27.56]) |
| *Ad-hoc* chat | 3,529 (49.71% [48.55, 50.87]) | 2,331 (60.59% [59.04, 62.13]) | 1,697 (65.34% [63.49, 67.15]) | 259 (86.91% [82.61, 90.28]) |
| Booked chat (no value of *p*) | 1 (0.01% [0.0, 0.08]) | 3 (0.08% [0.03, 0.23]) | 101 (3.89% [3.21, 4.7]) | 298 (100.0% [98.73, 100.0]) |
| Observational variables not used for clustering | | | | |
| Successful chat | 794 (11.18% [10.47, 11.94]) | 831 (21.6% [20.33, 22.93]) | 717 (27.61% [25.92, 29.36]) | 269 (90.27% [86.37, 93.14]) |
| Failed chat | 1,066 (15.02% [14.2, 15.87]) | 769 (19.99% [18.76, 21.28]) | 610 (23.49% [21.9, 25.16]) | 230 (77.18% [72.09, 81.58]) |

Reported values are absolute counts and percentages with 95% confidence intervals. Logistic regression models adjusted for demographic variables returned $p < 0.001$ for all variables unless otherwise specified. Cases where 'no value of *p*' is stated indicate where the algorithm did not converge.

service (27.93% from the full pre-COVID cohort and 35.04% from the full COVID cohort). Nevertheless, of age, gender and ethnicity, the study sample only differed from the whole population on signup age.

The digital mental health service examined moved through several product and service improvements, potentially influencing usage. This translates to unmeasured impact of changes which prevent us from establishing with complete certainty a relationship between the changes solely attributable to COVID-19 pandemic. Other potential mental health covariates were not available, like socioeconomic status or simultaneous engagement with other services, which has been shown to change during the pandemic (36, 37). The time period of data collection is also not consistent across cohorts, so there is a possibility that changes could be due to seasonality effects or other confounders.

Engagement with digital mental health services may also be subject to variation based on service availability, making it challenging to determine which digital behaviors are genuinely influenced by the service user and not by changes in the platform and resources to provide support. We made efforts to control for changes due to service availability in terms of messages sent to service users by practitioners. However, controlling for this factor becomes exceedingly difficult in an active, naturalistic environment where resource changes can occur at different times and in various regions where the service operates.

Our study was limited to a UK-only service, which restricted our ability to compare engagement data with similar services in other countries. Therefore, the generalizability of our findings is limited to the UK context, and caution should be exercised in extrapolating our results beyond the digital service examined.

## Future research

Future research into CYP engagement would benefit from incorporating mental health measures before and after engagements. This would allow us to explore if measure responses predict engagement with digital services when combined with age, ethnicity and gender. There are known barriers in access to mental health services, and therefore understanding engagement patterns with digital mental health services can provide an early look into engagement preferences or barriers. We had data on mental health measures associated with this study, but opportunities for completion of these measures within the system were based on engagement preferences and were therefore biased. Hence, we decided to exclude outcome measures with a view to investigating outcomes in a separate study.

This study focuses only on returning users. For Kooth users aged 14–25, 87.18% of users do not return to the site after initial signup which leaves a large portion of the service user population uninvestigated. This drop-off could be due to implementation barriers such as lack of personalization or human capacity (38), and it is similar to that reported by other digital platforms (39). Future research is needed to understand the difference between returning and non-returning users, and how to maximize the potential of brief engagement vs. more continuous and regular engagement.

Our main finding, that usage profiles are affected by major events, puts into question the stability of usage profiles using clustering methods of data based on engagement. Further analysis over periods without major catastrophic events is required to ascertain whether changes in usage profiles can also occur naturally (i.e., without the influence of major events), but this also highlights the importance of examining and accounting for such events when machine learning algorithms are used for cluster and designing products and services and optimization.

This study focuses on user engagement changes between pre-COVID and COVID cohorts. Future work should investigate the long-term impact of the pandemic on mental health. It would be beneficial to explore whether mental health issues during the pandemic have regressed back to their mean or if they have persisted, as well as the role of community support in mental health during the pandemic. This may involve longitudinal studies to track changes in service engagement and mental health outcomes over time.

Kooth is a standalone digital mental health platform that provides online support for children and young people – it is not specifically designed to be integrated into face-to-face mental health care or other health care systems. While Kooth can be accessed independently by users, it may also be used as part of a blended care approach, where digital interventions are combined with traditional face-to-face services. However, the extent to which Kooth is adaptable to a blended care approach is beyond the scope of this study and warrants further investigation.

## Conclusion

The study of the effect of the COVID-19 pandemic on digital mental health services is particularly relevant, as these remained uninterrupted, while face-to-face services paused or changed provision. We explored the user activity and engagement behavior within a digital mental healthcare service and identified changes in these digital profiles during the COVID-19 pandemic. This indicates that usage profiles are not suitable to inform service design or provide personalized interventions yet, as they are susceptible to change due to events like a pandemic. However, usage profiles can provide important insight into the analysis of such changes in digital behavior and can help us better understand digital mental health service user populations and contribute to future disaster management procedures (13).

While digital mental health interventions can be powerful support tools, particularly in periods when traditional face-to-face services lack capacity or space, a better understanding of user engagement with these systems and how it changes over time is needed to fully unlock their potential, alongside other important considerations such as effectiveness, usability, and equity of access.

## Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Under Kooth's privacy policy, data can only be shared with trusted partners for research studies dedicated to improving the service. Requests to access these datasets should be directed to AB, abernard@kooth.com.

## Ethics statement

The studies involving human participants were reviewed and approved by Swansea University Ethics Committee. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements. Study data covered only individuals that provided informed consent for their data to be utilized for research purposes.

## Author contributions

AB and MDP-B defined the original concept for the study, conducted the comparison of clusters, and wrote the original draft. AB prepared the data and conducted the cluster analysis. AB, SdOG, and MDP-B interpreted the results. All authors reviewed the original draft and contributed to the final version.

## Funding

AJ and MDP-B were funded by UKRI – Medical Research Council through the DATAMIND Hub (MRC reference: MR/W014386/1).

## Conflict of interest

AB, SdOG, and LS are employed and receive honorarium by Kooth plc. MDP-B was contracted by Kooth plc as a consultant for this work.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from Kooth Digital Health. The funder had the following involvement in the study: study design, data collection and analysis, interpretation of the results, and preparation of the manuscript.

## 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/fpsyt.2023.1143272/full#supplementary-material

## References

1. Hewlett E, Takino S, Nishina Y, Prinz C. *Tackling the mental health impact of the COVID-19 crisis: An integrated, whole-of-society response*. Paris: OECD (2021).

2. Mindel C, Salhi L, Oppong C, Lockwood J. Alienated and unsafe: Experiences of the first national UK COVID-19 lockdown for vulnerable young people (aged 11–24 years) as revealed in Web-based therapeutic sessions with mental health professionals. *Couns Psychother Res*. (2022) 14:708–24.

3. Elliot Major L, Eyles A, Machin S. Generation COVID: Emerging work and education inequalities. *Centre Econ Perform*. (2020) 14.

4. Young Minds. Coronavirus: impact on young people with mental health needs. *Young Minds*. (2021). Available at: www.youngminds.org.uk

5. Moynihan R, Sanders S, Michaleff ZA, Scott AM, Clark J, To EJ, et al. Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review. *BMJ Open*. (2021) 11:e045343. doi: 10.1136/BMJOPEN-2020-045343

6. Armitage R. Antidepressants, primary care, and adult mental health services in England during COVID-19. *Lancet Psychiatry*. (2021) 8:e3. doi: 10.1016/S2215-0366(20)30530-7

7. Bakolis I, Stewart R, Baldwin D, Beenstock J, Bibby P, Broadbent M, et al. Changes in daily mental health service use and mortality at the commencement and lifting of COVID-19 'lockdown' policy in 10 UK sites: a regression discontinuity in time design. *BMJ Open*. (2021). 11:e049721. doi: 10.1136/BMJOPEN-2021-049721

8. Martinez-Martin N, Dasgupta I, Carter A, Chandler JA, Kellmeyer P, Kreitmair K, et al. Ethics of Digital mental Health during COVID-19: crisis and opportunities. *JMIR Ment Health*. (2020) 7:e23776. doi: 10.2196/23776

9. Chien I, Enrique A, Palacios J, Regan T, Keegan D, Carter D, et al. A machine learning approach to understanding patterns of engagement with internet-delivered mental Health interventions. *JAMA Netw Open*. (2020) 3:e2010791. doi: 10.1001/JAMANETWORKOPEN.2020.10791

10. Matthews P, Topham P, Caleb-Solly P. Interaction and engagement with an anxiety management app: analysis using large-scale behavioral data. *JMIR Mental Health*. (2018) 5:9235. doi: 10.2196/MENTAL.9235

11. Sanatkar S, Baldwin PA, Huckvale K, Clarke J, Christensen H, Harvey S, et al. Using cluster analysis to explore engagement and e-attainment as emergent behavior in electronic mental Health. *J Med Internet Res*. (2019) 21:e14728. doi: 10.2196/14728

12. Fleming T, Merry S, Stasiak K, Hopkins S, Patolo T, Ruru S, et al. The importance of user segmentation for designing digital therapy for adolescent mental Health: findings from scoping processes. *JMIR Ment Health*. (2019) 6:e12656. doi: 10.2196/12656

13. Sheek-Hussein M, Abu-Zidan FM, Stip E. Disaster management of the psychological impact of the COVID-19 pandemic. *Int J Emerg Med*. (2021). doi: 10.1186/S12245-021-00342-Z/TABLES/2

14. North Central London Clinical Comissioning Group. (2021). Kooth: the UK's leading mental health & wellbeing platform for children and young people. North Central London GP Website Available at: https://gps.northcentrallondonccg.nhs.uk/education/video/kooth-the-uks-leading-mental-health-wellbeing-platform-for-children-and-young-people-islington (Accessed August, 2022)

15. Kooth Digital Health. (2022). Privacy policy – Kooth. Kooth.Com. Available at: https://www.kooth.com/privacy (Accessed August, 2022)

16. Griffith R. What is Gillick competence? *Hum Vaccin Immunother*. (2016) 12:244–7. doi: 10.1080/21645515.2015.1091548

17. Ghebreyesus T. (2020). WHO director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020. World Health Organisation. Available at: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

18. Python. (2021). Python release Python 3.9.2. Python.Org. Available at: https://www.python.org/downloads/release/python-392/ (Accessed April, 2022)

19. White P. (2021). Pahaz/sshtunnel: SSH tunnels to remote server. Github.Com. Available at: https://github.com/pahaz/sshtunnel/ (Accessed April, 2022)

20. Varrazzo D. Psycopg/psycopg2: PostgreSQL database adapter for the Python programming language. Github.Com (2021). Available at: https://github.com/psycopg/psycopg2 (Accessed April, 2022)

21. Naoki I. (2016). PyMySQL/PyMySQL: pure Python MySQL client. Github.Com. Available at: https://github.com/PyMySQL/PyMySQL (Accessed April, 2022)

22. Google. Releases…googleapis/python-bigquery. Github.Com (2021). Available at: https://github.com/googleapis/python-bigquery/releases (Accessed April 2022)

23. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. Scipy/scipy: SciPy library main repository. *Nature Methods*. (2020) 17:261–72. doi: 10.1038/s41592-019-0686-2

24. Seabold S, Perktold J. Statsmodels/statsmodels. Github.Com (2010). Available at: https://github.com/statsmodels/statsmodels (Accessed April 2022)

25. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*. (1947) 18:50–60. doi: 10.1214/AOMS/1177730491

26. Karl Pearson FRS. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Biometrika*. (2009) 50:157–75. doi: 10.1080/14786440009463897

27. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci*. (2001) 16:101–33. doi: 10.1214/SS/1009213286

28. Hodges JL. The significance probability of the mirnovv two-sample test. *Ark Mat*. 3:469–86. doi: 10.1007/BF02589501

29. Fan C, Chen M, Wang J, Huang B. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Front. Energy Res.* (2021) 9:652801. doi: 10.3389/fenrg.2021.652801

30. Jollife IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans R Soc A Math Phys Eng Sci*. (2016) 374:20150202. doi: 10.1098/rsta.2015.0202

31. Greenacre M, Blasius J. *Multiple correspondence analysis and related methods*. Boca Raton, FL: CRC Press (2006).

32. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. (1987) 20:53–65. doi: 10.1016/0377-0427(87)90125-7

33. Stephens MA. EDF statistics for goodness of fit and some comparisons. *J Am Stat Assoc*. (1974) 69:730–7. doi: 10.1080/01621459.1974.10480196

34. Sundqvist A, Hemberg J. Adolescents' and young adults' experiences of loneliness and their thoughts about its alleviation. *Int J Adolesc Youth*. (2021) 26:238–55. doi: 10.1080/02673843.2021.1908903

35. O'Connor RC, Wetherall K, Cleare S, McClelland H, Melson AJ, Niedzwiedz CL, et al. Mental health and well-being during the COVID-19 pandemic: longitudinal analyses of adults in the UK COVID-19 Mental Health & Wellbeing study. *Br J Psychiatry*. (2021) 218:326–33. doi: 10.1192/BJP.2020.212

36. Bécares L, Kneale D. Inequalities in mental health, self-rated health, and social support among sexual minority young adults during the COVID-19 pandemic: analyses from the UK millennium cohort study. *Soc Psychiatry Psychiatr Epidemiol*. (2022) 57:1979–86. doi: 10.1007/S00127-022-02291-1

37. Carr MJ, Steeg S, Webb RT, Kapur N, Chew-Graham CA, Abel KM, et al. Effects of the COVID-19 pandemic on primary care-recorded mental illness and self-harm episodes in the UK: a population-based cohort study. *Lancet Public Health*. (2021) 6:e124–35. doi: 10.1016/S2468-2667(20)30288-7

38. Balcombe L, De Leo D. Evaluation of the use of Digital mental Health platforms and interventions: scoping review. *Int J Environ Res Public Health*. (2023) 20:362. doi: 10.3390/ijerph20010362

39. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. *J Med Internet Res*. (2017) 19:e176. doi: 10.2196/jmir.7332

TYPE Original Research
PUBLISHED 11 December 2023
DOI 10.3389/fpsyt.2023.1217649

# Identifying features of risk periods for suicide attempts using document frequency and language use in electronic health records

Rina Dutta[1,2]*, George Gkotsis[1], Sumithra U. Velupillai[1], Johnny Downs[1,2], Angus Roberts[1], Robert Stewart[1,2] and Matthew Hotopf[1,2]

[1]King's College London, IoPPN, London, United Kingdom, [2]South London and Maudsley NHS Foundation Trust, London, United Kingdom

Check for updates

OPEN ACCESS

EDITED BY
Andreea Oliviana Diaconescu,
University of Toronto, Canada

REVIEWED BY
Xinyuan Yan,
University of Minnesota Twin Cities,
United States
Radwan Qasrawi,
Al-Quds University, Palestine

*CORRESPONDENCE
Rina Dutta
✉ rina.dutta@kcl.ac.uk

RECEIVED 05 May 2023
ACCEPTED 13 November 2023
PUBLISHED 11 December 2023

CITATION
Dutta R, Gkotsis G, Velupillai SU, Downs J, Roberts A, Stewart R and Hotopf M (2023) Identifying features of risk periods for suicide attempts using document frequency and language use in electronic health records. *Front. Psychiatry* 14:1217649.
doi: 10.3389/fpsyt.2023.1217649

**Background:** Individualising mental healthcare at times when a patient is most at risk of suicide involves shifting research emphasis from static risk factors to those that may be modifiable with interventions. Currently, risk assessment is based on a range of extensively reported stable risk factors, but critical to dynamic suicide risk assessment is an understanding of each individual patient's health trajectory over time. The use of electronic health records (EHRs) and analysis using machine learning has the potential to accelerate progress in developing early warning indicators.

**Setting:** EHR data from the South London and Maudsley NHS Foundation Trust (SLaM) which provides secondary mental healthcare for 1.8 million people living in four South London boroughs.

**Objectives:** To determine whether the time window proximal to a hospitalised suicide attempt can be discriminated from a distal period of lower risk by analysing the documentation and mental health clinical free text data from EHRs and (i) investigate whether the rate at which EHR documents are recorded per patient is associated with a suicide attempt; (ii) compare document-level word usage between documents proximal and distal to a suicide attempt; and (iii) compare n-gram frequency related to third-person pronoun use proximal and distal to a suicide attempt using machine learning.

**Methods:** The Clinical Record Interactive Search (CRIS) system allowed access to de-identified information from the EHRs. CRIS has been linked with Hospital Episode Statistics (HES) data for Admitted Patient Care. We analysed document and event data for patients who had at some point between 1 April 2006 and 31 March 2013 been hospitalised with a HES ICD-10 code related to attempted suicide (X60–X84; Y10–Y34; Y87.0/Y87.2).

**Findings:** $n = 8{,}247$ patients were identified to have made a hospitalised suicide attempt. Of these, $n = 3{,}167$ (39.8%) of patients had at least one document available in their EHR prior to their first suicide attempt. $N = 1{,}424$ (45.0%) of these patients had been "monitored" by mental healthcare services in the past 30 days. From 60 days prior to a first suicide attempt, there was a rapid increase in the monitoring level (document recording of the past 30 days) increasing from 35.1 to 45.0%. Documents containing words related to prescribed medications/drugs/overdose/poisoning/addiction had the highest odds of being a risk indicator used proximal to a suicide attempt (OR 1.88; precision 0.91 and recall 0.93), and documents with words citing a care plan were associated with the lowest risk for a suicide attempt (OR 0.22; precision 1.00 and recall 1.00). Function words, word sequence, and pronouns were most common in all three representations (uni-, bi-, and tri-gram).

**Conclusion:** EHR documentation frequency and language use can be used to distinguish periods distal from and proximal to a suicide attempt. However, in our study 55.0% of patients with documentation, prior to their first suicide attempt, did not have a record in the preceding 30 days, meaning that there are a high number who are not seen by services at their most vulnerable point.

KEYWORDS

**suicide, risk, electronic health records, language, assessment**

# Introduction

## Background

Individualising psychiatric care at times when patients are most at risk of suicide involves shifting research emphasis from static risk factors to those that may be modifiable with interventions (1, 2).

## Suicide risk assessment

Currently, risk assessment is based on a range of extensively reported risk factors gleaned from case–control studies using a psychological autopsy approach or nested within large register-based cohort studies (3). Critical to dynamic suicide risk assessment is an understanding of each individual patient's health trajectory over time.

## Electronic health records

Medical records provide a chronological account of healthcare and are designed to be updated by all members of the multidisciplinary team (4). With the adoption of Electronic Health Records (EHRs) in both outpatient and hospital-based care by many healthcare providers, there is an opportunity to generate artificial intelligence-based insights from the analysis of the entire patient record (5). There are of course potential challenges posed owing to the accuracy of data held, consistency of recording, and comprehensiveness of data completion (6). However, for clinicians, it can also be the metadata which is revealing. For example, little is reported about how EHR documentation changes prior to a suicide attempt or even the proportion of those known to services who have a recorded interaction in the time preceding a suicide attempt (7).

## Data-driven modelling

Recent studies using longitudinal EHRs to predict suicidal behaviour have moved away from traditional statistical analyses (which typically produce an algorithm of up to 20 factors (8) but often overfit to high-dimensional data). The move has been towards data-driven modelling approaches, such as the Naïve Bayesian classifier model (9), Random forests (10, 11) or ensemble learning, including combination predictions from elastic net penalised logistic regression, Random forests, gradient boosting, and neural networks (12).

## Natural language processing

Other approaches have analysed the text used in EHRs using natural language processing (NLP) to investigate whether it adds predictive value to existing suicide risk models, e.g., extracting clinical concepts that are then annotated with Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS) (13) or using a general-domain sentiment analysis tool to assess the utility of words conveying positive or negative emotion (i.e., valence) (14). To make the unstructured text computable, existing standard vocabularies [e.g., those used in healthcare and biomedical sciences for UMLS (13)] or curated lists of subjectively valence-conveying terms (e.g., an included lexicon of nearly 3,000 words annotated for polarity [negativity vs. positivity rated −1 to +1 (14)] are used.

## Scientific approaches in this study

We investigated whether the rate at which EHR documents are recorded per patient is associated with a suicide attempt. We hypothesised that by aligning to the first suicide attempt, it would be possible to identify an increasing trend in EHR documentation detecting the impending occurrence of a suicide attempt.

We realised one avenue that had not been explored in the field was domain experts themselves creating the categories based on available text to investigate whether there are differences in word usage between times proximal and distal to a suicide attempt.

As a complementary analysis to this "presence/absence" method, where the more local context around the word usage is, by definition, lost, and where very common words such as prepositions would not be captured, we also performed what we call an "*n*-gram frequency analysis." Changes in the length and frequency of sequential co-occurrence of words (n-grams) have been studied for other clinical use cases in the unstructured content of EHRs, e.g., oncology notes (15). We hypothesised that n-grams related to third-person pronoun use would emerge with increasing frequency nearer the date of attempted suicide as had been found in the clinical notes of veteran outpatients who died from suicide, compared to those who did not (16).

To overcome the challenges inherent to the way the data are locked in the free text of EHRs, we report on three measures to compare the proximal and distal periods from a suicide attempt: (i) rate of EHR documentation, (ii) categorisation of words used by clinicians in free text, and (iii) n-gram frequency related to third-person pronoun use.

## Materials and methods

We studied mental health service utilisation data and clinical free text data from 30 days time windows prior to suicide attempts and compared these to distal periods of lower risk. The selection of a 30 days window was based both on clinical knowledge of changes in mental health prior to an attempt and because 30 days windows have been used in other studies to train predictive models of suicide attempt risk (17).

The cohort of patients assessed in this study was assimilated from the South London and Maudsley NHS Foundation Trust (SLaM) Biomedical Research Centre (BRC) Clinical Record Interactive Search database: a case register system that provides de-identified information from electronic health records (EHRs) relating to secondary and tertiary mental healthcare services across 4 boroughs of South-East London and over 50 specialist services (18). SLaM provides secondary mental healthcare to a population of approximately 1·8 million residents of Lambeth, Southwark, Lewisham, and Croydon and national specialist services. EHRs have been used comprehensively across all SLaM services since 2006. CRIS was established in 2008 to allow searching and retrieval of full but de-identified clinical information for research purposes with permission for secondary data analysis, approved by the Oxfordshire Research Ethics Committee C (reference 08/H0606/71 + 5). As of 10 February 2017, CRIS contained clinical records on 277,700 patients, 176,242 of whom had contact with SLaM between April 1, 2006 and March 31, 2013, the period of interest for this study, for which there were data available, with at least one documented "event", or attachment, e.g., correspondence, in common word processed format. The event field of the EHR is used by clinicians to enter notes regarding a patient's history, mental state examination, progress, or risk in free text format.

CRIS has been linked with Hospital Episode Statistics (HES) data for Admitted Patient Care. HES is a national administrative database containing patient-level records of all admissions to NHS hospitals in England. Static extracts of HES data are linked to CRIS data within the Health and Social Care Information Centre and provided to the SLaM BRC with all identifiers removed. HES data are available within CRIS for all patients who have had any contact with SLaM services since 2006, regardless of where they were living at the time of their hospital use. Linked HES data were available up to 31 March 2013. Each record in HES corresponds to a finished consultant episode, during which a patient is under the care of an individual consultant. A hospital admission comprises a continuous time period of HES episodes.

### Identifying hospitalised suicide attempts

Our study included event and attachment data from $n = 8{,}247$ SLaM patients who had at some point between April 1, 2006 and March 31, 2013 been hospitalised with a HES ICD-10 code related to attempted suicide (X60–X84; Y10–Y34; Y87.0 / Y87.2; as described in http://www.ons.gov.uk/ons/dcp171778_351100.pdf). For these patients, all HES admission data were retrieved, even if they were unrelated to suicide. Some episodes formed part of a suicide-related admission or a completely different, non-suicide-related admission. Episodes were consolidated into hospital spells covering a patient's total length of stay in a hospital (i.e., a hospital admission) and from these only suicide-related admissions ($n = 12{,}798$) were retained for analysis (see Figure 1). We included 7,965 patients with at least one event or attachment available, about whom more than 1.5 million documents had been written.

### Rate of EHR documentation/monitoring level

We investigated whether the rate at which EHR documents are produced per patient is associated with a suicide attempt. For this analysis, we only considered the first HES-identified suicide attempt between April 1, 2006 and March 31, 2013 and aligned all patients by this date.

For any given date prior to the first suicide attempt, we defined monitoring level as the number of documents produced for each patient for a fixed time window. We also normalised this value by dividing it by the size of the time window considered. For example, in our approach, we considered a time window of 30 days; therefore, if 30 documents had been produced in the preceding 30 days, the average daily rate—denoted as $MonitoringLevel_{30}$—equalled 1. We only considered each patient as under monitoring at a given date if there was at least one document prior to that given date.

### Proximal and distal corpora selection and pre-processing

We aimed to compare documents entered by clinicians in two distinct time periods: (i) the *proximal* period comprising documents produced between 31 days and 1 day prior to a hospital admission linked to a suicide attempt and (ii) the *distal* period, including all documents created between 365 and 300 days prior to the first admission and all documents created between 365 and 300 days before any other admission, but *not less than* 300 days following their previous suicide-related admission (see Figure 2). For the proximal period, we retrieved 25,848 documents from 1,766 patients and for the distal period, we extracted 15,226 documents relating to 1,021 patients. 658 patients contributed documents to both the distal and proximal periods (see Figure 2). As we aimed to analyse the text in these documents, we only retained those documents that contained more than 100 characters.

### Word extraction for categorisation

We used standard corpus techniques to find the most discriminating words in the documents (19). We extracted the text from all 25,848 proximal documents and 15,226 distal documents.

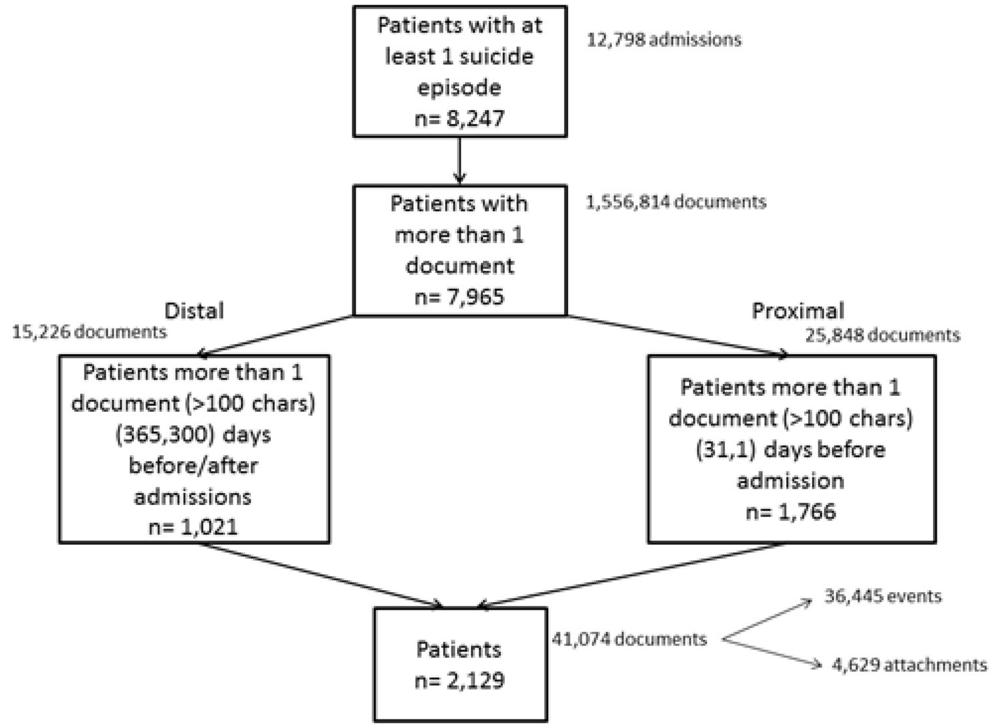

Patients with at least 1 suicide episode n= 8,247
12,798 admissions
Patients with more than 1 document n= 7,965
1,556,814 documents
Distal
15,226 documents
Proximal
25,848 documents
Patients more than 1 document (>100 chars) (365,300) days before/after admissions n= 1,021
Patients more than 1 document (>100 chars) (31,1) days before admission n= 1,766
Patients n= 2,129
41,074 documents
36,445 events
4,629 attachments

FIGURE 1
Derivation of the patient cohort and corpora of distal and proximal documents.

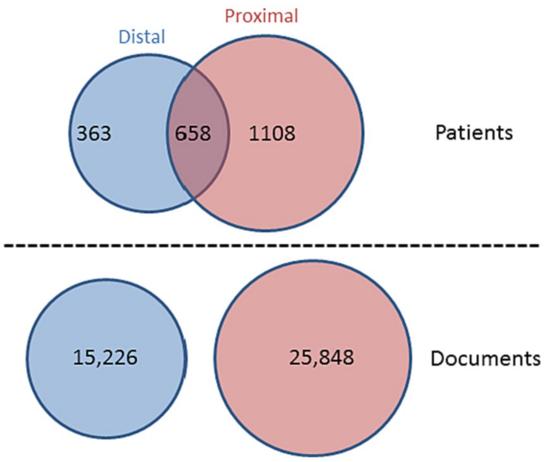


FIGURE 2
Venn diagram showing $n = 2,787$ patients contributing data to the analysis, with $n = 15,226$ documents pertaining to the distal period (between 365 and 300 days prior/following a suicide-related hospital admission) and $n = 25,848$ documents relating to the proximal period (between 31 days and 1 day prior to a suicide-related admission).

We applied Part-of-Speech (POS) tagging using spaCy[1] and replaced all words with their POS label except for the words identified as nouns, pronouns, or verbs. We applied lemmatisation to the words that were retained. We examined each word to see if their presence (or absence) yields any discriminative power. For instance, if the word “overdose” is more prominent in the proximal period, we expect that documents that used this word at least once would be present more frequently in the proximal period. To assess words for their discriminability, we considered odds ratios, where objects are documents and their class is the period from which they originated.

We examined all words retrieved from our corpus and retained those words ($n = 631$) that had $p$-value $\leq 0.05$ and odds ratio either lower than 0.66 or higher than 1.50. A senior clinician in mental health (RD) went through the list, excluding abbreviations (e.g., tc and pas), mentions of dates (e.g., 8th and 26th), times (e.g., 11 am and 5 pm), service-specific locations (e.g., Southwark and Ladywell), words of ambiguous meaning when not in context (e.g., paper—could be Mental Health Act (1983) Section paper, or paper used in Occupational Therapy activity; clean, clear—multiple meanings depending on context). $N = 390$ words were retained for human “topic modelling”.

We considered using computational topic modelling (20), but noted that computationally derived topics and representative terms are not always the same as the concepts used by clinicians (21). We therefore used a human-based topic model, in which clinician input was used to filter words and derive topics from those with discriminative power. We restricted the words considered for modelling to nouns and verbs as these are more likely to make a semantic contribution to the text. We also manually filtered out discriminative words that did not contribute to clinical interpretation.

RD curated the initial list by manually grouping them into clusters of similar meaning. RD formulated structural descriptions of each category based on empirical observations of the data (see Appendix 1).

[1] https://spacy.io

A second senior clinician (JD) was then given the precompiled list of categories and asked to assign all $n = 390$ words to them, without introducing any additional categories. The odds ratios for each group were then calculated as we had done previously for individual words. We considered a document as exposed if it contained at least one word from a given group.

## N-gram frequency analysis

We applied a machine-learning classification algorithm to the corpus, to classify each document as either distal or proximal (binary classification) and extracted the most informative *n*-gram features as found by the classifier.

An *n*-gram is a sequence of *n*-words in a text. For instance, for the word sequence, "*the patient is not suicidal.*" a uni-, bi- and tri-gram (1, 2, and 3) representation would be ["*the,*" "*patient,*" "*is,*" "*not,*" "*suicidal,*" "*.*"], ["*the patient,*" "*patient is,*" "*is not,*" "*not suicidal,*" "*suicidal.*"], and ["*the patient is,*" "*patient is not,*" "*is not suicidal,*" "*not suicidal.*"], respectively. This is a common model for representing text content in NLP classification tasks (22). We lemmatized the corpora using SpaCy and then applied the Naïve Bayes classification algorithm as implemented in the Python scikit-learn toolkit (23) using each of the three representations and then extracted the top 30 most informative features from each classification model. Informative features were those that contributed the most to discerning whether a document is distal or proximal.

The $n = 90$ resultant uni-, bi-, and tri-grams were then analysed and sorted with respect to their ORs in relation to their mean frequency of occurrence per document in the entire corpus. In this way, we were also able to analyse each feature with respect to whether it was informative for discerning a document as distal or proximal.

We analysed the n-grams in the following way: (i) an overall analysis of word types (part-of-speech and content), (ii) an analysis with respect to the feature's OR, and (iii) an analysis with respect to *n*-gram content, e.g., whether or not similar words/word sequences were consistently scored as informative in the three representations.

# Results

Of the 8,247 SLaM patients who had at some time between April 1, 2006 and March 31, 2013 been hospitalised with a HES ICD-10 code related to attempted suicide (X60–X84; Y10–Y34; Y87.0/Y87.2), $n = 4,607$ (55.9%) were female, and the median age at first admission was 33 years (IQR 22–44; mean: 34.6 years and SD: 15.4 years).

## Documentation level prior to the first suicide attempt

Only 3,167 (39.8%) of patients who had made a suicide attempt had at least one document available in their EHR prior to their first suicide attempt. $N = 1,424$ (45.0%) of these patients had been monitored by mental healthcare services in the past 30 days. Yet the majority ($n = 1,743$; 55.0%) of patients with documentation prior to their first suicide attempt did not have an EHR in the preceding 30 days.

The percentage of patients with more than one document in the preceding 30 days is generally within the range of 32.1–36.9%. However, from 60 days prior to a first suicide attempt, there is an exponential-like increase in the monitoring level in the past 30 days (increasing from 35.1 to 45.0%) (Figure 3).

## Comparison of document-level word categorisation between proximal and distal data

The list of $n = 390$ words retained for topic modelling was categorised into 17 groups (7 "protective" [PROT-A to PROT-G] for suicide attempt with OR < 0.66 (no. of exposed docs = 9,801); 10 "risk-related" [RISK-H to RISK-Q] with OR > 1.50 (no. of exposed docs = 62,118). (Refer to the Appendix 1 for comprehensive descriptions of each category, the number of words in each category, and the numbers of documents analysed with examples of words used in the EHR free text. The complete list of $n = 390$ words may be obtained from the authors upon request).

The groups vary in size: the smallest group containing $n = 3$ words (senior healthcare professional roles) and the largest $n = 91$ words (suicide "risk" terms and formal clinical distancing language). The odds ratios for each group were calculated, and these are summarised, along with precision, recall, and F1 scores, in Table 1.

The clear diagonal shown in the confusion matrix indicates the overall high level of agreement between the two annotators (Cohen's kappa coefficient (κ) 0.82). There was more disagreement between Risk-I to Risk-N and Risk Q, which were the most challenging categories to define and also had the highest prevalence of words per group (Figure 4).

## N-gram frequency analysis

The majority of words used in both proximal and distal time windows are function words and pronouns. In all three representations (uni-, bi-, and tri-grams), function words (e.g., *to*, *by*, *on the*, *to the*, and *there be no*) and pronouns (e.g., *he*, *she*, *he have*, *she do*, and *that she would*) were most common.

A few verbs and nouns were also found to be informative. Reporting verbs such as say, state, and report were identified, e.g., *she say that—say that—she say she—she say—say that she—state that—state that she—report that she.* Other verbs included *feel* and *want*, e.g., *want to—do not want—not want to—she want to.* Nouns were only found as parts of bi- or tri-grams, e.g., *the ward*, *self-harm—of self-harm.*

In relation to their odds ratios, the features most informative for the distal documents were male pronouns—*his, he, he be*, while *self-harm* and female pronouns were more informative for the proximal documents. For function words such as *to and of*, the number of times they need to be present in a document for them to be distinctive is $n = 7$. The highest proximal scores are generally bi- and tri-grams, while uni- and bigrams are generally related to distal periods.

When comparing the content of the *n*-grams, many features were captured in all three representations, such as pronouns and function words. This confirmed the informativeness of unique words. The

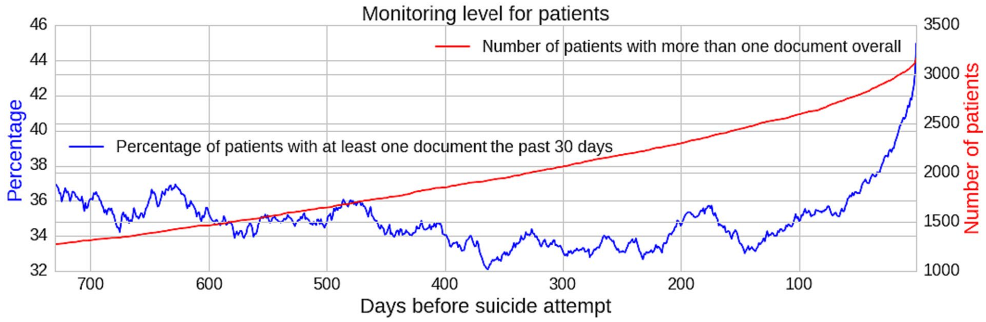


FIGURE 3
Graph showing monitoring level for patients indicating an increase in monitoring level proximal to a suicide attempt. Monitoring level numerator in blue and denominator in red.

TABLE 1 17 categories of words used by clinicians in free text with ORs of proximal to distal use, with precision, recall, and F1 scores (for fuller descriptions of categories refer to Appendix 1).

| Group | Category description | Odds ratio | Precision | Recall | F1 score |
|---|---|---|---|---|---|
| PROT-A | Care plan | 0.22 | 1.00 | 1.00 | 1.00 |
| PROT-B | Senior healthcare professional role | 0.32 | 0.60 | 1.00 | 0.75 |
| PROT-C | Chronic physical comorbidity / symptom | 0.50 | 0.67 | 0.44 | 0.53 |
| PROT-D | Treatment for drug addiction or depot treatment | 0.55 | 1.00 | 0.80 | 0.89 |
| PROT-E | Food/meals/activities | 0.57 | 0.95 | 0.88 | 0.91 |
| PROT-F | Positive connotations | 0.58 | 0.56 | 0.75 | 0.64 |
| PROT-G | Items used on ward | 0.58 | 0.80 | 0.89 | 0.84 |
| RISK-H | Items of clothing | 1.54 | 1.00 | 0.89 | 0.94 |
| RISK-I | Subheadings of clerking/diagnosis/psychiatric symptoms | 1.62 | 0.77 | 0.80 | 0.79 |
| RISK-J | Interventions | 1.62 | 0.75 | 0.62 | 0.68 |
| RISK-K | Time- or life event- or person/relationship-related | 1.63 | 0.89 | 0.93 | 0.91 |
| RISK-L | Suicide “risk” terms and formal clinical distancing language | 1.64 | 0.79 | 0.79 | 0.79 |
| RISK-M | Implement/mechanism of self-harm or suicide attempt | 1.72 | 0.79 | 0.93 | 0.86 |
| RISK-N | Negative connotations/judgemental language | 1.72 | 0.86 | 0.63 | 0.73 |
| RISK-O | Physical symptom or sign | 1.84 | 0.60 | 0.67 | 0.63 |
| RISK-P | Junior or multidisciplinary healthcare professional role | 1.85 | 1.00 | 0.78 | 0.88 |
| RISK-Q | Prescribed medications/drugs/overdose/poisoning/addiction | 1.88 | 0.91 | 0.93 | 0.92 |

bi- and tri-grams gave a “richer picture” of why some unigrams are found informative by the classifier.

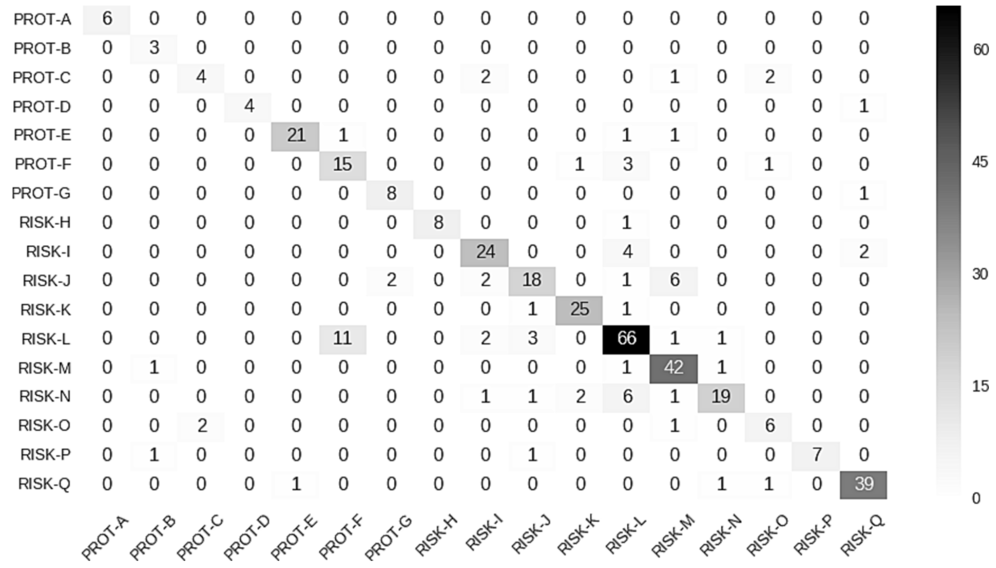
For the proximal period, the most distinctive n-grams were “self harm,” “she want to,” “of self harm,” also the distancing phrases “report that she,” “not want to.” For the distal period, “his,” “he,” “he be” and “he have” were the most informative features.

# Discussion

## Proportion of patients with documentation prior to the first suicide attempt

Our finding of approximately 40% of patients having at least one document available in their EHR prior to their first suicide attempt was congruent with a recent analysis of national trends in suicide attempts and mental health service use for adults in the US, where only approximately 40% had documented service use in the prior 12 months (24). This is of interest given they were both population-based samples but with widely different healthcare systems (25). An earlier US study conducted on an insured sample had a much higher proportion (95%) of mental healthcare contact prior to a suicide attempt (26), yet the national study by Bommersbach et al. (24) of all people who attempted suicide, regardless of insurance or treatment-seeking behaviour, paralleled our findings in the UK where we were studying the population served by the National Health Service.

Similarly, although we did not have access to primary care or Emergency Department notes, our finding of 45.0% of patients with prior records having a record documented by mental healthcare services in the past 30 days was in keeping with the frequently quoted 50% of all adults who die by suicide visiting a healthcare professional in the 4 weeks before their death (27).



| | PROT-A | PROT-B | PROT-C | PROT-D | PROT-E | PROT-F | PROT-G | RISK-H | RISK-I | RISK-J | RISK-K | RISK-L | RISK-M | RISK-N | RISK-O | RISK-P | RISK-Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PROT-A | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PROT-B | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PROT-C | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| PROT-D | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| PROT-E | 0 | 0 | 0 | 0 | 21 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| PROT-F | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 1 | 0 | 0 |
| PROT-G | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| RISK-H | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| RISK-I | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 24 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 2 |
| RISK-J | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 18 | 0 | 1 | 6 | 0 | 0 | 0 | 0 |
| RISK-K | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 25 | 1 | 0 | 0 | 0 | 0 | 0 |
| RISK-L | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 2 | 3 | 0 | 66 | 1 | 1 | 0 | 0 | 0 |
| RISK-M | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 42 | 1 | 0 | 0 | 0 |
| RISK-N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 6 | 1 | 19 | 0 | 0 | 0 |
| RISK-O | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 0 | 0 |
| RISK-P | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| RISK-Q | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 39 |

FIGURE 4
Confusion matrix showing the degree of interannotator agreement across the 17 categories.

## Documentation increases prior to the first suicide attempt

Our study confirmed the specific characteristics of the time period proximal to a suicide attempt which discriminates it from a distal period of lower risk. Firstly, recognition of the large increase in documentation level in the past 30 days, detectable from 60 days prior to a first suicide attempt, was only possible because of the format of recording in EHRs as opposed to paper records.

## Word categories associated with the highest and lowest risk of suicide

The category associated with the highest risk of suicide was that incorporating prescribed medications/drugs/overdose/poisoning/addiction, whereas terms associated with treatment for drug addiction or depot treatment were associated with a lower risk of suicide.

This accords with the current literature. For example, when using administrative data to predict suicide after psychiatric hospitalisation in the Veterans Health Administration System, 2 of the top 10 predictors in the Super Learner ensemble machine-learning model created were associated with drug dependence (28). Using predictive structured–unstructured interactions in EHR models, Bayramli et al. (29) showed that drug abuse or specifically named illicit drugs were the structured feature associated with greater suicide risk for many feature pairs (29).

Interestingly, this same study is one of the only published articles to specifically study apparent "protective factors" against suicide as we did (29). Concepts such as mammograms for malignant neoplasm of the breast, osteoporosis, and haemorrhoids were associated with lower risk which was analogous to our "protective" chronic physical comorbidity/symptom category. Of course, there are issues of confounding with older age, which is protective of suicide attempt risk.

Mention of a senior healthcare professional was associated with a lower risk of suicide, contrary to what was found for junior healthcare professionals. However, this is most likely confounding by indication, i.e., junior healthcare professionals being more involved in healthcare provision (30) and their roles cited in the EHR proximal to a suicide attempt, rather than being directly linked to suicide risk. Similarly, word categories which were ascribed as "protective" according to their odds ratios may simply be incidental words used to describe patients' activities at times of low risk (e.g., food/meals/activities and items used on the ward).

A particularly interesting category was the clothing one which was associated with higher risk. Items of clothing can be used for ligatures (31), or comments can be recorded in the EHR regarding items of clothing patients bring in as property. Where a term was ambiguous, e.g., tie [which could be assigned to "implement / mechanism of self-harm or suicide attempt" (M) or "clothing" (H)], the consensus was to assign to the category conveying the highest potential risk (M). In the end, the "clothing" category was similarly categorised as of increased risk.

It was interesting that two of the categories were directly related to valence: terms with negative connotation/judgemental language being associated with increased risk, and words imparting positive connotation being "protective". In our previous research studying six general-purpose sentiment lexicons for suicide risk assessment in EHRs (32), we found that many of the most representative keywords in the suicide-related subcorpus were not identified by any of the lexicons. The corpus word frequencies for the proximal and distal periods could be used as a guide to the inclusion of words in a novel lexicon, merging healthcare terminology as another source.

## Contextual language proximal and distal to a suicide attempt

The complementary aspect of using the n-gram method was that it allowed us to analyse word usage that captured common words/word sequences and contextual information, meaning that the proximal and distal periods could be compared based upon a contiguous sequence of n-words rather than single words (33).

Although single-word frequencies are associated with patient status and can therefore provide useful indicators of risk, single words suffer from a lack of this contextual information. For example, the same word can be used in both an affirmative and a negative context or contexts describing people other than the patient. By including surrounding context, n-grams allowed us to increase the predictive value of the textual indicators used. There is, however, some loss of sensitivity as the length of the n-grams increase: given the variable nature of language, long text sequences are less likely to provide generalisable descriptions of clinical status.

Using U.S. Veterans Administration medical records, Poulin et al. (34) generated datasets of single keywords and multi-word phrases and constructed prediction models using a machine-learning algorithm. They showed that methodologically word pairs were more useful than single words for suicide predictive model construction (34). Basic NLP features, including n-gram features, have also been used for psychiatric stressor recognition from clinical notes to study the association with suicidal behaviours (35).

Whereas gender differences in psychosocial and clinical determinants of suicide risk have been studied using EHRs (36), differences in language used have not been researched to a large extent. In our study, female pronouns being more informative for the proximal documents and male for the distal documents do not merely reflect the numbers of female and male patients, given only a slightly higher proportion of patients (55.9%) were female. Further study is needed to investigate whether clinicians document differently for female and male patients in the time leading to a suicide attempt. One study reported quoting "he/she says" is increased in records of clinician–patient interactions that involve the communication of bad news between doctor and patient (37), and this would be worth further investigation to see whether reporting styles become more formal or defensive (38) when clinicians are concerned about risk. Clinician narrative style in EHRs, e.g., use of quoted patients' speech, has not been investigated in detail to date (39).

## Potential improvements to current EHR systems

In a review of 40 studies of the impact of EHRs on information practices in mental health contexts, Kariotis et al. (40) found that EHRs improved the amount of information documented. However, if EHRs do not include search functions or data visualisation strategies, navigating the amount of data contained in clinical notes can be challenging (40). Visualising source data from multiple domains (e.g., using Cogstack (41) or NeuroBlu (42)) can enable dynamic monitoring of risk over time, and the rate of documentation could be one aspect of this for risk of a suicide attempt. Natural language processing techniques, either rule-based, machine learning-based, or deep learning-based, can be used to extract information from clinical narratives (43). The next stage is then to build automated alerting systems with all predictive features to ensure that clinicians are notified of patients at risk so that appropriate actions can be pursued.

## Strengths and limitations

As a proxy for hospitalised suicide attempts and to study the more severe end, we purposively used HES admission data knowing that is more reliable than HES emergency department data but misses non-admitted episodes of self-harm (44, 45). Identifying suicide, self-harm, or even suicidal ideation using NLP would allow the analysis to be conducted on a broader group (46, 47).

The novel approach in this analysis was to move away from a case–control study design to consider whether it was possible to discriminate between EHR documentation proximal and distal to suicide attempts using three features of free text documentation. The main limitation was not using the features studied and other predictors in a predictive model. However, our aim was to analyse what aspects of EHR documentation and language used by clinicians change nearer to the time of a suicide attempt.

A drawback of concentrating on EHRs from a mental health trust was that we were unable to link with notes made in primary care, the general hospital, or Emergency departments as these are on separate systems.

## Conclusion

Despite its importance, clinical record keeping is often given a low priority and there is inconsistency between the entries by different healthcare professionals, yet patterns emerge in changes in documentation level, topic categories of words, and n-grams prior to a suicide attempt. More automated means of leveraging unstructured data from daily clinical practice is crucial as access to individual-level health information increases. The widespread use of EHRs has the potential to accelerate progress in developing both healthcare and research. Adopting clinical dashboards to visualise change may be particularly helpful to understand changes in suicide risk for individual patients over time.

## Data availability statement

The data analysed in this study is subject to the following licenses/restrictions: Data are owned by a third-party South London and Maudsley Biomedical Research Centre Clinical Record Interactive Search tool that provides access to anonymised data derived from electronic medical records of the South London and Maudsley National Health Service Foundation Trust. These data can only be accessed by permitted individuals from within a secure firewall (i.e., remote access is not possible, and the data cannot be sent elsewhere) in the same manner as the authors. Requests to access these datasets should be directed to rina.dutta@kcl.ac.uk.

## Ethics statement

Ethical approval/written informed consent was not required for the study of human electronic health record data in accordance with the local legislation and institutional requirements.

## Author contributions

RD was lead author and involved in conceptualising the study, methodology, analysing data, and writing and finalising the draft. GG and SV provided data curation, methodology support, and support with reviewing and editing the final draft. JD and AR provided support with methodology and analysis. RS and MH provided support with resources, and review and edit of the final draft. All authors contributed to the article and approved the submitted version.

## Funding

This study utilised the Clinical Record Interactive Search (CRIS) platform funded and developed by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, and a joint infrastructure grant from Guy's and St Thomas' Charity and the Maudsley Charity (grant number BRC-2011-10035). RD and GG were funded by a Clinician Scientist Fellowship (project e-HOST-IT) from the Health Foundation in partnership with the Academy of Medical Sciences. RD was also funded by an MRC award (grant code: MR/S020365/1). SV was funded by the Swedish Research Council (2015-00359), Marie Skłodowska-Curie Actions, COFUND, Project INCA 600398. JD received support from an NIHR Clinician Scientist Fellowship (CS-2018-18-ST2-014) and Psychiatry Research Trust Peggy Pollak Research Fellowship in Developmental Psychiatry. AR is part-supported by Health Data Research UK, an initiative funded by UK Research and Innovation, the Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. RS is part-funded by an NIHR Senior Investigator Award and by the NIHR Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust. This study represents independent research part-funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. For the purposes of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Accepted Author Manuscript version arising from this submission.

## Conflict of interest

MH leads the RADAR-CNS consortium, a private–public pre-competitive collaboration on mobile health, through which King's College London receives in-kind and cash contributions from Janssen, Biogen, UCB, Merck, and Lundbeck.

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.

## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1217649/full#supplementary-material

## References

1. APA. *The American Psychiatric Association practice guidelines for the psychiatric evaluation of adults: guideline III. Assessment of suicide risk*. Arlington, VA: American Psychiatric Association (2016).

2. Velupillai S, Hadlaczky G, Baca-Garcia E, Gorrell GM, Werbeloff N, Nguyen D, et al. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. *Front Psych*. (2019) 10:36. doi: 10.3389/fpsyt.2019.00036

3. Pirkis J, Nicholas A, Gunnell D. The case for case-control studies in the field of suicide prevention. *Epidemiol Psychiatr Sci*. (2019) 29:e62. doi: 10.1017/S2045796019000581

4. Abdelrahman W, Abdelmageed A. Medical record keeping: clarity, accuracy, and timeliness are essential. *BMJ*. (2014) 348:f7716. doi: 10.1136/bmj.f7716

5. Suryanarayanan P, Epstein EA, Malvankar A, Lewis BL, DeGenaro L, Liang JJ, et al. Timely and efficient AI insights on EHR: system design. *AMIA Annu Symp Proc*. (2020) 2020:1180–9.

6. Holmes JH, Beinlich J, Boland MR, Bowles KH, Chen Y, Cook TS, et al. Why is the electronic health record so challenging for research and clinical care? *Methods Inf Med*. (2021) 60:32–48. doi: 10.1055/s-0041-1731784

7. Metzger MH, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. *Int J Methods Psychiatr Res*. (2017) 26:26. doi: 10.1002/mpr.1522

8. Fazel S, Wolf A, Larsson H, Mallett S, Fanshawe TR. The prediction of suicide in severe mental illness: development and validation of a clinical prediction rule (OxMIS). *Transl Psychiatry*. (2019) 9:98. doi: 10.1038/s41398-019-0428-3

9. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, et al. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry*. (2017) 174:154–62. doi: 10.1176/appi.ajp.2016.16010077

10. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *J Child Psychol Psychiatry*. (2018) 59:1261–70. doi: 10.1111/jcpp.12916

11. Gradus JL, Rosellini AJ, Horváth-Puhó E, Street AE, Galatzer-Levy I, Jiang T, et al. Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. *JAMA Psychiatry*. (2020) 77:25–34. doi: 10.1001/jamapsychiatry.2019.2905

12. Chen Q, Zhang-James Y, Barnett EJ, Lichtenstein P, Jokinen J, D'Onofrio BM, et al. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: a machine learning study using Swedish national registry data. *PLoS Med*. (2020) 17:e1003416. doi: 10.1371/journal.pmed.1003416

13. Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. *JAMIA Open*. (2021) 4:ooab011. doi: 10.1093/jamiaopen/ooab011

14. McCoy TH Jr, Castro VM, Roberson AM, Snapper LA, Perlis RH. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA Psychiatry*. (2016) 73:1064–71. doi: 10.1001/jamapsychiatry.2016.2172

15. Rahimian M, Warner JL, Jain SK, Davis RB, Zerillo JA, Joyce RM. Significant and distinctive n-grams in oncology notes: a text-mining method to analyze the effect of OpenNotes on clinical documentation. *JCO Clin Cancer Inform*. (2019) 3:1–9. doi: 10.1200/CCI.19.00012

16. Leonard Westgate C, Shiner B, Thompson P, Watts BV. Evaluation of veterans' suicide risk with the use of linguistic detection methods. *Psychiatr Serv*. (2015) 66:1051–6. doi: 10.1176/appi.ps.201400283

17. Walsh CG, Johnson KB, Ripperger M, Sperry S, Harris J, Clark N, et al. Prospective validation of an electronic health record–based, real-time suicide risk model. *JAMA Netw Open*. (2021) 4:e211428–8. doi: 10.1001/jamanetworkopen.2021.1428

18. Perera G, Broadbent M, Callard F, Chang CK, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust biomedical research Centre (SLaM BRC) case register: current status and recent enhancement of an electronic mental health record-derived data resource. *BMJ Open*. (2016) 6:e008721. doi: 10.1136/bmjopen-2015-008721

19. Brezina V. *Statistics in corpus linguistics* Cambridge University Press (2018). doi: 10.1017/9781316410899

20. Blei DM, Ng AY, Jordan M. Latent dirichlet allocation. *J Mach Learn Res*. (2003) 3:993–1022.

21. Miner AS, Stewart SA, Halley MC, Nelson LK, Linos E. Formally comparing topic models and human-generated qualitative coding of physician mothers' experiences of workplace discrimination. *Big Data Soc*. (2023) 10:205395172211491. doi: 10.1177/20539517221149106

22. Broder A, Glassman SC, Manasse MS, Zweig G (1997). Syntactic clustering of the web. In Computer networks and ISDN systems 298 sixth international world wide web Conference. pp. 1157–1166. Available at: https://cadmo.ethz.ch/education/lectures/FS18/SDBS/papers/broder.pdf

23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. (2011) 12:2825–30.

24. Bommersbach TJ, Rosenheck RA, Rhee TG. National Trends of mental health care among US adults who attempted suicide in the past 12 months. *JAMA Psychiatry*. (2022) 79:219–31. doi: 10.1001/jamapsychiatry.2021.3958

25. Cusick M, Velupillai S, Downs J, Campion TR Jr, Sholle ET, Dutta R, et al. Portability of natural language processing methods to detect suicidality from clinical text in US and UK electronic health records. *J Affect Disord Rep*. (2022) 10:100430. doi: 10.1016/j.jadr.2022.100430

26. Ahmedani BK, Stewart C, Simon GE, Lynch F, Lu CY, Waitzfelder BE, et al. Racial/ethnic differences in health care visits made before suicide attempt across the United States. *Med Care*. (2015) 53:430–5. doi: 10.1097/MLR.0000000000000335

27. Luoma JB, Martin CE, Pearson JL. Contact with mental health and primary care providers before suicide: a review of the evidence. *Am J Psychiatry*. (2002) 159:909–16. doi: 10.1176/appi.ajp.159.6.909

28. Kessler RC, Bauer MS, Bishop TM, Demler OV, Dobscha SK, Gildea SM, et al. Using administrative data to predict suicide after psychiatric hospitalization in the veterans health administration system. *Front Psych*. (2020) 11:390. doi: 10.3389/fpsyt.2020.00390

29. Bayramli I, Castro V, Barak-Corren Y, Madsen EM, Nock MK, Smoller JW, et al. Predictive structured-unstructured interactions in EHR models: a case study of suicide prediction. *NPJ Digit Med*. (2022) 5:15. doi: 10.1038/s41746-022-00558-0

30. Awenat Y, Peters S, Shaw-Nunez E, Gooding P, Pratt D, Haddock G. Staff experiences and perceptions of working with in-patients who are suicidal: qualitative analysis. *Br J Psychiatry*. (2017) 211:103–8. doi: 10.1192/bjp.bp.116.191817

31. Gunnell D, Bennewith O, Hawton K, Simkin S, Kapur N. The epidemiology and prevention of suicide by hanging: a systematic review. *Int J Epidemiol*. (2005) 34:433–42. doi: 10.1093/ije/dyh398

32. Bittar A, Velupillai S, Roberts A, Dutta R. Using general-purpose sentiment lexicons for suicide risk assessment in electronic health records: Corpus-based analysis. *JMIR Med Inform*. (2021) 9:e22397. doi: 10.2196/22397

33. Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Comput Math Methods Med*. (2016) 2016:8708434–8. doi: 10.1155/2016/8708434

34. Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, et al. Predicting the risk of suicide by analyzing the text of clinical notes. *PLoS One*. (2014) 9:e85733. doi: 10.1371/journal.pone.0085733

35. Zhang Y, Zhang OR, Li R, Flores A, Selek S, Zhang XY, et al. Psychiatric stressor recognition from clinical notes to reveal association with suicide. *Health Informatics J*. (2019) 25:1846–62. doi: 10.1177/1460458218796598

36. McQuaid RJ, Nikolitch K, Vandeloo KL, Burhunduli P, Phillips JL. Sex differences in determinants of suicide risk preceding psychiatric admission: an electronic medical record study. *Front Psych*. (2022) 13:892225. doi: 10.3389/fpsyt.2022.892225

37. Van De Mieroop D. The quotative "he/she says" in interpreted doctor–patient interaction. *Interpreting*. (2012) 14:92–117. doi: 10.1075/intp.14.1.05mie

38. Irving K, Treacy M, Scott A, Hyde A, Butler M, MacNeela P. Discursive practices in the documentation of patient assessments. *J Adv Nurs*. (2006) 53:151–9. doi: 10.1111/j.1365-2648.2006.03710.x

39. Jayasinghe L, Bittar A, Dutta R, Stewart R. Clinician-recalled quoted speech in electronic health records and risk of suicide attempt: a case–crossover study. *BMJ Open*. (2020) 10:e036186. doi: 10.1136/bmjopen-2019-036186

40. Kariotis TC, Prictor M, Chang S, Gray K. Impact of electronic health records on information practices in mental health contexts: scoping review. *J Med Internet Res*. (2022) 24:e30405. doi: 10.2196/30405

41. Wang T, Oliver D, Msosa Y, Colling C, Spada G, Roguski L, et al. Implementation of a real-time psychosis risk detection and alerting system based on electronic health records using CogStack. *J Vis Exp*. (2020) 159:e60794. doi: 10.3791/60794-v

42. Patel R, Wee SN, Ramaswamy R, Thadani S, Tandi J, Garg R, et al. NeuroBlu, an electronic health record (EHR) trusted research environment (TRE) to support mental healthcare analytics with real-world data. *BMJ Open*. (2022) 12:e057227. doi: 10.1136/bmjopen-2021-057227

43. Negro-Calduch E, Azzopardi-Muscat N, Krishnamurthy RS, Novillo-Ortiz D. Technological progress in electronic health record system optimization: systematic review of systematic literature reviews. *Int J Med Inform*. (2021) 152:104507. doi: 10.1016/j.ijmedinf.2021.104507

44. Clements C, Turnbull P, Hawton K, Geulayov G, Waters K, Ness J, et al. Rates of self-harm presenting to general hospitals: a comparison of data from the multicentre study of self-harm in England and hospital episode statistics. *BMJ Open*. (2016) 6:e009749. doi: 10.1136/bmjopen-2015-009749

45. Polling C, Bakolis I, Hotopf M, Hatch SL. Differences in hospital admissions practices following self-harm and their influence on population-level comparisons of self-harm rates in South London: an observational study. *BMJ Open*. (2019) 9:e032906. doi: 10.1136/bmjopen-2019-032906

46. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. *AMIA Annu Symp Proc*. (2012) 2012:1244–53.

47. Cliffe C, Seyedsalehi A, Vardavoulia K, Bittar A, Velupillai S, Shetty H, et al. Using natural language processing to extract self-harm and suicidality data from a clinical sample of patients with eating disorders: a retrospective cohort study. *BMJ Open*. (2021) 11:e053808. doi: 10.1136/bmjopen-2021-053808

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 04 January 2024
DOI 10.3389/fpsyt.2023.1256571

Check for updates

OPEN ACCESS

EDITED BY
Ann John,
Swansea University Medical School,
United Kingdom

REVIEWED BY
Saeid Komasi,
Mind GPS Institute, Iran
Ebrahim Ghaderpour,
Sapienza University of Rome, Italy

*CORRESPONDENCE
Dai-Jin Kim
✉ kdj922@catholic.ac.kr

[†]These authors have contributed equally to this work and share first authorship

RECEIVED 11 July 2023
ACCEPTED 13 December 2023
PUBLISHED 04 January 2024

CITATION
Oh J, Lee T, Chung ES, Kim H, Cho K, Kim H, Choi J, Sim HH, Lee J, Choi IY and Kim DJ (2024) Development of depression detection algorithm using text scripts of routine psychiatric interview. *Front. Psychiatry* 14:1256571. doi: 10.3389/fpsyt.2023.1256571

# Development of depression detection algorithm using text scripts of routine psychiatric interview

Jihoon Oh[1†], Taekgyu Lee[2†], Eun Su Chung[2], Hyonsoo Kim[3], Kyongchul Cho[3], Hyunkyu Kim[3], Jihye Choi[1], Hyeon-Hee Sim[1], Jongseo Lee[2], In Young Choi[4] and Dai-Jin Kim[1,4*]

[1]Department of Psychiatry, College of Medicine, Seoul St. Mary's Hospital, The Catholic University of Korea, Seoul, Republic of Korea, [2]College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea, [3]Acryl, Seoul, Republic of Korea, [4]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

**Background:** A psychiatric interview is one of the important procedures in diagnosing psychiatric disorders. Through this interview, psychiatrists listen to the patient's medical history and major complaints, check their emotional state, and obtain clues for clinical diagnosis. Although there have been attempts to diagnose a specific mental disorder from a short doctor-patient conversation, there has been no attempt to classify the patient's emotional state based on the text scripts from a formal interview of more than 30 min and use it to diagnose depression. This study aimed to utilize the existing machine learning algorithm in diagnosing depression using the transcripts of one-on-one interviews between psychiatrists and depressed patients.

**Methods:** Seventy-seven clinical patients [with depression ($n = 60$); without depression ($n = 17$)] with a prior psychiatric diagnosis history participated in this study. The study was conducted with 24 male and 53 female subjects with the mean age of 33.8 ($\pm$ 3.0). Psychiatrists conducted a conversational interview with each patient that lasted at least 30 min. All interviews with the subjects between August 2021 and November 2022 were recorded and transcribed into text scripts, and a text emotion recognition module was used to indicate the subject's representative emotions of each sentence. A machine learning algorithm discriminates patients with depression and those without depression based on text scripts.

**Results:** A machine learning model classified text scripts from depressive patients with non-depressive ones with an acceptable accuracy rate (AUC of 0.85). The distribution of emotions (surprise, fear, anger, love, sadness, disgust, neutral, and happiness) was significantly different between patients with depression and those without depression ($p < 0.001$), and the most contributing emotion in classifying the two groups was disgust ($p < 0.001$).

**Conclusion:** This is a qualitative and retrospective study to develop a tool to detect depression against patients without depression based on the text scripts of psychiatric interview, suggesting a novel and practical approach to understand the emotional characteristics of depression patients and to use them to detect the diagnosis of depression based on machine learning methods. This model could assist psychiatrists in clinical settings who conduct routine conversations with patients using text transcripts of the interviews.

KEYWORDS

**machine learning, depression, emotions, psychological interview, sentiment analysis**

## Introduction

Depression is the most prevalent mental health issue that affects hundreds of millions of people and is considered one of the leading causes of burden globally (1, 2). It is estimated that the lifetime prevalence of depression among adults was 10.8% from 1994 to 2014 (3), and the burden due to mental disorders has not been reduced despite evidence-based interventions (1). In addition, the prevalence of depression in South Korea shows an increasing trend (4).

The diagnosis and evaluation of major depressive disorder (MDD) are based on diagnostic criteria based on DSM-5, which requires a clinical judgment of a trained clinician on listed symptoms, including depressed mood, markedly diminished interest or pleasure, significant weight loss, slowing down of thought, a reduction of physical movement, fatigue or loss of energy, reduced ability to think or concentrate, and recurrent thoughts of death (5). The screening for these symptoms mainly depends on diagnostic questionnaires such as Patient Health Questionnaire-9 (PHQ-9), Beck Depression Inventory (BDI) (6), and the Hamilton Depression Rating Scale (HDRS) (7). This questionnaire-based diagnostic approach necessitates an interview with clinicians, but it can be prone to biases as they are either self-reported by patients or administered by clinicians (8).

It is important to start treatment earlier for patients with MDD because the time to treatment is correlated with the prognosis (9). A diverse range of barriers, such as education, income, and accessibility, contribute to the underdiagnosis of depression (10). As an early diagnosis of depression may reduce the severe depressive symptoms and improve the prognosis, there is a need for an objective method that can diagnose patients' emotional and depressive states.

Recent AI-based approaches have gained attraction to provide additional information on diagnosing depression. Physiological signals such as electroencephalogram (11, 12) and features from eye-blinking (13) were captured upon audio-visual stimuli to classify emotions by utilizing deep neural networks. More common approaches include applying deep learning models on audio and visual data from clinical patients and public datasets (14, 15), where widely used datasets classified facial expressions into emotional labels such as anger, disgust, fear, happiness, sadness, surprise, and neutral (16). Symptom severity of depression was measured based on the speech and 3D facial scan data in DAIC-WOZ dataset, and the convolutional neural network (CNN) model was reported to demonstrate reliable results in detecting MDD (14). Potential depression risk was tried to be identified on the video recordings of depression patients in China conducting structured tasks with a deep belief network (DBN) based model (15). There was also an audio-focused approach where patients' low-level and high-level audio features were used to estimate depression severity scores and detect depression (17).

A series of studies have focused mainly on the acoustic and text features from the conversations (18, 19). Acoustic features in spontaneous speech were used to recognize depression against the normal control, with improvement was reported in the performance using the first few sentences (18). Indirect text features from the patients, such as the total number of sentences, average words spoken in each sentence, frequency of laughter, and depression-related words, were fed into the model in addition to audio and visual features (19). However, the nature of audio and video data requires much preparation for consistent recording quality across the samples (20), and even the laboratory setup to collect audiovisual data still requires extensive pre-processing to guarantee the quality of input into the model (21).

In addition, there have not been many attempts to measure symptom severity or identify depression by directly collecting data from the psychiatric interviews between the psychiatrists and the patients, where structured psychiatric interviews are essential in making an accurate diagnosis to satisfy the categorical conditions listed in DSM-5. The interviews are still often encouraged to induce free-of-context, unstructured conversations that can illicit subjective experiences from the patients (22), as such interviews are often the single most important source of information in obtaining clinical cues for psychiatrists.

In this study, we utilized XGBoost algorithm to identify depression based on the actual psychiatric interviews between the psychiatrists and the patients. We aimed to identify patients with depression against the psychiatric patients without depression based on the text scripts of routine psychotherapy sessions to overcome burdensome requirements in collecting and pre-processing the audiovisual data that have been widely used to analyze the depression patients with machine learning methods. We classified emotional characteristics of the text scripts from the interviews on the back of the improved accuracy of text emotion recognition applications (23–25). Transcripts from psychiatric interviews are easy to collect and require minimal pre-processing, whereas audio and visual data are more complex in nature and data processing perspective. It is one of the first attempts to identify depression using text emotion recognition based on routine psychiatric interviews in the clinical setting.

The rest of this paper is organized as follows: the data acquisition process from the clinical patients and the machine learning model were presented in Materials and Methods; results of depression classification is presented in Results; summary, future works, and limitations are discussed in Discussion; and lastly Conclusions.

# Materials and methods

## Participants

Seventy-seven clinical patients (24 male, 53 female) between 20 and 65 years old participated in this qualitative and retrospective study to develop a tool to detect depression. The dialogue data were acquired in a consecutive manner from all inpatients and outpatients who agreed to record their interview during the treatment. Participants were diagnosed with depression or anxiety, with or without a current episode, established through DSM-5. The clinical diagnosis was provided by the agreement of two or more psychiatrists at Seoul St. Mary's Hospital by assessing the patients in person. Interviews with the participants were conducted from August 2021 to November 2022. All participants were required to provide informed consent forms to be considered as the subjects, and the Institutional Review Board of Seoul St. Mary's Hospital approved this study (KC21ONSI0387).

Inclusion criteria included (1) adults aged 18–65 years; (2) individuals who have received a primary diagnosis of depression (ICD codes: F32, F33, F34) from the Department of Psychiatry and have undergone treatment; (3) for the control group, individuals who have not received the diagnoses or treatment mentioned in (2); and (4) individuals who have received sufficient explanation of this clinical trial, have understood it, voluntarily decided to participate, and provided written consent to adhere to precautions.

Exclusion criteria included any current or lifetime axis I psychiatric disorders, such as schizophrenia, schizoaffective disorder, other psychotic and substance-related disorders, organic mental disorders, neurological disorders (e.g., epilepsy, dementia), and cardiovascular disorders. A total of 10 people were excluded due to intake of prohibited substances such as alcohol and psychostimulant ($n = 3$), change in diagnosis ($n = 5$), and voluntary withdrawal of consent ($n = 2$).

## Patient characteristics

Among the 77 participants, 60 subjects were diagnosed with depression, and 17 subjects had other psychiatric illnesses (Table 1). The with-depression group included 16 males (26.7%) and 44 females (73.3%), whereas the without-depression group consisted of 8 males (47.1%) and nine females (52.9%). The mean age was 33.2 (±3.3) for the with-depression group and 35.9 (±6.9) for the control group. There were no significant differences in gender and age between the two groups ($p > 0.05$, Table 1).

## Data acquisition

A psychiatrist performed a psychiatric interview with each subject in a quiet psychiatric consultation room. The interviews were conducted as part of psychotherapy, in the form of semi-structured format which included typical attributes such as daily lives, chief complaints, thought contents, cognitions, judgments, and insights. The interviews lasted 30 min or longer. All interviews were recorded under the subjects' consent, and text scripts were produced by a separate scripter for the first 15 to 20 min of the voice recordings after each interview.

Then, sentences from psychiatrist were removed from the text scripts so that only the sentences from the subjects could be left in the scripts. Emotional classification of each sentence was conducted by Emotional Analysis Module patented by Acryl Inc. at the Republic of Korea Intellectual Property Office (26), where the input is a single sentence, and the output is a list of probabilities of 8 emotions of the corresponding sentence, namely surprise, fear, anger, love, disgust, sadness, neutral, and happiness. For each transcript, probabilities of eight emotions were derived for the first 250 sentences, resulting in 2,000 probability data. The average probability value for each emotion was calculated and appended as statistics in front of the 2,000 data. As a result, 2,008 probability data were formed as vectors and became the input vector for the machine learning model.

The transcription and feeding of the input vectors into the machine learning model was conducted until the model to detect depression was believed to perform with adequate accuracy.

## Machine learning model to detect depression

Boosting is an ensemble method to create a strong learner by combining multiple weak learners. A weak learner indicates a model that performs slightly better than a randomized prediction. In contrast, a strong learner suggests a model that performs well, significantly better than a randomized prediction. A model is iteratively modified to minimize a loss function by evaluating errors from the previous model and adjust the weights to "boost" the accuracy, but overfitting can remain as a problem (27).

XGBoost is an algorithm that combines multiple decision trees to make predictions (28) based on Gradient Boosting Model (GBM) to overcome the overfitting problem by adopting Classification and Regression Tree (CART) model for regression. It also makes predictions extremely fast by parallel processi3ng of the data. In addition, a weighted quantile sketch was used to handle missing data.

The 166 scripts were split in training and test sets using scikit-learn package, which uses the stratified random sampling method, into an 80/20 ratio. 4-fold cross-validation was conducted on the training set to prevent overfitting (29). Hyperparameters, including learning rate, maximum depth, regularization factor (lambda), early stopping, and evaluation metric, were optimized using grid search (30).

The performance of a model was evaluated with Accuracy and F1 score. Accuracy is the percentage of correct predictions made, but it can sometimes be misleading when the dataset is unbalanced. The F1 score is a harmonic mean of precision and recall, reflecting the imbalance of the dataset. In addition, Area Under the Curve (AUC) was also evaluated, where in general, AUC under 0.7 indicates less reliable, AUC between 0.7 and 0.8 shows somewhat reliable, and more than 0.8 means highly reliable.

RStudio 2022.12.0 + 353 was used for the statistical analysis of the data collected.

**TABLE 1** Patient characteristics.

| | | No. (%) of patients | | | |
|---|---|---|---|---|---|
| | | Total (*n* = 77) | With- depression (*n* = 60) | Without-depression (*n* = 17) | *p*-value |
| Sex[a] | | | | | 0.192 |
| | Male | 24 (31.2) | 16 (26.7) | 8 (47.1) | |
| | Female | 53 (68.8) | 44 (73.3) | 9 (52.9) | |
| Age[b] | | | | | 0.947 |
| | Mean | 33.8 | 33.2 | 35.9 | |
| | (95% CI) | (30.8–36.8) | (29.9–36.5) | (29.0–42.7) | |
| | 20–29 | 42 (54.5) | 33 (55.0) | 9 (52.9) | |
| | 30–39 | 16 (20.8) | 13 (21.7) | 3 (17.6) | |
| | 40–49 | 5 (6.5) | 4 (6.7) | 1 (5.9) | |
| | 50–59 | 9 (11.7) | 6 (10.0) | 3 (17.6) | |
| | 60+ | 5 (6.5) | 4 (6.7) | 1 (5.9) | |
| | Minimum | 20 | 20 | 20 | |
| | Maximum | 64 | 64 | 63 | |
| Diagnosis | | | | | |
| | Adjustment disorder with depressed mood | 3 (3.9) | 3 (5.0) | | |
| | Bipolar disorder (currently depression) | 11 (14.3) | 11 (18.3) | | |
| | Major depressive disorder | 22 (28.6) | 22 (36.7) | | |
| | Persistent depressive disorder | 22 (28.6) | 22 (36.7) | | |
| | Other specified depressive disorder | 2 (2.6) | 2 (3.3) | | |
| | Anxiety disorder | 1 (1.3) | | 1 (5.9) | |
| | Anorexia nervosa | 1 (1.3) | | 1 (5.9) | |
| | Acute stress disorder | 1 (1.3) | | 1 (5.9) | |
| | Alochol use disorder | 3 (3.9) | | 3 (17.6) | |
| | Bipolar and related disorder | 1 (1.3) | | 1 (5.9) | |
| | Intermittent explosive disorder | 1 (1.3) | | 1 (5.9) | |
| | Post-traumatic stress disorder | 6 (7.8) | | 6 (35.3) | |
| | Somatic symptom disorder | 1 (1.3) | | 1 (5.9) | |
| | Substance use disorder | 1 (1.3) | | 1 (5.9) | |
| | Trichotillomania | 1 (1.3) | | 1 (5.9) | |

[a] Chi-squared test on with-depression group vs. without-depression group.
[b] Fisher's exact test on with-depression group vs. without-depression group.

# Results

## Characteristics of extracted sentences

A total of 451 scripts were originally collected from the 77 subjects. The scripts were pre-processed in the form appropriate for learning the model. To avoid overweighting a particular diagnosis or subject, the emotion vectors collected from the first five scripts from each subject were selected in the sequential order and used for analysis to avoid oversampling, as the average number of scripts collected from the subjects was 5.8. As a result, 166 scripts were eventually fed into the model to detect depression.

As a result, a total of 20,405 sentences were split from the 166 scripts, and an emotion with the highest probability was considered as the representative emotion of each sentence in comparing emotional characteristics of the two groups. In the with-depression group, there were 15,223 sentences with an average of 2,184 words consisting of 8,072 characters on each script. There were 5,182 sentences with an average of 2,171 words and 8,156 characters on each script in the without-depression group.

TABLE 2 Emotions counts from the scripts.

| | | No. (%) of sentences | | | |
|---|---|---|---|---|---|
| | | Total | With- depression | Without- depression | $p$-value[a] |
| Emotions classified | | | | | <0.001 |
| | Surprise | 146 (0.7) | 113 (0.7) | 33 (0.6) | |
| | Fear | 1,519 (7.4) | 1,112 (7.3) | 407 (7.9) | |
| | Anger | 153 (0.7) | 107 (0.7) | 46 (0.9) | |
| | Love | 45 (0.2) | 39 (0.3) | 6 (0.1) | |
| | Sadness | 3,304 (16.2) | 2,487 (16.3) | 817 (15.8) | |
| | Disgust*** | 2,357 (11.6) | 1,626 (10.7) | 731 (14.1) | |
| | Neutral** | 12,069 (59.1) | 9,110 (59.8) | 2,959 (57.1) | |
| | Happiness | 812 (4.0) | 629 (4.1) | 183 (3.5) | |
| Total | | 20,405 (100.0) | 15,223 (100.0) | 5,182 (100.0) | |

[a] Chi-squared test on with-depression group vs. without-depression group.
**$p$ < 0.01 based on post-hoc analysis of Pearson's Chi-squared test.
***$p$ < 0.001 based on post-hoc analysis of Pearson's Chi-squared test.

## Distribution of emotions

The frequently represented emotions in the with-depression group were neutral (59.8%), sadness (16.3%), disgust (10.7%), fear (7.3%), and happiness (4.1%). The without-depression group had a similar order of the frequently represented emotions, namely neutral (57.1%), sadness (15.8%), disgust (14.1%), fear (7.9%), and happiness (3.5%). The distribution of eight emotions represented by the sentences significantly differed between the two groups based on the Chi-squared test of homogeneity ($p$ < 0.001, Table 2).

Disgust ($p$ < 0.001) and neutral ($p$ < 0.01) were identified as the emotions that contributed to the significant difference in the distributions between the two groups based on the *post hoc* analysis of the residuals of the chi-squared test (31).

## Classification results of with-depression and without-depression groups

The ROC curve of the machine learning model which used the original probability vectors showed an AUC of 0.85 (Figure 1) upon the hyperparameters optimized with grid search (32). The model classified patients with depression against those without depression with a sensitivity of 0.96, specificity of 0.25, an accuracy of 0.79, and an F1 score of 0.88 (Table 3).

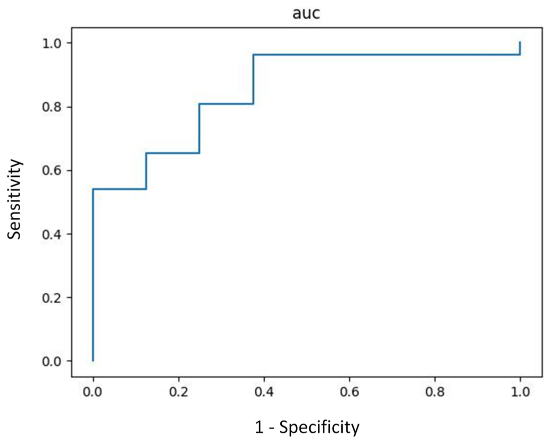


FIGURE 1
ROC curve on the test set.

TABLE 3 Confusion matrix on the test set.

| | | Ground truth | |
|---|---|---|---|
| | | With- depression | Without- depression |
| Model output | With-depression | 73.5% (25) | 17.6% (6) |
| | Without-depression | 2.9% (1) | 5.9% (2) |

* Key metrics of classification include accuracy of 79.4%, F1 score of 87.7%, sensitivity of 96.2%, and specificity of 25.0%. F1 score is defined by 2 * Precision * Recall / (Precision + Recall).

# Discussion

Our text emotion recognition algorithm revealed the difference in emotion distributions between the patients with depression and the control group. The distribution of emotions extracted from the sentences showed significant differences between the two groups, mainly due to less frequent expressions of disgust in the with-depression group. The machine learning model could classify patients with depression against the without-depression control with good reliability based on the emotional profiles extracted from the transcripts.

Among eight emotional labels (surprise, fear, anger, love, disgust, sadness, neutral, and happiness), the most contributing emotion that discriminates between depression and the control was disgust. Patients with depression were known to have problems recognizing facial expressions showing disgust (33, 34). Functional MRI signals responded in higher intensity among patients with depression to disgust (33), suggesting impaired functioning in the basal ganglia (34).

Depression is associated with self-disgust (35), presumably due to altered emotion regulation strategies (36). In addition to the recognition of and response to external stimuli, the findings of this study concur with the association of expression of disgust and depression. Neutral was one of the contributing factors in discriminating the two groups.

Depression is typically characterized by a depressed mood or sadness, but its contribution in discriminating the two groups in this study was not significant. The mood or the sadness is generally determined by physicians from the general atmosphere throughout the conversation. Meanwhile in this study, "sadness" as an emotion was derived from specific sentences, indicating sentiments at certain moments during the conversation. Sadness was not significant in our study due to the difference of the time-interval between the clinical cue and our method. In addition, we compared the with-depression group against the patients without depression, not against the normal control group. Some patients in the control group such as those with PTSD and somatic symptoms disorder, might have expressed sadness as much as depression group under the influence of accompanying symptoms.

The probability vectors of emotions derived from sentences were fed to train the machine learning model, and the model discriminated depression from the control group with an AUC of 0.85, indicating a high reliability of the model. Feature importance analysis revealed that the model did not depend solely on any single emotion in detecting the depression, and the probability vectors of the sentences from the early part of the interviews were considered more important by the model compared to the latter part of the interviews (Table 4). Feature importance represents the contribution of each input feature in making branches in the decision tree. It is evaluated by the change in the model performance given the exclusion of a certain input feature.

Previous studies have normally used audio and visual dataset as inputs to detect depression and its severity (14, 15, 17, 18), but the nature of audiovisual data poses hurdles in contemplating clinical applications for psychiatrists (20, 21). In contrast, text data in the form of transcripts of conversations based on the recordings of routine psychiatric interviews, as collected in this study, is incomparably easier to obtain upon the subject's consent. An ordinary voice recorder in the office and a mean to transcribe of the conversation would suffice the setting for the data collection and the audio-to-text pre-processing. Such a simple requirement to generate the model input suggests a great advantage in applying to clinical situations.

Considering the objective of this study to assist psychiatrists in the actual clinical situations, the model should be able to detect subtlety of depression that psychiatrists might have missed. Currently, the model provides relatively low specificity compared to its very high sensitivity. While we recognize the need to demonstrate improved overall performance of the model, we also believe that the advantage of high sensitivity outweighs any disadvantage posed by the low specificity, as early recognition and proper intervention are important in treating depression with better outcomes (37).

There are several limitations to this study. First, psychotherapy sessions are semi-structured and conducted by multiple psychiatrists of the hospital depending on the availability. This would have allowed flexibility to explore deeper into the thoughts and emotions brought up by the patients depending on the flow of the conversation. Such less standardized interviews were thus considered more suitable for this study. However, psychotherapy sessions are less standardized and more difficult to quantify, and the questions and contents may vary depending on the interviewers. Structured interviews could have improved the credibility of the probability vectors of the emotions derived from the interviews.

TABLE 4 Feature importance analysis.

| | | Importance |
|---|---|---|
| By emotions | Neutral | 0.187 |
| | Love | 0.155 |
| | Fear | 0.153 |
| | Surprise | 0.134 |
| | Anger | 0.120 |
| | Disgust | 0.094 |
| | Happiness | 0.079 |
| | Sadness | 0.077 |
| By location of sentences ($n^{th}$ sentence) | 1–20 | 0.317 |
| | 81–100 | 0.186 |
| | 21–40 | 0.151 |
| | 61–80 | 0.110 |
| | 41–60 | 0.103 |
| | 101–120 | 0.054 |
| | 141–160 | 0.043 |
| | 121–140 | 0.036 |
| | 161 and later | 0.000 |

Also, the random split of input data by scikit learn package might have resulted in the scripts from the same person being put into both the training and test set, considering the dataset size for this study. The model could have been trained in a way that classifies depression based on the person's traits rather than the traits of the depression itself. A larger dataset could improve the model, not only in terms of the overall performance, including sensitivity, but also by minimizing the possibility of learning any individual's trait so that the model ultimately identifies the depression solely based on the emotional features of depression.

There are a couple of factors that might have affected the external validity of this study. The number of data is limited due to the retrospective nature of the study, and the model's performance along with statistical power could have improved further by feeding model inputs. Also, the control group consisted of psychiatric patients without depression, rather than non-clinical samples without any psychiatric diagnosis. It would have been valuable if such non-clinical samples were also recruited to compare against the with-depression group. However, we believe that it is more difficult to detect patients with depression against the patients with other psychiatric diagnosis, as conducted in this study. In addition, the subjects in the with-depression group and the without-depression control group were not exactly matched due to the retrospective nature of this study. We plan to test the detection algorithm on non-clinical subjects in the future in a prospective manner.

The number of scripts collected for this study was originally much larger than that of the input scripts fed into the model. We decided to use a maximum of 5 scripts for each subject to avoid potential bias due to oversampling. For example, we collected more than 40 scripts from five subjects, three from the with-depression group and the rest from the without-depression control group. It could have improved the performance metrics of the model when the entire data collection was used, but the risk associated with depending on a few subjects should be avoided. Collecting an evenly distributed number of scripts from the subjects would improve the model's performance and avoid bias arising from the oversampling.

Acryl's Emotional Analysis Module, which was used to derive probability vectors assigned to the sentences of the text scripts, did not consider any context or meanings of the sentence. Large Language Models (LLM) has been increasingly used recently in many applications which can consider textual contexts based on the parameters and datasets much larger than the conventional models in analyzing text data. It remains as a future work to incorporate LLM in the process of classifying emotions from the text scripts.

## Conclusion

This study suggests a novel approach to detect depression with conversational scripts with patients based on text emotion recognition and a machine learning model. Emotional distribution significantly differed between the depression and the control group, and the model showed a reliable performance in classifying patients with depression from those without depression. Our results could assist clinicians in the initial diagnosis and follow-up of depressive patients with conventional diagnostic tools. Further studies would improve the performance, potentially detecting depression alongside the psychiatrists in the clinics and hospitals.

## 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 humans were approved by the Institutional Review Board of Seoul St. Mary's Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

JO: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing, Supervision, Investigation. TL: Formal analysis, Validation, Writing – original draft, Writing – review & editing, Investigation, Methodology. EC: Writing – original draft, Investigation. HyoK: Formal analysis, Visualization, Writing – review & editing. KC: Formal analysis, Visualization, Writing – review & editing. HyuK: Formal analysis, Visualization, Writing – review & editing. JC: Data curation, Writing – review & editing. H-HS: Data curation, Writing – review & editing. JL: Writing – review & editing. IC: Writing – review & editing. D-JK: Conceptualization, Data curation, Funding acquisition, Supervision, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National IT Industry Promotion Agency (NIPA) grant funded by the Ministry of Science and ICT (No. S0252-21-1001). The funder had no role in the design, data collection, analysis, interpretation of results, and manuscript drafting.

## Conflict of interest

HyoK, KC, and HyuK were employed by Acryl.

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.

## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1256571/full#supplementary-material

## References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. *Lancet Psychiatry*. (2022) 9:137–50. doi: 10.1016/S2215-0366(21)00395-3

2. Sinyor M, Rezmovitz J, Zaretsky A. Screen all for depression. *BMJ*. (2016) 352:i1617. doi: 10.1136/bmj.i1617

3. Lim GY, Tam WW, Lu Y, Ho CS, Zhang MW, Ho RC. Prevalence of depression in the community from 30 countries between 1994 and 2014. *Sci Rep*. (2018) 8:2861. doi: 10.1038/s41598-018-21243-x

4. Kim GE, Jo M-W, Shin Y-W. Increased prevalence of depression in South Korea from 2002 to 2013. *Sci Rep*. (2020) 10:16979. doi: 10.1038/s41598-020-74119-4

5. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders. 5th* ed. Washington, D.C: American Psychiatric Association (2013).

6. Maurer DM, Raymond TJ, Davis BN. Depression: screening and diagnosis. *Am Fam Physician*. (2018) 98:508–15.

7. Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton depression rating scale: has the gold standard become a lead weight? *AJP*. (2004) 161:2163–77. doi: 10.1176/appi.ajp.161.12.2163

8. Thombs BD, Kwakkenbos L, Levis AW, Benedetti A. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. *CMAJ*. (2018) 190:E44–9. doi: 10.1503/cmaj.170691

9. Ghio L, Gotelli S, Cervetti A, Respino M, Natta W, Marcenaro M, et al. Duration of untreated depression influences clinical outcomes and disability. *J Affect Disord*. (2015) 175:224–8. doi: 10.1016/j.jad.2015.01.014

10. Faisal-Cury A, Ziebold C, Rodrigues DMO, Matijasevich A. Depression underdiagnosis: prevalence and associated factors. A population-based study. *J Psychiatr Res*. (2022) 151:157–65. doi: 10.1016/j.jpsychires.2022.04.025

11. Liu J, Wu G, Luo Y, Qiu S, Yang S, Li W, et al. EEG-based emotion classification using a deep neural network and sparse autoencoder. *Front Syst Neurosci*. (2020) 14:43. doi: 10.3389/fnsys.2020.00043

12. Ahmed MZI, Sinha N, Ghaderpour E, Phadikar S, Ghosh R. A novel baseline removal paradigm for subject-independent features in emotion classification using EEG. *Bioengineering*. (2023) 10:54. doi: 10.3390/bioengineering10010054

13. Korda AI, Giannakakis G, Ventouras E, Asvestas PA, Smyrnis N, Marias K, et al. Recognition of blinks activity patterns during stress conditions using CNN and Markovian analysis. *Signals*. (2021) 2:55–71. doi: 10.3390/signals2010006

14. Haque A, Guo M, Miner AS, Fei-Fei L. Measuring depression symptom severity from spoken language and 3D facial expressions. *arXiv [Preprint]*. (2018). doi: 10.48550/arXiv.1811.08592

15. Guo W, Yang H, Liu Z, Xu Y, Hu B. Deep neural networks for depression recognition based on 2D and 3D facial expressions under emotional stimulus tasks. *Front Neurosci*. (2021) 15:609760. doi: 10.3389/fnins.2021.609760

16. Almeida J, Vilaça L, Teixeira IN, Viana P. Emotion identification in movies through facial expression recognition. *Appl Sci*. (2021) 11:6827. doi: 10.3390/app11156827

17. Rejaibi E, Komaty A, Meriaudeau F, Agrebi S, Othmani A. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomed Signal Process Control*. (2022) 71:103107. doi: 10.1016/j.bspc.2021.103107

18. Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. Detecting depression: a comparison between spontaneous and read speech. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE (2013). p. 7547–7551.

19. Dham S, Sharma A, Dhall A. Depression scale recognition from audio, visual and text analysis. arXiv [Preprint]. (2017). Available at: http://arxiv.org/abs/1709.05865

20. Kumar A, Kaur A, Kumar M. Face detection techniques: a review. *Artif Intell Rev*. (2019) 52:927–48. doi: 10.1007/s10462-018-9650-2

21. Albahra S, Gorbett T, Robertson S, D'Aleo G, Kumar SVS, Ockunzzi S, et al. Artificial intelligence and machine learning overview in pathology & laboratory medicine: a general review of data preprocessing and basic supervised concepts. *Semin Diagn Pathol*. (2023) 40:71–87. doi: 10.1053/j.semdp.2023.02.002

22. Nordgaard J, Sass LA, Parnas J. The psychiatric interview: validity, structure, and subjectivity. *Eur Arch Psychiatry Clin Neurosci*. (2013) 263:353–64. doi: 10.1007/s00406-012-0366-z

23. Kratzwald B, Ilic S, Kraus M, Feuerriegel S, Prendinger H. Deep learning for affective computing: text-based emotion recognition in decision support. *Decis Support Syst*. (2018) 115:24–35. doi: 10.1016/j.dss.2018.09.002

24. Calefato F, Lanubile F, Novielli N. EmoTxt: a toolkit for emotion recognition from text. Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). San Antonio, TX: IEEE (2017). p. 79–80

25. Batbaatar E, Li M, Ryu KH. Semantic-emotion neural network for emotion recognition from text. *IEEE Access*. (2019) 7:111866–78. doi: 10.1109/ACCESS.2019.2934529

26. Oh SS, Lee HH, Park WJinventors; Acryl Inc., assignee. Emotion recognition method and computer program for executing the method, emotion recognizer generation method and computer program for executing the method. Republic of Korea Patent 10-2110393-0000. (2020).

27. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot*. (2013) 7:21. doi: 10.3389/fnbot.2013.00021

28. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM (2016). p. 785–794.

29. Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans Pattern Anal Mach Intell*. (2010) 32:569–75. doi: 10.1109/TPAMI.2009.187

30. Hutter F, Kotthoff L, Vanschoren J. *Automated machine learning: methods, systems, challenges*. Cham: Springer International Publishing (2019).

31. Beasley TM, Schumacker RE. Multiple regression approach to analyzing contingency tables: post hoc and planned comparison procedures. *J Exp Educ*. (1995) 64:79–93. doi: 10.1080/00220973.1995.9943797

32. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*. (2006) 27:861–74. doi: 10.1016/j.patrec.2005.10.010

33. Surguladze SA, El-Hage W, Dalgleish T, Radua J, Gohier B, Phillips ML. Depression is associated with increased sensitivity to signals of disgust: a functional magnetic resonance imaging study. *J Psychiatr Res*. (2010) 44:894–902. doi: 10.1016/j.jpsychires.2010.02.010

34. Douglas KM, Porter RJ. Recognition of disgusted facial expressions in severe depression. *Br J Psychiatry*. (2010) 197:156–7. doi: 10.1192/bjp.bp.110.078113

35. Gao S, Zhang L, Yao X, Lin J, Meng X. Associations between self-disgust, depression, and anxiety: a three-level meta-analytic review. *Acta Psychol*. (2022) 228:103658. doi: 10.1016/j.actpsy.2022.103658

36. Ypsilanti A, Lazuras L, Powell P, Overton P. Self-disgust as a potential mechanism explaining the association between loneliness and depression. *J Affect Disord*. (2019) 243:108–15. doi: 10.1016/j.jad.2018.09.056

37. Kraus C, Kadriu B, Lanzenberger R, Zarate CA Jr, Kasper S. Prognosis and improved outcomes in major depression: a review. *Transl Psychiatry*. (2019) 9:127. doi: 10.1038/s41398-019-0460-3

frontiers | Frontiers in Psychiatry

TYPE Opinion
PUBLISHED 18 January 2024
DOI 10.3389/fpsyt.2024.1347358

OPEN ACCESS

EDITED BY
Steven Fernandes,
Creighton University, United States

REVIEWED BY
Rikinkumar S. Patel,
Duke University, United States

*CORRESPONDENCE
Jan Speechley
✉ datamind@swansea.ac.uk
Michael McTernan
✉ datamind@swansea.ac.uk

[†]These authors share first authorship

RECEIVED 30 November 2023
ACCEPTED 04 January 2024
PUBLISHED 18 January 2024

CITATION
Speechley J and McTernan M (2024) How will AI make sense of our messy lives and improve our mental health?
*Front. Psychiatry* 15:1347358.
doi: 10.3389/fpsyt.2024.1347358

# How will AI make sense of our messy lives and improve our mental health?

Jan Speechley*[†] and Michael McTernan*[†]

DATAMIND - The Hub for Mental Health Informatics Research Development, Swansea, United Kingdom

KEYWORDS

**artificial intelligence, mental health, patient and public engagement, trust, data**

## Introduction

There is a growing belief that Artificial Intelligence (AI) will play a major part in mental health research and the development and delivery of new services. We are told that AI could provide all of us with access to fast, effective, and personalized healthcare. We hear stories about how AI is more effective at diagnosis than healthcare professionals (HCP) (1), yet there is a lack of trust amongst the public in AI. Would you rely on artificial intelligence (AI) to help you with your mental health issues?

Jan and Michael, lived experience experts, share their opinions on how AI could connect, or not, with their mental health issues and sometimes "messy" lives.

## Can you map our messy lives in discrete, tidy data sets?

Our mental health is complicated, you could say it is messy, and the factors that influence it create our life story. Our mental health is ours; it is unique, it is personal and precious to us as individuals. We want healthcare professionals to understand our experiences, how they affect us, as this articulates how our mental health issues have developed and also how they could be improved.

AI is built and trained on the data it receives. More data, from more sources with more detail can result in better outcomes. Could sharing our mental health data really enable the creation of personalized, tailored care for those with mental health issues and help them make sense of their messy lives and the impact on their mental health?

Jan reflects on her "messy life."

Consider one person's life. Their earliest care can shape their lives, parents with complex mental health issues, a childhood punctuated with parental mental health symptoms and behaviors. Basic things like mealtimes never at the same time or even certain.

Add in life – school – exams - work – relationships – children - friendships – physical health, money and the 'messy life' takes shape with so many external factors including the weather! My mental health is always better in the Summer and the shorter days and cold weather of Winter exacerbates all my anxiety and depression. Can all of these variables be captured in a way that is useful? More importantly, can you persuade people to make this information available?

The stigma of this, the coping strategies that we employ, the feeling of being weak, of needing to hide how we feel. All this adds to the strain of just living a messy life with all its component parts and its demands on time and energy. The fear of stigma and shame means we are less likely to want to share our messy lives.

The care and treatments we receive can impact our mental health in a negative way and add to our messy life. Jan says, "I was told that my depression was "difficult to treat," It made me feel it was my fault and "I did not want to get better." How language is used in the collection of data and the provision of service is vitally important. Labels can be the start of health inequalities and increase stigma. Jan says, "I know now none of it was my fault, but it took me many years and therapies to successfully reach and live with that conclusion." The term "treatment resistant depression" is a less judgemental and more positive sounding phrase.

Generative AI tools don't "understand" mental health and can deliver inaccurate and misleading answers. So, how can the personalized healthcare services promised by AI would be developed to use the right language for each situation?

Could AI be used to develop new approaches to delivering mental health care that offer alternatives to medication, addressing an individual's messy life, recommending lifestyle changes, and tailored talking therapies? For patients like Jan labeled with "treatment resistant depression" could there be an alternative route to treatment.

## Trust and transparency

Jan and Michael consider the decision to share their data.

If we are considering sharing all our deeply personal, messy life data to improve mental health care services and treatment we have to trust those who use it and this now includes AI and machine learning tools. Right now, most people donot trust AI.

A new study from the BSI describes that while half of us support the use of AI in healthcare to reduce waiting times, there is still a significant lack of trust in AI (2). Almost two-thirds of respondents in the UK believe that "that patients should be informed if an AI tool is being used during the diagnostic process" (2). We think there are a few reasons for this mistrust.

Firstly, the use of AI and Machine Learning (ML) in healthcare seems futuristic, uncertain, and risky for patients. The news is full of individuals telling us how dangerous AI could be for society. So, it is little wonder that we are skeptical about AI being used to provide our health services. Aligned to this there is a lack of understanding by the public about what AI is and how it would be used in healthcare. It's just not been on our radar. The public needs a better understanding of how AI and ML can be used in healthcare, the pros and cons, and the impact that will have on them.

If you believe the hype, AI has the potential to make healthcare more accessible, triaging patients to the right treatments and therapies to meet their needs. AI could provide individuals with a personalized treatment plan based on their symptoms, history and lifestyle, without seeing a healthcare professional.

We value seeing a clinician, we build usually trusting (but not always), relationships with healthcare professionals. Will we build similar trusting relationships with Healthcare AI agents? Can AI replace the relationship that we have with a psychologist or a GP?

We would welcome the advances in diagnosis and treatment that AI and ML could bring. To radically improve mental health care, we would allow access to our healthcare data, but we also need to know that our data is safe and secure. For the public to be comfortable with sharing their data we need to overcome the stigma and personal shame associated with mental health issues.

There are good examples of altruistic giving in healthcare. For example, Michael gives blood, he doesn't know what happens to the blood that he gives but trusts the Blood Transfusion Service (BTS) to use it appropriately. How can people working in mental health research and development gain and maintain our trust?

## Are we motivated to share our mental health data with researchers?

Jan says, "I have so many questions about my options, I would like to help others but is it safe for me and helpful for them?"

The public needs a better understanding of how AI and ML will be used in healthcare. For many people AI is a scary concept and terms like Machine Learning are meaningless. This needs to be articulated and delivered in terms that we understand. Make your messaging about AI and ML accessible and relevant to the public. But donot patronize us, we are experts in the mental health issues that we face and have spent much time and effort understanding our situation and how best to manage it.

Jan asks, "who cares about me and my privacy? Will my data be safe and protected, could it be sold or appear on social media platforms. What rights do I have if it all goes wrong?"

If you want our data then we need to know that it will be used by researchers whose credentials and purpose are checked by gatekeepers, including members of the public. At the same time, we don't want our data locked away and never used. We want to make our data easily accessible to researchers to allow them to make good use of it.

## How will our data be used to help others have better mental health?

Can our data be used to stop others developing poor mental health and the issues we have experienced? Can our data be used to help people learn from our experiences? Could AI unpick our messy lives and create a personalized treatment for me?

Researchers need to make sure that we understand their big vision, tell us that "our data saves lives" and tell us how. Help all of us understand how AI and ML built on our data can change lives by creating a better understanding of the causes of mental health issues, the strategies for prevention and better treatments for issues and symptoms.

## What impact will sharing my data have?

Could real life improvements to care treatment and services in the NHS change our lives? Will there be small gradual changes or a big bang that changes everything? Will AI help or hinder, slow or increase the pace of improvements in mental health care?

Researchers need to help us feel like we are making a difference, give us feedback on how our data has helped, maybe even an annual newsletter. Make it part of the process that researchers accessing our data must report, in a public friendly way, on their research and the impact that it could create. You could go even further and ask people providing access to their healthcare data to vote on priorities, suggest areas for research, or be public contributors or participants in a research study.

## Data bias and inclusivity

How will we guard against bias in the data that is used in research? Underserved and hard to reach communities struggle to access services and their data is often excluded from research. What safeguards can we put in place to make sure that AI creates equitable and accessible data?

## Final thoughts

People with mental health issues deserve, need, and want improved, more personalized health care, treatment and services. They know their own lives and are experts in its detail and content.

There is much work to be done by governments, law, and policy makers and all involved in research using data, AI and machine learning, to encourage us to share our precious personal information, to make us understand they can be trusted, to keep it safe and use it to help us and others in the future.

We need to understand the difference that making our mental health data available will make, we need to feel valued and be shown how it has helped and the improvements that will happen for people with mental health issues.

Change can be frightening and can take a lot of getting used to, we have to see the point of the change and that it will make things better - the use of AI and Machine Learning is rapidly changing our lives. In healthcare this could create better, accessible, personalized services, but we cannot hope to accept them without question. We need to:

- Help the public understand what AI and Machine Learning is.
- Describe to them the potential impact that their data can have on research and treatment and care services development.
- Gain and maintain their trust in how their data will be used.
- Keep the public informed of how their data is helping and the impact it is having.
- Make sure that the public understands how much their input is valued and make them feel part of a process of positive change.

Michael and Jan are members of the DATAMIND Super Advisory Group, they are lived experience experts. Take a look at our work and the public and patient facing resources we have created at DATAMIND (https://datamind.org.uk/), including a glossary of terms used in mental health data science (3) and a data literacy course to help people understand how their healthcare data is used and stored (4).

## Author contributions

JS: Conceptualization, Writing—original draft, Writing—review & editing. MM: Conceptualization, Writing—original draft, Writing—review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. JS and MM were funded by the Medical Research Council and Health Data Research UK (Grant DATAMIND: Data Hub for Mental Health INformatics research Development, with Ref.: MR/W014386/1).

## 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.

## References

1. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ.* (2023) 23:689. doi: 10.1186/s12909-023-04698-z

2. Morris AH, Horvat C, Stagg B, Grainger DW, Lanspa M, Orme J, et al. Computer clinical decision support that automates personalized clinical care: a challenging but needed healthcare delivery strategy. *J Am Med Inform Assoc.* (2023) 30:178–94. doi: 10.1093/jamia/ocac143

3. DATAMIND Glossary. Available online at: https://datamind.org.uk/glossary/

4. DATAMIND Data Literacy Course. Available online at: https://datamind.org.uk/patients-and-public/data-literacy-short-course-2/

frontiers | Frontiers in Digital Health

TYPE Methods
PUBLISHED 02 February 2024
DOI 10.3389/fdgth.2024.1322555

OPEN ACCESS

EDITED BY
Marcos Del Pozo Banos,
Swansea University Medical School,
United Kingdom

REVIEWED BY
Sudipta Chowdhury,
Swansea University Medical School,
United Kingdom
Bahaa Eddine Elbaghazaoui,
Ibn Tofail University, Morocco

*CORRESPONDENCE
Roger Garriga
✉ roger.garrigacalleja@koahealth.com

RECEIVED 16 October 2023
ACCEPTED 17 January 2024
PUBLISHED 02 February 2024

CITATION
Garriga R, Gómez V and Lugosi G (2024) Individualized post-crisis monitoring of psychiatric patients via Hidden Markov models. Front. Digit. Health 6:1322555.
doi: 10.3389/fdgth.2024.1322555

# Individualized post-crisis monitoring of psychiatric patients via Hidden Markov models

Roger Garriga[1,2]*, Vicenç Gómez[2] and Gábor Lugosi[3,4,5]

[1]Koa Health, Barcelona, Spain, [2]Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, [3]ICREA, Barcelona, Spain, [4]Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain, [5]Barcelona School of Economics, Barcelona, Spain

**Introduction:** Individuals in the midst of a mental health crisis frequently exhibit instability and face an elevated risk of recurring crises in the subsequent weeks, which underscores the importance of timely intervention in mental healthcare. This work presents a data-driven method to infer the mental state of a patient during the weeks following a mental health crisis by leveraging their historical data. Additionally, we propose a policy that determines the necessary duration for closely monitoring a patient after a mental health crisis before considering them stable.
**Methods:** We model the patient's mental state as a Hidden Markov Process, partially observed through mental health crisis events. We introduce a closed-form solution that leverages the model parameters to optimally estimate the risk of future mental health crises. Our policy determines a patient should be closely monitored when their estimated risk of crisis exceeds a predefined threshold. The method's performance is evaluated using both simulated data and a real-world dataset comprising 162 anonymized psychiatric patients.
**Results:** In the simulations, 96.2% of the patients identified by the policy were in an unstable state, achieving a F1 score of 0.74. In the real-world dataset, the policy yielded an F1 score of 0.79, with a sensitivity of 79.8% and specificity of 88.9%. Under this policy, 67.3% of the patients should undergo close monitoring for one week, 21.6% during 2 weeks or more, while 11.1% do not need close monitoring.
**Discussion:** The simulation results provide compelling evidence that the method is effective under the specified assumptions. When applied to actual psychiatric patients, the proposed policy showed significant potential for providing an individualized assessment of the required duration for close and automatic monitoring after a mental health crisis to reduce the relapse risks.

KEYWORDS

probabilistic modeling, mental health crisis, Hidden Markov model, mental health, psychiatry, machine learning, predictive analytics

## 1 Introduction

A mental health crisis is any situation in which a person's behavior puts them at risk of hurting themselves or others and/or prevents them from being able to care for themselves or function effectively in the community (1). Those situations include self-harm, delusions or suicide attempts, often requiring hospitalization, and are very detrimental to the patient's mental and social wellbeing. Mental health crises are commonly suffered by patients diagnosed with psychotic, personality or severe mood disorders. However, they also occur to patients diagnosed with less severe disorders or even non diagnosed individuals under stressful situations (1). The patient usually undergoes four phases in the process of a crisis (2), (i) an initial threat when the patient is stable, (ii) an escalation phase, (iii) the crisis (iv) resolution and return to stability or personality disorganization if the problem does not get resolved. Once the patient destabilizes, they remain unstable for some period of time during which they might suffer one or multiple mental health crises. In order to avoid further escalation and prevent subsequent crises, patients should be kept under close monitoring and treatment until they stabilize (3–6). However, it is difficult to ascertain when the patient has become stable.

In this work, we present a data-driven method to infer the mental state of a patient given their history of mental health crises and propose a policy to determine for how many weeks the patient needs to receive close attention before being deemed stable. This method is based on modeling the mental state of the patient as a Hidden Markov Model (HMM) (7), a probabilistic framework in which the observed data is generated by one or multiple hidden states. This allows one to infer whether the patient is stable or unstable and make a prediction of the risk that the patient is going to suffer a mental health crisis. Under this modeling framework, our method is implemented in the following way:

1. **Learning the model parameters of each patient:** Initially, the model parameters for the average patient are determined by maximizing the likelihood of the observed sequence of mental health crises experienced by all patients. These parameters are assigned to patients with a relatively short history at the hospital (3 months or less in this study). For patients with a longer history, the model parameters of each patient are estimated from their individual observed sequence of mental health crises.
2. **Estimating the risk of mental health crisis at each week:** For a given patient, the risk is estimated based on the patient's model parameters, taking into account the time elapsed since their last mental health crisis.
3. **Selecting the patients to be closely monitored:** Identify those patients whose predicted risk of a mental health crisis exceeds a predefined threshold. Patients that do not reach the threshold are considered stable due to their low risk to suffer a mental health crisis.

Probabilistic models are very well suited to uncover hidden phenotypes or internal states in healthcare settings and to build policies based on partial observability of internal states (8). The use of HMM's to infer the mental state of an individual has been explored in the past for detecting depressive states or schizophrenic episodes (9–11) and identifying mental disorders (12, 13). A similar probabilistic model called maximum-entropy Markov model was used to predict emergency psychiatric states (14) from biometric sensors and questionnaires. However, these studies rely on external data sources such as sensor data, questionnaires or other user inputs. In contrast, our work proposes a method that relies solely on past crises to infer the mental state of the patient, which is accessible in any hospital and does not require external data collection.

Previous research has demonstrated the feasibility of predicting mental health crisis relapses when patients appear stable utilizing a machine learning model based on Electronic Health Records (EHR) (15, 16). However, these studies assumed that all patients achieved stability after just one week without crisis events. In reality, certain patients might necessitate prolonged and vigilant monitoring to ascertain their stability accurately and avoid readmission. The present study complements the existing literature by introducing a method to determine the optimal duration of monitoring required for each individual patient before they can be confidently deemed stable. By adopting this data-driven approach, clinicians can make informed decisions that facilitate personalized care. This approach follows the principle of Precision Medicine, a field that has been implemented across various healthcare domains and is now gaining traction within the field of psychiatry, promising enhanced patient outcomes and more effective interventions (17, 18).

# 2 Materials and methods

In this section, we detail the steps required to implement our proposed method. In Section 2.1, we formalize the problem and the mental state model upon which our method is built, and discuss the assumptions. In Section 2.2, we present an optimal solution to predict the risk that a patient will suffer a mental health crisis within the next week given the model parameters, and how the risk evolves over time. In Section 2.3, we describe the process to estimate the model parameters from a sequence of weeks with and without mental health crisis events. Finally, in Section 2.4, we propose a policy for determining the duration a hospital should closely monitor patients before deeming them stable.

## 2.1 Mental health state probabilistic model

We consider a hospital with $N$ patients, with each patient $n$ having an associated mental health state $X_{t,n} \in \{S, U\}$ at each week $t = 0, \ldots, T$ and a binary random variable $Y_{t,n}$ that denotes whether the patient had a mental crisis at week $t$ ($Y_{t,n} = 1$) or not ($Y_{t,n} = 0$). Every week $t$ of the patient $n$ is characterised by the $(X_{t,n}, Y_{t,n})$ pair and we denote by $H_{t,n} = \{(X_{0,n}, Y_{0,n}), \ldots, (X_{t,n}, Y_{t,n})\}$ the entire history of the patient up to week $t$. We use $x_{t,n}$ and $y_{t,n}$ to denote the realizations of $X_{t,n}$ and $Y_{t,n}$, and introduce the notation $Y_{a,n}^{b} = (Y_{a,n}, \ldots, Y_{b,n})$, $a < b \in \mathbb{Z}$ (similarly for other random variables $X_{a,n}^{b}$ and realizations $x_{a,n}^{b}$).

### 2.1.1 Assumptions

We consider the following set of assumptions associated with our problem:

- There are two possible mental states that a patient $n$ can have at any week $t$, stable (S) or unstable (U), thus $X_{t,n} \in \{S, U\}$, $\forall t = 1, \ldots, T$.
- The mental state of the patient evolves following a Markov Chain (i.e., $P(X_{t,n} = x_{t,n} | X_{0,n} = x_{0,n}, \ldots, X_{t-1,n} = x_{t-1,n}) = P(X_{t,n} = x_{t,n} | X_{t-1,n} = x_{t-1,n})$). We denote by $q = P(X_{t,n} = U | X_{t-1,n} = S)$ the transition probability from state $S$ to state $U$ and by $r = P(X_{t,n} = U | X_{t-1,n} = U)$ the transition probability

from state $U$ to state $U$, with $q \neq r$. This results in the following transition matrix:

$$P_X = \begin{pmatrix} p_{SS} & p_{SU} \\ p_{US} & p_{UU} \end{pmatrix} = \begin{pmatrix} 1-q & q \\ 1-r & r \end{pmatrix}$$

- The probability that the patient $n$ has a mental health crisis at time $t$ depends solely on the state $X_{t,n}$. In particular, we assume that the patient cannot suffer a mental health crisis when the patient is at state $S$ and when the patient is at state $U$ the probability of crisis if $0 < p < 1$, that is,

$$P(Y_{t,n} = 1 | X_{t,n} = x_{t,n}) = \begin{cases} 0 & \mathrm{if x_{t,n} = S} \\ p & \mathrm{if x_{t,n} = U} \end{cases}$$

- The actual mental state of the patient is hidden and only partially observed through the crisis variable. We denote by $\mathcal{O}_{t,n} = \{Y_{0,n}, \ldots, Y_{t,n}\}$ the observed history up to time $t$.

We present two cases that depend on whether the patients of the hospital are characterized by homogeneous or diverse model parameters.

- Case 1: Each patient has a different set of model parameters.
- Case 2: All patients have the same set of model parameters.

## 2.2 Mental health crisis prediction

The purpose of the method is to predict whether a patient $n$ is going to have a mental health crisis at time $t$ given the observed history up to time $t-1$, $\mathcal{O}_{t-1,n}$. In particular, we want to estimate the probability that $Y_{t,n} = 1$ given $\mathcal{O}_{t-1,n}$. Considering that the model parameters are known (or have been estimated as we will see in the next section), we can make inference on the current state of the patient given the observed history $\mathcal{O}_{t-1,n}$ and use it to estimate the probability that the patient is going to have a crisis at time $t$. Since the prediction is done for each patient independently to the rest of the patients, we simplify the notation in this section by removing the subscript $n$.

First, we consider the case in which the last state $X_{t-1}$ is observed ($X_{t-1} \in \mathcal{O}_{t-1}$). Using the Markov property we obtain that

$$\begin{aligned} P(Y_t = 1 | \mathcal{O}_{t-1}) &= P(Y_t = 1 | X_t = U) P(X_t = U | \mathcal{O}_{t-1}) \\ &= P(Y_t = 1 | X_t = U) P(X_t = U | X_{t-1}) \\ &= \begin{cases} pr & \mathrm{if X_{t-1} = U} \\ pq & \mathrm{if X_{t-1} = S} \end{cases}. \end{aligned}$$

A priori, we assumed that the state is never observed. However, there might be cases in which it is possible to observe the state either directly or indirectly. For instance, by monitoring the patient, the clinical teams can infer whether the patient is at stable state or not. Importantly, there are two consequences that follow from the proposed model: first, when a mental health crisis is observed at a time $t-1$ ($Y_{t-1} = 1$), we can infer that $X_{t-1} = U$ because a patient can only suffer a mental health crisis when they are unstable; second, since the state transition is Markov, if the state is known at $s < t$ ($X_s = x_s$) and there are no other known states between $s+1$ and $X_{t-1}$, then the distribution over states at $X_{t-1}$ only depends on the observations between $s$ and $t-1$ and $X_s = x_s$. Therefore, without loss of generality, we can assume that the last observed state is at $s = 0$ because the observations prior to $s$ do not influence the probability distribution of states beyond $s$ conditioned on $X_s = x_s$ (by the Markov property) - we could redefine a new $t' = t - s$.

The following theorem presents a function to estimate the risk of mental health crisis at each week $t$ after the last observed state. This is particularly relevant because the occurrence of a mental health crisis reveals that the patient is in an unstable state, and the theorem enables the determination of the number of weeks until the patient likely regains stability.

> **Theorem 1:** Let $\mathcal{O}_{t-1} = \{X_0 = x_0, Y_0 = y_0, Y_1 = 0, \ldots, Y_{t-1} = 0\}$ be the observed history of a patient $n$ up to the week $t$ and $p, q, r$ the model parameters associated with the patient. Then, the probability that patient $n$ suffers a mental health crisis at week $t$ is given by
>
> $$P(Y_t = 1 | \mathcal{O}_{t-1}) = \\ 1 - (1 - pr + r - q) \frac{(y_0 - y_-) y_+^{t+1} - (y_+ - y_0) y_-^{t+1}}{(y_0 - y_-) y_+^{t} - (y_+ - y_0) y_-^{t}}, \quad (1)$$
>
> with
>
> $$y_+ = \frac{1 + \sqrt{1 - 4 \frac{(r-q)(1-p)}{(1-pr+r-q)^2}}}{2}, \\ y_- = \frac{1 - \sqrt{1 - 4 \frac{(r-q)(1-p)}{(1-pr+r-q)^2}}}{2}, \\ y_0 = \frac{2Rw_{x_0} - R}{2R + w_{x_0} - 1},$$
>
> where $R = \frac{(r-q)(1-p)}{(1-pr+r-q)^2}$, $w_U = \frac{1-pr}{1-pr+r-q}$ (when $x_0 = U$) and $w_S = \frac{1-pq}{1-pr+r-q}$ (when $x_0 = S$).

To create a policy that works for all patients on a weekly basis, we need to understand how the estimated risk of a patient $n$ experiencing a mental health crisis changes over time. The analytical solution from the theorem is particularly useful for this purpose, as it allows us to study how the risk evolves and converges. The following corollary demonstrates the convergence of the solution.

> **Corollary 1.1:** The optimal solution to estimate the risk that a patient with model parameters $q, r, p$ converges to $1 - (1 - pr + r - q) y_+$ when $t$ grows.

Due to the exponential convergence primarily driven by the $y_+$ term, the convergence of the solution is expected to be rapid. This rapid convergence guarantees that the estimated risk does not oscillate indefinitely but rather quickly stabilizes at a steady value. Together, the results of Theorem 1 and the Corollary 1.1 show that we can estimate the risk of mental health crisis analytically and determine the week when a patient is likely to reach a stable state.

The proofs of Theorem 1 and Corollary 1.1 can be found in Supplementary Appendix A.

## 2.3 Estimation of the model parameters

The HMM has 3 parameters $(p, q, r)$, specifically:

- $p$: the probability of mental health crisis given that the patient is at state $U$.
- $q$: the transition probability from state $S$ to state $U$.
- $r$: the transition probability from state $U$ to state $U$.

To simplify the notation, in this section we introduce $p(a) = P(A = a)$ to denote the probability that a random variable $A$ takes the value $a$ (e.g., $p(x_{t,n}) = P(X_{t,n} = x_{t,n})$). Similarly, we use the same notation for joint probabilities and conditional probabilities (e.g., $p(y_{0,n}^t, x_{t,n}) = P(Y_{0,n}^t = y_{0,n}^t, X_{t,n} = x_{t,n})$ or $p(y_{t+1,n}^T|x_{t,n}) = p(Y_{t+1,n}^T = y_{t+1,n}^T|X_{t,n} = x_{t,n})$).

We use the Baum–Welch algorithm (19) to estimate the model parameters from the observed history of the patients. This method, is a standard algorithm that uses an Expectation–Maximization approach to find the parameters that maximize the expected likelihood of the observed data given the model HMM. The Baum–Welch algorithm is guaranteed to converge to a local optimum (20) and consists of the following steps:

1. **Initialization:** The parameters of the model are initialized either randomly or using some reasonable estimates. In this case, we initialize the parameters $(p, q, r)$ at random.
2. **Expectation step:** In this step, the probabilities of being in each hidden state at each time step $t$ given the current model parameters and the observed sequence $\mathcal{O}_{t,n}$ are calculated. These probabilities are computed using the Forward-Backward algorithm, that consist of a forward function $\alpha(x_{t,n}) = p(y_{0,n}^t, x_{t,n})$ defined as the joint probability of the observed data up to time $t$, and a backward function $\beta(x_{t,n}) = p(y_{t+1,n}^T|x_{t,n})$ defined as the conditional probability of the observed data from time $t+1$ given the hidden state at $t$. Here, we abuse notation in $\alpha(x_{t,n})$ and $\beta(x_{t,n})$ by omitting the dependence on $y_{0,n}^t$ and $y_{t+1,n}^T$ respectively.
3. **Maximization step:** In this step, the probabilities calculated in the Expectation step are used to update the model parameters to maximize the expected log-likelihood of the observed data. This involves adjusting the probability $p$ of mental health crisis when the patient is at state $U$ and the transition probabilities between hidden states $(q, r)$.
4. **Iterate:** Steps 2 and 3 are repeated iteratively until a convergence criterion is met. In this case, convergence criteria is set to stop when the change between two consecutive iterations is below a certain tolerance $(10^{-5})$ or until a maximum number of iterations are completed (100).

### 2.3.1 Case 1: Parameter estimation per patient

To estimate the parameters of the model for a patient $n$ $(p_n, q_n, r_n)$, we want to find the values $q_n^*, r_n^*, p_n^*$ that maximize the likelihood of the observed history of the patient. Since we are estimating the parameters of the model, the likelihood and all the probability distributions are conditioned to the value of the parameters, i.e.,

$$\begin{aligned}\mathcal{L}(y_{1,n}^T|q_n, r_n, p_n) &= \log p(y_{0,n}^T|q_n, r_n, p_n)\\ &= \log \sum_{x_{0,n}^T \in \{S,U\}^T} p(x_{0,n}^T, y_0^T|q_n, r_n, p_n).\end{aligned}$$

For simplicity, we drop $q_n$, $r_n$, $p_n$ and the subscript $n$ from the notation in the following equations. Let's start with the joint probability of each state $x_t$ for $t < T$ given a set of parameters $q_n$, $r_n$, $p_n$ and the observed data $\mathcal{O}_n = y_{0,n}^T$.

$$\begin{aligned}p(x_t, y_0^T) &= p(y_0^T|x_t)p(x_t) = p(y_0^t|x_t)p(x_t)p(y_{t+1}^T|x_t)\\ &= p(y_0^t, x_t)p(y_{t+1}^T|x_t) = \alpha(x_t)\beta(x_t),\end{aligned}$$

with $\alpha(x_t) = p(y_0^t, x_t)$ and $\beta(x_t) = p(y_{t+1}^T|x_t)$ being the forward and backward functions repectively. Both $\alpha(x_t)$ and $\beta(x_t)$ can be computed iteratively.

The process of computing the $\alpha(x_0), \ldots, \alpha(x_T)$ is called forward step and can be derived as follows:

$$\alpha(x_t) = p(y_0^t, x_t) = \sum_{x_{t-1} \in \{S,U\}} \alpha(x_{t-1})p(x_t|x_{t-1})p(y_t|x_t),$$

with $\alpha(x_0) = p(x_0)p(y_0|x_0)$.

$\beta(x_0), \ldots, \beta(x_T)$ are computed iteratively starting backwards, this process is called the backward step:

$$\beta(x_t) = p(y_{t+1}^T|x_t) = \sum_{x_{t+1} \in \{S,U\}} \beta(x_{t+1})p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t),$$

with $\beta(x_T) = 1$ and $\beta(x_{T-1}) = \sum_{x_T \in \{S,U\}} p(y_T|x_T)p(x_T|x_{T-1})$.

Through these expressions, $\alpha(x_t)$ and $\beta(x_t)$ can be computed for all $t = 0, \ldots, T$ (21). From $\alpha(x_t)$ and $\beta(x_t)$ we can compute the probability distribution of the hidden states given the observations as

$$\begin{aligned}\gamma(x_t) = p(x_t|y_0^T) &= \frac{p(x_t, y_0^T)}{p(y_0^T)} = \frac{\alpha(x_t)\beta(x_t)}{p(y_0^T)}\\ &= \frac{\alpha(x_t)\beta(x_t)}{\sum_{x_t \in \{S,U\}} \alpha(x_t)\beta(x_t)}.\end{aligned} \tag{2}$$

To finish the Expectation step, we need to compute the probability distribution of the transitions given the observations:

$$\xi(x_t, x_{t+1}) = p(x_t, x_{t+1}|y_0^T) = \frac{p(x_t, x_{t+1}, y_0^T)}{p(y_0^T)}$$
$$= \frac{\alpha(x_t)p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})\beta(x_{t+1})}{\sum\limits_{x_t, x_{t+1} \in \{S,U\}^2} \alpha(x_t)\beta(x_{t+1})p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})}. \quad (3)$$

In both Equations 2 and 3 the denominators are computed by regularizing the numerator to convert it to probabilities. Observe that we abused notation by omitting the dependence on $y_0^T$ when we defined $\gamma(x_t)$ and $\xi(x_t, x_{t+1})$.

Once the probability distribution of the hidden states and the transition probabilities given the observed data are computed, we can use them in order to estimate the new set of parameters in the Maximization step. In particular, the estimated value for the parameter $q_n$, $\hat{q}_n$, can be calculated as the expected number of transitions from state $S$ to state $U$ divided by the expected number of transitions starting at state $S$

$$\hat{q}_n = \frac{\sum\limits_{t=0}^{T-1} \xi(x_{t,n} = S, x_{t+1,n} = U)}{\sum\limits_{t=0}^{T-1} \gamma(x_{t,n} = S)}.$$

The estimated value $\hat{r}_n$ for the parameter $r_n$, can be computed as the expected number of transitions from state $U$ to state $U$ divided by the expected number of transitions starting at state $U$,

$$\hat{r}_n = \frac{\sum\limits_{t=0}^{T-1} \xi(x_{t,n} = U, x_{t+1,n} = U)}{\sum\limits_{t=0}^{T-1} \gamma(x_{t,n} = U)},$$

and the estimated value for the parameter $p_n$, $\hat{p}_n$, can be estimated as the expected number of times at state $U$ and observing a crisis divided by the expected number of times at state $U$,

$$\hat{p}_n = \frac{\sum\limits_{t=0}^{T-1} 1_{y_{t,n}=1} \gamma(x_{t,n} = U)}{\sum\limits_{t=0}^{T-1} \gamma(x_{t,n} = U)}.$$

By iterating over the Expectation and Maximization step the algorithm converges to a local maximum on the likelihood function.

### 2.3.2 Case 2: Single parameter estimation for all patients

To estimate the parameters of the model assuming that all the patients have the same parameter values $(p, q, r)$, we want to find the values $p^*$, $q^*$ and $r^*$ that maximize the likelihood of the observed history of all the patients. Since the observations of each patient are independent of the rest of the patients, we have

$$\mathcal{L}(y_{0,1}^T, \ldots, y_{1,N}^T|q, r, p) = \log p(y_{0,1}^T, \ldots, y_{0,N}^T|q, r, p)$$
$$= \sum_{n=1}^{N} \log \sum_{x_{0,n}^T \in \{S,U\}^T} p(x_{0,n}^T, y_0^T|q, r, p).$$

Furthermore, the joint probabilities, given the parameters $p$, $q$ and $r$, can be computed per patient independently. Therefore, the results of the expectation step derived in the previous section can be used to compute the probability distribution of the hidden states and the probability distribution of the transitions given the observations of the patient. In this case, we define $\alpha_n(x_{t,n})$, $\beta_n(x_{t+1,n})$, $\gamma_n(x_{t,n})$ and $\xi_n(x_{t,n}, x_{t+1,n})$ for each patient $n$ like in Case 1, which are computed as

$$\alpha_n(x_{t,n}) = \sum_{x_{t-1,n} \in \{S,U\}} \alpha_n(x_{t-1,n})p(x_{t,n}|x_{t-1,n})p(y_{t,n}|x_{t,n}),$$

$$\beta_n(x_{t,n}) = \sum_{x_{t+1,n} \in \{S,U\}} \beta_n(x_{t+1,n})p(y_{t+1,n}|x_{t+1,n})p(x_{t+1,n}|x_{t,n}),$$

$$\gamma_n(x_{t,n}) = \frac{\alpha_n(x_{t,n})\beta_n(x_{t,n})}{\sum\limits_{x_{t,n} \in \{S,U\}} \alpha_n(x_{t,n})\beta_n(x_{t,n})}.$$

$$\xi_n(x_{t,n}, x_{t+1,n}) = \frac{\alpha_n(x_{t,n})p(x_{t+1,n}|x_{t,n})p(y_{t+1,n}|x_{t+1,n})\beta_n(x_{t+1,n})}{\sum\limits_{x_{t,n}, x_{t+1,n} \in \{S,U\}^2} \alpha_n(x_{t,n})\beta_n(x_{t+1,n})p(x_{t+1,n}|x_{t,n})p(y_{t+1,n}|x_{t+1,n})}.$$

In the maximization step, the new set of parameters can be estimated in a similar way as shown in the previous section. In this case, the expected values are computed using the distributions of all patients. As a result, we obtain the following formulas for the maximization step:

$$\hat{q} = \frac{\sum\limits_{n=1}^{N} \sum\limits_{t=0}^{T-1} \xi_n(x_{t,n} = S, x_{t+1,n} = U)}{\sum\limits_{n=1}^{N} \sum\limits_{t=0}^{T-1} \gamma_n(x_{t,n} = S)},$$

$$\hat{r} = \frac{\sum\limits_{n=1}^{N} \sum\limits_{t=0}^{T-1} \xi_n(x_{t,n} = U, x_{t+1,n} = U)}{\sum\limits_{n=1}^{N} \sum\limits_{t=0}^{T-1} \gamma_n(x_{t,n} = U)},$$

$$\hat{p} = \frac{\sum\limits_{n=1}^{N} \sum\limits_{t=0}^{T-1} 1_{y_{t,n}=1} \gamma_n(x_{t,n} = U)}{\sum\limits_{n=1}^{N} \sum\limits_{t=0}^{T-1} \gamma_n(x_{t,n} = U)}.$$

## 2.4 Patient monitoring policy

The final stage of this method is to devise a policy to decide when the patient is unstable or at a high enough risk of mental health crisis to require close monitoring from the clinical teams. For this, we define a threshold $\tau$ above which the patient has a high risk of crisis and needs to be followed closely. To generate our results we used $\tau = 0.35$, which maximizes the F1 score in the simulation. This threshold implies an estimated risk of crisis of 35%, but this threshold is adjustable depending on the capacity of the hospital. In order to implement this monitoring policy in clinical practice, we may follow the next steps:

1. Estimate the model parameters for each patient. First, the model parameters for an average patient are estimated using the data from all the patients in the hospital. These model parameters are assigned to all patients that have less than 3 months of data, as they have limited history with the hospital (the minimum number of months is configurable per hospital). The model parameters for the patients with more than 3 months of data are estimated using their individual history of data.
2. Compute the probability of mental health crisis. We can use the estimated parameters of each patient together with the time since their last observed crisis to compute the risk that the patient is going to suffer a mental health crisis during the current week.
3. Decide whether the patient needs close monitoring. If the risk to suffer a mental health crisis is higher than the threshold $\tau$ the patient is given close monitoring. Otherwise, the patient is deemed stable and they can be followed through less intensive means.

## 2.5 Data source

The results shown in this paper are based on simulations and an anonymised dataset. This anonymised dataset comprises 4,871 mental health crises from 162 psychiatric patients from the Birmingham and Solihull Mental Health Foundation Trust. The methods described in this manuscript are general and can be applied on similar datasets.

The programming language used to make the simulations, estimate the model parameters and produce the results was Python 3.9.8.

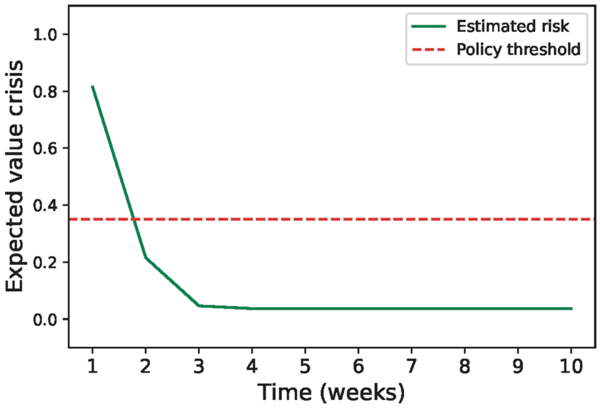


FIGURE 1
Evolution of the expected value of mental health crisis after the last crisis based on the parameters estimated using all the patients ($q^* = 0.037$, $r^* = 0.86$ and $p^* = 0.95$). The green line shows how the risk of crisis decreases with the number of weeks without crisis and the red line shows the value below which the patient is considered stable.

# 3 Results

## 3.1 Policy evaluation with data from a psychiatric hospital

To evaluate the performance of our method with actual data, we applied the steps described in Section 2.4 to a cohort of 162 patients that suffered mental health crises between September of 2012 and August 2016. We divided the dataset into the parameter learning set and the evaluation set. We used the data

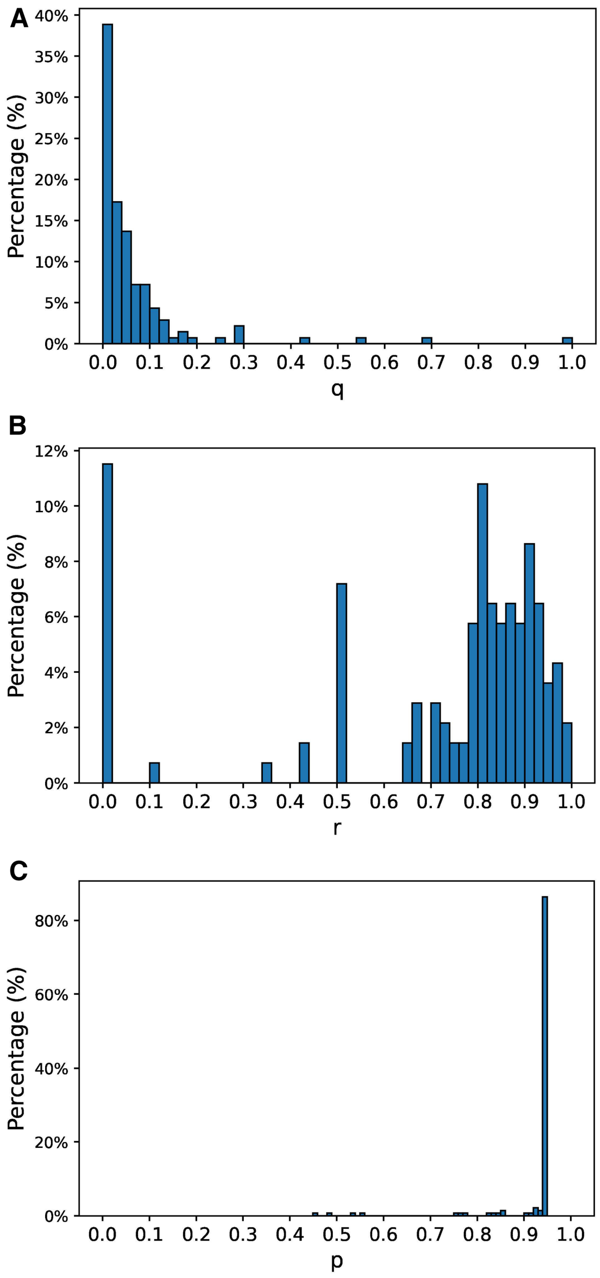


FIGURE 2
Estimated model parameters' distributions. (A) Distribution of estimated parameter $q$, (B) distribution of estimated parameter $r$, (C) distribution of estimated parameter $p$.

from 2012 until the end of 2015 to learn the model parameters (parameter learning set) and evaluated the performance of the model using the data from 2016 (evaluation set). The evaluation set corresponds roughly 30% of the data (note that not all patients had their first record during September 2012). The policy threshold was set to 0.35.

We started by estimating the parameters of the average patient following the procedure described at Section 2.3.2. We run the Baum–Welch algorithm with 100 different initial conditions, obtaining $p^* = 0.95$, $q^* = 0.037$ and $r^* = 0.86$ in all of them. This suggests that we reached the global optimum because all the initializations converged to the same model parameters.

As shown in Figure 1, with these parameters the risk of mental health crisis starts at 0.81 during the first week after a crisis and decreases each week until the 5th one, when the risk stabilises to 0.037. Under this setting and with the policy threshold at 0.35, the patient should be monitored during the week following a mental health crisis and would be considered to be stable starting the second week after their last crisis.

Then, we estimated the parameters for each patient separately. There were 23 patients (14.1%) that had less than 3 months of

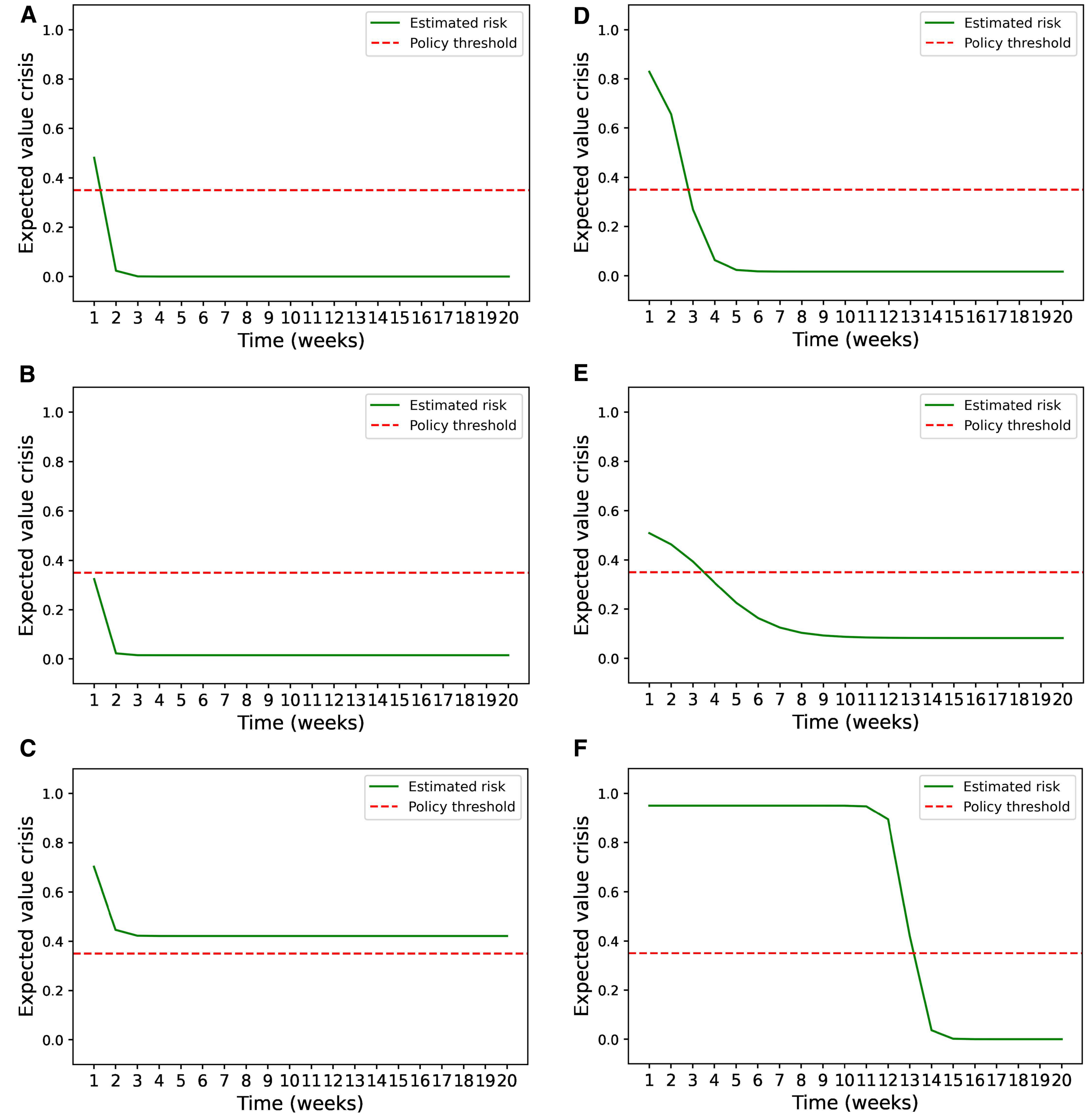


FIGURE 3
Estimated risk evolution. (A–F) Examples of how the estimated risk of mental health crisis evolves over the weeks following a mental health crisis for 6 patients that had different model parameters.

data at the end of 2015. The parameters estimated using the complete set of patients were assigned to these patients. For the remaining patients, we individually estimated the model parameters based on their particular observed histories. The distribution of the estimated parameters is shown in Figure 2. The distribution of estimated $q$ is skewed towards 0, denoting that most of the patients have a low probability of relapsing once they are stable. By looking at the distribution of estimated $r$ we see that a large portion of patients (85%) have a probability of staying in an unstable state higher than 0.5. This indicates that most patients tend to stay unstable for more than one week. A significant portion of patients (11.5%) have $r = 0$, which means that these patients usually have isolated crisis and stabilize quickly. Finally, the distribution of estimated $p$ is very skewed towards 1, which means that most patients experience a crisis when they are unstable.

The evolution of the estimated risk of mental health crisis over time depends on the patient's model parameters. Figure 3 shows some examples of how this risk evolves after the patient's last crisis. Most patients display a pattern similar to $A$, $B$ and $C$, having a fast decrease on the estimated risk during the second week after the crisis and reaching convergence to a certain risk level in 3 or 4 weeks. Some other patients, such as examples $D$, $E$, had a slower convergence rate that required more than 7 iterations to converge. There were three patients that did not display a significant decrease until 13 weeks after the patient had their last crisis, a representative example is shown in $F$. These patients exhibit the estimated parameters $r$ and $q$ close to 1 and 0 respectively.

In Figure 4, we show the distribution of the number of weeks that the patient needs close monitoring before is considered stable. The proposed policy established that 67.3% of the patients should be closely monitored only one week after their last mental health crisis, 16.7% during the following two weeks and 4.9% for 3 weeks or more (including 2.5% that should be always monitored). The remaining 11.1% were patients whose risk of mental health crisis was lower than the policy threshold at all weeks.

By following this policy, 56 patients would be closely monitored each week on average (corresponding to 34.3% of the patients) -close to the 55 (33.7%) mental health crisis that occur on average-, among which 78.6% would be patients that suffer a mental health crisis (precision). This policy detects 79.8% of the crises (recall) with a false positive rate of 11.1%, corresponding to a F1-score (22) of 0.79. Figure 5 shows the confusion matrix.

## 3.2 Model learning and policy validation in a hospital simulation

To test how well our method performs given that our assumptions hold, we simulated 5 years of data from a cohort of 3,000 random patients. Each patient has a different set of model parameters generated at random. We chose the distributions to generate the model parameters to resemble the distribution of the estimated model parameters with the data from the psychiatric hospital (see Figure 2). Specifically, we sampled $\tilde{p}$, $\tilde{q}$ and $\tilde{r}$ from a lognormal distribution according to

$$\tilde{p} \sim \log\mathcal{N}(0, 1.2)$$
$$\tilde{q} \sim \log\mathcal{N}(0, 1)$$
$$\tilde{r} \sim \log\mathcal{N}(0, 0.3).$$

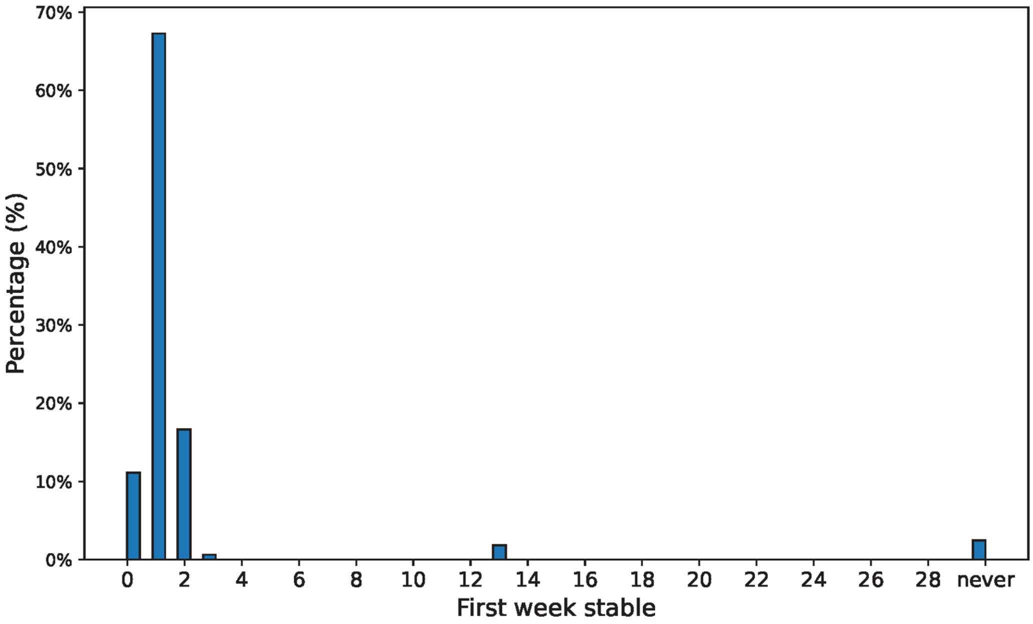


FIGURE 4
Distribution of the number of weeks that a patient needs to be closely monitored before deemed stable.

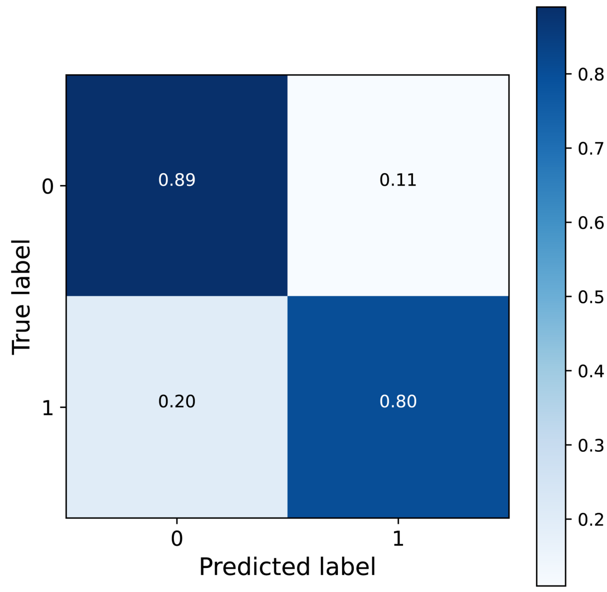


FIGURE 5
Confusion matrix of the detected crisis by following the policy with a threshold of 0.35.

Then, the values generated by $\tilde{p}$, $\tilde{q}$ and $\tilde{r}$ were scaled and transformed to get the $p$, $q$ and $r$ to lie within the range $(0.02, 0.98)$, $(0.02, 0.5)$ and $(0.02, 0.98)$ respectively. Since we observed that the estimated value of $r$ for around 10% of the patients in the real hospital was 0, we selected 300 patients from the simulation at random and assigned them $r = 0$. The distribution of $p$, $q$ and $r$ are shown in Figure 6.

For each of the patients, we estimated the model parameters by following the steps described in Section 2.3.1 using the data from the first 4 years (parameter learning set). We executed the Expectation and Maximization steps iteratively until convergence or until 1000 iterations were completed. Convergence was defined as the point at which the difference in log-likelihood between two consecutive iterations was less than $10^{-5}$. Remarkably, convergence was achieved in 98.5% of the patients, requiring no more than 50 iterations in 91.4% of the cases. The distribution of iteration counts leading to convergence is shown in Figure 7. The parameter estimation process had a mean absolute error of 0.03 for the parameter $q$, 0.06 for the parameter $r$ and 0.08 for parameter $p$. The distribution of the errors is shown in Figure 8.

The estimated model parameters were then used to produce the predicted risk of mental health crisis using the Equation 1 from Theorem 1. We computed the predictions for every patient and every week of the last year of the simulation (evaluation set). For the purpose of this simulation, we decided that the policy threshold from which the patients would be considered to be at a high risk of suffering a mental health crisis was 0.35, which corresponded to the maximum F1-score (22) (0.74) in the evaluation set. With this threshold, 77.3% of the patients were considered stable after a week of not having a mental health crisis, 9.0% required 2 weeks to be deemed stable, while 12.7% were estimated to not need close attention even the first week after the mental health crisis occurred. The rest of the patients (1.0%) required 3 or more weeks without a mental health crisis before they are considered to be stable (0.2%) or were considered always unstable (0.8%). Figure 9A shows the distribution of the number of weeks without crisis before a patient is deemed stable. Under this policy there are on average 608 patients at risk of suffering a mental health crisis that should be closely monitored each week, corresponding to 20.3% of the total number of patients. In the same period of time, there were 602 patients on average that suffered a mental health crisis (corresponding to

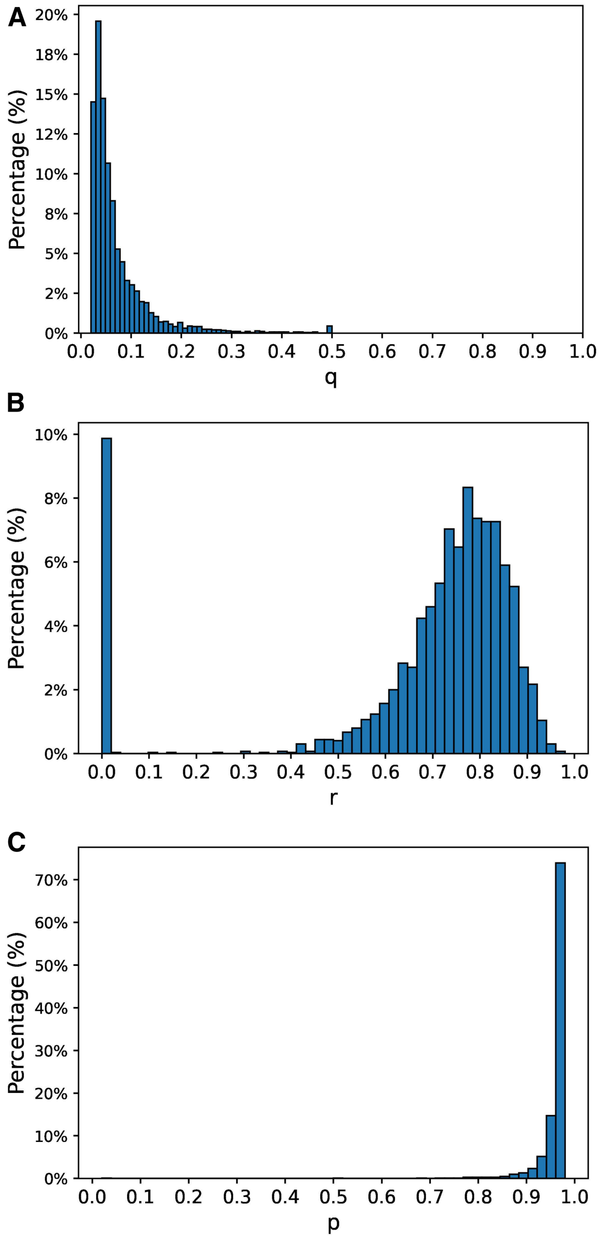


FIGURE 6
Distribution of the model parameters used in the simulation. (A) Distribution of parameter $q$, (B) distribution of parameter $r$, (C) distribution of parameter $p$.

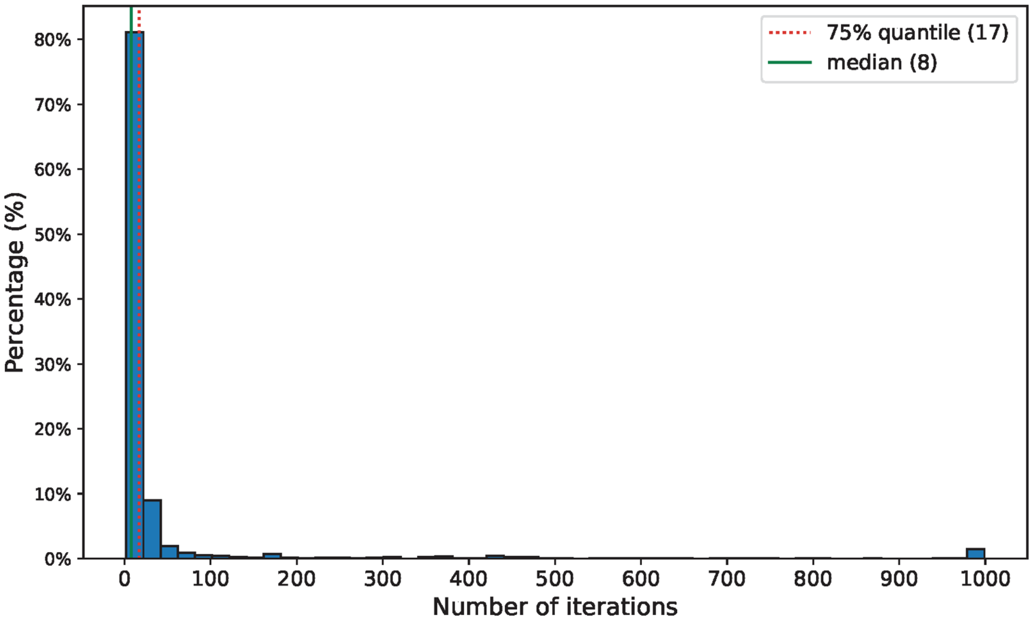


FIGURE 7
Distribution of the number of iterations required before convergence.

20.0%), which is close to the number of flagged patients. Among the cases in which a patient was flagged to be monitored closely, 96.2% were patients at state $U$. In comparison, the patient was at state $U$ only in 25.1% of the cases at the week that the patient was deemed stable and 6.5% of the instances during the following 4 weeks. Figure 9B shows the percentage of patients at state $U$ at each week after the one they were considered to be stable.

## 4 Discussion

In this work, we introduced a novel method to determine the optimal monitoring duration for a psychiatric patient following a mental health crisis before being considered stable. Our method leverages a probabilistic framework utilizing a HMM that solely relies on the historical record of observed crises. To estimate the parameters of the HMM we employed the Baum–Welch algorithm, a well-established technique that remains as the preferred choice to optimize the parameters of a HMM. These parameters can be used to infer the probability that a patient is unstable during the weeks following a mental health crisis and to estimate the risk of a new crisis occurring at each week. Through the resolution of a Ricatti difference equation we demonstrated the existence of a closed-form solution that exhibits exponential convergence and estimated the probability of a mental health crisis at each week following the last occurrence. These results enable the development of a policy for determining the point at which a patient can be deemed stable, with a minimal risk of experiencing a new mental health crisis.

When defining the probabilistic model for mental health states, we established four assumptions. First, we assumed that patients can be in one of two possible mental states during any given week: stable (S) or unstable (U). These states evolve following a Markov Chain, providing a simplified framework that reduces the complexity of the model and makes it more tractable for analysis. Another key assumption is that the patient's mental state remains hidden and is only partially observable through the crisis variable. This assumption reflects the practical constraints associated with directly measuring a patient's mental state. It aligns with real-world scenarios, where a patient's state is indirectly inferred through interactions with the hospital. Furthermore, we assumed that patients in state S cannot experience a crisis, while those at state U have a non-zero probability of suffering a crisis. This assumption aligns with reality and simplifies the modeling of crisis events by directly linking them to the patient's current mental state. Finally, we assumed that $p < 1$ and $q \neq r$ as it enables the solution of the Riccati equation presented in Section 2.2. This is a reasonable assumption because patients in an unstable state do not experience crises continuously until they stabilize, and patients are often more likely to remain in their current state than to switch (typically, $r > q$).

The assumptions we made serve the purpose of simplifying a complex problem, making it tractable for analysis, all while maintaining consistency with real-world scenario. However, it is essential to acknowledge potential limitations. Firstly, the evolution of a patient's mental state is a complex process and our model may not fully capture the spectrum of mental states a patient can experience or the intricacy of their transitions. To address this, the model could be extended by introducing a broader range of possible states and considering a higher order Markov Chain (23), which accounts for the influence of past mental states. Although this would yield different analytical

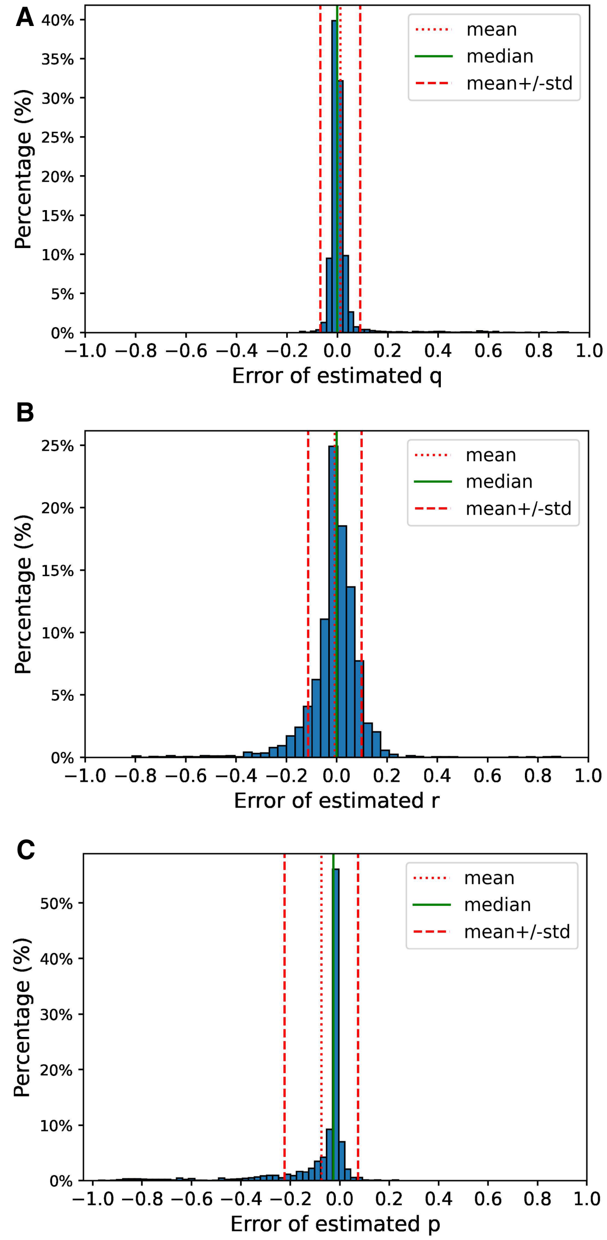


FIGURE 8
Distribution of the errors during the parameter estimation using the data from the simulation. The error for each of the parameters is computed as estimated parameter minus the actual value of the parameter. (A) Distribution of estimation error for parameter $q$, (B) distribution of estimation error for parameter $r$, (C) distribution of estimation error for parameter $p$.

results in Section 2.2, similar steps could be taken, and an adapted version of the Baum–Welch algorithm could be applied to estimate the model parameters in higher order HMM (24). However, the introduction of additional parameters to the model would make the parameter estimation harder. Secondly, we assumed that the probability that a patient in state U suffers a mental health crisis remains constant, yet this probability might increase or decrease over time in state U. Introducing a time dependence to the variable $p$ would not alter the solution from Theorem 1, but the parameter estimation would change based on the chosen family of functions used to define this time dependence. Finally, while mental health crises are the sole observed signal in our model, clinicians may directly or indirectly observe the mental state of a patient during regular visits. Theorem 1 provides the solution when an S state is observed, and the inclusion of these observations would enhance the estimation of the model parameters. However, the incorporation of additional relevant information, such as data from routine visits between crises, diagnosed disorders, or prescribed medications, would require further research.

We presented two sets of results. The first set, based on actual data collected at a psychiatric hospital, aimed to assess the performance of our method in a real-world scenario. When we estimated the model parameters assuming uniform model parameters for all patients, the predicted risk of mental health crisis dropped substantially between the first and the second week after the last mental health crisis, from 0.81 to 0.21. This suggests that by employing a one-size-fits-all approach, patients can generally be considered stable after just one week, aligning with previous literature assumptions (15, 16). However, when we estimated the parameters individually for each patient, significant variations emerged. In most cases, $p^*$ exceeded 0.9, but $r^*$ ranged from 0 in 11.5% of the cases to skewing towards 1 in the remainder, while $q^*$ predominantly skewed towards 0. Each patient's risk of crisis exhibited distinct patterns based on their estimated parameters. Under our policy, 67.3% of the patients required close monitoring only during the first week after their last mental health crisis, 16.7% for two weeks, and 4.9% for three weeks or more. A small percentage of patients (2.5%) maintained a risk above 0.35 even after convergence, requiring continuous close monitoring. In contrast, 11.1% of patients never exceeded and did not necessitate monitoring according to our policy. The application of this policy yielded an F1 score of 0.79 in the evaluation set, detecting 79.8% of the crises with a false positive rate of 11.1%. While these outcomes underscore the method's strong performance with real-world data and its potential to determine the optimal monitoring duration for each patient before deeming them stable, it is essential to note that these conclusions are drawn from a sample size of 162. This limitation suggests the importance of further validation with larger patient cohorts.

The second set of results, based on a simulation designed to validate the performance of our method when our underlying assumptions are met. In this analysis, we observed rapid convergence in the parameter estimation step across nearly all the cases, with a mean absolute error below 0.1 for all three estimated parameters. However, it is important to note that in a small number of instances, substantial differences emerged between our estimated parameters and their actual values. Leveraging these estimated model parameters, we predicted the risk of mental health crisis as outlined in Theorem 1 and created a policy to identify those patients with a risk exceeding a defined threshold (0.35 in this particular case). By applying this policy,

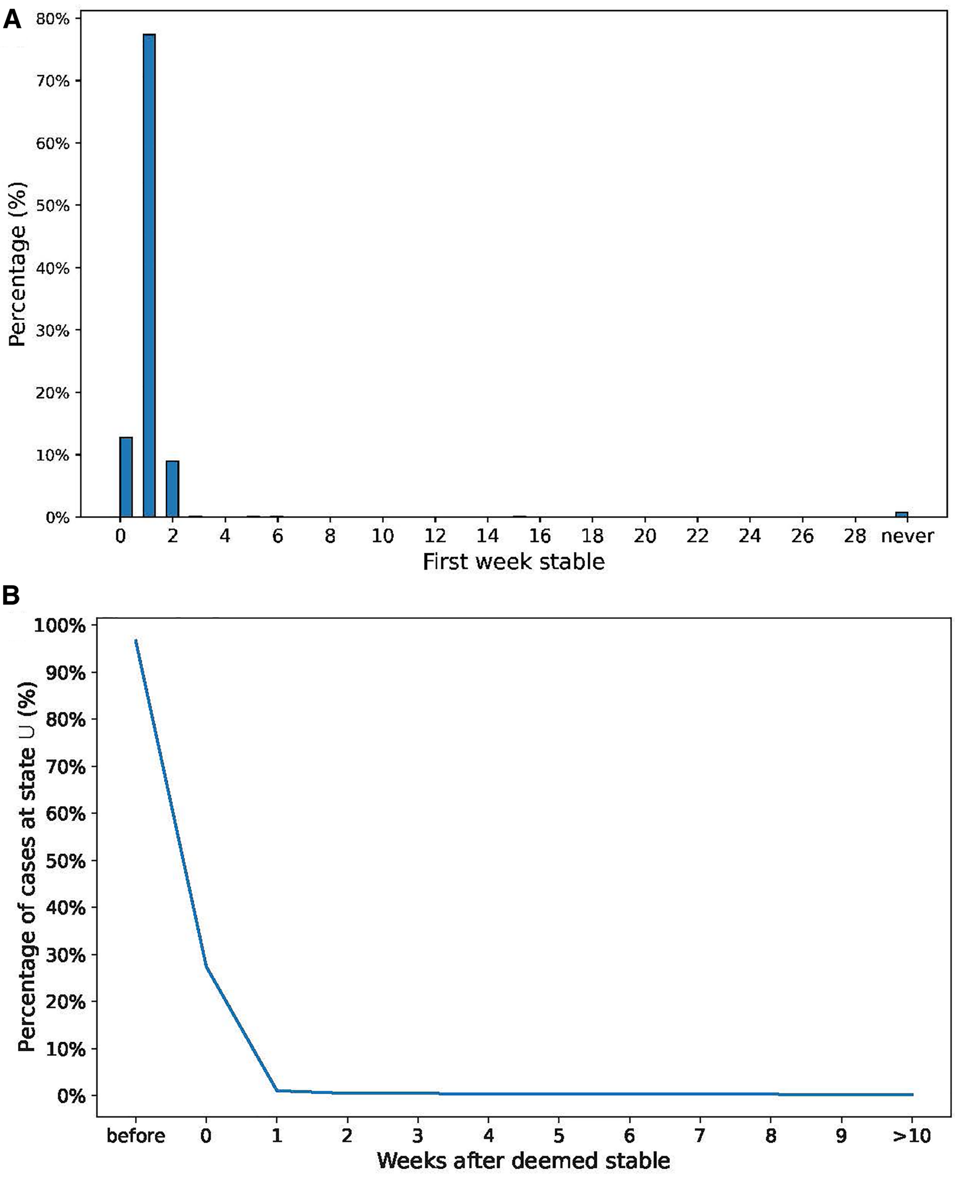


FIGURE 9
(A) Distribution of weeks required to consider the patient "stable". (B) Percentage of *U* state at each week after the patient is "stable".

we obtained an F1 score of 0.74, with 96.2% of the flagged patients in an unstable state. These results provide compelling evidence that our method performs as intended when our model assumptions hold.

Recurring mental health crises pose a profound threat to both the individual patient's mental and social well-being, and by extension, that of their family. With each hospital admission, a substantial allocation of resources becomes imperative, imposing a considerable financial burden on mental healthcare facilities. While hospitals typically implement one-size-fits-all policies shaped by the needs of the majority of patients, these policies often overlook the nuanced variations in mental health disorders and among individual patients. Our innovative data-driven approach offers a bespoke assessment that meticulously considers these variations, paving the way for more personalized and effective mental healthcare interventions.

Recurring mental health crises pose a profound threat to both the individual patient's mental and social well-being, and by extension, that of their family. With each hospital admission, a large quantity of resources needs to be allocated to treat the patient, imposing a considerable financial burden on mental healthcare facilities. While hospitals typically implement one-size-fits-all policies driven by the needs of the majority of patients, these policies often overlook the nuanced variations in mental health disorders and among individual patients. Our data-driven approach provides an individualized assessment that considers these variations, paving the way for more personalized and effective mental healthcare interventions.

## Data availability statement

The datasets presented in this article are not readily available because hospital data cannot be shared publicly due to the risk of violating privacy. All the other sources including the code used to generate the simulated data, estimate the model parameters, evaluate the models and generate the results presented in this study can be found at the open source repository https://github.com/Icedgarr/post_crisis_monitoring. Requests to access the datasets should be directed to Roger Garriga, roger.garrigacalleja@koahealth.com.

## Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

## Author contributions

RG: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing; VG: Supervision, Validation, Writing – review & editing, Methodology; GL: Conceptualization, Formal Analysis, Methodology, Supervision, Validation, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article.

This work has been co-funded by MCIN/AEI/10.13039/501100011033 under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M). This publication is part of the action CNS2022-136178 financed by MCIN/AEI/10.13039/501100011033 and by the European Union Next Generation EU/PRTR. The authors would like to thank Koa Health for supporting and funding this study. Koa Health was not involved in the study design, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## Acknowledgments

The authors would like to thank Birmingham and Solihull Mental Health Foundation NHS Trust for providing the anonymized dataset of mental health crises used to evaluate the performance of this method.

## Conflict of interest

RG was employed by Koa 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

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.2024.1322555/full#supplementary-material

## References

1. National Alliance of Mental Illness. *Navigating a Mental Health Crises*. Arlington: National Alliance of Mental Illness (2018).

2. Caplan G, *Principles of Preventive Psychiatry*. New York: Basic Books (1964).

3. Heyland M, Johnson M. Evaluating an alternative to the emergency department for adults in mental health crisis. *Issues Ment Health Nurs*. (2017) 38:557–61. doi: 10.1080/01612840.2017.1300841

4. Zeller SL. Treatment of psychiatric patients in emergency settings. *Prim Psychiatry*. (2010) 17:35–41. Available at: https://www.researchgate.net/publication/283157460_Treatment_of_psychiatric_patients_in_emergency_settings (Accessed January 29, 2024).

5. Miller V, Robertson S. A role for occupational therapy in crisis intervention, prevention. *Aust Occup Ther J*. (1991) 38:143–6. doi: 10.1111/j.1440-1630.1991.tb01710.x

6. Morrice JKW, *Crisis Intervention: Studies in Community Care*. Oxford: Pergamon Press (1976).

7. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. *Ann Math Stat*. (1966) 37:1554–63. doi: 10.1214/aoms/1177699147

8. Chen IY, Joshi S, Ghassemi M, Ranganath R. Probabilistic machine learning for healthcare. *Annu Rev Biomed Data Sci*. (2021) 4:393–415. doi: 10.1146/annurev-biodatasci-092820-033938. PMID: 34465179

9. Ansari H, Vijayvergia A, Kumar K. Dcr-hmm: depression detection based on content rating using hidden Markov model. In: *2018 Conference on Information, Communication Technology (CICT)*. IEEE (2018). p. 1–6.

10. Jiang X, Chen Y, Ao N, Xiao Y, Du F. A depression-risk mental pattern identified by hidden Markov model in undergraduates. *Int J Environ Res Public Health*. (2022) 19:14411. doi: 10.3390/ijerph192114411

11. Hulme WJ, Martin GP, Sperrin M, Casson AJ, Bucci S, Lewis S, et al. Adaptive symptom monitoring using hidden Markov models—an application in ecological momentary assessment. *IEEE J Biomed Health Inform*. (2021) 25:1770–80. doi: 10.1109/JBHI.2020.3031263

12. Chen Y, Oyama-Higa M, Pham TD. Identification of mental disorders by hidden Markov modeling of photoplethysmograms. In: *International Conference on Biomedical Informatics and Technology*. Springer (2013). p. 29–39

13. Boeker M, Hammer HL, Riegler MA, Halvorsen P, Jakobsen P. Prediction of schizophrenia from activity data using hidden Markov model parameters. *Neural Comput Appl*. (2023) 35:5619–30. doi: 10.1007/s00521-022-07845-7

14. Alam MGR, Haw R, Kim SS, Azad MAK, Abedin SF, Hong CS. Em-psychiatry: An ambient intelligent system for psychiatric emergency. *IEEE Trans Ind Inform*. (2016) 12:2321–30. doi: 10.1109/TII.2016.2610191

15. Garriga R, Mas J, Abraha S, Nolan J, Harrison O, Tadros G, et al. Machine learning model to predict mental health crises from electronic health records. *Nat Med*. (2022) 28(6):1240–8.

16. Garriga R, Buda TS, Guerreiro J, Omana Iglesias J, Estella Aguerri I, Matic A. Combining clinical notes with structured electronic health records enhances the prediction of mental health crisis. *Cell Rep Med*. (2023) 4(11). doi: 10.1016/j.xcrm.2023.101260

17. Shandhi MMH, Dunn JP. Ai in medicine: Where are we now, where are we going?. *Cell Rep Med*. (2022) 3:100861. doi: 10.1016/j.xcrm.2022.100861

18. Manchia M, Pisanu C, Squassina A, Carpiniello B. Challenges, future prospects of precision medicine in psychiatry. *Pharmgenomics Pers Med*. (2020) 13:127–40. doi: 10.2147/PGPM.S198225

19. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat*. (1970) 41:164–71. doi: 10.1214/aoms/1177697196

20. Yang F, Balakrishnan S, Wainwright MJ. Statistical and computational guarantees for the Baum–Welch algorithm. In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA*. (2015). p. 658–65. doi: 10.1109/ALLERTON.2015.7447067

21. Barber D, *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press (2012).

22. Chinchor N. MUC-4 evaluation metrics. In: *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16–18, 1992* (1992).

23. Ching WK, Ng MK. Higher-order Markov chains. In: *Markov Chains: Models, Algorithms and Applications. International Series in Operations Research & Management Science*, vol 83. Boston, MA: Springer US (2006). p. 111–39. doi: 10.1007/0-387-29337-X_6.

24. Seifert M. *Extensions of hidden Markov models for the analysis of DNA microarray data* (Ph.D. thesis). Halle (Saale), Martin-Luther-Universität Halle-Wittenberg, Diss., (2010)

frontiers | Frontiers in Psychiatry

TYPE Opinion
PUBLISHED 11 March 2024
DOI 10.3389/fpsyt.2024.1347100

Check for updates

OPEN ACCESS

EDITED BY
Hugo Schnack,
Utrecht University, Netherlands

REVIEWED BY
Seyed Mostafa Kia,
TilburgUniversity, Netherlands

*CORRESPONDENCE
Marcos DelPozo-Banos
m.delpozobanos@swansea.ac.uk

RECEIVED 30 November 2023
ACCEPTED 22 February 2024
PUBLISHED 11 March 2024

CITATION
DelPozo-Banos M, Stewart R and John A (2024) Machine learning in mental health and its relationship with epidemiological practice. *Front. Psychiatry* 15:1347100. doi: 10.3389/fpsyt.2024.1347100

# Machine learning in mental health and its relationship with epidemiological practice

Marcos DelPozo-Banos[1]*, Robert Stewart[2,3] and Ann John[1]

[1]Swansea University Medical School, Swansea, United Kingdom, [2]King's College London, Institute of Psychiatry, Psychology and Neuroscience, London, United Kingdom, [3]South London and Maudsley National Health Service (NHS) Foundation Trust, London, United Kingdom

KEYWORDS

**mental health, epidemiology, machine learning, research methods, challenges and opportunities**

## Introduction

It is fair to say that the application of machine learning (ML) in healthcare has not been smooth. The field of ML has let down the medical community and the wider public in many respects: from research that is clinically irrelevant (1) or applying flawed methodologies (2), to non-transparent sharing of data with industry (3, 4). Success stories do exist across a range of physical health specialties, but they currently remain a minority (5, 6).

In hindsight, the pitfalls for ML in medical research are hardly surprising. Epidemiology (medicine's own approach) is underpinned by statistics and hypothesis testing, designed to maintain ethical etiquette, ensure robust, unbiased results, and produce strong evidence and knowledge in measured phenomena – at least in principle (7). It therefore aims to understand the 'true' mechanisms connecting exposures and outcomes (features and targets in ML jargon) and naturally gravitates towards simpler, easier to interpret models. ML has its own established methodology (8), but one that is fundamentally different, geared towards solving problems and developing applications (9). It therefore pursues maximum accuracy at predicting the outcome and naturally prefers complex, more powerful models. The different use of logistic regression by both fields illustrates this. While epidemiology takes special care with correlated independent variables and directs its attention to the estimated coefficients, ML mostly disregards these and focusses on predictive power. Overall, while both epidemiology and ML rely on data to obtain their results, their core principles are at odds. Nevertheless, appropriately introducing ML elements into epidemiological research is possible and guidelines have been published (10).

Mental health has been a target for ML, with the number of ML mental health publications increasing dramatically since, 2017 (Figure 1), and the research community is rightly expectant of its impact. However, the challenges are amplified: (1) losing sight of mental health objectives, over-promising on data processing and problem-solving (9); (2) technical hurdles of multiple underlying biases and often heightened privacy requirements (11); and (3) difficulties building, validating and approving ML-enabled clinical devices for diseases with insufficiently clear underlying mechanisms (12). Overall, it is the

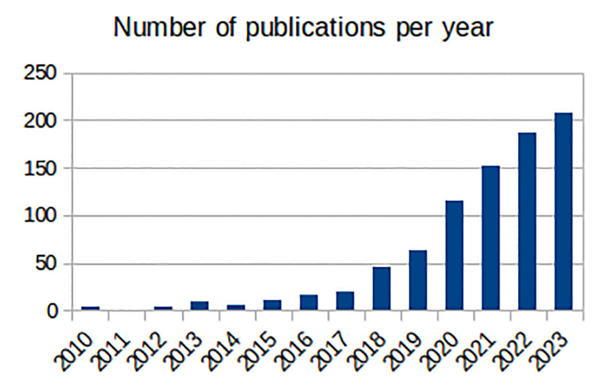


FIGURE 1
Number of annual publications found by PubMed (https://pubmed.ncbi.nlm.nih.gov/) between, 2010 and, 2023 with the terms "machine learning" and one of "mental health", "mental illness", "depression", "anxiety", "bipolar disorder, "schizophrenia", "psychotic disorder", "ADHD" or "autism" in the title.

responsibility of individual researchers and institutions alike to demonstrate the value of ML for mental health. Here, we reflect on these ideas and their corresponding steps within the workflow of ML mental health research (Figure 2), in the hope of bringing awareness to the field and to elicit further conversations.

## The ideal target for ML

Factors affecting an individual's mental health extend far beyond the clinical setting and are numerous, with complex interactions. Social, demographic, and economic factors and people's psychological make-up carry as much or more weight in estimating risk of mental health outcomes as medical symptoms, biological factors, and previous health (e.g., the effect of loneliness in suicidal thoughts and self-harm) (13). The complexity of these relationships is typified in suicide research, where a meta-analysis of risk factors identified little progress in prevention over a span of 50 years (14). Consequently, the heavy reliance of classical statistics on prior expert knowledge and model assumptions is another important limiting factor in mental health research. Progress in data provision and data linkage has addressed some of the challenges of mental health research (i.e., providing better population coverage and a wider range of risk factors) but has brought additional challenges such as larger volumes of data, lower data quality, increased missing data and unstandardised phenotypes (15, 16). Furthermore, the field of mental health is evolving, and expert consensus is lacking on the taxonomy of psychiatric diagnosis (17) or on preferred 'transdiagnostic' clinical phenotypes (18).

The complexity and wide reach of its disease models are why mental health might particularly benefit from ML. ML is better equipped than classical statistics to deal with large numbers of factors, complex (i.e., non-linear) interactions, and noise (i.e., low quality or missing data, unstandardised phenotypes) (19). A data-driven approach is of particular value (20), such as deep learning techniques (21), and ML could be pivotal in evidence provision for diagnostic taxonomies or clinical phenotypes. However, this requires demonstrable evidence on applied clinical validity.

## Keeping sight of mental health aims and objectives

Single studies of ML predicting an outcome from a given dataset, and therefore only presenting performance results of these models, are of limited interest for mental health research (22). More valuable applications seek to improve our understanding of the disease (e.g., risk factors or time trends) and/or identify intervention opportunities. Therefore, researchers working on ML mental health should strive to: (1) extract new clinical insights from their models; (2) validate such insights with supplementary statistical analyses, and (3) contextualise their findings in the existing clinical literature. Completing all three objectives in full is not always possible, but researchers should make an honest effort on each of them and, when unsuccessful, acknowledge it as a limitation of their research.

This is not to say that research aiming at developing new ML algorithms and methodologies to process data with similar characteristics to those from mental health data (outlined below) are unimportant. Such research may naturally rely on mental health data, but the focus is on the fundamental characteristics of the data, not its mental health content – indeed, the research could have been completed using any other (non-mental health) data with the same fundamental characteristics. In this scenario, researchers should recognise that their work is about ML and not mental health, and this should be reflected in the focus of their papers and their targeted audience.

## ML challenges when using mental health data

Data curation is a critical part in developing ML models for healthcare. Some of the steps involved in this process are identical to those seen in epidemiological research: determining the sample size through power calculations; assessing the quality of the variables; studying bias in the patterns of missing data and recording practices; and evaluating the representation of the study population by the study sample. Other data curation steps are more specific to ML: the need for larger volumes of data, especially for complex models (23); comprehensive evaluation of outcome variable quality (24); data partition strategies for model building and validation (in ML jargon *training* and *testing*; *cross-validation*, often done repeatedly to improve robustness and generalizability of the results) (25); and considering additional security measures to prevent data inference from the ML model itself (in ML jargon *membership inference attacks*) (26).

Many of the data curation steps described above are potentially more complex in mental health research. Recall and reporting biases are common in self-reported mental health data, and can lead to under- or overestimation of underlying associations (27). When these biases affect the outcome variable, the entire validity of the model can be compromised. With ML being a "data driven" approach, these biases can be especially damaging in ML applications. They should therefore be reduced as much as

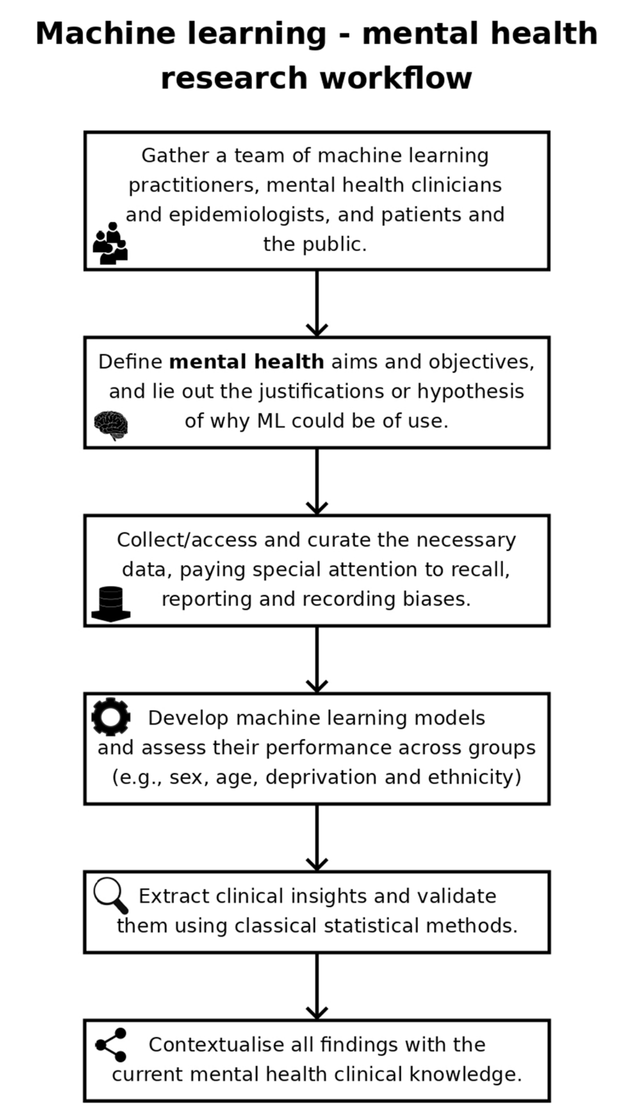

Machine learning - mental health research workflow
Gather a team of machine learning practitioners, mental health clinicians and epidemiologists, and patients and the public.
Define **mental health** aims and objectives, and lie out the justifications or hypothesis of why ML could be of use.
Collect/access and curate the necessary data, paying special attention to recall, reporting and recording biases.
Develop machine learning models and assess their performance across groups (e.g., sex, age, deprivation and ethnicity)
Extract clinical insights and validate them using classical statistical methods.
Contextualise all findings with the current mental health clinical knowledge.

FIGURE 2
Typical workflow of machine learning – mental health research.

possible to improve the model's performance and clinical validity, with the remaining bias carefully considered when assessing the results (24). Additionally, achieving participation and retention of participants in mental health research may also be challenging (28). The use of routinely collected electronic health records alleviates these issues to some extent; however, many important constructs of interest are subjective and can only be self-reported. Furthermore, minority groups (often the most affected by mental health inequalities) and those with more severe syndromes are frequently excluded and underserved (29). In addition, outcomes such as self-harm are known to be under-recorded in electronic health records (30). More generally, mental health data are viewed as relatively sensitive, partly due to the personal nature of the questions asked in a typical clinical assessment but also due to the stigma surrounding mental health conditions and consequent heightened privacy concerns – the public is slightly less inclined to share their mental health data for research compared to their physical health (31). This results in additional ethical and legal hurdles for mental health research (32), and more so for the application of ML due to its need of large volumes of data and the risk of models inadvertently carrying these data (26).

There is no *easy fix* for these problems, and, compared to most other medical specialties, mental health researchers, especially those applying ML, often need to: (1) focus more resources on their data curation strategy; (2) address bias in their data with statistical tools such as inverse probability weighing, which can be applied to both epidemiology (33) and ML (34) methods; and (3) have a stronger patient and public involvement and engagement plan (35).

## ML enabled clinical mental health devices with unknowns

The path from the lab to the clinical setting for medical innovations is not simple. This is especially true for ML-enabled devices, and still under discussion (5, 36) with regulatory frameworks evolving (37). In fact, only a small proportion of the published clinical ML research has been focused on deployment (5); as of October 19, 2023, the United States Food and Drug Agency reports approving less than 700 ML-enabled medical devices (based on their summary descriptions) (38), although this is likely an underestimation due to bias in explicit reporting of ML methods (39).

The situation is exacerbated for clinical mental health devices, with less real-world deployments (40) and fewer FDA approved devices (6). This may reflect the currently restricted scope of such devices as a consequence of our limited knowledge of the mechanisms underlying mental disorders, at least relative to other specialties (12). Without such knowledge, ML models are often fed a wide range of risk factors suspected to be related to the outcome (or in the hope that they will be of value during prediction). The assumption here is that if a model accurately predicts the outcome, it must be a true representation of the real-world phenomena described by the data. However, the data may contain variables that are confounders or act as proxies to latent variables, thus rendering the assumption unfair. When the *potential* risk factors fed to the ML algorithm lack evidence supporting and explaining their relationship with the outcome (as it is often the case), the clinical validity of the resulting ML-enabled mental health device remains to be proven, regardless of its accuracy. However, with the clinical knowledge laid down, healthcare professionals and patients will be more likely to accept the *black box* quality of ML models (41), and ML will have a clearer path to developing mental health solutions.

## Individual and collective responsibility

Researchers have a responsibility to demonstrate that, when correctly applied, ML can lead to improved knowledge and care of mental health disorders. To achieve this, ML practitioners must work in close collaboration with mental health epidemiologists and clinicians, and actively seek their input to protocol design and data interpretation. Crucially, they need to acknowledge that data fed into ML models represent personal experiences, to be aware of the particular sensitivities of mental health data, and to learn to handle these data responsibly above and beyond legislated privacy and security requirements. Conversely, mental health researchers seeking to engage with ML must avoid being blinded by the hype. Instead, they must continue to adhere to the main methodological principles of epidemiology and mental health research, and scrutinise any ML models generated (42). They should also be cautious of utilizing easy-to-use ML libraries and tools without the appropriate training, as these have led to the abuse and misuse of ML by non-experts (43).

Organisations and large projects could play a key role in ensuring that the fields of mental health and ML interact as described here. For example, DATAMIND (the MRC funded, UK Hub for Mental Health Data Science; www.datamind.org.uk) brings the issues outlined above to the attention of the field of mental health research at large, holding regular meetings and conferences with a wide range of stakeholders, and providing mental health data science workshops for early career researchers. DATAMIND is also developing a set of standardised mental health phenotypes to be used by the scientific community (44) and contributing to the cataloguing of available mental health data resources to improve discoverability and accessibility (45). Crucially, DATAMIND achieves this in close collaboration with academics, healthcare professionals, industry, and, most importantly, patients and people with lived experiences.

## Concluding remarks

Overall, the opportunity of using ML in mental health is not cost-free. As described, it introduces complexity, especially in mental health research, and additional workflow steps. Therefore, its application in healthcare generally, and in mental health particularly, needs to be justified. Ideally, this should be done at the planning stage, evidencing why the use of ML is needed to solve an existing problem that is hindering research: for example, to reduce an original set of available measurements to a size that is more manageable for traditional statistical regression (46). Alternatively, the benefits of using ML over conventional statistical methods can be treated as a hypothesis to be tested as part of the research project: for example, by comparing how well ML and statistical models fit the used data.

Beyond the hype, ML can genuinely play a central role in the future of psychiatry and mental healthcare. However, this depends on researchers applying ML responsibly and avoiding the mistakes seen in its application to other medical specialties.

## Author contributions

MDPB: Writing – review & editing, Writing – original draft, Conceptualization. RS: Writing – review & editing. AJ: Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. All authors were funded by the Medical Research Council and Health Data Research UK (Grant DATAMIND: Data Hub for Mental Health INformatics research Development, with Ref.: MR/W014386/1). RS is additionally part-funded by: i) the NIHR Maudsley Biomedical Research Centre at the South London and Maudsley NHS Foundation Trust and King's College London; ii) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust; iii) the UK Prevention Research Partnership (Violence, Health and Society; MR-VO49879/1), an initiative funded by UK Research and Innovation Councils, the Department of Health and Social Care (England) and the UK devolved administrations, and leading health research charities.

## Conflict of interest

AJ chairs the National Advisory Group on Suicide and Self-harm Prevention to Welsh Government. RS declares research funding/support from Janssen, GSK and Takeda in the last 3 years.

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

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. (2021) 3:199–217. doi: 10.1038/s42256-021-00307-0

2. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital Med*. (2022) 5:48. doi: 10.1038/s41746-022-00592-y

3. NewScientist. Revealed: Google AI has access to huge haul of NHS patient data (2016). Available online at: https://www.newscientist.com/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/.

4. BBC News. Project Nightingale: Google accesses trove of US patient data (2019). Available online at: https://www.bbc.co.uk/news/technology-50388464.

5. Drysdale E, Dolatabadi E, Chivers C, Liu V, Saria S, Sendak M, et al. (2019). Implementing AI in healthcare, in: *Vector-SickKids Health AI Deployment*

*Symposium*, Toronto. Canada: Vector Institute and the Hospital for Sick Children

6. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ digital Med*. (2020) 3:118. doi: 10.1038/s41746-020-00324-0

7. Shy CM. The failure of academic epidemiology: witness for the prosecution. *Am J Epidemiol*. (1997) 145:479–84. doi: 10.1093/oxfordjournals.aje.a009133

8. Lones MA. How to avoid machine learning pitfalls: a guide for academic researchers - v3. *arXiv preprint arXiv:2108.02497*. (2023). doi: 10.48550/arXiv.2108.02497

9. Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype? *Jama*. (2019) 321:2281–2. doi: 10.1001/jama.2019.4914

10. Hamilton AJ, Strauss AT, Martinez DA, Hinson JS, Levin S, Lin G, et al. Machine learning and artificial intelligence: Applications in healthcare epidemiology. *Antimicrobial Stewardship Healthcare Epidemiol*. (2021) 1:e28. doi: 10.1017/ash.2021.192

11. Morgan K, Page N, Brown R, Long S, Hewitt G, Del Pozo-Banos M, et al. Sources of potential bias when combining routine data linkage and a national survey of secondary school-aged children: a record linkage study. *BMC Med Res Method*. (2020) 20:1–13. doi: 10.1186/s12874-020-01064-1

12. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biol Psychiatry: Cogn Neurosci Neuroimaging*. (2021) 6:856–64. doi: 10.1016/j.bpsc.2021.02.001

13. John A, Lee SC, Solomon S, Crepaz-Keay D, McDaid S, Morton A, et al. Loneliness, coping, suicidal thoughts and self-harm during the COVID-19 pandemic: A repeat cross-sectional UK population survey. *BMJ Open*. (2021) 11:e048123. doi: 10.1136/bmjopen-2020-048123

14. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *psychol Bull*. (2017) 143:187. doi: 10.1037/bul0000084

15. Lee CH, Yoon HJ. Medical big data: Promise and challenges. *Kidney Res Clin Pract*. (2017) 36:3. doi: 10.23876/j.krcp.2017.36.1.3

16. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: A systematic review. *Br J Clin Pharmacol*. (2010) 69:4–14. doi: 10.1111/j.1365-2125.2009.03537.x

17. Kim YK, Park SC. Classification of psychiatric disorders. In: *Frontiers in Psychiatry: Artificial Intelligence, Precision Medicine, and Other Paradigm Shifts* USA: Springer. (2019). p. 17–25. doi: 10.1007/978-981-32-9721-0

18. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to mental health problems: Current status and future directions. *J consulting Clin Psychol*. (2020) 88:179. doi: 10.1037/ccp0000482

19. Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. In: *MEDINFO*. Amsterdam: IOS Press (2004). p. 736–40.

20. Liang Y, Zheng X, Zeng DD. A survey on big data-driven digital phenotyping of mental health. *Inf Fusion*. (2019) 52:290–307. doi: 10.1016/j.inffus.2019.04.001

21. Wang X, Zhao Y, Pourpanah F. Recent advances in deep learning. *Int J Mach Learn Cybernetics*. (2020) 11:747–50. doi: 10.1007/s13042-020-01096-5

22. Tornero-Costa R, Martinez-Millana A, Azzopardi-Muscat N, Lazeri L, Traver V, Novillo-Ortiz D. Methodological and quality flaws in the use of artificial intelligence in mental health research: Systematic review. *JMIR Ment Health*. (2023) 10:e42045. doi: 10.2196/42045

23. Riley RD, Ensor J, Snell KI, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *Bmj*. (2020) 368. doi: 10.1136/bmj.m441

24. Chen PHC, Liu Y, Peng L. How to develop machine learning models for healthcare. *Nat materials*. (2019) 18:410–4. doi: 10.1038/s41563-019-0345-0

25. Tougui I, Jilbab A, El Mhamdi J. Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthcare Inf Res*. (2021) 27:189–99. doi: 10.4258/hir.2021.27.3.189

26. Shokri R, Stronati M, Song C, Shmatikov V. (2017). Membership inference attacks against machine learning models, in: *2017 IEEE symposium on security and privacy (SP)*, USA: IEEE. pp. 3–18.

27. Rhodes AE, Fung K. Self-reported use of mental health services versus administrative records: care to recall? *Int J Methods Psychiatr Res*. (2004) 13:165–75. doi: 10.1002/mpr.172

28. Granero Pérez R, Ezpeleta L, Domenech JM. Features associated with the non-participation and drop out by socially-at-risk children and adolescents in mental-health epidemiological studies. *Soc Psychiatry Psychiatr Epidemiol*. (2007) 42:251–8. doi: 10.1007/s00127-006-0155-y

29. Rees S, Fry R, Davies J, John A, Condon L. Can routine data be used to estimate the mental health service use of children and young people living on Gypsy and Traveller sites in Wales? A feasibility study. *PLoS One*. (2023) 18:e0281504. doi: 10.1371/journal.pone.0281504

30. Arensman E, Corcoran P, McMahon E. The iceberg model of self-harm: new evidence and insights. *Lancet Psychiatry*. (2018) 5:100–1. doi: 10.1016/S2215-0366(17)30477-7

31. Jones LA, Nelder JR, Fryer JM, Alsop PH, Geary MR, Prince M, et al. Public opinion on sharing data from health services for clinical and research purposes without explicit consent: an anonymous online survey in the UK. *BMJ Open*. (2022) 12:e057579. doi: 10.1136/bmjopen-2021-057579

32. Ford T, Mansfield KL, Markham S, McManus S, John A, O'reilly D, et al. The challenges and opportunities of mental health data sharing in the UK. *Lancet Digital Health*. (2021) 3:e333–6. doi: 10.1016/S2589-7500(21)00078-9

33. John A, Lee SC, PuChades A, Del Pozo-Baños M, Morgan K, Page N, et al. Self-harm, in-person bullying and cyberbullying in secondary school-aged children: A data linkage study in Wales. *J Adolescence*. (2023) 95:97–114. doi: 10.1002/jad.12102

34. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. (2010) 29:337–46. doi: 10.1002/sim.3782

35. Banerjee S, Alsop P, Jones L, Cardinal RN. Patient and public involvement to build trust in artificial intelligence: A framework, tools, and case studies. *Patterns*. (2022) 3(6). doi: 10.1016/j.patter.2022.100506

36. Rouger M. AI improves value in radiology, but needs more clinical evidence. Healthcare IT News (2020). Available at: https://www.healthcareitnews.com/news/emea/ai-improves-value-radiology-needs-more-clinical-evidence.

37. United Kingdom Medicines and Healthcare products Regulatory Agency. Software and AI as a Medical Device Change Programme – Roadmap (2023). Available online at: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap.

38. United States Food and Drug Agency. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices (2024). Available online at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

39. Muehlematter UJ, Bluethgen C, Vokinger KN. FDA-cleared artificial intelligence and machine learning-based medical devices and their 510 (k) predicate networks. *Lancet Digital Health*. (2023) 5:e618–26. doi: 10.1016/S2589-7500(23)00126-7

40. Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M. From promise to practice: Towards the realisation of AI-informed mental health care. *Lancet Digital Health*. (2022) 4:e829–40. doi: 10.1016/S2589-7500(22)00153-4

41. Castelvecchi D. Can we open the black box of AI? *Nat News*. (2016) 538:20. doi: 10.1038/538020a

42. Al-Zaiti SS, Alghwiri AA, Hu X, Clermont G, Peace A, Macfarlane P, et al. A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML). *Eur Heart Journal-Digital Health*. (2022) 3:125–40. doi: 10.1093/ehjdh/ztac016

43. Riley P. Three pitfalls to avoid in machine learning. *Nature*. (2019) 572:27–9. doi: 10.1038/d41586-019-02307-y

44. DATAMIND collection of phenotypes in HDR-UK Phenotype Library. Available online at: https://phenotypes.healthdatagateway.org/phenotypes/?collections=27.

45. Catalogue of mental health measures. Available online at: https://www.cataloguementalhealth.ac.uk/.

46. Dipnall JF, Pasco JA, Berk M, Williams LJ, Dodd S, Jacka FN, et al. Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. *PLoS One*. (2016) 11:e0148195. doi: 10.1371/journal.pone.0148195

TYPE Original Research
PUBLISHED 19 March 2024
DOI 10.3389/fpsyt.2024.1282026

# Insights from the Twittersphere: a cross-sectional study of public perceptions, usage patterns, and geographical differences of tweets discussing cocaine

Consuelo Castillo-Toledo[1,2], Oscar Fraile-Martínez[2,3]*, Carolina Donat-Vargas[4,5], F. J. Lara-Abelenda[2,6], Miguel Angel Ortega[2,3], Cielo Garcia-Montero[2,3], Fernando Mora[1,7], Melchor Alvarez-Mon[2,3,8], Javier Quintero[1,7] and Miguel Angel Alvarez-Mon[1,2,3]

[1]Department of Psychiatry and Mental Health, Hospital Universitario Infanta Leonor, Madrid, Spain, [2]Department of Medicine and Medical Specialities, Faculty of Medicine and Health Sciences, University of Alcala, Alcala de Henares, Spain, [3]Ramón y Cajal Institute of Sanitary Research (IRYCIS), Madrid, Spain, [4]Cardiovascular and Nutritional Epidemiology, Institute of Environmental Medicine, Karolinska Institute, Stockholm, Sweden, [5]IMDEA-Food Institute, Universidad Autónoma de Madrid, Consejo Superior de Investigaciones Científicas, Madrid, Spain, [6]Departamento Teoria de la Señal y Comunicaciones y Sistemas Telemáticos y Computación, Escuela Tecnica Superior de Ingenieria de Telecomunicación, Universidad Rey Juan Carlos, Fuenlabrada, Spain, [7]Department of Legal Medicine and Psychiatry, Complutense University, Madrid, Spain, [8]Service of Internal Medicine and Immune System Diseases-Rheumatology, University Hospital Príncipe de Asturias, (CIBEREHD), Alcalá de Henares, Spain

Check for updates

OPEN ACCESS

EDITED BY
Robert Stewart,
King's College London, United Kingdom

REVIEWED BY
Raymond Spiteri,
University of Saskatchewan, Canada
Catherine Woodstock Striley,
University of Florida, United States

*CORRESPONDENCE
Oscar Fraile-Martínez
oscarfra.7@hotmail.com

RECEIVED 23 August 2023
ACCEPTED 27 February 2024
PUBLISHED 19 March 2024

CITATION
Castillo-Toledo C, Fraile-Martínez O, Donat-Vargas C, Lara-Abelenda FJ, Ortega MA, Garcia-Montero C, Mora F, Alvarez-Mon M, Quintero J and Alvarez-Mon MA (2024) Insights from the Twittersphere: a cross-sectional study of public perceptions, usage patterns, and geographical differences of tweets discussing cocaine. *Front. Psychiatry* 15:1282026. doi: 10.3389/fpsyt.2024.1282026

**Introduction:** Cocaine abuse represents a major public health concern. The social perception of cocaine has been changing over the decades, a phenomenon closely tied to its patterns of use and abuse. Twitter is a valuable tool to understand the status of drug use and abuse globally. However, no specific studies discussing cocaine have been conducted on this platform.

**Methods:** 111,508 English and Spanish tweets containing "cocaine" from 2018 to 2022 were analyzed. 550 were manually studied, and the largest subset underwent automated classification. Then, tweets related to cocaine were analyzed to examine their content, types of Twitter users, usage patterns, health effects, and personal experiences. Geolocation data was also considered to understand regional differences.

**Results:** A total of 71,844 classifiable tweets were obtained. Among these, 15.95% of users discussed the harm of cocaine consumption to health. Media outlets had the highest number of tweets (35.11%) and the most frequent theme was social/political denunciation (67.88%). Regarding the experience related to consumption, there are more tweets with a negative sentiment. The 9.03% of tweets explicitly mention frequent use of the drug. The continent with the highest number of tweets was America (55.44% of the total).

**Discussion:** The findings underscore the significance of cocaine as a current social and political issue, with a predominant focus on political and social denunciation in the majority of tweets. Notably, the study reveals a concentration of tweets from the United States and South American countries, reflecting the high prevalence of cocaine-related disorders and overdose cases in these regions. Alarmingly, the study highlights the trivialization of cocaine consumption on Twitter, accompanied by a misleading promotion of its health benefits, emphasizing the urgent need for targeted interventions and antidrug content on social media platforms. Finally, the unexpected advocacy for cocaine by healthcare professionals raises concerns about potential drug abuse within this demographic, warranting further investigation.

KEYWORDS

cocaine, Twitter, social perception, infodemiology, drug use/abuse, geolocalization

## 1 Introduction

Cocaine abuse represents a significant public health concern with relevant medical and socioeconomic consequences worldwide (1, 2). The United Nations Office on Drugs and Crime (UNODC) claims that cocaine is the third type of illicit drug most consumed in the world, just after opiates and cannabis (3). According to their last World Drug Report, approximately 21.5 million people are estimated to have used cocaine in 2020, representing 0.4% of the global population aged between 15 to 64 years old (3). Moreover, the escalating annual trend in cocaine consumption since 2010 underscores the increasing level of concern associated with its use.

Understanding the public´s perception of a drug is essential, as both factors are directly related to its consumption and legislation (4, 5). For instance, previous studies have linked increased cannabis consumption to a perception of low associated risks, influenced partly by varying legislation on medical cannabis use and exposure to related advertising (6–9). Cocaine was first isolated in the middle of the 19th century and gained popularity in the early 1900s (10). However, due to its addictive properties, widespread abuse and related health issues it was banned in the United States in 1914 (11). The public perception of cocaine underwent shifts, notably in the 1970s leading to increased abuse (12). Subsequently, in the 1980s and early 1990s, it became linked to crime, violence, and racial concerns, influencing public policies on its regulation (10). Therefore, analyzing the public perception of drugs, especially cocaine, is crucial for comprehending its current global use/abuse status and the impact of related public policies.

An increasing body of research advocates for the use of social networks as a valuable tool in drug research. They facilitate the understanding and collection of data on social perception, misinformation, and pharmacovigilance (13–15). Twitter is seen as a safe and non-judgmental platform for sharing honest experiences, including sensitive topics like drug use and abuse (16). Previous studies have successfully utilized Twitter as a public health tool to analyze and study drug-related issues (17–19). Artificial intelligence (AI), enables the processing and analysis of vast amounts of data (20). Within AI, Machine Learning (ML) has become a prominent field, focusing on extracting knowledge from data through computational models. A subset of ML known as Deep Learning (DL) employs neural networks inspired by the human brain to process information (21). These neural networks find applications in various domains related to substance use, enabling detection of abuse patterns (22) and related harms (23), also allowing researchers to understand public perceptions and opinions of a drug (5) while exploring potential differences in these points across regions and countries (24). Another essential application is Natural Language Processing (NLP), which extensively utilizes neural networks to analyze text, facilitate conversations, and extract key ideas (25). Most studies conducted on Twitter have focused on cannabis and opioids (5, 18, 26, 27). Currently, some preliminary results related to cocaine use have been obtained from different social media by the use of AI and ML (28, 29) and previous works in Twitter analysis have considered cocaine use in the context of polysubstance use (30, 31). Nevertheless, there is a notable gap in the literature concerning detailed studies collecting information on the use/abuse of cocaine on Twitter through these techniques.

Given the existing gap in detailed studies on cocaine discussions on Twitter, we propose the following hypotheses: First, we hypothesize that through the use of AI and ML, it is possible to find geographical differences in the opinions and concerns expressed about cocaine that reflect unique regional dynamics and social attitudes. Second, we hypothesize that there are distinct considerations related to cocaine

**Abbreviations:** AI, Artificial intelligence; CUD, Cocaine use disorder; CDC, Centers for Disease Control and prevention; ML, Machine Learning; UNODC, United Nations Office on Drugs and Crime.

based on user profile. Specifically, we anticipate differing opinions among different user groups, such as general or non-identifiable individuals, healthcare professionals, the media, or celebrities language, thereby contributing to a nuanced understanding of the diversity of discourse Finally, we hypothesized that individuals' personal experiences with cocaine would correlate with their assessment of the risks involved when discussing the substance on Twitter, and that the platform would also collect different frequencies and consumption patterns. This correlation will influence the nature and tone of their contribution to the platform. By addressing these multifaceted aspects, this study aims to provide valuable insights into the complex dynamics of public discourse on cocaine in the digital sphere, providing a comprehensive understanding about the factors that form and differentiate views on this quality.

# 2 Methods

## 2.1 Data collection

This mixed-method, quantitative and qualitative analysis focused on the content of tweets related to cocaine posted on the social media platform Twitter. Our study included tweets that met specific criteria: they had to be public, contain the words "cocaine" or "cocaina," be published between January 1, 2018, and April 30, 2022, and be in English or Spanish, with a minimum of 10 retweets. These criteria were chosen to ensure a comprehensive and representative sample of social media discussions on the topic. We employed Tweet Binder, a widely used tool in previous research (32–35), to collect the tweets, providing essential information such as retweet and like counts, publication date, tweet context link, user description, and geolocation. The number of retweets and likes served as indicators of user engagement and interest in the tweeted content (36, 37).

## 2.2 Content analysis process

Using the previously mentioned search criteria, we collected 57,192 tweets in Spanish and 54,316 tweets in English. Next, with the remaining tweets, the content was analyzed using a mixed inductive-deductive approach to develop a codebook for classifying the tweets into key thematic categories. A manual classification of a small subset of tweets (n = 100) was conducted by two members of the research team, who later convened to discuss the different categories analyzed. We created a codebook based on our research questions, our previous experience in analyzing tweets, and what we determined to be the most common themes. After discussing discrepancies and reaching a consensus on the codebook, an additional 450 tweets were analyzed. This process also provided a larger sample for training the Machine Learning model. Finally, an automated and computerized classification was performed on the remaining and larger subset of tweets (n = 111,508).

The tweets were classified as classifiable or non-classifiable. A tweet was considered non-classifiable if it was written in a way that made its meaning uncertain, too brief to contain relevant information, if its content was purely political, if the information was not relevant to the objectives of this study, or if it was a joke. In each of the classifiable tweets, the content was analyzed according to the following themes: 1) Tweet topic; 2) Evaluation of the effect; 3) Sentiment regarding consumption; 4) Type of consumption. Finally, the users were classified into four categories: 1) General Twitter users; 2) Media outlets; 3) Public figures; and 4) Healthcare professionals. The classification criteria and examples of tweets are shown in (Table 1).

## 2.3 Machine-learning classifier

The methodology followed in this project has been validated in prior research studies (38, 39). First, a preprocessing of the database should be executed. This preprocessing involves a translation of the non-English tweets to English using Google Translator and a normalization of the tweets by removing special characters, splitting negative contractions, and removing repetitions. Then, we employ a pre-trained network called BERTWEET, trained on 850 million English tweets (40), to classify cocaine-related tweets. Since BERTWEET was not initially designed for the specific classification categories, fine-tuning was performed. Manually classified tweets were randomly divided into an 80% training subset and a 20% testing subset. The training subset was used to fine-tune the network, while the testing subset was used to validate its performance. Additionally, to address some imbalanced categories (where certain options had a higher number of tweets compared to others), text augmentation was performed using the library called textattack (41). Furthermore, emotion analysis was conducted using a pretrained neural network called emotion-english-distilroberta-base (42). This network is capable of detecting six basic emotions according to Ekman's theory (43) along with neutral sentiment. The emotion analysis was applied to the 71,884 tweets categorized as classifiable.

## 2.4 Statistical analysis

The results were presented in tables or figures, showing the percentage of tweets or the median of likes and retweets in each category. To compare the proportions of tweets between categories, Pearson's chi-square test was utilized, yielding a p-value indicating statistical significance.

To evaluate the relationships between tweet content, user type, and other tweet characteristics with the number of likes and retweets, linear regression models were employed. The individual beta coefficients were adjusted for the remaining tweet characteristics. Choropleth maps were generated as a visualization tool to depict the global distribution of tweets. Additionally, these maps were used to illustrate the geographic distribution of tweets expressing support for the legislation and exhibiting a sentiment favorable to cocaine.

The statistical analyses were performed using the software packages STATA v16 (StataCorp) and MS Excel.

TABLE 1 Category, definitions and examples of classification.

| Category | Examples |
|---|---|
| **Effect assessment**<br>***(Whether consumption is perceived as beneficial or a health risk.)***<br>**1. Health benefit**<br>**2. Harmful to health** | 1. I'm just going to say that cocaine use is destroying a friend and we can't get him out of there. Stop fucking around. Legal or illegal kills the same. |
| **Topic**<br>***1.* Claim (Refers to both police/social/ political complaint/ claim (for or against))**<br>**2. General information *(Refers to when talking about more scientific issues).***<br>**3. Sale/advertising *(Tobacco is advertised).***<br>***4.* Testimonials *(Regarding consumption, experience, more from the opinion of drug users or families/ friends).***<br>**5. Trivialization. *(Minimization of the consequences of consumption, stigmatization, humorous tweets)*** | 6. The Departmental Anti-drug Brigade arrested a Colombian citizen who was making pink cocaine in an apartment located on Paysandú and Ejido streets in downtown Montevideo.<br>7. Finally published the analysis I did of 19,000 admissions to mental health hospitalization. There are more and more problems related to cannabis, cocaine and other stimulants and we still do not have a care plan for dual pathology in Andalusia. I want one! |
| **Personal experience with drugs.**<br>***(Personal experience with cocaine, whether through acquaintances, friends, family members, or personal use, or related to social events associated with its consumption.)*** | 1. Impossible to talk to people without culture, mostly high school and called "truckers" enough of "filthy broken" in good Chilean Urgent Railroad PLAN to regulate this plague that HURT the country. BEWARE MANY OF THESE GUYS DRIVE DRUGGED ADDICTED TO COCAINE!!!! |
| **Consumption type.**<br>***(Whether it's about using cocaine frequently, only occasionally, or in binges, not only personal use but also when discussing the consumption of family members or friends.)*** | 2. Sigrid Alegría confessed in "De tú a tú" about her addiction: "I used cocaine to avoid gaining weight." |
| **User type**<br>***(Refers to the person sharing the tweet.)***<br>**3. Health professionals.**<br>**4. Undetermined. *(General population or it is not possible to identify)***<br>**5. Media.**<br>**6. Celebrity. (*Any famous person; singers, actors, politicians, influencers…*).** | 7. Cocaine has vasoconstrictive properties which along with other secondary effects lead to ischemia and subsequent perforation of the hard palate (the roof of the mouth).<br>8. She found that "media reports on crack cocaine frequently referenced African Americans and depicted the drug in conjunction with violent crime. However articles on methamphetamine were more likely to reference poor Whites and associate this drug as a public health problem."<br>9. In a single enforcement action #CBP officers at Laredo Port of Entry seize a poly-drug load of black tar heroin brown heroin and cocaine valued at $400K.<br>10. Cocaine is now legal in Oregon but now straws are illegal. Damn that must be mighty frustrating. |

Usernames and personal names were removed.

# 3 Results

## 3.1 Content themes

The study involved analyzing the frequency distribution of tweets across various categories based on tweet characteristics. According to the codebook, a total of 71,844 classifiable tweets were obtained. Among these, 15.95% of users discussed the harm of cocaine consumption to health. Although tweets expressing some health benefits of cocaine receive a higher number of likes, 50% of the tweets have 121.5 likes or more (Table 2). Of the total number of users that could be defined, media outlets had the highest number of tweets, with 25,228 tweets (35.11%). The most frequent theme is social or political claims, with 48,768 tweets published, accounting for 67.88% of the total. The least frequent theme is trivialization, but it has a higher number of likes and retweets. Regarding the experience related to consumption, there are more tweets with a negative sentiment compared to a positive sentiment. Approximately 37.07% of the tweets (26,597) display a negative sentiment. Regarding the discourse on cocaine consumption, 9.03% of tweets explicitly mention frequent use of the drug, and they also receive a higher number of likes compared to other subcategories.

In terms of emotional expression, the most frequent response from Twitter users is to remain neutral in the majority of their posts, as depicted in Figure 1.

The continent with the highest number of tweets is America, with 39,830 tweets published, accounting for 55.44% of the total. Among the top 5 countries with the highest number of tweets, the first four are from this continent, in descending order: United States, Colombia, Venezuela, and Argentina, representing 41.82% of the total tweets (Figure 2).

## 3.2 Geographical analysis

Content analysis by continents reveals that out of the 59,725 geolocated tweets analyzed as shown in (Table 3), the most frequent theme across all continents, similar to the overall analysis, is the expression of social/political denunciation, particularly prevalent in America, accounting for 73.05% of the tweets. Regarding the evaluation of the effects, Europe has the highest percentage of tweets discussing the harm caused by cocaine, at 21.63%. Additionally, Asia has the highest proportion of tweets expressing negative sentiment related to consumption, with 41.65% of the tweets falling into this category. Lastly, Africa exhibits the highest content about frequent cocaine use, comprising 13.18% of the tweets.

## 3.3 User type

If we examine each item concerning types of Twitter users (Table 4), it is observed that, about the assessment of cocaine's effects, the majority of users, excluding healthcare professionals,

TABLE 2 Descriptive characteristics of the tweets are considered classifiable in the content analysis.

| | Tweets | | Median likes | Median retweet |
|---|---|---|---|---|
| | n | % | - | - |
| **Overall** | 71,844 | 100 | – | – |
| *Effect assessment* | | | | |
| **No mention** | 59,282 | 82.51 | 65 | 34 |
| **Health benefit** | 1,102 | 1.53 | 121.5 | 29 |
| **Harmful for health** | 11,460 | 15.95 | 88 | 32 |
| *User type* | | | | |
| **Health professionals** | 2,030 | 2,83 | 87 | 28 |
| **Undetermined** | 37,381 | 52.03 | 83 | 37 |
| **Media** | 25,228 | 35.11 | 50 | 30 |
| **Celebrity** | 7,205 | 10.03 | 85 | 36 |
| *Topic* | | | | |
| **Claim** | 48,768 | 67.88 | 64 | 35 |
| **General information** | 2,230 | 3.10 | 60 | 28 |
| **Sale/advertising** | 6,441 | 8.97 | 56 | 36 |
| **Testimonials** | 13,316 | 18.53 | 98 | 30 |
| **Trivialization** | 1,089 | 1.52 | 172 | 37 |
| *Personal experience with drugs* | | | | |
| **No mention** | 43,257 | 60.21 | 60 | 33 |
| **Positive** | 1,990 | 2.77 | 174.5 | 32 |
| **Negative** | 26,597 | 37.02 | 83 | 36 |
| *Consumption type* | | | | |
| **No mention** | 63,488 | 88.31 | 66 | 34 |
| **Frequent consumption** | 6,491 | 9.03 | 101 | 29 |
| **Occasional/binge consumption** | 1,905 | 2.62 | 73 | 27 |

refrain from mentioning it in their tweet content. Nonetheless, healthcare professionals indicate it as a detriment to health in 82.07% of instances. Additionally, healthcare professionals exhibit the highest percentage (77.04%) of expressing negative experiences related to consumption, followed by public figures, where this aspect appears in 66.74% of the tweets. Finally, regarding the type of consumption, a notably high percentage (81.72%) of healthcare professionals share their perspective on frequent cocaine use.

## 3.4 The assessment of cocaine's effects and individual experience, related to the type of consumption

If we relate the evaluation of the effect by Twitter users with those who talk about consumption, it has been observed that 64.34% of the tweets that mention health benefits also mention frequent consumption, nearly double the percentage of those who mention harm to health and frequent consumption, which is 39.73% (Table 4).

Regarding individual experiences with the substance, it has been found that almost half (48.14%) of the tweets that speak positively also mention frequent consumption. However, only 17.01% of those who mention harm relate it to frequent consumption (Table 5).

# 4 Discussion

In the present work, we have collected and classified 71,844 tweets discussing cocaine according to the content of the message, geolocation, type of user, and consumption frequency reported. The results obtained in this article go hand in hand with previous results reported in the Twittersphere in which this type of detail has been studied in other drugs such as opioids or cannabis (30, 44, 45); however, as far as we know this article is the first to deeply explore this type of data about cocaine on this platform.

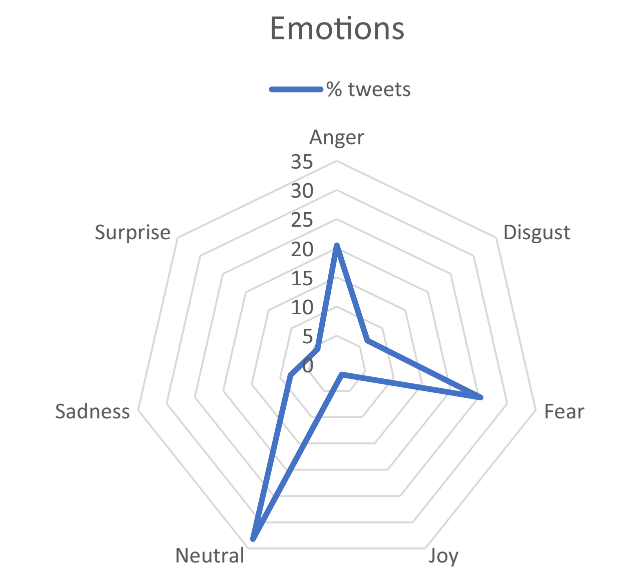


FIGURE 1
Sentiment analysis (the emotional tone expressed in text). the percentage of tweets associated with each specific emotion.

The majority of analyzed tweets (67.88%) focused on political and social denunciation. Media sources accounted for 35.11% of the tweets, with 55.44% originating from American users who predominantly expressed political and social denunciation (73.05%). These findings highlight evidence cocaine consumption is a significant current social and political issue, particularly in the United States and South American countries. The United States has experienced the highest number of cocaine-related disorders and overdose mortality cases globally (46–48). Recent data from the Centers for Disease Control and prevention (CDC) showed a 54% increase in cocaine-involved deaths, rising from 15,883 in 2019 to 24,486 deaths in 2021 (47). Given these statistics, it is understandable that many tweets from the United States focus on denouncing cocaine abuse from a political and social perspective, emphasizing the need for inclusive public policy reforms (49). In the case of South American countries, a broad number of tweets were identified from Colombia, Venezuela, and Argentina. Colombia in particular has a long history of cocaine trade and continues to be involved in its production and cultivation (10). Twitter and scientific articles discuss the complex sociopolitical context of cocaine crops in this country, analyzing the problem comprehensively (50, 51).

Tweets from Europe and Africa primarily focused on the detrimental health effects of cocaine and the frequent consumption of this drug. In the European Union, 14.4 million people have consumed cocaine at least once in their lives, accounting for 5% of the population (52). Among adults aged 15 to 64, 3.5 million reported cocaine use in the last year, with 2.2 million between the ages of 15 and 34. Cocaine ranked as the second most problematic drug for first-time treatment seekers and the second most commonly reported substance for acute toxicity by Euro-DEN Plus hospitals in 2020 (52). In the same manner, various studies conducted in different European countries have found an increase in cocaine consumption and cocaine-related deaths, also highlighting the multiple health complications related such as psychiatric and psychotic disorders, neurological maladies and cardiovascular diseases (53–55). Thus, our results seem to support

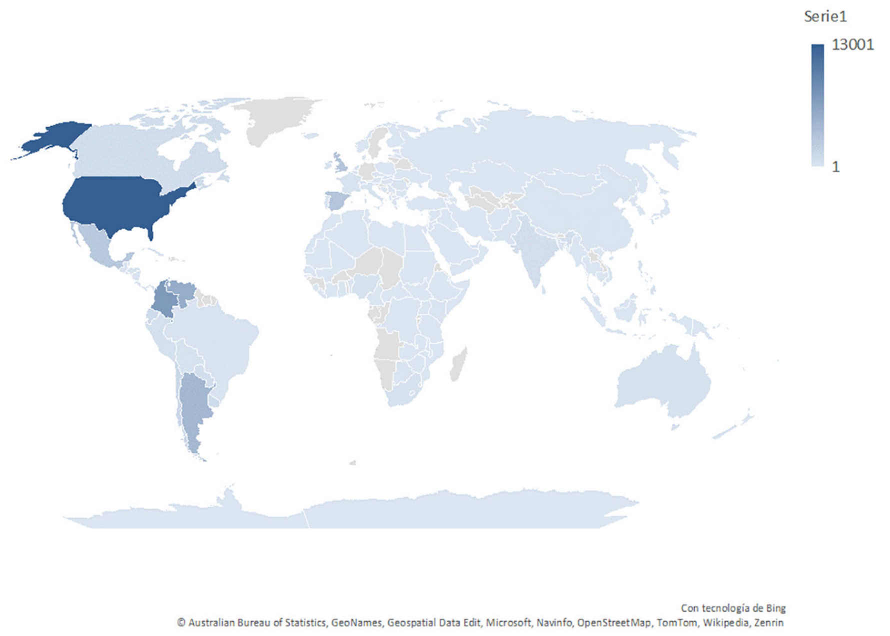

Serie1
13001
1
Con tecnología de Bing
© Australian Bureau of Statistics, GeoNames, Geospatial Data Edit, Microsoft, Navinfo, OpenStreetMap, TomTom, Wikipedia, Zenrin

FIGURE 2
Distribution of the number of tweets worldwide. The area with the highest number of tweets about cocaine is represented with a darker blue color, and the color tone decreases as the number of tweets decreases.

TABLE 3 Number of tweets by continent and category of the codebook.

| | AMERICA | EUROPE | AFRICA | ASIA | OCEANIA |
|---|---|---|---|---|---|
| | n (%) | n (%) | n (%) | n (%) | n (%) |
| *Effect assessment* | | | | | |
| **No mention** | 34,233 (85.95) | 5,602 (76.51) | 986 (77.33) | 1,256 (80.10) | 630 (86.30) |
| **Health benefit** | 437 (1.10) | 136 (1.86) | 29 (2.27) | 41 (2.61) | 12 (1.64) |
| **Harmful for health** | 5,160 (12.96) | 1,584 (21.63) | 260 (20.39) | 271 (17.28) | 88 (12.05) |
| | P<0.001 | | | | |
| *User type* | | | | | |
| **Health professionals** | 804 (2.02) | 308 (4.21) | 44 (3.45) | 50(3.19) | 24 (3.28) |
| **Undetermined** | 29,984(52.68) | 3,802 (42.09) | 541 (42.43) | 556 (35.46) | 311 (42.60) |
| **Media** | 14,711(36.93) | 3,139 (42.87) | 564 (44.23) | 616 (39.29) | 305 (41.78) |
| **Celebrity** | 3,311(8.36) | 793 (10.83) | 126 (9.88) | 346 (22.07) | 90 (12.32) |
| | P<0.001 | | | | |
| *Topic* | | | | | |
| **Claim** | 29,097 (73.05) | 4,323 (59.04) | 809 (63.45) | 1,003 (63.97) | 474 (64.93) |
| **General information** | 945 (2.37) | 346 (4.73) | 31 (2.43) | 40 (2.55) | 34 (4.66) |
| **Sale/advertising** | 3,653 (9.17) | 483 (6.60) | 68 (5.33) | 109(6.95) | 49 (6.71) |
| **Testimonials** | 5,715 (14.35) | 2,060 (28.13) | 326 (25.57) | 376(23.98) | 163 (22.33) |
| **Trivialization** | 420 (1.05) | 110(1.5) | 41 (3.22) | 40 (2.55) | 10 (1.37) |
| | P<0.001 | | | | |
| *Personal experience with drugs* | | | | | |
| **No mention** | 26,561 (66.69) | 4,093 (55.90) | 743 (58.27) | 854 (54.46) | 404 (55.34) |
| **Positive** | 813 (2.04) | 204 (2.79) | 70 (5.49) | 61 (3.89) | 14 (1.94) |
| **Negative** | 12,456 (31.27) | 3,025 (41.31) | 462 (36.24) | 653 (41.65) | 312 (42.74) |
| | P<0.001 | | | | |
| *Consumption type* | | | | | |
| **No mention** | 36,139 (90.73) | 5,985 (81.74) | 1,061 (83.22) | 1,397 (89.09) | 641 (87.81) |
| **Frequent consumption** | 2,841 (7.13) | 929 (12.69) | 168 (13.18) | 150 (9.57) | 72 (9.86) |
| **Occasional/binge consumption** | 850 (2.13) | 408 (5.57) | 46 (3.61) | 21 (1.34) | 17 (2.33) |
| | P<0.001 | | | | |

that Twitter is seen as a valuable tool to raise awareness about the real problem of cocaine in Europe and its overall negative effects on health. On the other hand, fewer studies are available in the literature regarding cocaine use in Africa. However, different platforms like the Africa Organized Crime Index (56) have evidenced the problem of cocaine trade and abuse in some countries like Guinea-Bissau, Cabo Verde or Guinea, as well as in South Africa or the sub-Saharan countries (57, 58). According to the literature, despite Africa being neither a major producer nor a major consumer of cocaine, the evidence of cocaine's destabilizing impact has been considered an emerging problem for the last decade (59). The sheer value of the cocaine trade in this region from South American and Caribbean countries poses not only security threats, but also risks distorting the region's economy, investment flows, development and democracy. Therefore, Twitter can be used as a platform to denounce the habitual consumption of cocaine in this region and the detrimental health effects derived in this region. However, additional efforts in this platform are warranted, particularly in light of our results.

Despite the trivialization of cocaine consumption being the less discussed topic on Twitter, it accumulated almost double the interactions with other Twitter users (172 likes and 37 retweets versus 64 likes and 35 retweets), as well as those reporting positive versus negative effects. In addition, when considering the type of cocaine consumption on Twitter, frequent consumption was more common than occasional use (9.03% versus 2.62%), also receiving

TABLE 4 Number of tweets by user type and category of the codebook.

| | User Type | | | |
|---|---|---|---|---|
| | Health Professional | Undetermined | Media | Celebrity |
| | n (%) | n (%) | n (%) | n (%) |
| *Effect assessment* | | | | |
| **No mention** | 115 (5.67) | 30,418 (81.37) | 22,777 (90.28) | 5,972 (82.89) |
| **Health benefit** | 249 (12.27) | 674 (1.80) | 97 (0.38) | 82 (1.14) |
| **Harmful for health** | 1,666 (82.07) | 6,289 (16.82) | 2,354 (9.33) | 1,151 (15.98) |
| | P<0.001 | | | |
| *Personal experience with drugs* | | | | |
| **No mention** | 242 (11.92) | 19,769 (52.89) | 21,290 (84.39) | 1,956 (27.15) |
| **Positive** | 224 (11.03) | 1,414 (3.78) | 56 (0.22) | 296 (4.11) |
| **Negative** | 1,564 (77.04) | 16,198 (43.33) | 3,882 (15.39) | 4,953 (68.74) |
| | P=<.001 | | | |
| *Consumption type* | | | | |
| **No mention** | 345 (17.00) | 33,126 (88.62) | 23,030 (91.29) | 6,947 (96.42) |
| **Frequent consumption** | 1,659 (81.72) | 3,440 (9.20) | 1,150 (4.56) | 242 (3.36) |
| **Occasional/binge consumption** | 26 (1.28) | 815 (2.18) | 1,048 (4.15) | 16 (0.22) |
| | P<0.001 | | | |

more interactions. Previous research has indicated that drugs are often discussed positively on social media platforms like Twitter, and the lack of antidrug content may contribute to the normalization and justification of drug use, highlighting the importance of addressing this issue (60). Furthermore, the dissemination of trivialization may contribute to an increase in hospitalizations due to cocaine consumption, even in the pediatric population (61). In agreement with previous works (62, 63), our results support the notion that social media like Twitter can serve as valuable resources for understanding drug patterns, prevailing attitudes, monitoring and intervening in drug abuse and addiction problems.

TABLE 5 Number of tweets by consumption type and category of the codebook.

| | Consumption type | | |
|---|---|---|---|
| | No mention | Frequent consumption | Occasional/ binge consumption |
| | n (%) | n (%) | n (%) |
| *Effect assessment* | | | |
| **No mention** | 57,427 (96.87) | 1,229 (2.07) | 626 (1.06) |
| **Health benefit** | 392 (35.57) | 709 (64.34) | 1 (0.09) |
| **Harmful for health** | 5,626 (49.12) | 4,553 (39.73) | 1,278 (11.15) |
| | P<0.001 | | |
| *Personal experience with drugs* | | | |
| **No mention** | 41,662 (96.31) | 1,010 (2.33) | 585 (1.35) |
| **Positive** | 1,025 (51.51) | 958 (48.14) | 7 (0.35) |
| **Negative** | 20,761 (78.05) | 4,526 (17.01) | 1,313 (4.94) |
| | P<0.001 | | |

We found a small proportion of tweets promoting the supposed health benefits of cocaine use, which received significant engagement. This is an important issue to address, as there are no safe ways to consume cocaine. Misconceptions regarding the health benefits of cocaine may stem from historical events and practices, such as its traditional use in South America for over 5,000 years as a stimulant in the form of teas or by chewing the leaves of the *Erythroxylon coca* plant (64). Additionally, influential figures like Sigmund Freud, as well as the incorporation of cocaine in beverages like Coca-Cola and coca wine during the late 19th and early 20th centuries, contributed to its popularity (11). As previously mentioned, despite being banned in the USA in 1914, during the 1970s, cocaine regained a positive image, fueled by perceptions of glamour and media influence. Even the Ford White House in 1975 released a white paper stating that cocaine was not physically addictive and generally did not have serious consequences (12). Conversely, cocaine use leads to a wide range of harmful effects including tachycardia, hypertension, acute coronary syndrome, stroke, and even death (65). Mixing cocaine with substances like sugar, talc, and cornstarch exacerbates these adverse effects (66). Factors such as high drug purity, frequent or binge consumption and polydrug use (particularly with alcohol and fentanyl/heroin) contribute to toxicity and overdose risks (67–69). Previous Twitter analyses have shown that polysubstance use involving cocaine and other drugs is a common topic in discussions about overdose and drug-related concerns (18, 30, 31, 70). Although our study did not focus on polydrug use, it is important to consider these findings, as the low perception of risks associated with cocaine use obtained in our study may even be more concerning in such contexts. Furthermore, long-term consumption of cocaine is associated with significant brain changes in the dopaminergic reward system, resulting in addiction, persistent cravings and a high risk of relapse, even with treatment (71). Cocaine use disorder (CUD) represents a serious global health concern, and while psychosocial and pharmacological interventions can assist in the medical management of this condition, the efficacy is limited and ineffective for most patients (72). Moreover, despite some specific clinical cases in the 20th century, the risks of cocaine use outweigh any potential benefits, and there are safer alternatives for various purposes attributed to this substance (10).

Therefore, it is crucial to address and intervene in the content on Twitter that trivializes or supports the alleged health benefits of cocaine use.

Intriguingly, our study shows that healthcare professionals on Twitter were among the strongest advocates for the health benefits, frequent use and positive experiences related to cocaine (12.27%, 81.72% and 11.03%, respectively). This could be relevant considering previous studies that have identified drug abuse among healthcare professionals as a concern (73), especially when considering certain risk factors such as certain medical specialties, psychopathological or social factors, positive attitudes toward drugs, unhealthy lifestyle habits and so on (73). Although we could not explore all contributing factors, further investigation is needed to understand the relationship between drug abuse and healthcare professionals on social media platforms like Twitter, as our findings imply that they may use it to share personal experiences and concerns related to drug use and abuse.

Finally, we also observed a notable proportion of tweets (8.97%) showing sale/advertising content. This is not a novel issue as previous works have also identified social media like Twitter as a conduit for the sale and supply of illicit drugs like opioids (74, 75). We encourage the regulation of this type of illegal cocaine sale, proposing the inclusion and use of possible programs implicated in the detection, classification and reporting of illicit online sale tweets, as promoted in previous works (76).

## 5 Limitations

This research has some notable limitations. Firstly, Twitter users' social, economic, and demographic attributes do not accurately mirror the entire society. Second, just like practically all qualitative investigations, the construction of the codebook and the analysis of the tweets involve certain subjectivity. Third, there is a chance that we overlooked tweets that made reference to cocaine but did so in slang or contractions like "coke", "C", "snow", "flake" and "blow". Similarly, it is also possible that bots or fake accounts have to some extent affected our data. Finally, the inclusion of tweets with 10 or more retweets could also be a limitation of the study, as it might have overlooked relevant tweets for this article.

## Data availability statement

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

## Ethics statement

This study was approved by the Research Ethics Committee of Universidad de Alcalá and is compliant with the ethical principles from the World Medical Association Declaration of Helsinki (7th revision, 2013).

## Author contributions

CCT: Data curation, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. OF-M: Conceptualization, Validation, Writing – original draft, Writing – review & editing. CD-V: Formal analysis, Investigation, Methodology, Writing – review & editing. FL-A: Formal analysis, Investigation, Software, Writing – review & editing. MO: Conceptualization, Supervision, Validation, Writing – original draft, Writing – review & editing. CG-M: Supervision, Validation, Writing – original draft, Writing – review & editing. FM: Conceptualization, Resources, Visualization, Writing – review & editing. MA-M: Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. JQ: Methodology, Supervision, Visualization, Writing – review & editing. MAA-M: Conceptualization, Investigation, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study (FIS-PI22/00653) was supported by the Instituto de Salud Carlos III (grant no. Estatal de I + D + I 2020–2027).

## 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.

## References

1. Castaldelli-Maia JM, Wang YP, Brunoni AR, Faro A, Guimarães RA, Lucchetti G, et al. Burden of disease due to amphetamines, cannabis, cocaine, and opioid use disorders in South America, 1990-2019: a systematic analysis of the Global Burden of Disease Study 2019. *Lancet Psychiatry*. (2023) 10:85–97. doi: 10.1016/S2215-0366(22)00339-X

2. Crepalde RDS, Bonadiman CSC, Malta DC, Naghavi M, Melo APS. The burden of mental disorders attributable by cocaine use: Global Burden of Diseases Study in Brazil, 1990 and 2019. *Rev Soc Bras Med Trop*. (2022) 55:320–2021. doi: 10.1590/0037-8682-0320-2021

3. United Nations: Office on Drugs and Crime. World drug report 2021. Nueva York, NY, Estados Unidos de América: United Nations; 2022.

4. Alhyas L, Al Ozaibi N, Elarabi H, El-Kashef A, Wanigaratne S, Almarzouqi A, et al. Adolescents' perception of substance use and factors influencing its use: a qualitative study in Abu Dhabi. *JRSM Open*. (2015) 6:205427041456716. doi: 10.1177/2054270414567167

5. Carabot F, Fraile-Martínez O, Donat-Vargas C, Santoma J, Garcia-Montero C, da Costa MP, et al. Understanding public perceptions and discussions on opioids through twitter: cross-sectional infodemiology study. *J Med Internet Res*. (2023) 25. doi: 10.2196/50013

6. De Luca MA, Di Chiara G, Cadoni C, Lecca D, Orsolini L, Papanti D, et al. Cannabis; epidemiological, neurobiological and psychopathological issues: an update. *CNS Neurol Disord Drug Targets*. (2017) 16. doi: 10.2174/1871527316666170413113246

7. Harper S, Strumpf EC, Kaufman JS. Do medical marijuana laws increase marijuana use? Replication study and extension. *Ann Epidemiol*. (2012) 22:207–12. doi: 10.1016/J.ANNEPIDEM.2011.12.002

8. Cerdá M, Wall M, Keyes KM, Galea S, Hasin D. Medical marijuana laws in 50 states: investigating the relationship between state legalization of medical marijuana and marijuana use, abuse and dependence. *Drug Alcohol Depend*. (2012) 120:22–7. doi: 10.1016/J.DRUGALCDEP.2011.06.011

9. D'Amico EJ, Miles JNV, Tucker JS. Gateway to curiosity: medical marijuana ads and intention and use during middle school. *Psychol Addict Behav*. (2015) 29:613. doi: 10.1037/ADB0000094

10. Drake LR, Scott PJH. DARK classics in chemical neuroscience: cocaine. *ACS Chem Neurosci*. (2018) 9:2358. doi: 10.1021/ACSCHEMNEURO.8B00117

11. Das G. Cocaine abuse in North America: a milestone in history. *J Clin Pharmacol*. (1993) 33:296–310. doi: 10.1002/j.1552-4604.1993.tb04661.x

12. Miech R. The formation of a socioeconomic health disparity: the case of cocaine use during the 1980s and 1990s. *J Health Soc Behav*. (2008) 49:352. doi: 10.1177/002214650804900308

13. Crosier BS, Marsch LA. Harnessing social media for substance use research and treatment. *J Alcohol Drug Depend*. (2016) 4. doi: 10.4172/2329-6488.1000238

14. van Stekelenborg J, Ellenius J, Maskell S, Bergvall T, Caster O, Dasgupta N, et al. Recommendations for the use of social media in pharmacovigilance: lessons from IMI WEB-RADR. *Drug Saf*. (2019) 42:1393–407. doi: 10.1007/S40264-019-00858-7

15. Al Khaja KAJ, AlKhaja AK, Sequeira RP. Drug information, misinformation, and disinformation on social media: a content analysis study. *J Public Health Policy*. (2018) 39:343–57. doi: 10.1057/S41271-018-0131-2

16. Berry N, Lobban F, Belousov M, Emsley R, Nenadic G, Bucci S. #WhyWeTweetMH: understanding why people use twitter to discuss mental health problems. *J Med Internet Res*. (2017) 19. doi: 10.2196/JMIR.6173

17. Meng HW, Kath S, Li D, Nguyen QC. National substance use patterns on Twitter. *PloS One*. (2017) 12. doi: 10.1371/JOURNAL.PONE.0187691

18. Tofighi B, Aphinyanaphongs Y, Marini C, Ghassemlou S, Nayebvali P, Metzger I, et al. Detecting illicit opioid content on Twitter. *Drug Alcohol Rev*. (2020) 39:205–8. doi: 10.1111/DAR.13048

19. Tofighi B, El Shahawy O, Segoshi A, Moreno KP, Badiei B, Sarker A, et al. Assessing perceptions about medications for opioid use disorder and Naloxone on Twitter. *J Addict Dis*. (2021) 39:37. doi: 10.1080/10550887.2020.1811456

20. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. *Curr Rev Musculoskelet Med*. (2020) 13:69–76. doi: 10.1007/S12178-020-09600-8

21. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature*. (2015) 52:436–44. doi: 10.1038/nature14539

22. Hu H, Phan NH, Geller J, Iezzi S, Vo H, Dou D, et al. An ensemble deep learning model for drug abuse detection in sparse twitter-sphere. *Stud Health Technol Inform*. (2019) 264:163–7. doi: 10.3233/SHTI190204

23. Fisher A, Young MM, Payer D, Pacheco K, Dubeau C, Mago V. Automating detection of drug-related harms on social media: machine learning framework. *J Med Internet Res*. (2023) 25. doi: 10.2196/43630

24. Severson MA, Onanong S, Dolezal A, Bartelt-Hunt SL, Snow DD, McFadden LM. Analysis of wastewater samples to explore community substance use in the United States: pilot correlative and machine learning study. *JMIR Form Res*. (2023) 7. doi: 10.2196/45353

25. Deng L, Liu Singapore Y. Deep Learning in Natural Language Processing, edited by Li Deng and Yang Liu. Singapore: Springer, 2018. ISBN 9789811052088. XVII + 329 pages. *Nat Lang Eng*. (2021) 27:373–5. doi: 10.1017/S1351324919000597

26. Najafizada M, Rahman A, Donnan J, Dong Z, Bishop L. Analyzing sentiments and themes on cannabis in Canada using 2018 to 2020 Twitter data. *J Cannabis Res*. (2022) 4:1–14. doi: 10.1186/s42238-022-00132-1

27. van Draanen J, Tao HD, Gupta S, Liu S. Geographic differences in cannabis conversations on twitter: infodemiology study. *JMIR Public Health Surveill*. (2020) 6. doi: 10.2196/18540

28. Bergman BG, Wu W, Marsch LA, Crosier BS, DeLise TC, Hassanpour S. Associations between substance use and instagram participation to inform social network–based screening models: multimodal cross-sectional study. *J Med Internet Res*. (2020) 22. doi: 10.2196/21916

29. Miliano C, Margiani G, Fattore L, De Luca MA. Sales and advertising channels of new psychoactive substances (NPS): internet, social networks, and smartphone apps. *Brain Sci*. (2018) 8. doi: 10.3390/BRAINSCI8070123

30. Allem JP, Escobedo P, Dharmapuri L. Cannabis surveillance with twitter data: emerging topics and social bots. *Am J Public Health*. (2020) 110:357–62. doi: 10.2105/AJPH.2019.305461

31. Tassone J, Yan P, Simpson M, Mendhe C, Mago V, Choudhury S. Utilizing deep learning and graph mining to identify drug use on Twitter data. *BMC Med Inform Decis Mak*. (2020) 20. doi: 10.1186/S12911-020-01335-3

32. de Anta L, Alvarez-Mon MA, Ortega MA, Salazar C, Donat-Vargas C, Santoma-Vilaclara J, et al. Areas of interest and social consideration of antidepressants on english tweets: A natural language processing classification study. *J Pers Med*. (2022) 12. doi: 10.3390/jpm12020155

33. Alvarez-Mon MA, Donat-Vargas C, Santoma-Vilaclara J, de Anta L, Goena J, Sanchez-Bayona R, et al. Assessment of antipsychotic medications on social media: machine learning study. *Front Psychiatry*. (2021) 12:737684. doi: 10.3389/FPSYT.2021.737684

34. Alvarez-Mon MA, Llavero-Valero M, Del Barco AA, Zaragozá C, Ortega MA, Lahera G, et al. Areas of interest and attitudes toward antiobesity drugs: thematic and quantitative analysis using twitter. *J Med Internet Res*. (2021) 23. doi: 10.2196/24336

35. Alvarez-Mon MA, Fernandez-Lazaro CI, Llavero-Valero M, Alvarez-Mon M, Mora S, Martínez-González MA, et al. Mediterranean diet social network impact along 11 years in the major US media outlets: thematic and quantitative analysis using twitter. *Int J Environ Res Public Health*. (2022) 19. doi: 10.3390/IJERPH19020784

36. Alvarez-Mon MA, del Barco AA, Lahera G, Quintero J, Ferre F, Pereira-Sanchez V, et al. Increasing interest of mass communication media and the general public in the distribution of tweets about mental disorders: observational study. *J Med Internet Res*. (2018) 20. doi: 10.2196/JMIR.9582

37. Alvarez-Mon MA, Donat-Vargas C, Llavero-Valero M, Gea A, Alvarez-Mon M, Martinez-Gonzalez MA, et al. Analysis of media outlets on women's health: thematic and quantitative analyses using twitter. *Front Public Health*. (2021) 9:644284. doi: 10.3389/FPUBH.2021.644284

38. Butt S, Sharma S, Sharma R, Sidorov G, Gelbukh A. What goes on inside rumour and non-rumour tweets and their reactions: A psycholinguistic analyses. *Comput Hum Behav*. (2022) 135:107345. doi: 10.1016/J.CHB.2022.107345

39. de Anta L, Alvarez-Mon MA, Donat-Vargas C, Lara-Abelanda FJ, Pereira-Sanchez V, Gonzalez Rodriguez C, et al. Assessment of beliefs and attitudes about electroconvulsive therapy posted on Twitter: An observational study. *Eur Psychiatry*. (2023) 66. doi: 10.1192/J.EURPSY.2022.2359

40. Nguyen DQ, Vu T, Tuan Nguyen AT. *BERTweet: A pre-trained language model for English Tweets*. (2020) pp. 9–14. Available at: http://arxiv.org/abs/2005.10200

41. Morris JX, Lifland E, Yoo JY, Grigsby J, Jin D, Qi Y. A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.16.

42. Hartmann J. Emotion English DistilRoBERTa-base (2022). Available online at: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base (Accessed July 4, 2023).

43. Ekman P. Basic Emotions. In: *Handbook of Cognition and Emotion* (2005). p. 45–60. doi: 10.1002/0470013494.CH3

44. Al-Rawi A. The convergence of social media and other communication technologies in the promotion of illicit and controlled drugs. *J Public Health (Oxf)*. (2022) 44:E153–60. doi: 10.1093/PUBMED/FDAA210

45. Black JC, Margolin ZR, Olson RA, Dart RC. Online conversation monitoring to understand the opioid epidemic: epidemiological surveillance study. *JMIR Public Health Surveill*. (2020) 6:e17073. doi: 10.2196/17073

46. Cano M, Oh S, Salas-Wright CP, Vaughn MG. Cocaine use and overdose mortality in the United States: Evidence from two national data sources, 2002-2018. *Drug Alcohol Depend*. (2020) 214. doi: 10.1016/J.DRUGALCDEP.2020.108148

47. Drug Overdose Death Rates | National Institute on Drug Abuse (NIDA). Available online at: https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates (Accessed June 23, 2023).

48. Ritchie H, Arriagada P, Roser M. Opioids, cocaine, cannabis and illicit drugs. *Our World Data*. (2022).

49. Cohen A, Vakharia SP, Netherland J, Frederique K. How the war on drugs impacts social determinants of health beyond the criminal legal system. *Ann Med*. (2022) 54:2024. doi: 10.1080/07853890.2022.2100926

50. Rincón-Ruiz A, Correa HL, León DO, Williams S. Coca cultivation and crop eradication in Colombia: The challenges of integrating rural reality into effective anti-drug policy. *Int J Drug Policy*. (2016) 33:56–65. doi: 10.1016/J.DRUGPO.2016.06.011

51. Gutiérrez-Sanín F. Tough Tradeoffs: Coca crops and agrarian alternatives in Colombia. *Int J Drug Policy*. (2021) 89. doi: 10.1016/J.DRUGPO.2021.103156

52. Cocaine – the current situation in Europe (European Drug Report 2023). Available online at: www.emcdda.europa.euhttps://www.emcdda.europa.eu/publications/european-drug-report/2023/cocaine_en (Accessed June 23, 2023).

53. Rooney B, Sobiecka P, Rock K, Copeland C. From bumps to binges: overview of deaths associated with cocaine in England, Wales and Northern Ireland (2000-2019). *J Anal Toxicol*. (2023) 47:207–15. doi: 10.1093/JAT/BKAD002

54. Sabe M, Zhao N, Kaiser S. A systematic review and meta-analysis of the prevalence of cocaine-induced psychosis in cocaine users. *Prog Neuropsychopharmacol Biol Psychiatry*. (2021) 109. doi: 10.1016/J.PNPBP.2021.110263

55. Eiden C, Vincent M, Serrand C, Serre A, Richard N, Picot MC, et al. Health consequences of cocaine use in France: data from the French Addictovigilance Network. *Fundam Clin Pharmacol*. (2021) 35:455–65. doi: 10.1111/FCP.12603

56. Countries with the Highest Cocaine Trade rate in Africa - The Organized Crime Index | ENACT. Available online at: https://africa.ocindex.net/rankings/cocaine_trade (Accessed July 4, 2023]).

57. Acuda W, Othieno CJ, Obondo A, Crome IB. The epidemiology of addiction in Sub-Saharan Africa: a synthesis of reports, reviews, and original articles. *Am J Addict*. (2011) 20:87–99. doi: 10.1111/j.1521-0391.2010.00111.x

58. Peltzer K, Ramlagan S, Johnson BD, Phaswana-Mafuya N. Illicit drug use and treatment in South Africa: a review. *Subst Use Misuse*. (2010) 45:2221–43. doi: 10.3109/10826084.2010.481594

59. Cocaine and Instability in Africa: Lessons from Latin America and the Caribbean - GSDRC. Available online at: https://gsdrc.org/document-library/cocaine-and-instability-in-africa-lessons-from-latin-america-and-the-caribbean/ (Accessed January 30, 2024).

60. Stevens RC, Brawner BM, Kranzler E, Giorgi S, Lazarus E, Abera M, et al. Exploring substance use tweets of youth in the United States: mixed methods study. *JMIR Public Health Surveill*. (2020) 6. doi: 10.2196/16191

61. Eiden C, Roy S, Malafaye N, Lehmann M, Peyrière H. Ten-year trends in hospitalizations related to cocaine abuse in France. *Fundam Clin Pharmacol*. (2022) 36:1128–32. doi: 10.1111/FCP.12815

62. Scott K R, Nelson L, Meisel Z, Perrone J. Opportunities for exploring and reducing prescription drug abuse through social media. *J Addict Dis*. (2015) 34:178–84. doi: 10.1080/10550887.2015.1059712

63. Kim SJ, Marsch LA, Hancock JT, Das AK. Scaling up research on drug abuse and addiction through social media big data. *J Med Internet Res*. (2017) 19. doi: 10.2196/JMIR.6426

64. Stolberg VB. The use of coca: prehistory, history, and ethnography. *J Ethn Subst Abuse*. (2011) 10:126–46. doi: 10.1080/15332640.2011.573310

65. Richards JR, Le JK. *Cocaine Toxicity*. Treasure Island (FL): StatPearls Publishing (2022). Available at: https://www.ncbi.nlm.nih.gov/books/NBK430976/.

66. Goldstein RA, DesLauriers C, Burda A, Johnson-Arbor K. Cocaine: history, social implications, and toxicity: a review. *Semin Diagn Pathol*. (2009) 26:10–7. doi: 10.1053/J.SEMDP.2008.12.001

67. Villar Núñez M de los Á, Sánchez Morcillo J, Ruíz Martínez MA. Purity and adulteration in cocaine seizures and drug market inspection in Galicia (Spain) across an eight-year period. *Drug Test Anal*. (2018) 10:381–91. doi: 10.1002/DTA.2216

68. Roque Bravo R, Faria AC, Brito-Da-costa AM, Carmo H, Mladěnka P, Dias da Silva D, et al. Cocaine: an updated overview on chemistry, detection, biokinetics, and pharmacotoxicological aspects including abuse pattern. *Toxins (Basel)*. (2022) 14. doi: 10.3390/TOXINS14040278

69. Park JN, Rashidi E, Foti K, Zoorob M, Sherman S, Alexander GC. Fentanyl and fentanyl analogs in the illicit stimulant supply: Results from U.S. drug seizure data, 2011-2016. *Drug Alcohol Depend*. (2021) 218. doi: 10.1016/J.DRUGALCDEP.2020.108416

70. Calac AJ, McMann T, Cai M, Li J, Cuomo R, Mackey TK. Exploring substance use disorder discussions in Native American communities: a retrospective Twitter infodemiology study. *Harm Reduct J*. (2022) 19:141. doi: 10.1186/s12954-022-00728-z

71. Nestler EJ. The neurobiology of cocaine addiction. *Sci Pract Perspect*. (2005) 3:4. doi: 10.1151/SPP05314

72. Kampman KM. The treatment of cocaine use disorder. *Sci Adv*. (2019) 5. doi: 10.1126/SCIADV.AAX1532

73. Baldisseri MR. Impaired healthcare professional. *Crit Care Med*. (2007) 35. doi: 10.1097/01.CCM.0000252918.87746.96

74. Mackey TK, Kalyanam J, Katsuki T, Lanckriet G. Twitter-based detection of illegal online sale of prescription opioid. *Am J Public Health*. (2017) 107:1910–5. doi: 10.2105/AJPH.2017.303994

75. Mackey TK, Kalyanam J. Detection of illicit online sales of fentanyls via Twitter. *F1000Res*. (2017) 6. doi: 10.12688/f1000research

76. Mackey T, Kalyanam J, Klugman J, Kuzmenko E, Gupta R. Solution to detect, classify, and report illicit online marketing and sales of controlled substances via twitter: using machine learning and web forensics to combat digital opioid access. *J Med Internet Res*. (2018) 20. doi: 10.2196/10029

frontiers | Frontiers in Psychiatry

TYPE Review
PUBLISHED 26 March 2024
DOI 10.3389/fpsyt.2024.1319522

OPEN ACCESS

EDITED BY
Robert Stewart,
King's College London, United Kingdom

REVIEWED BY
Janie Busby Grant,
University of Canberra, Australia
Tahsin Saira,
Swansea University Medical School,
United Kingdom

*CORRESPONDENCE
Mascha Kurpicz-Briki
mascha.kurpicz@bfh.ch

RECEIVED 11 October 2023
ACCEPTED 05 March 2024
PUBLISHED 26 March 2024

CITATION
Merhbene G, Puttick A and Kurpicz-Briki M (2024) Investigating machine learning and natural language processing techniques applied for detecting eating disorders: a systematic literature review. *Front. Psychiatry* 15:1319522. doi: 10.3389/fpsyt.2024.1319522

# Investigating machine learning and natural language processing techniques applied for detecting eating disorders: a systematic literature review

Ghofrane Merhbene, Alexandre Puttick and Mascha Kurpicz-Briki*

Applied Machine Intelligence, Bern University of Applied Sciences, Biel/Bienne, Switzerland

Recent developments in the fields of natural language processing (NLP) and machine learning (ML) have shown significant improvements in automatic text processing. At the same time, the expression of human language plays a central role in the detection of mental health problems. Whereas spoken language is implicitly assessed during interviews with patients, written language can also provide interesting insights to clinical professionals. Existing work in the field often investigates mental health problems such as depression or anxiety. However, there is also work investigating how the diagnostics of eating disorders can benefit from these novel technologies. In this paper, we present a systematic overview of the latest research in this field. Our investigation encompasses four key areas: (a) an analysis of the metadata from published papers, (b) an examination of the sizes and specific topics of the datasets employed, (c) a review of the application of machine learning techniques in detecting eating disorders from text, and finally (d) an evaluation of the models used, focusing on their performance, limitations, and the potential risks associated with current methodologies.

KEYWORDS

**natural language processing, machine learning, eating disorders, mental health, artificial intelligence, anorexia, bulimia, binge eating**

## 1 Introduction

Recent reports in broad media about the latest conversational chatbots, which can generate human-like texts in response to user questions have made natural language processing (NLP) famous to the broad public. Yet the possibilities of this field go far beyond text generation and chatbots. Classifying texts into two (or more) groups and automatically extracting indicators that suggest that a text snippet belongs to either of the groups is also a common task. In particular, when using machine learning, this allows the identification of patterns that might differ from what a human might detect that are nonetheless effective in separating the two groups.

Meanwhile, in clinical practice in mental health, inventories with scaling questions are often used for diagnosis. Such inventories have limitations, including for example defensiveness (the denial of symptoms) or social bias that can influence the results of the questionnaires (1). In these cases, an automated text analysis applied to specific open questions or interview transcripts can provide further source of information indicating the patient's condition that is more resistant to manipulations such as those arising from defensiveness.

Defensiveness is common amongst those afflicted with eating disorders (EDs). Respondents to a survey investigating the denial and concealment of EDs (2) reported a variety of attempts to hide the respective ED. Furthermore, the authors of the study state that such methods were described as deliberate strategies. This makes it challenging to use clinical instruments where an inventory item contains obvious indications for which options to choose in order to obtain a specific result.

EDs generally occur in the form of unhealthy eating habits, disturbances in behaviors, thoughts, and attitudes towards food, causing in some cases extreme weight loss or gain. These disorders not only impact mental health but also have physical effects (3). EDs are classified in the category F50 of the ICD-10 and can refer to different disorders including anorexia, bulimia or overeating[1]. A study conducted by Mohler-Kuo et al. (4) in Switzerland discovered that the lifetime prevalence for any ED is 3.5%. Another survey investigating the lifetime prevalence of EDs in English and French studies from 2000 to 2018 found that the weighted means were 8.4% for women, and 2.2% for men (5).

The power of natural language processing (NLP) has already been applied to the field of mental health, especially in research. Feelings and written expression are closely correlated: An analysis of student essays has shown that students suffering from depression use more negatively valenced[2] words and more frequently use the word "I" (6). Different approaches have been applied to explore how to use automated text analysis on tasks such as the detection of burnout (7), depression (8, 9), the particular case of post-partum depression (10, 11), anxiety (12), and suicide risk assessment (13), (14). Often, such methods are based on anonymized publicly available online data. Only little work makes use of clinical data. Furthermore, the English language has been the primary focus, even though these methods can be highly language-dependent, meaning that data and methods should be carefully reviewed when adapting to local languages. This is relevant, as it has been shown that adapting to the patient's language is beneficial in mental health diagnostics and treatment (15). In our view, one aim of such technologies should be to explore ways to support clinical practitioners in their daily work, and provide them with additional sources of information to consider. Therefore, we often refer to such solutions as Augmented Intelligence[3], rather than Artificial Intelligence, as they aim to empower humans rather than replacing them.

Despite existing work in the field of ML and NLP for depression, anxiety or suicide risk assessment, there has been a lack of a detailed systematic literature comparison on the automatic detection of EDs using NLP technologies for both clinical and non-clinical data. A recent survey (16) investigated the use of natural language processing applied to mental illness detection. The majority of the identified results (45%) had worked on depression, whereas only 2% were about eating disorders in general and 3% about anorexia. Whereas the broad scope of the survey provides a generous overview of the research landscape, it does not compare the case of eating disorders in detail.

In this paper, we have undertaken a systematic literature review to address this research gap, following the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (17) to ensure a well-structured and transparent methodology.

We contribute to the field by (a) analyzing the metadata of published papers to understand the current trends and methodologies, (b) examining the sizes and targeted topics of the datasets used in these studies, (c) reviewing how machine learning techniques are applied to detect eating disorders from textual data, and (d) evaluating the performance, limitations, and potential risks of the models deployed in this domain.

Our research is guided by specific questions, structured around four distinct perspectives, which collectively form the core of our investigative approach.

- Demographical Questions (DemRQ): Focus on metadata aspects of the paper:
  - DemRQ1: When was the paper published?
  - DemRQ2: From which countries were the contributors of the papers included in this study?
- Input Questions (InputRQ): Focus on the format and topic of the input data:
  - InputRQ1: Which languages were taken into consideration?
  - InputRQ2: What was the size of the dataset used?
  - InputRQ3: Which data sources were used for data collection in the case of both clinical and non-clinical data?
  - InputRQ4: What types of eating disorders were addressed in these studies?
- Architectural Questions (ArchRQ): Focus on the experimental architecture:
  - ArchRQ1: Which feature extraction technique was used?
  - ArchRQ2: Which machine learning techniques in the field of NLP have been used for ED detection?
- Evaluation Questions (EvalRQ): Focus on the evaluation aspects of the trained model:
  - EvalRQ1: How did the model perform?
  - EvalRQ2: What are the limitations and risks of the existing methods, and how can they be improved?

[1] https://icd.who.int/browse10/2019/en#/F50

[2] Valence is a measure of the emotional intensity or positivity/negativity associated with a word.

[3] See e.g., https://digitalreality.ieee.org/publications/what-is-augmented-intelligence

The article is structured as follows: First, we describe our methodology such as the study design and the paper selection process. We then describe the results of the literature search and describe the findings of our review. Finally, we summarize our results and describe perspectives for future research in the field.

# 2 Methods

## 2.1 Study design

To answer our research questions, we conducted a structured literature review (SLR) following the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (17). This includes standards for literature search strategies and setting criteria for the inclusion or exclusion of gathered works in the final review.

## 2.2 Literature search strategy

In accordance with PRISMA standards, we have set an 8-year time span for searching for documents (2014-2022) related to our research scope. We consider the year 2014 mainly because Bellows et al. (18) conducted a study on automatically detecting binge Eating disorder using clinical data, which we deem to be the initial research in the field. We then compiled a list of all databases to be searched. The list included the following databases:

- Google Scholar
- IEEE Xplore
- Pubmed

In addition, in order to efficiently conduct our database search we have compiled a list of keywords and conditions. These keywords are relevant to the research topic of EDs and their detection using NLP and machine learning techniques. Furthermore, the list included specific terms related to social media and online social networks in order to enable the identification of studies that explore the use of social media for the early detection of EDs, which is an ongoing research interest. The final query is presented below:

(eating disorder OR anorexia OR binge eating OR bulimia OR overeating) *AND* (natural language processing OR NLP OR text mining OR inventories OR machine learning OR artificial intelligence OR automatic detection OR early detection OR social media OR online social network OR clinical).

Using the aforementioned search keywords and conditions, we retrieved research articles where NLP techniques have been used for the detection of EDs from clinical and non-clinical data. The detailed workflow is depicted in Figure 1, and the corresponding PRISMA flow diagram for this SLR is shown in Figure 2.

With the initially proposed search query, a large number of papers was identified. With manual analysis we explored options to define a more restrictive query, still making sure to capture the relevant papers, which turned out challenging. We therefore adapted our method to consider the first 100 elements returned by the search query on each database, sorted by relevance. This furthermore allowed to apply the same methodology for all three data sources, including especially Google Scholar, where the search functionalities are limited compared to databases like PubMed, and thus we had to make a selection on the number of items to be reviewed. Given the interidisciplinarity of our approach, we wanted to include Google Scholar to target a vast number of sources and ensure the most relevant work can be included.

A Python script was used to screen the articles for duplicates. As a result, 1 article was excluded from further consideration, leaving a total of 299 articles for further analysis (see Figure 2). To refine the results further, a manual title scan was performed to exclude articles that were not pertinent to the research topic. This resulted in the exclusion of 237 articles, leaving a total of 62 for further analysis. Additionally, a manual scan of the abstracts from the remaining 62 articles was performed to exclude any that were not relevant to the study. This process resulted in the exclusion of an additional 30 articles, leaving a total of 32 for inclusion in the final analysis. After thoroughly reading and evaluating 32 articles, 27 were selected as relevant for the researched topic (according to the criteria from Table 1). These chosen articles were deemed to possess high relevance and reliability for this SLR. Finally, we scanned the references section of the articles included in our survey and identified any relevant literature that may have been missed in the initial database search. This added n=18 articles to the studies that were finally included in the review (n=45). The process is illustrated in Figure 2.

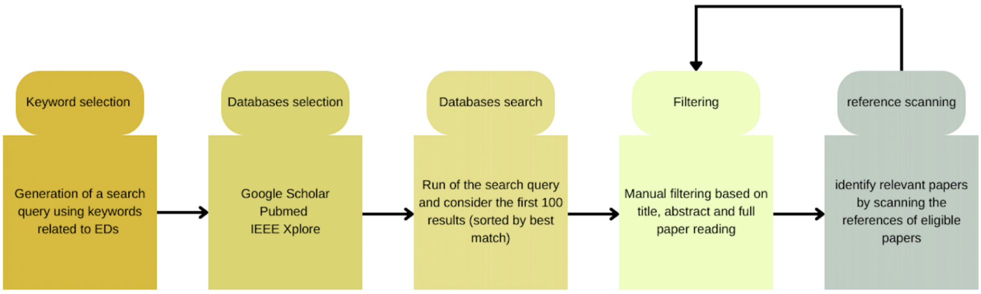

Keyword selection
Generation of a search query using keywords related to EDs
Databases selection
Google Scholar
Pubmed
IEEE Xplore
Databases search
Run of the search query and consider the first 100 results (sorted by best match)
Filtering
Manual filtering based on title, abstract and full paper reading
reference scanning
identify relevant papers by scanning the references of eligible papers

FIGURE 1
Methodology for document collection.

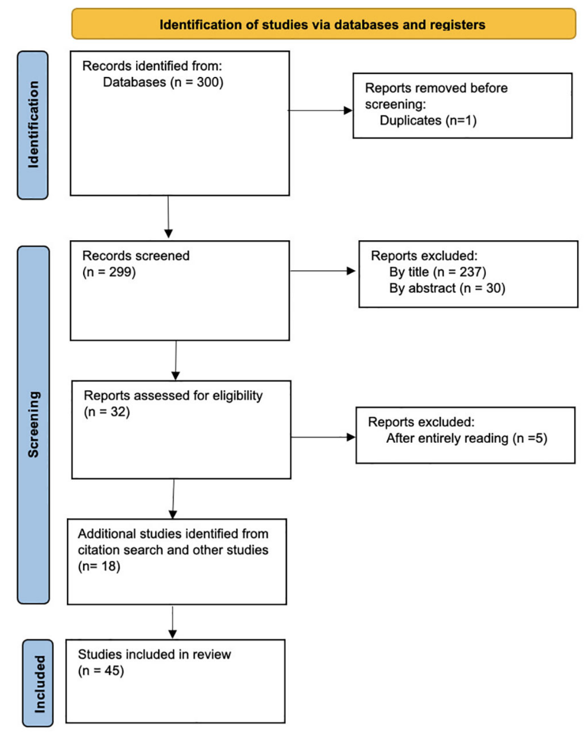

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
Databases (n = 300)
Reports removed before screening:
Duplicates (n=1)
Records screened
(n = 299)
Reports excluded:
By title (n = 237)
By abstract (n = 30)
Reports assessed for eligibility
(n = 32)
Reports excluded:
After entirely reading (n =5)
Additional studies identified from citation search and other studies
(n= 18)
Studies included in review
(n = 45)

FIGURE 2
PRISMA Flow diagram. Based on: Page et al., (17).

## 2.3 Inclusion and exclusion criteria

Table 1 outlines the predefined exclusion and inclusion criteria that were used to guide the selection of related studies for the review. These criteria were established in advance to help simplify the process of identifying and selecting relevant papers. In particular, papers that focused solely on the psychological aspects of EDs and did not consider the use of automated text analysis technologies were excluded from the review. By adhering to these criteria, we were able to more effectively and efficiently select the relevant papers.

TABLE 1 SLR study selection of literature using inclusion and exclusion criteria.

| Criteria | Decision |
|---|---|
| When the predefined keywords exist in title, keywords or abstract section of the paper. | Inclusion |
| The paper should be written in the English language | Inclusion |
| When the paper targets other languages | Inclusion |
| Papers that are duplicated within the search documents | Exclusion |
| Papers that don't make use of automated text analysis | Exclusion |
| Papers that deal with other types of data (than textual) | Exclusion |
| Papers that got published before 2014 | Exclusion |

# 3 Results

In this section, we provide a thorough review and analysis of the research studies included in this systematic literature review.

## 3.1 Terminology

- Bag of Words (BoW) is a fundamental technique used in NLP for text representation. It involves representing text data by counting the frequency of occurrence of each word in a document.
- Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic used to evaluate the importance of a word in a document within a collection or corpus. It combines two metrics: term frequency (TF), which measures the frequency of a word in a document, and inverse document frequency (IDF), which penalizes words that are common across the entire corpus.
- Bidirectional Encoder Representations from Transformers (BERT) (19) is a pretrained deep learning model introduced by Google in 2018. It belongs to the Transformer architecture and is designed to understand the context of words in a sentence by considering both left and right context simultaneously
- Word2Vec (20) is a technique for learning word embeddings. Word2Vec represents each word as a vector, with similar words having vectors that are closer together in the vector space.
- Global Vectors for Word Representation (GloVe) (21) is another technique for learning word embeddings. GloVe also generates vector representations of words based on their co-occurrence statistics in a corpus. However, GloVe considers the global context of the entire corpus to learn word embeddings, unlike Word2Vec, which focuses on local context.
- Embeddings from Language Models (ELMO) (22) is a deep contextualized word representation model. It generates word embeddings by considering the entire input sentence and capturing its contextual information.
- Doc2Vec (23) also known as Paragraph Vector, is an unsupervised learning algorithm to generate vector representations for pieces of texts like sentences and documents, it extends the Word2Vec methodology to larger blocks of text, capturing the context of words in a document.
- Bidirectional Long Short-Term Memory (Bi-LSTM) (24) is a type of Recurrent Neural Network (RNN) that processes data in both forward and backward directions. This architecture is particularly effective in understanding the context in sequence data like text or time series, as it captures information from both past (backward) and future (forward) states.
- Linguistic Inquiry and Word Count (LIWC) (25) is a text analysis program that counts words in psychologically meaningful categories.

## 3.2 Demographical research questions

Figure 3 shows the yearly distribution of the selected research work (DemRQ1). The data suggests a growing interest in this topic in recent years. This is in line with the findings of Zhang et al. (16) that found that there has been an upward trend over the last years in using NLP and machine learning methods to detect mental health problems. Notably, we highlight a prominent peak in 2018 and 2019, which coincides with the emergence of tasks related to EDs in eRisk competitions.

We also observed the geographical distribution of the authors' affiliations of the selected studies (DemRQ2). As visualized in the heat-map in Figure 4, 7 of the selected studies were from the USA and Spain, 5 from Mexico and France.

From the 45 selected studies, 24 were results from the eRisk lab[4], hosted by the CLEF Conference since 2017. This academic research competition focuses on the development and evaluation of text-based risk prediction models for social media. Each year, the lab provides a shared task framework where teams of participants are tasked with developing NLP techniques to automatically identify and predict the risk of different mental illness behaviors from social media data, including Eating Disorders. Participants are provided with a training dataset and a test dataset, and the performance of their models is evaluated based on two categories: performance and latency. The eRisk lab provides a unique opportunity for researchers to collaborate and innovate in the field of NLP and mental health, aiming to improve the detection and prevention of mental health issues in online communities. The datasets used in the eRisk lab are primarily sourced from the social media platform Reddit.

Since 2017, the challenge has included two tasks pertaining to the early detection of Eating Disorders. In both 2018 and 2019, the task involved the early detection of signs of anorexia [see e.g., Losada et al. (26)]. In contrast, the 2022 iteration introduced a novel task centered on measuring the severity of eating disorders (27). This task diverged from the previous ones in that no labeled training data was supplied to participants, meaning that participants could not evaluate the quality of their models' predictions until test time. The task objective was to assess a user's level of eating disorder severity through analysis of their Reddit posting history. In order to achieve this, participants were required to predict users' responses to a standard eating disorder questionnaire (EDE-Q)[5] (28).

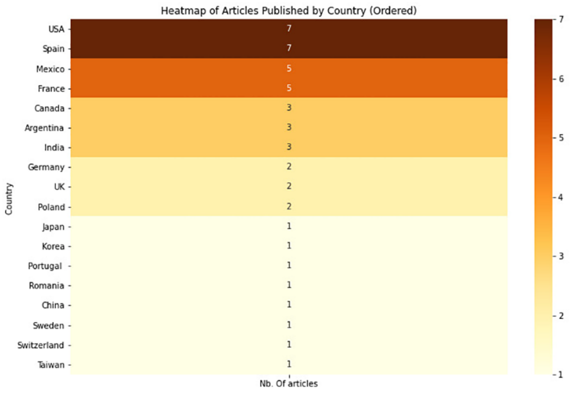


FIGURE 4
Geographic distribution of all institutions involved in the selected research articles.

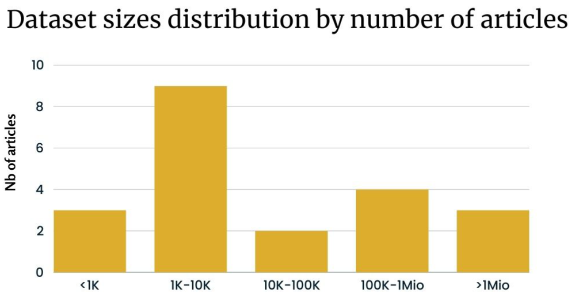


FIGURE 5
Dataset sizes distribution based on Table 2 excluding articles from eRisk.

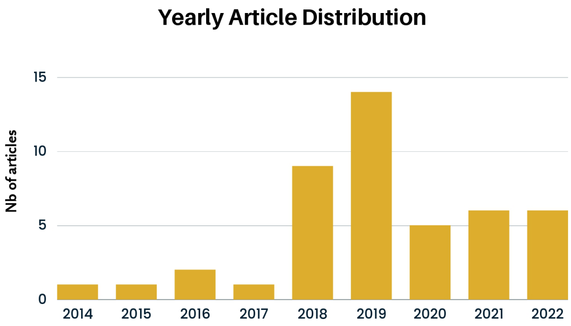


FIGURE 3
Yearly distribution of all research articles.

## 3.3 Input research questions

Our first input research question (InputRQ1) investigates the different languages that are considered in the studies included in this SLR. Research has shown that only a small number of the over 7000 languages used worldwide are represented in recent technologies from the field of natural language processing (29). We wanted to investigate whether this is also the case for the detection of eating disorders. Text analysis, naturally, depends on the specific language and can typically not be transferred from one language to another without specific adaptions.

Table 2 gives indication about the language of data used, its size, its source, and the type of eating disorder that was investigated in the

[4] https://erisk.irlab.org/

[5] https://www.corc.uk.net/media/1273/ede-qquesionnaire.pdf

TABLE 2 Datasets characteristics.

| Paper | Language | Dataset Size | Data Source | Targeted ED |
|---|---|---|---|---|
| Choudhury (30) | English | 10K-100K (55'334) | Social Media (Tumblr) | Anorexia |
| Yan et al. (31) | English | 1K-10K (4'812 collected, 53 labelled by specialists) | Social Media (Reddit) | ED |
| BeníCheck that all equations and special characters are displayed correctly.tez-Andrades et al. (32) | English | 1K-10K (1'085'957 collected, 2'000 manually labelled) | Social Media (Twitter) | ED |
| López Úbeda et al. (33) | Spanish | 1K-10K (5'707) | Social Media (Twitter) | Anorexia |
| Zhou et al. (34) | English | 1K-10K (123'977 collected, 2'219 manually labelled) | Social Media (Twitter) | ED |
| Aguilera et al. (35) | English | 100k-1Mio (Dataset from 2018-2019 editions of eRisk shared tasks) | Social Media (Reddit) | Anorexia |
| Spinczyk et al. (36) | Polish | <1K (96 written statements about the body image: 44 Anorexia females, 52 Healthy females) | Clinical Data | Anorexia |
| Aragon et al. (37) | English | <1K (Dataset from CLEF eRisk 2018 shared task) | Social Media (Reddit) | Anorexia |
| Bellows et al. (18) | English | 1K-10K (1'000 Narrative Electronic Health Records) | Clinical Data | Binge Eating |
| Benítez-Andrades et al. (38) | English | 1K-10K (1'085'957 collected, 2000 manually labelled) | Social Media (Twitter) | ED |
| Ramiandrisoa and Mothe (39) | English | 100k-1Mio (Sequence of writings in chronological order of 472 users (eRisk 2019 data)) | Social Media (Reddit) | Anorexia |
| Wang et al. (40) | English | >1Mio (119'825'361) | Social media (Twitter) | ED |
| He and Luo (41) | English | 1K-10K (Tumblr 5'965 manually labeles)<br>100k-1000k (Twitter labeled based on hashtags) | Social Media (Tumblr and Twitter) | ED |
| Tébar and Gopalan (42) | English | 100k-1Mio (253'341) | Social Media (Reddit) | ED |
| Dinu and Moldovan (43) | English | 10k-100k (50'000) | Social Media [Reddit : Sample data from SMHD dataset from Cohan et al. (2018)] | MD[6] |
| Jiang et al. (44) | English | >1Mio (17.5m) | Social Media (Reddit) | MD |
| Zhang et al. (45) | English | 1K-10K (8'554) | Social Media (Reddit) | MD |
| Hwang et al. (46) | English | 1K-10K (3'714'057, 5'126 labelled) | Social Media (Reddit) | ED |
| Rojewska et al. (47) | Polish | <1K (51 written statements) | Clinical Data | Anorexia |
| Villegas et al. (48) | English | 100k-1Mio (253'752) | Social Media (Reddit) | Anorexia |
| Chancellor et al. (49) | English | >1Mio (2'416'272) | Social Media (Instagram) | ED |

[6]Mental disorders including EDs.

selected studies (excluding studies from eRisk). 18 of the 21 studies used English data, 2 used Polish and 1 Spanish data. The 24 papers from the eRisk lab challenges all relied on English data from the platform Reddit. Overall, only 3 out of 45 studies used a language other than English (7%). This confirms the need for further work in applying the latest technological developments to non-English texts.

The dataset size is another crucial factor we took into account in our analysis ((InputRQ2). As depicted in Figure 5, the distribution of dataset sizes used in the studies reveals that datasets ranging from 1k to 10k instances are the most frequently used.

The distribution of dataset sizes across different research topics, as illustrated in Figure 6, offers insightful perspectives. Notably, Anorexia research displays the most significant variance in dataset sizes, spanning from less than 1K to over 1 million data points. In contrast, binge eating research predominantly employs datasets within a narrower range of 1K to 10K data points. For broader Eating Disorders, 6 studies leverage datasets between 10K and 100K, while 3 others operate with datasets in the 100K to 1 million range. Finally, research on Mental Disorders encompasses datasets varying from 1K to more than 1 million data points.

Table 2 also gives an overview of the data sources (InputRQ3). From the 45 studies, the used datasets can be classified as follows in four groups:

- eRisk lab datasets: 24 studies
- Other online forums and social media: 17
- Medical data: 3
- SMHD dataset (50): 1

The distribution of the primary focus of these studies is illustrated in Figure 7 (InputRQ4) The majority of the studies (n=29) we collected focused on anorexia, while 12 studies conducted a broader investigation of EDs in general rather than focusing on a specific type. Additionally, three studies had a more extensive scope, delving into various mental disorders, including but not limited to EDs, while one study focused on binge eating.

## 3.4 Architectural and evaluations research questions

### 3.4.1 eRisk challenge

Table 3 summarizes all the papers that we identified following our strategy, including the ones from eRisk. In 2018 and 2019, the eRisk papers focused on a text classification task aimed at developing an early detection system for eating disorders on social media using the history of users' writings data. The aim was to train a text classifier that could effectively identify and flag potential cases of anorexia based on users' social media content. For the eRisk challenge resulting in papers from 2022, the task was different. Participants were provided with the social media history of specific users and had to predict their answers to questions 1-12 and 19-28 from the Eating Disorder Examination Questionnaire (EDE-Q)[7] (28).

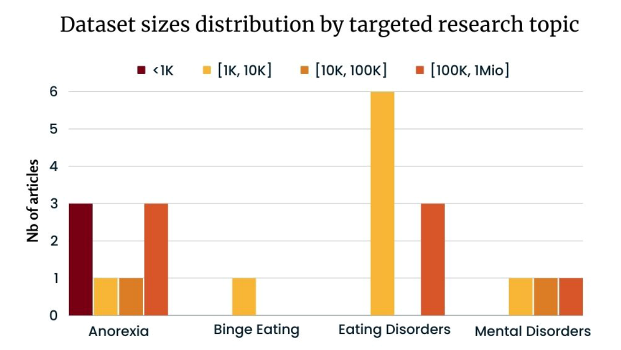


FIGURE 6
Dataset sizes distribution by targeted ED based on Table 2 excluding articles from eRisk.

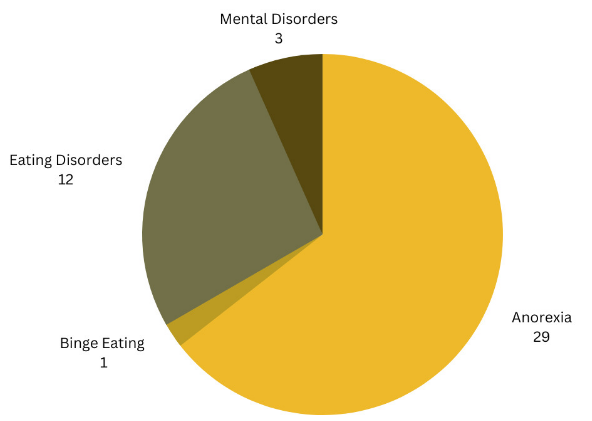


FIGURE 7
Research distribution of all research articles.

(ArchRQ1) The complexity of this task, along with the development in the field of NLP over the years 2019 to 2022, explains the choice of word2vec, GloVe (72) or transformer-based models (62, 66, 73) for vectorization/feature representation. For the remaining entries, very different approaches were used, ranging from anorexia specific vocabulary and LIWC (58) to more general approaches like Bag of Words (BoW) (52, 53) or TF-IDF (51, 57). (ArchRQ2) The choices of methods for prediction were also heterogeneous, ranging from cosine similarity (72) to linear models (52, 54, 58, 66, 71), to neural networks (51, 53, 56).

(EvalRQ1) For the 2018-2019 eRisk papers, we report F1 values corresponding to the binary classification task, whereas for the 2022 paper we report mean average error (MAE), corresponding to the average deviation between user's predicted questionnaire responses and the ground truth responses.

### 3.4.2 Non-eRisk studies

Table 3 shows the feature representation, tasks studied, machine learning techniques, and performance metrics of all studies included in this SLR. In this section we focus on Non-eRisk studies. We grouped these studies into the following categories with regard to the feature extraction techniques they apply (ArchRQ1):

- Bag of Words (BoW)
- Word embeddings
- TF-IDF
- BERT representations
- and other feature representations

Furthermore, it is worth noting that the machine learning methods used in these studies span various categories (ArchRQ2), including:

- Classical machine learning (ML) methods such as Support Vector Machine (SVM), Naive Bayes, Logistic Regression, etc.
- Deep learning (DL) methods, e.g., recurrent neural networks.
- Combination of different methods from classical ML and DL.
- Large language models (LLMs), e.g., BERT.
- Other approaches.

[7] https://www.corc.uk.net/media/1273/ede-qquesionnaire.pdf

TABLE 3 Overview of machine learning methods and performance metrics of the studies included in this systematic literature review.

| Paper | Feature Extraction | Studied task | ML Techniques | Performance |
|---|---|---|---|---|
| Wang et al. (51) | **TF-IDF** for keyword selection and sentences encoded using the CNNbased sentence encoder | **Classification** (eRisk 2018) | Convolutional neural networks (CNN) | F1 score = 0.67 |
| Paul et al. (52) | **BoW, UMLS** (Unified Medical Language System), and a combination of both | **Classification** (eRisk 2018) | SVM | F1 score = 0.67 with BoW |
| Trotzek et al., (53) | **Other** (Different techniques:**BoW/GloVe embeddings/ fastText embeddings)** | **Classification** (eRisk 2018) | CNN | F1 score = 0.85 |
| Ramiandrisoa et al. (54) | **Other** (Text vectorization using **doc2vec** (Two separate models were trained: 1- Distributed BOW with 100d output. 2Distributed Memory model wi9th 100-dimensional output)) | **Classification** (eRisk 2018) | Logistic Regression | F1 score = 0.76 |
| Ortega-Mendoza et al. (55) | **Other** (Discriminative personal purity (DPP), and a term weighting scheme called exponential reward of personal information (EXPEI)) | **Classification** (eRisk 2018) | IG-EXPEI (a supervised classification model based on information gain and a term weighting scheme) | F1 score = 0.67 |
| Ragheb et al. (56) | **Other** (Bi-LSTM Encoder) | **Classification** (eRisk 2018) | Bayesian inversion and Multi-layer Perceptron classifier | F1 score = 0.54 |
| Liu et al. (57) | **TF-IDF** | **Classification** (eRisk 2018) | SVM, CNN+LSTM and a simple keyword model | F1 score = 0.36 for CNN+LSTM |
| Ramírez-Cifuentes and Freire (58) | **Other** (LIWC, anorexia vocabulary: 9 features and 1 weighted feature) | **Classification** (eRisk 2018) | Linear Regression | F1 = 0.73 |
| Funez et al. (59) | **Other** (Sequential Incremental Classification (SIC)) | **Classification** (eRisk 2018) | Sequential Incremental Classification (SIC) | F1 score= 0.60 |
| Aragon et al. (60) | **Other** (Bag of Sub-emotions (BoSe)) | **Classification** (eRisk 2019) | SVM | F1 score= 0.68 |
| Burdisso et al. (61) | **Other** (Dictionary with a confidence value assigned to each work) | **Classification** (eRisk 2019) | SS3 (Burdisso et al., 2019a) | F1 score= 0.55 |
| Ragheb et al. (62) | **Other** (Bi-LSTM Encoder) | **Classification** (eRisk 2019) | a Universal Language Model Fine-tuning for text classification with an additional attention layer | F1 score = 0.68 |
| Fano et al. (63) | **Other** (GloVe) | **Classification** (eRisk 2019) | a Multilayer perceptron | F1 score = 0.68 |
| Masood et al. (64) | **Other** (Term-frequency transformer + feature selection using chi-squared test to select the most significant 500 terms) | **Classification** (eRisk 2019) | SVM | F1 score = 0.61 |
| Naderi et al. (65) | **TF-IDF** | **Classification** (eRisk 2019) | SVM | F1 score = 0.54 |
| Mohammadi et al. (66) | **Other** (GloVe and ELMO (Both were used as submodels for an ensemble model for generating embeddings)) | **Classification** (eRisk 2019) | SVM | F1 score = 0.71 |
| Del Arco et al. (67) | **Other** (UMLS) | **Classification** (eRisk 2019) | SVM | F1 score = 0.30 |
| Ranganathan et al. (68) | **Other** (Rapid automated keyword extraction (RAKE)) | **Classification** (eRisk 2019) | CNN-LSTM (2-layer LSTM with normed-bahdanau attention) | F1 score = 0.34 |
| Ferdowsi et al. (69) | **TF-IDF** | **Classification** (eRisk 2019) | CNN | F1 score = 0.17 |
| Trifan and Oliveira (70) | **BoW** and **TF-IDF** | **Classification** (eRisk 2019) | SVM with SGD classifier | F1 score = 0.37 |

*(Continued)*

TABLE 3 Continued

| Paper | Feature Extraction | Studied task | ML Techniques | Performance |
|---|---|---|---|---|
| Ortega-Mendoza et al. (71) | **Other** (DPP-EXPEI (55)) | **Classification** (eRisk 2019) | Linear SVM with L2 norm | F1 score= 0.58 |
| Hosseini Saravani et al. (72) | **Other** (22 feature sets developed with expert knowledge and 300-dimensional word2vec and GloVe vectors of different sizes) | **Answer prediction** (eRisk 2022) | Cosine similarity | MAE = 3.15 |
| Mármol-Romero et al. (73) | **Other** (RoBERTa contextualized word embeddings) | **Answer prediction** (eRisk 2022) | RoBERTa | MAE = 2.60 |
| Srivastava et al. (74) | **Other** (Cosine Similarity) | **Answer prediction** (eRisk 2022) | BERT | MAE = 2.18 |
| 30 | **Other** (Each data point is represented as a vector of four categories of measures: social, affective, linguistic style, and cognitive processes) | **Classification** (Binary: Detect anorexia content, differentiate between two online communities) | Binary SVM | F1 score= 0.818 |
| Yan et al. (31) | **TF-IDF** (Bag of Bigram with TF-IDF reweighting) for trial 1-2, Word Embeddings (Word Mover's Distance) for trial 3 | **Classification** (Binary: Identify posts that require intervention as positive or negative) | Logistic Regression and Word Mover's distance | Error Rate= 0.04 |
| Benítez-Andrades et al. (32) | **BERT representations** | **Classification** (Binary: People that suffer(ed) from ED Vs. People that do/did not) | 5 BERT based models | Accuracy= 0.875 for RoBERTa |
| López Úbeda et al. (33) | **TF-IDF** | **Classification** (people that suffer (ed) from anorexia vs. people that do/did not) | 5 Different supervised learning models including: SVM, Multilayer Perceptron classifier, Naive Bayes, Decision Tree and Logistic Regression | F1 score= 0.91 for SVM |
| Zhou et al. (34) | **Word Embeddings** (Global Vectors for Word Representation pretrained 200-dimension Twitter word embeddings) | **Classification** (ED irrelevant, promotional information ED amd laypeople discussion ED) | Convolutional neural network (CNN), long short-term memory (LSTM), support vector machine, and Na¨ıve Bayes and CorEx for topic modelling | F1 score=0.90 for CNNLSTM and Coherence rate= 0.771 for topic modelling |
| Aguilera et al. (35) | **BoW** (1000 terms and TF weights) and average of the following word embeddings: 200- dimensions **GloVe** vectors trained on Twitter data, 300-dimensions **Word2Vec** vectors trained on the Google News dataset and 300- dimensions **FastText** trained on Wikipedia and on the UMBC and statmt.org news dataset | **Classification** (anorexia 1-class classification: The focus is only on instances that belong to the anorexia class). | One-class Classification kstrongest Strengths (OCCkSS) and Global Strength Classifier (gSC) both built based on the K-Strongest Strengths algorithm | F1 score= 0.671 with gSC |
| Spinczyk et al. (36) | **Word2Vec** 100-dimensions vectors | **General sentiment analysis** from patient statements about their body images | Recurrent Neural Network (RNN) and Dictionary-based methods | F1 score= 0.70 for RNN and F1 score= 0.65 for Dictionary-based methods |
| Bellows et al. (18) | **Other** (Rule-based approach) | **Classification** (Identify binge eating Disorder Patients from EHR) | Not precise | Accuracy= 0.918 |
| Benítez-Andrades et al. (38) | **Other** (Not precise) | **Classification** (Binary categories in 4 categorization tasks (People suffering from ED Vs. Rest, Tweets promoting ED Vs. Rest, Informative VS. Noninformative, Scientific tweets Vs. Rest) | Random forest, Recurrent neural networks, Bidirectional long short-term memory networks, Bidirectional encoder representations from transformer-based models | F1 score= 0.864 with RoBERTa |
| Ramiandrisoa and Mothe (39) | Method 1: **Other**: Feature-based text representation (Based on features extracted by the authors) Method 2: text vectorization using doc2vec. | **Classification** (Early detection of signs of anorexia) | Random Forest, Logistic Regression combined with word embedding text representation | F1 score= 0.71 for Random Forest and F1 score= 0.73 for Logistic regression |

*(Continued)*

TABLE 3 Continued

| Paper | Feature Extraction | Studied task | ML Techniques | Performance |
|---|---|---|---|---|
| Wang et al. (40) | **Other** (Each user in the dataset was represented as a vector of 97 features obtained from the following measures:<br>6 social-status features, 11 behavioral features, and 80 psychometric features) | Snowball Sampling for Identifying Eating Disorder Communities on Twitter and a **Classification** (Binary: ED vs. NoED) | SVM | F1 score= 0.975 |
| He and Luo (41) | **Other** (ADTree, a decision tree algorithm used to rank hashtags, the top 10 ranked hashtags were used as features) | **Classification** (Identify pro-ED posts on Tumblr and pro-ED users on Twitter) | CMAR (75). | Accuracy = 0.68 for identification of pro-ED posts on Tumblr and Accuracy= 0.92 for identification of pro-ED posts on Twitter |
| Tébar and Gopalan (42) | **Other** (Used topic modeling to get topics as features, frequency of ED-related words, and writing features (Nb. of words per post, time gap and Weekday/ weekend posts and time of the day)) | **Classification** (Early detection of signs of EDs) | Feature fusion Multimodal model | F1 score= 0.82 with BoSEunigrams |
| Aragon et al. (37) | **Other** (Used BoSE-based representations, and contrasted them against BoE and BoW schemes | **Classification** (Anorexia or depression vs. Control group) | SVM with a linear kernel | F1 score= 0.97 |
| Dinu and Moldovan (43) | **Other** (used Naïve Bayes Classifier in order to find out the most informative features from each category in the dataset) | **Classification** of different mental illnesses including EDs | BERT, RoBERTa and XLNET | F1 score= 0.81 for BERT |
| Jiang et al. (44) | **Other** (LIWC (Used with logistic regression) and BERT representations (Used with an Attentionbased model) | **Classification** of different mental illnesses including EDs | BERT and REALM (76) | F1 score= 0.736 for BERT (post level classification) |
| Zhang et al. (45) | **BERT representations** | Build an annotated dataset for mental illnesses and **Classification** of these illnesses | BERT and MBERT (77). | F1 score= 0.51 for BERT |
| Hwang et al. (46) | **TF-IDF** | **Topic Modeling** (Analyze behavioral patterns of Emotional Eaters) | Stochastic gradient descent based ML model and LDA (Latent Dirichlet Allocation) | F1 = 0.91 |
| Rojewska et al. (47) | **BoW** and Nencki Affective Word List | **Sentiment Analysis** and Emotion Detection | Recurrent Neural Network | – |
| Villegas et al. (48) | K-TVT, BoW, Word2Vec, GloVe and BERT representations | **Classification** (Early detection of signs of anorexia) | Naïve Bayes, Random Forest, Logistic Regression and SVM | F1 = 0.76 for BERT and Naïve Bayes |
| Chancellor et al. (49) | **Other** (Not precise) | **Topic Modeling** (Analyze the lexical variations and changes in pro-ED tags, and perform topic modeling on these tags) | Spectral Clustering algorithm | – |

Additionally, the tasks addressed in these studies can be broadly grouped into categories such as:

- Classification
- Topic modeling
- Sentiment analysis

In terms of feature extraction techniques employed across the 21 studies, a variety of methods were utilized. Among these, three studies (33, 46, 78) relied on TF-IDF. Four studies, including Zhang et al. (16) Benítez-Andrades et al. (38) Villegas et al. (48), and Jiang et al. (44), opted for BERT representations. Notably, Jiang et al. (44) combined BERT with LIWC.

Moreover, Bag of Words (BoW) and various types of Word Embeddings, including GloVe (35, 48), FastText (35), and Word2Vec (35, 36), were widely employed as feature extraction techniques in these studies.

It is pertinent to note that some studies, like Chancellor et al. (79) and Benítez-Andrades et al. (38), did not provide comprehensive details on this aspect in their papers. Conversely, other articles adopted a more personalized approach to construct their features. For instance, some represented each data point as a vector within certain categories (39, 40), while others used rule-based methods (18) or leveraged algorithms like decision trees (41) and topic modeling (42) to determine feature selection.

Our results show that from the 21 studies, 8 make use of classical machine learning methods, 1 uses deep learning, 5 use a combination of classical ML and DL, 4 use large-language models and 3 use other approaches.

When using classical machine learning, some studies compare different methods. For example, López Úbeda et al. (33) apply 5 different supervised machine learning models: SVM, multilayer peceptron classifier, naive bayes, decision tree and logistic regression, and Villegas et al. (48) compare naive bayes, random forest, logistic regression and SVM. Along with the classical machine learning methods, the studies apply different feature representations ranging from Bag of Words (BoW) to TF-IDF (33, 78), up to contextualized embeddings such as BERT (48).

Other studies compared both classical machine learning as well as deep learning methods. For example, in the case of Tébar and Gopalan (42), a so-called feature fusion model that includes both deep learning (a convolutional neural network (CNN) and a BiGRU model), as well as a classical machine learning model (logistic regression classifier with handcrafted features) is used.

For the studies using transformer-based large language models, different models including the BERT (19) model and its variations have been used. For example, Benítez-Andrades et al. (32) applied five variations of the BERT model. The paper from Dinu and Moldovan (43) uses BERT, RoBERTa and XLNET, whereas Jiang et al. (44) use BERT and REALM. The work from Zhang et al. (45) focusing on different mental illnesses used the BERT model, as well as the MBERT variation.

(EvalRQ1) The performance of each study is also reported in Table 3.

(EvalRQ2) Finally, we investigated the limitations of the proposed studies (RQ4) in order to provide a structured outlook for future work in the field.

In many cases, there were limitations in terms of the datasets. For example, Yan et al. (78) cites the limited availability of labeled data. They used a dataset of 50 posts, which they expect to be labeled correctly. Also Zhou et al. (34) mention that their study is limited by the number of collected tweets, which may result in some irrelevant topics arising from noise for their topic modeling task.

In many studies, social media data is used. The nature of such data is seen as a potential limitation for the resulting methods (37). Other studies indicated as a limitation that only one social media platform was used to gather their data (38, 42). For example, a study from (35) points out that their work did not take into account the potential biases in the data that may exist, such as underrepresented population or lack of diverse perspectives. In addition, one of the notable constraints arises from the fundamental disparity between social media data and traditional clinical text data, often used in healthcare and medical research. Clinical records encompass detailed information on patients' medical histories, diagnoses, treatments, and outcomes, rendering them fundamentally distinct from the informal, user-generated content prevalent on social media platforms. Several studies point out that the involvement of clinical professionals would be beneficial. For example, Choudhury (30) states that their method could be more successful with the involvement of clinicians.

Different studies rely on anonymous data, which makes it difficult to ensure a good distribution within the training data over different populations and underrepresented groups. For example, Ragheb et al. (62) sees potential to optimize the model for different use cases and populations. Manual labeling by humans is also considered a source of bias since limited information about the users writing them is available to the annotators. This limited information may not encompass the full context of the users' lives, beliefs, or backgrounds. Annotators may make subjective judgments based solely on the content of the post, which can be influenced by their own biases and interpretations. Thus, limited context can lead to misinterpretations or mislabeling, potentially distorting the research results (38).

In the limitations, it is also discussed how texts written by laypeople and ED promotional[8] and educational materials can be hard to classify (34). This can be partly explained by the short length of texts, for example in the case of tweets, and the semantic similarity of the two types of texts.

Whereas many studies achieved good performance in terms of accuracy or f1-scores, they see a potential limitation in this matter. For example, Wang et al. (40) discusses that the validation was done only with a small sample of the data, and thus further validation is required with larger samples. In another study, the authors were concerned about the problem of overfitting (52).

## 4 Discussion

In this systematic literature survey we have discussed the use of machine learning and natural language processing methods for the detection of eating disorders. Our survey was conducted using the PRISMA framework (17). Our results have shown that many studies focus on the detection of anorexia, or eating disorders in general (see Figure 7). We have also seen that there was more work over the last couple of years, indicating a growing interest in the topic (as shown in Figure 3). Whereas most publications were from institutions in the USA and Spain, work from other countries including Mexico, France and Canada was also identified, as shown in Figure 4. Nevertheless, our work has shown that most research efforts have only been applied to the English language. Given the relevance of local languages for mental health diagnostics and treatment (15), it is thus necessary for future research to address other languages. With regard to the machine learning and feature extraction methods being applied, a comparison turned out to be challenging due to the diverse nature of the datasets and approaches used. The proposed approaches were classified into different categories, including classical machine learning, deep learning, a combination of classical and deep learning, the use of large language models, as well as other approaches. Several studies used f1-score as a common measure, reaching different performances ranging from 0.67 to 0.93. Overall, having a sufficient data quality and quantity was often seen as a major limitation of the approaches. Since 2017, the eRisk challenge has included two tasks pertaining to the early detection of Eating

[8] A content or an activity that promotes or encourages eating disorders (EDs).

Disorders. In both 2018 and 2019, the task involved the early detection of signs of anorexia [see e.g., Losada et al. (26)]. In contrast, the 2022 iteration introduced a novel task centered on measuring the severity of eating disorders (27). This task diverged from the previous ones in that no labeled training data was supplied to participants, meaning that participants could not evaluate the quality of their models' predictions until test time. The objective task was to assess a user's level of eating disorder severity through analysis of their Reddit posting history.

Given the composition of both the eRisk lab and the SMHD dataset (50) predominantly with social media data, it is notable that an overwhelming majority (93%) of the studies in our analysis employ this data type. This underscores the widespread reliance on social media sources in modern research methodologies. This finding confirms the results of Zhang et al. (16) who found that among 399 papers applying NLP methods for the identification of mental health problems, 81% consisted of social media data.

It is worth mentioning that we came across two types of use cases in the studies. Many studies focus on the individual's expression of their behavior and feelings with regard to eating disorders. Some studies, namely Choudhury (30) and Chancellor et al. (49), investigate the wording of pro-anorexia or pro-eating disorders communities on social media and online forums. Such communities promote disordered eating habits as acceptable alternative lifestyles (49). Whereas in many of the studies the technologies target support for clinical professionals, in these cases other applications such as content moderation are in the foreground.

In the realm of data collection for eating disorder research, manual labeling of datasets has been a common approach, with various strategies employed. For instance, Zhang et al. (45) relied on the voluntary efforts of 31 individuals to meticulously annotate 8554 data points encompassing 38 symptoms related to MD (Mental Disorders). Other studies took different routes, combining expert knowledge with input from non-expert annotators[9] (38), or solely relying on domain experts (46). In some cases, researchers have employed machine learning algorithms to automatically annotate their datasets and subsequently validated the results with input from human labelers (44). The majority of datasets underwent annotation by non-expert human annotators, as seen in studies conducted by (79, 40, 34, 41).

Our review revealed few instances of Large Language Models (LLMs) application (10, 11, 19, 30, 38, 43, 44, 45, 49, 50, 61, 67, 73, 74, 79, 80). Despite this, the rising adoption of technologies like MentalBERT (77) and MentaLLama (81), alongside traditional machine and deep learning approaches, is notable. This trend, driven by the impressive efficacy of LLMs in natural language processing, is expected to continue on. As these technologies evolve and become more accessible, we anticipate their increased utilization in this field of research, enhancing computational model accuracy and efficiency.

Based on the identified limitations in the selected studies, we infer the following focus topics that we suggest for future work in the field of using natural language processing and machine learning in ED research:

- Data Quantity and Quality: how can more high-quality data be created and shared, while respecting the ethical and privacy limitations of such sensitive data?
- Involvement of Clinical Professionals: how can machine learning engineers and clinical professionals work together more closely?
- More Diversity in Data: How can the diversity of the population in the used datasets be increased to avoid bias in the classification?
- Local Languages: How can the proposed methods be extended to local languages other than English?

In conclusion, based on the studies investigated in this literature survey, there is potential for further development and in the long-term a novel tool support for clinical professionals based on text data.

## Author contributions

GM: Formal analysis, Writing – review & editing, Writing – original draft, Visualization, Investigation, Data curation. AP: Formal analysis, Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization. MK-B: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors gratefully acknowledge the support of the Inventus Bern Foundation for our research in the field of augmented intelligence for the detection of eating disorders.

## 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.

[9] individuals who lack specialized domain knowledge or expertise in the subject matter.

# References

1. Williams CL, Butcher JN, Paulsen JA. 13 - overview of multidimensional inventories of psychopathology with a focus on the mmpi-2. In: Goldstein G, Allen DN, DeLuca J, editors. *Handbook of Psychological Assessment*, *4th ed*. Academic Press, San Diego (2019). 397–417. doi: 10.1016/B978-0-12-802203-0.00013-4

2. Vandereycken W, Van Humbeeck I. Denial and concealment of eating disorders: a retrospective survey. *Eur Eating Disord Review: Prof J Eating Disord Assoc.* (2008) 16:109–14.

3. Smink FR, van Hoeken D, Hoek HW. Epidemiology, course, and outcome of eating disorders. *Curr Opin Psychiatry*. (2013) 26:543–8. doi: 10.1097/yco.0b013e328365a24f

4. Mohler-Kuo M, Schnyder U, Dermota P, Wei W, Milos G. The prevalence, correlates, and help-seeking of eating disorders in Switzerland. *psychol Med.* (2016) 46:2749–58. doi: 10.1017/S0033291716001136

5. Galmiche M, Déchelotte P, Lambert G, Tavolacci MP. Prevalence of eating disorders over the 2000–2018 period: a systematic literature review. *Am J Clin Nutr.* (2019) 109:1402–13.

6. Rude S, Gortner E-M, Pennebaker J. Language use of depressed and depression-vulnerable college students. *Cogn Emotion*. (2004) 18:1121–33.

7. Merhbene G, Nath S, Puttick AR, Kurpicz-Briki M. Burnoutensemble: Augmented intelligence to detect indications for burnout in clinical psychology. *Front Big Data.* (2022) 4.

8. Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, et al. Feeling bad on facebook: Depression disclosures by college students on a social networking site. *Depress. Anxiety*. (2011) 28:447–55. doi: 10.1002/da.20805

9. Schwartz HA, Eichstaedt J, Kern ML, Park G, Sap M, Stillwell D, et al. (2014). "Towards assessing changes in degree of depression through facebook," In: Resnik P, Resnik R, Mitchell M. editors. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: from linguistic Signal to Clinical Reality*, (Baltimore, Maryland, USA).

10. De Choudhury M, Counts S, Horvitz E. "Predicting postpartum changes in emotion and behavior via social media," In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery. (2013). pp. 3267–76. doi: 10.1145/2470654.2466447

11. De Choudhury M, Counts S, Horvitz EJ, Hoff A. "Characterizing and predicting postpartum depression from shared facebook data". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. New York, NY, USA: Association for Computing Machinery). 2014). pp. 626–38. doi: 10.1145/2531602.2531675

12. Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in reddit social media forum. *IEEE Access*. (2019) 7:44883–93. doi: 10.1109/ACCESS.2019.2909180

13. Morales M, Dey P, Theisen T, Belitz D, Chernova N. "An investigation of deep learning systems for suicide risk assessment." In: Niederhoffer K, Hollingshead K, Resnik P, Resnik R, Loveys K editors *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Minneapolis, Minnesota: Association for Computational Linguistics (2019). p. 177–81. doi: 10.18653/v1/W19-3023

14. Just MA, Pan L, Cherkassky VL, McMakin DL, Cha C, Nock MK, et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat Hum Behav*. (2017) 1:911–9. doi: 10.1038/s41562-017-0234-y

15. Griner D, Smith TB. Culturally adapted mental health intervention: A meta-analytic review. *Psychotherapy: Theory research practice Training.* (2006) 43:531.

16. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital Med*. (2022) 5:46.

17. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*. (2021) 372. doi: 10.1136/BMJ.N71

18. Bellows BK, LaFleur J, Kamauu AWC, Ginter T, Forbush TB, Agbor S, et al. Automated identification of patients with a diagnosis of binge eating disorder from narrative electronic health records. *J Am Med Inform Assoc.* (2014) 21(e1):e163–8. doi: 10.1136/amiajnl-2013-001859

19. Devlin J, Chang M, Lee K, Toutanova K. "BERT: pre-training of deep bidirectional transformers for language understanding." In: Burstein J, Doran C, Solorio T, editors. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019*, vol. 1 (Long and Short Papers). Association for Computational Linguistics (2019). p. 4171–86. doi: 10.18653/V1/N19-1423

20. Mikolov T, Chen K, Corrado G, Dean J. "Efficient estimation of word representations in vector space." In: *International Conference on Learning Representations*. (2013).

21. Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Moschitti A, Pang B, Daelemans W, editors. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar (2014). p. 1532–43. doi: 10.3115/v1/D14-1162

22. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. "Deep contextualized word representations." In Walker M, Ji H, Stent A editors. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics (2018) 2227–37. doi: 10.18653/v1/N18-1202

23. Le QV, Mikolov T. "Distributed representations of sentences and documents." In: Xing EP, Jebara T, editors. *Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research*. Beijing, China: PMLR (2014) 32(2):1188–96. Available at: http://proceedings.mlr.press/v32/le14.pdf.

24. Schuster M, Paliwal K. Bidirectional recurrent neural networks. *IEEE Trans Signal Process.* (1997) 45:2673–81. doi: 10.1109/78.650093

25. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* (2001) 71(2001):2001.

26. Losada DE, Crestani F, Parapar J. "Overview of eRisk: Early Risk Prediction on the Internet." In: Bellot P, Trabelsi C, Mothe J, Murtagh F, Nie JY, Soulier L, et al editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer International Publishing. (2018) 343–61.

27. Parapar J, Martín-Rodilla P, Losada DE, Crestani F. Overview of eRisk 2022: Early risk prediction on the internet. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*. Springer (2022). p. 233–56.

28. Fairburn CG, Beglin SJ. Eating disorder examination questionnaire (ede-q) Database record, APA PsycTests. (1994).

29. Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M. "The state and fate of linguistic diversity and inclusion in the nlp world." In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020) 6282–6293. doi: 10.18653/v1/2020.acl-main.560

30. De Choudhury M. Anorexia on tumblr: A characterization study. In: *Proceedings of the 5th International Conference on Digital Health 2015*. New York, NY, USA: Association for Computing Machinery (2015) 43–50. doi: 10.1145/2750511.2750515

31. Yan H, Fitzsimmons-Craft E, Goodman M, Krauss M, Das S, Cavazos-Rehg P. Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention. *Int J Eating Disord*. (2019) 52:1150–6. doi: 10.1002/eat.23148

32. Benítez-Andrades JA, Alija-Pérez JM, García-Rodríguez I, Benavides C, Alaiz-Moretón H, Vargas RP, et al. BERT model-based approach for detecting categories of tweets in the field of eating disorders (ED). In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)* (New York, USA: IEEE) (2021). p. 586–90. Available at: https://api.semanticscholar.org/CorpusID:236095644.

33. López Úbeda P, Plaza del Arco FM, Díaz Galiano MC, Urena Lopez LA, Martin M. "Detecting anorexia in Spanish tweets." In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd. (2019) 655–63. doi: 10.26615/978-954-452-056-4077

34. Zhou S, Zhao Y, Bian J, Haynos AF, Zhang R. Exploring eating disorder topics on twitter: Machine learning approach. *JMIR Med Inform.* (2020) 8(10):e18273. doi: 10.2196/18273

35. Aguilera J, Hernández Farías DI, Ortega-Mendoza RM, Montes-y-Gómez M. Depression and anorexia detection in social media as a one-class classification problem. *Applied Intelligence.* (2021) 51:6088–103. doi: 10.1007/s10489-020-02131-2

36. Spinczyk D, Bas M, Dzieciatko M, Maćkowski M, Rojewska K, Maćkowska S. Computer-aided therapeutic diagnosis for anorexia. *BioMed Eng OnLine* (2020) 19:53. doi: 10.1186/s12938-020-00798-9

37. Aragón ME, López-Monroy AP, González-Gurrola LC, Montes-y-Gómez M. Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. *EEE Transactions on Affective Computing.* (2021) 14(1):211–22. doi: 10.1109/TAFFC.2021.3075638

38. Benítez-Andrades JA, Alija-Pérez J-M, Vidal M-E, Pastor-Vargas R, Vidal ME, García-Ordás T. Traditional machine learning models and bidirectional encoder representations from transformer (BERT)-based automatic classification of tweets about eating disorders: Algorithm development and validation study. *JMIR Medical Informatics* (2022) 10(2):e34492. doi: 10.2196/34492

39. Ramiandrisoa F, Mothe J. Early Detection of Depression and Anorexia from Social Media: A Machine Learning Approach. In: Cantador I, Chevalier M, Melucci M, Mothe J, editors. *Circle 2020*, vol. 2621 . Proceedings of the Conference CIRCLE 2020, Samatan, France (2020).

40. Wang T, Brede M, Ianni A, Mentzakis E. "Detecting and characterizing eating-disorder communities on social media." In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. New York, NY, USA: Association for Computing Machinery (2017) 91–100. doi: 10.1145/3018661.3018706

41. He L, Luo J. "What makes a pro eating disorder hashtag: Using hashtags to identify pro eating disorder tumblr posts and Twitter users." In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE (2016). 3977–9. doi: 10.1109/BigData.2016.7841081

42. Tébar B, Gopalan A. "Early Detection of Eating Disorders using Social Media." *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, Washington, DC, USA (2021). pp. 193–8. doi: 10.1109/CHASE52844.2021.00042.

43. Dinu A, Moldovan A-C. "Automatic detection and classification of mental illnesses from general social media texts." In: Mitkov R, Angelova G *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd. (2021). p. 358–66. Available at: https://aclanthology.org/2021.ranlp-1.41

44. Jiang Z, Levitan SI, Zomick J, Hirschberg J. Detection of mental health from Reddit via deep contextualized representations. In: Holderness E, Jimeno Yepes A, Lavelli A, Minard A-L, Pustejovsky J, Rinaldi F, et al editors. *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Online: Association for Computational Linguistics (2020) 147–56. doi: 10.18653/v1/2020.louhi-1.16

45. Zhang Z, Chen S, Wu M, Zhu KQ. Symptom identification for interpretable detection of multiple mental disorders. In: Goldberg Y, Kozareva Z, Zhang Y, editos. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics (2022) 9970–85. doi: 10.18653/v1/2022.emnlp-main.677

46. Hwang Y, Kim H, Choi H, Lee J. Exploring abnormal behavior patterns of online users with emotional eating behavior: Topic modeling study. *J Med Internet Res*. (2020) 22:e15700. doi: 10.2196/15700

47. Rojewska K, Maćkowska S, Maćkowski M, Różańska A, Barańska K, Dzieciatko M, et al. Natural language processing and machine learning supporting the work of a psychologist and its evaluation on the example of support for psychological diagnosis of anorexia. *Appl Sci*. (2022) 12. doi: 10.3390/app12094702

48. Villegas MP, Errecalde ML, Cagnina LC. "A comparison of text representation approaches for early detection of anorexia." In: *Memorias del Congreso Argentino en Ciencias de la Computación - CACIC 2021*, Workshop: WBDMD - Base de Datos y MinerÍa de Datos. (2021). 301–10.

49. Chancellor S, Pater JA, Clear T, Gilbert E, De Choudhury M. #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities (New York, NY, USA: Association for Computing Machinery). *CSCW*. (2016) 16:1201–13. doi: 10.1145/2818048.2819963

50. Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In: Bender EM, Derczynski L, Isabelle P editors. *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics (2018). 1485–97. Available at: https://aclanthology.org/C18-1126

51. Wang Y-T, Huang H-H, Chen H-H. "A neural network approach to early risk detection of depression and anorexia on social media text." In: *Conference and Labs of the Evaluation Forum (CLEF)*. Aachen, Germany: CEUR-WS.org (2018). Available at: https://api.semanticscholar.org/CorpusID:51940589

52. Paul S, Jandhyala SK, Basu T. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In: *Conference and Labs of the Evaluation Forum (CLEF)* (2018). Available at: https://api.semanticscholar.org/CorpusID:51942457.

53. Trotzek M, Koitka S, Friedrich CM. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In: *Conference and Labs of the Evaluation Forum (CLEF)*. (2018). Available at: https://api.semanticscholar.org/CorpusID:51939971.

54. Ramiandrisoa F, Mothe J, Benamara F, Moriceau V. IRIT at e-Risk 2018. In: *9th Conference and Labs of the Evaluation Forum, Living Labs (CLEF 2018)*, Avignon, France: CEUR-WS.org. (2018). pp. 1–12. Available at: https://hal.science/hal-02290007.

55. Ortega-Mendoza RM, López-Monroy AP, Franco-Arcega A, Montes-y-Gómez M. PEIMEX at eRisk2018: Emphasizing personal information for depression and anorexia detection. In: *Conference and Labs of the Evaluation Forum (CLEF)*. (2018). Available at: https://api.semanticscholar.org/CorpusID:51939864.

56. Ragheb W, Moulahi B, Azé J, Bringay S, Servajean M. Temporal mood variation: at the CLEF eRisk-2018 tasks for early risk detection on the internet. In: *CLEF 2018 - Conference and Labs of the Evaluation Forum*. Avignon, France. Aachen, Germany: CEUR Workshop Proceedings (2018) 2125(78). Available at: https://hal-lirmm.ccsd.cnrs.fr/lirmm-01989632/file/paper_78.pdf.

57. Liu N, Zhou Z, Xin K, Ren F. TUA1 at eRisk 2018. In: Cappellato L, Ferro N, Nie J, Soulier L, editors. *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018* Aachen, Germany: CEUR-WS.org. CEUR Workshop Proceedings (2018). Available at: https://ceur-ws.org/Vol-2125/paper_121.pdf.

58. Ramírez-Cifuentes D, Freire A. UPF's participation at the clef eRisk 2018: Early risk prediction on the internet. In: *Conference and Labs of the Evaluation Forum (CLEF)*. (2018).

59. Funez DG, Ucelay MJG, Villegas MP, Burdisso SG, Cagnina LC, Montes-y-Gómez M, et al. UNSL's participation at eRisk 2018 lab. In: *Conference and Labs of the Evaluation Forum (CLEF)*. Aachen, Germany: CEUR-WS.org (2018). Available at: https://api.semanticscholar.org/CorpusID:198489135.

60. Aragón ME, Lopez-Monroy AP, Montes-y-Gómez M. (2019). "INAOE-CIMAT at eRisk 2019: Detecting Signs of Anorexia using Fine-Grained Emotions," in: *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. Aachen, Germany: CEUR-WS.org. Available at: https://api.semanticscholar.org/CorpusID:198489135.

61. Burdisso SG, Errecalde ML, Montes-y-Gómez M. UNSL at eRisk 2019: a Unified Approach for Anorexia, Self-harm, and Depression Detection in Social Media. In: Cappellato L, Ferro N, Losada DE, Muller H. *Conference and Labs of the Evaluation Forum (CLEF)*. Aachen, Germany: CEUR-WS.org (2019). Available at: https://api.semanticscholar.org/CorpusID:198490018.

62. Ragheb W, Azé J, Bringay S, Servajean M. Attentive Multi-stage Learning for Early Risk Detection of Signs of Anorexia and Self-harm on Social Media. In: Cappellato L, Ferro N, Losada DE, Müller H, editors. *CLEF 2019 - Conference and Labs of the Evaluation Forum*, vol. 2380 . CEUR Workshop Proceedings, Lugano, Switzerland (2019).

63. Fano E, Karlgren J, Nivre J. "Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019." Vol. 2380. Cappellato L, Ferro N, Losada DE, Muller H. editors. Vol. 2380. Aachen, Germany: CEUR Workshop Proceedings (2019).

64. Masood R, Ramiandrisoa F, Aker A. "UDE at eRisk 2019: Early risk prediction on the internet." In: Cappellato L, Ferro N, Losada DE, Müller H, editors. *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019* Aachen, Germany: CEUR Workshop Proceedings (2019). 2380.

65. Naderi N, Gobeill J, Teodoro D, Pasche E, Ruch P. A baseline approach for early detection of signs of anorexia and self-harm in reddit posts, in: *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*, Lugano, Switzerland: CEUR-WS.org. (2019). CEUR Workshop Proceedings.

66. Mohammadi E, Amini H, Kosseim L. Quick and (maybe not so) easy detection of anorexia in social media posts. *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes* (2019). CEUR Workshop Proceedings.

67. Plaza del Arco FM, López-Úbeda P, Díaz-Galiano MC, Ureña López LA, Martín Valdivia MT. "Integrating UMLS for Early Detection of Signs of Anorexia," In: *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. Aachen, Germany: CEUR-WS.org. (2019). Available at: https://api.semanticscholar.org/CorpusID:198489706.

68. Ranganathan A, Haritha A, Thenmozhi D, Aravindan C. "Early detection of anorexia using rnn-lstm and svm classifiers," In: *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. Aachen, Germany: CEUR-WS.org. (2019). Available at: https://api.semanticscholar.org/CorpusID:198488874.

69. Ferdowsi S, Knafou J, Borissov N, Vicente Alvarez D, Mishra R, Amini P, et al. Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study. *Patterns*. (2023) 4:100689. doi: 10.1016/j.patter.2023.100689

70. Trifan A, Oliveira JL. (2019). "BioInfo@UAVR at eRisk 2019: Delving into Social Media Texts for the Early Detection of Mental and Food Disorders," In: *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. Aachen, Germany: CEUR-WS.org. Available at: https://api.semanticscholar.org/CorpusID:198488663.

71. Ortega-Mendoza RM, Irazú D, Farías H, Montes-Y-Gómez M. "LTL-INAOE's Participation at eRisk 2019: Detecting Anorexia in Social Media through Shared Personal Information." In: Cappellato L, Ferro N, Losada DE, Müller H. editors. *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings* Vol. 2380. Lugano, Switzerland, September 9-12, 2019. CEUR-WS.org (2019). Available at: https://ceur-ws.org/Vol-2380/paper_75.pdf.

72. Hosseini Saravani SH, Normand L, Maupomé D, Rancourt F, Soulas T, Besharati S, et al. Measuring the severity of the signs of eating disorders using similarity-based models. *CLEF (Working Notes)* (2022). 936–46.

73. Mármol-Romero AM, Jiménez-Zafra SM, Plaza-Del-Arco FM, Molina-González MD, Martín-Valdivia M-T, Montejo-Ráez A. "SINAI at eRisk@CLEF 2022: Approaching Early Detection of Gambling and Eating Disorders with Natural Language Processing." In: Faggioli G, Ferro N, Hanbury A, Potthast M editors. *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings* (2022) . Bologna, Italy. September 5th - to - 8th, 2022. CEUR-WS.org. 3180:961–71. Available at: https://ceur-ws.org/Vol-3180/paper-76.pdf.

74. Srivastava H, Lijin NS, Sruthi S, Basu T. "Exploring the potential of bag of words, document embeddings and transformer based framework for early prediction of eating disorder, depression and pathological gambling over social media." In: Faggioli G, Ferro N, Hanbury A, Potthast M, editors. *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th* Aachen, Germany: CEUR Workshop Proceedings (2022). p. 972–86.

75. Li W, Han J, Pei J. CMAR: accurate and efficient classification based on multiple classassociation rules, in: *Proceedings 2001 IEEE International Conference on Data Mining*. San Jose, CA, USA: IEEE. (2001). 369–76. doi: 10.1109/ICDM.2001.989541

76. Guu K, Lee K, Tung Z, Pasupat P, Chang M-W. REALM: retrieval-augmented language model pre-training. In: *Proceedings of the 37th International Conference on Machine Learning. ICML'20. JMLR.org*. (2020).

77. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, et al editors. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association (2022). 7184–90. Available at: https://aclanthology.org/2022.lrec-1.778.

78. Yan H, Phd EEF-C, Goodman M, Krauss M, Das S, Cavazos-Rehg P. (2019). doi: 10.1002/eat.23148

79. Chancellor S, Kalantidis Y, Pater JA, De Choudhury MD, Shamma DA. Multimodal Classification of Moderated Online Pro-Eating Disorder Content, In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery. (2017). 3213–26. doi: 10.1145/3025453.3025985

80. Burdisso SG, Errecalde M, Montes-y-Gómez M. A text classification framework for simple and effective early depression detection over social media streams. *Expert Syst Appl*. (2019) 133:182–97. doi: 10.1016/j.eswa.2019.05.023

81. Yang K, Zhang T, Kuang Z, Xie Q, Ananiadou S. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv*. (2023) arXiv preprint arXiv:2309.13567.

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 10 September 2024
DOI 10.3389/fpsyt.2024.1181739

Check for updates

OPEN ACCESS

EDITED BY
Farhad Soleimanian Gharehchopogh,
Hacettepe University, Türkiye

REVIEWED BY
Zhihao Ma,
Nanjing University, China
André Luiz Monezi Andrade,
Pontifical Catholic University of Campinas,
Brazil

*CORRESPONDENCE
Ava J. C. Mason
✉ ava.mason.20@ucl.ac.uk

RECEIVED 07 March 2023
ACCEPTED 14 August 2024
PUBLISHED 10 September 2024

CITATION
Mason AJC, Bhavsar V, Botelle R, Chandran D, Li L, Mascio A, Sanyal J, Kadra-Scalzo G, Roberts A, Williams M and Stewart R (2024) Applying neural network algorithms to ascertain reported experiences of violence in routine mental healthcare records and distributions of reports by diagnosis. *Front. Psychiatry* 15:1181739.
doi: 10.3389/fpsyt.2024.1181739

# Applying neural network algorithms to ascertain reported experiences of violence in routine mental healthcare records and distributions of reports by diagnosis

Ava J. C. Mason[1]*, Vishal Bhavsar[1,2], Riley Botelle[1], David Chandran[1], Lifang Li[1], Aurelie Mascio[1], Jyoti Sanyal[2], Giouliana Kadra-Scalzo[1], Angus Roberts[1], Marcus Williams[2,3] and Robert Stewart[1,2]

[1]King's College London Institute of Psychiatry, Psychology and Neuroscience, De Crespigny Park, London, United Kingdom, [2]Biomedical Research Centre, South London and Maudsley National Health Service (NHS) Foundation Trust, London, United Kingdom, [3]Sandwell and West Birmingham Hospitals National Health Service (NHS) Trust, West Bromwich, United Kingdom

**Introduction:** Experiences of violence are important risk factors for worse outcome in people with mental health conditions; however, they are not routinely collected be mental health services, so their ascertainment depends on extraction from text fields with natural language processing (NLP) algorithms.

**Methods:** Applying previously developed neural network algorithms to routine mental healthcare records, we sought to describe the distribution of recorded violence victimisation by demographic and diagnostic characteristics. We ascertained recorded violence victimisation from the records of 60,021 patients receiving care from a large south London NHS mental healthcare provider during 2019. Descriptive and regression analyses were conducted to investigate variation by age, sex, ethnic group, and diagnostic category (ICD-10 F chapter sub-headings plus post-traumatic stress disorder (PTSD) as a specific condition).

**Results:** Patients with a mood disorder (adjusted odds ratio 1.63, 1.55-1.72), personality disorder (4.03, 3.65-4.45), schizophrenia spectrum disorder (1.84, 1.74-1.95) or PTSD (2.36, 2.08-2.69) had a significantly increased likelihood of victimisation compared to those with other mental health diagnoses. Additionally, patients from minority ethnic groups (1.10 (1.02-1.20) for Black, 1.40 (1.31-1.49) for Asian compared to White groups) had significantly higher likelihood of recorded violence victimisation. Males were significantly less likely to have reported recorded violence victimisation (0.44, 0.42-0.45) than females.

**Discussion:** We thus demonstrate the successful deployment of machine learning based NLP algorithms to ascertain important entities for outcome prediction in mental healthcare. The observed distributions highlight which sex, ethnicity and diagnostic groups had more records of violence victimisation. Further development of these algorithms could usefully capture broader experiences, such as differentiating more efficiently between witnessed, perpetrated and experienced violence and broader violence experiences like emotional abuse.

KEYWORDS

natural language processing, victimisation, mental health records, CRIS, violence

## Introduction

Interpersonal violence is defined as threatened or actual use of physical force or power against another person, involving one or more perpetrators and victims (1). Violence can be categorised in a variety of ways (e.g., physical, sexual, emotional, domestic) but all cause significant physical and mental morbidity within general populations (2–4). Individuals with a severe mental illness have been found to be significantly more likely to experience domestic, physical, and sexual violence compared to the general population (5–8). Despite this, data on violence (all forms) has been inadequately available from healthcare records. This is partly due to the lack of routine enquiry by professionals at points of clinical contact, and partly because instances of violence are difficult to identify in healthcare data in the absence of specific coding systems (9, 10).

Inconsistencies are also present between different mental health services. For instance, individuals from inpatient settings are more likely to have structured data collected on violent incidents, although the form of data collection also varies depending on the type of violence experienced (11). Electronic healthcare records data could help researchers and clinicians understand the occurrence of interpersonal violence (when disclosed), its risk factors, and the level of treatment and support provided. However, research has focused mainly on recorded incidents within inpatient settings, such as using specific violence definitions to examine the prevalence of recorded experiences of physical assault (12). Because most instances are likely to be recorded as unstructured text data, violence experiences across mental healthcare settings cannot be adequately captured without natural language processing (NLP).

A general challenge for using health records data for research is that the most valuable and granular information is frequently contained in text fields (e.g., routine case notes, clinical correspondence) rather than in pre-structured fields; this includes mentions of violence whether experienced as a victim or perpetrated. NLP has been used increasingly to extract information automatically from unstructured text in electronic health records, particularly in mental healthcare, on clinical entities such as diagnosis, symptoms, and treatment (11–14). However, few of these studies have applied NLP to investigate mentions of violence across different clinical samples. One study using NLP reported greater odds of physical victimisation within groups who had an ICD-10 diagnosis of F2x (schizophrenia, schizotypal and delusional disorder), F6x (disorders of adult personality and behaviour), F7x (mental retardation) and F3x (mood disorders) diagnostic groups vs those with an organic syndrome. However, this was specifically examined within an inpatient setting, where victimisation would be expected to be mentioned more regularly than in outpatient samples (11). Another study using NLP found individuals with victimisation to be most commonly diagnosed with psychotic disorders (20.4%) or mood disorders (16.3%) (12). However, this study specifically investigated physical victimisation, rather than other types of experienced victimisation. From these findings, it could be suggested that physical victimisation may be more prevalent in individuals with a diagnosis of a psychotic or mood disorder.

An NLP approach was previously developed to ascertain violence according to its presence, agent (i.e., patient as perpetrator or victim) and certain subtypes (physical, domestic, sexual) (15). This method provided a potential way of furthering research on how professionals and services respond to violence, as well as provide opportunities for monitoring recorded violence victimisation in different groups (16). For example, one application of these NLP algorithms included their use in a study investigating associations of victimisation with adverse mental healthcare outcomes during the early stages of the COVID-19 pandemic (17). Having run these previously developed algorithms across a large mental healthcare data resource, we sought to describe the distribution of interpersonal violence ascertained in this way across different psychiatric settings and diagnostic groups. The output presented here examined the distribution of any recorded violence victimisation, with secondary analysis examining the distribution of specific victimisation types: physical, domestic, and sexual violence. This was a descriptive study testing victimisation seeking primarily to estimate the prevalence of recorded victimisation using the aforementioned NLP algorithm across a large mental health resource. Therefore, we did not have specific hypotheses relating to which diagnostic groups would have higher prevalence of specific victimisation types. However, it was anticipated from the previous studies mentioned, that physical victimisation may be higher in patients with an ICD-10 diagnosed psychotic or mood disorder (F2x, F3x).

## Materials and methods

The study reported in this paper analysed information about violence extracted from the English language text portions of a de-identified secondary care psychiatric electronic health record (EHR), from the South London and Maudsley NHS Foundation Trust. The text consisted of a mix of document types from several EHR fields, including correspondence between clinicians, event notes written by clinicians in day-to-day clinical care, and discharge summaries (18, 19).

### Extracting violence from mental health records text using NLP algorithms

The method by which violence information was extracted from the text is in routine at the UK's National Institute for Health Research Maudsley Biomedical Research Centre, where it is regularly run over the dataset. The full method, and its evaluation, has been previously reported (15). We provide a summary here for convenience.

As a first step, a list of violence-related keywords based on literature, clinical experience and informatics expertise was created. Seventeen keywords were assembled in this respect. Next, a technique called sequence classification, a common sub-task in NLP, was implemented. This involves obtaining text sequences that contain one of the listed violence-related key words, and manually labelling them as being indicative or not of five binary classes (a mention of victimisation, perpetration of violence, or general mention (as victim, witness or perpetrator) of domestic violence, physical violence or sexual violence) by multiple annotators. Guidelines were then developed on how to annotate further text sequences based on discussions with these annotators of their experiences (e.g., what text sequences would be more indicative of victimisation vs other text sequences). Inter-annotator agreement was estimated on a subset of the data labelled, giving agreements in this case of 82%-96%, and Cohen's kappa coefficients of 60%-85% (15). As previously stated, the selection of keywords, labelling guidelines, characteristics of the labelled text and the labelling process are fully described in a previous (open access) publication (15). After measuring inter-annotator agreement on what would be classed as one of the five binary classes, separate binary classification models were trained from the labelled data, one for each of these five classes. Models were built by adapting a widely used transformer model (a type of neural network model), BioBERT (20). BioBERT was adapted using the Hugging Face Bert-For-Sequence-Classification interface (21) adding a single classification layer to the standard transformer model. Cross entropy loss with custom weight parameters was used to account for dataset imbalance. Each model created in this way classifies a text sequence as being a member or not a member of a class, such as physical violence or domestic violence. We refer to these as "instance level" mentions of violence. These instance level text sequences are derived from documents, such as clinician notes. The final algorithm labels any document that contains one or more text sequence instance of a given class with that same class, thus creating a "document level" label. For example, if a document contains two sequences labelled as being in the physical violence class, and three sequences in the domestic violence class, then the document will be labelled as being in the physical violence and domestic violence classes. As documents are written about and linked to patients, we are then able to draw conclusions about those patients. Blind testing of the final NLP algorithms on 1411 random documents gave document level F1 statistics of 0.90, 0.85, 0.98, 0.93, and 0.93 for victimisation, perpetration, physical, domestic, or sexual attributes respectively (15).

### Data resource

As with the NLP development, data for the analyses presented here were extracted from the case register of the South London and Maudsley NHS Foundation Trust (SLaM). SLaM is a large secondary care mental healthcare provider, serving around 1.3 million residents of a defined catchment of four London boroughs (Croydon, Lambeth, Lewisham, and Southwark). SLaM care covers all specialist mental health care, including liaison and crisis teams, community and inpatient services and early intervention services. Electronic health records (EHRs) have been used for all SLaM services since 2006, and the Maudsley Clinical Record Interactive Search (CRIS) platform was established in 2008 in order to retrieve de-identified data from records of patients previously or currently receiving SLaM care (18). The EHR source includes structured fields coding demographic information (e.g., ethnicity, sex, age), and unstructured free text fields from case notes, mental health examinations, personal histories, management plans and correspondence. Within the last decade, a range of NLP algorithms have been developed, whose detailed performance data and descriptions can be found in an open-access catalogue (22). CRIS has a robust, patient-led governance and data security model and has approval as a data resource for secondary analysis (Oxford Research Ethics Committee C, reference 18/SC/0372).

### Analysed sample

For the analyses within this paper, data were extracted for all individuals receiving SLaM services at any point during 2019, defining their demographic and diagnostic status on or as closest as possible to an index date of 1$^{st}$ July 2019 and ascertaining any recorded violence victimisation from the full record up to the end of 2019. The NLP algorithm can assess for a mention of violence victimisation, but it cannot accurately indicate the frequency with which that victimisation has occurred (e.g., three mentions of victimisation in different documents highlighted by the NLP could refer to the same event). As this study is interested in whether individuals have a mention of recorded victimisation in general, patients were classified within two groups based on whether they had one or more mention of recorded violence victimisation in any free text fields occurring within the study period. Records describing the violence victimisation were then further evaluated for the presence or not of physical, domestic, or sexual violence. Because the violence app in its current version does not identify the intersection of violence type and violence victimisation specifically at instance level, performance was re-checked by extracting documents to analyse accordance of each recorded victimisation and type combination. Based on 50 randomly selected positive instances for each, evaluated for the analyses presented in this report, the precision statistics for victimisation for physical violence, domestic violence, and sexual violence were 0.72, 0.72 and 0.62 respectively.

## Measurements

Demographic variables extracted were age, sex, and ethnicity. Age at the index date was categorised and entered in 10-year increments. Ethnicity was categorised into six groups for analysis compiled using census categories (23): ‘Asian’ (Indian, Bangladeshi, Pakistani, Chinese or any other Asian background), ‘Black’ (Caribbean, African or any other black background), ‘White British’ (British), ‘White other’ (Irish or any other white background), ‘Other/mixed’ (White and Asian, White and Black Caribbean, White and Black African, any other ethnic group) and ‘Not stated’. Diagnoses are coded in structured fields in the source record according to the International Classification of Diseases, 10th Edition (ICD-10). Participants were categorised by ICD-10 codes (24) for primary diagnosis (recorded closest to 01.07.2019) as follows: F0x (organic mental disorders), F1x (psychoactive substance use), F2x (schizophrenia, schizotypal and delusional disorders), F3x (mood disorders), F4x (neurotic, stress-related and somatoform disorders), F5x (behaviour syndromes associated with physiological and physical factors), F6x (disorders of adult personality and behaviour), F7x (mental retardation), F8x (disorders of psychological development), F9x (behavioural and emotional disorders with onset during childhood and adolescence), ‘unspecified’ and ‘no axis 1’. In addition, post-traumatic stress disorder (PTSD; F43.1) was ascertained as an individual disorder of interest.

## Statistical analysis

All analysis was conducted in R (version 4.1.2) using various packages (readr (25); dplyr (26); ggplot2 (27);). Descriptive statistics (means, standard deviations, frequencies, and percentages) of age, sex, ethnicity, and victimisation mentions were provided. Patients without any of the sociodemographic data were excluded from analysis. Chi square tests were also conducted to investigate victimisation differences between different demographic groups (age, sex, ethnicity) and diagnostic groups, supplemented by Cramér’s V effect sizes. These results were reported for any recorded violence victimisation, as well as specifically for domestic, physical, and sexual victimisation. Logistic regression analysis was conducted to investigate whether being part of a specific diagnostic group predicted mention of any recorded violence victimisation. Diagnostic groups were defined as separate binary variables for each diagnosis, (e.g., F0x diagnoses vs all other categories). Unadjusted models assessed age, sex, ethnicity, and each binary diagnostic group comparison in relation to presence or not of recorded experiences of recorded violence victimisation. Each of these models (for each sociodemographic variable and separate diagnostic group comparison) was then adjusted for age, sex, and ethnicity. The adjusted models were also conducted within males and female subsamples independently. For secondary analysis, unadjusted and adjusted regressions were conducted to measure whether being part of a specific diagnostic group predicted mention of physical, domestic, and sexual victimisation specifically. Bonferroni correction was used to adjust for multiple comparisons, whereby the alpha value was lowered to account for the number of comparisons performed (0.05 divided by number of tests conducted). P values from the regression analysis were considered significant if they were lower than the adjusted value. Multicollinearity tests using the R function vif() within the [car package] were undertaken to avoid issues with overlapping predictor variables. The predictor of being of age 91-100 was not added to the adjusted regressions, as it was highly correlated with other predictor age groups (with a VIF value above five (28)).

# Results

We present results of the violence prevalence analysis based on information extracted using NLP. A full evaluation of the NLP itself can be found in the previously published paper (15). The cohort comprised 60,021 individuals: 56,482 with a F0-F9 diagnosis, 3527 with an unspecified disorder and 12 with no axis 1 disorder recorded. Of the 56,482 individuals with a F0-F9 diagnosis, there were 27,191 (46.3%) with at least one victimisation mention: 26,038 (46.1%) with a mention of physical violence, 22,396 (39.7%) with domestic violence, and 13,558 (24.0%) with sexual violence. The mean (SD) age of the cohort was 37.6 (20.4) years. Distribution frequencies and Chi squared test results for associations with demographic variables and diagnostic group can be found in Table 1. Age, sex, ethnicity, and diagnostic group were all significantly associated with any victimisation, physical, domestic, and sexual victimisation mentions. For age groups, violence prevalence showed an inverted-U-shaped pattern of association with highest proportions in the 41-60y groups for all types. All victimisation types were more commonly recorded in women than men. For ethnicity, the highest prevalence of overall victimisation was within the Black ethnic group (62.3%), which was also observed for recorded physical and sexual violence victimisation specifically, but the highest prevalence of domestic victimisation was in the Other/Mixed group. For diagnostic groups, overall recorded violence victimisation prevalence was highest in patients with schizophrenia and related disorders (F2x) or personality disorders (F6x), the same being observed for physical violence. Recorded domestic and sexual violence victimisation prevalence were highest in those with personality disorder diagnoses. Considering effect sizes, as quantified by Cramér’s V statistic, these were moderate (0.2-0.6) for ethnicity and diagnosis and small (<0.2) for age and sex. Most did not vary substantially by violence category apart from sex which had higher effect sizes for domestic and sexual than physical violence, and ethnicity which was strongest for physical violence.

TABLE 1 Distribution frequencies (N(%)) and chi square test statistics measuring group differences in recorded violence victimization or specific physical, domestic, or sexual victimisation in 2019 for each age category, sex, ethnicity, and diagnostic group.

| Predictor variable | All patients | Any victimisation | | Physical | | Domestic | | Sexual | |
|---|---|---|---|---|---|---|---|---|---|
| | | N (%) | $X^2$ (V) | N (%) | $X^2$ (V) | N (%) | $X^2$ (V) | N (%) | $X^2$ (V) |
| **Age** | | | 2118.6* (0.19) | | 2023* (0.18) | | 1603.9* (0.16) | | 1928.8* (0.18) |
| 0-10 years | 3220 | 981(30.47) | | 875(27.17) | | 825(25.62) | | 241(7.48) | |
| 11-20 years | 11457 | 5289(46.16) | | 5020(43.81) | | 4493(39.22) | | 2154(18.80) | |
| 21-30 years | 11039 | 5115(46.34) | | 4908(44.46) | | 4365(39.54) | | 2754(24.95) | |
| 31-40 years | 10232 | 5265(51.46) | | 4990(48.77) | | 4543(44.40) | | 2871(28.06) | |
| 41-50 years | 8225 | 4600(55.93) | | 4450(54.10) | | 3713(45.14) | | 2526(30.71) | |
| 51-60 years | 7317 | 4249(58.07) | | 4073(55.66) | | 3325(45.44) | | 2231(30.49) | |
| 61-70 years | 3456 | 1654(47.86) | | 1619(46.85) | | 1234(35.71) | | 805(23.29) | |
| 71-80 years | 2808 | 814(28.99) | | 783(27.88) | | 596(21.23) | | 298(10.61) | |
| 81-90 years | 2267 | 440(19.41) | | 441(19.45) | | 321(14.16) | | 86(3.79) | |
| **Sex** | | | 563.45* (0.10) | | 397.02* (0.08) | | 1980.7* (0.18) | | 1640.1* (0.17) |
| Female | 29823 | 15567(52.20) | | 14710(49.32) | | 14294(47.93) | | 9036(30.29) | |
| Male | 30198 | 12840(42.52) | | 12449(41.22) | | 9121(30.21) | | 4930(16.33) | |
| **Ethnic group** | | | 4097.4* (0.26) | | 4231.1* (0.27) | | 2870.3* (0.22) | | 2058.2* (0.19) |
| White British | 22317 | 11413(51.14) | | 10896(48.83) | | 9570(42.88) | | 5799(25.98) | |
| White Other | 4586 | 2302(50.19) | | 2209(48.17) | | 1940(42.30) | | 1108(24.16) | |
| Black | 11433 | 7120(62.28) | | 6967(60.94) | | 5596(48.95) | | 3706(32.41) | |
| Asian | 2955 | 1586(53.67) | | 1536(51.98) | | 1270(42.98) | | 699(23.65) | |
| Other/Mixed | 5100 | 2954(57.92) | | 2802(54.94) | | 2574(50.47) | | 1485(29.12) | |
| Not Stated | 12630 | 3032(24.01) | | 2749(21.77) | | 2456(19.45) | | 1169(9.26) | |
| **Diagnostic group** | | | 6182.4* (0.32) | | 6365.3* (0.33) | | 4548.6* (0.28) | | 5168* (0.29) |
| F0-F09 | 4287 | 842(19.65) | | 839(19.57) | | 593(13.83) | | 209(4.88) | |
| F10-F19 | 5560 | 2527(45.45) | | 2360(42.45) | | 1938(34.86) | | 1081(19.44) | |
| F20-F29 | 7212 | 5467(75.81) | | 5433(75.33) | | 3980(55.19) | | 3003(41.64) | |
| F30-F39 | 7166 | 4119(57.48) | | 3932(54.87) | | 3737(52.15) | | 2182(30.45) | |
| F40-F49 | 8296 | 3997(48.18) | | 3814(45.97) | | 3486(42.02) | | 2042(24.61) | |
| F50-F59 | 1878 | 721(38.39) | | 664(35.36) | | 666(35.46) | | 368(19.60) | |
| F60-F69 | 2381 | 1972(82.82) | | 1909(80.18) | | 1808(75.93) | | 1442(60.56) | |
| F70-F79 | 787 | 412(52.35) | | 423(53.75) | | 271(34.43) | | 188(23.89) | |
| F80-F89 | 3273 | 1225(37.43) | | 1163(35.53) | | 953(29.12) | | 388(11.85) | |
| F90-F98 | 15642 | 5909(37.78) | | 5501(35.17) | | 4964(31.74) | | 2655(16.97) | |
| Unspecified | 3527 | 1211(34.34) | | 1114(31.59) | | 1013(28.72) | | 404(11.45) | |
| No axis 1 | 12 | 5(41.67) | | 7(58.33) | | 6(50.00) | | 4(33.33) | |

*All p-values <.01; Cramér's V effect size provided.

For overall recorded violence victimisation, results from unadjusted and adjusted logistic regression models are displayed in Table 2. In adjusted models, the same mid-life peaks in age distribution were observed as in unadjusted analyses, as were associations with female sex and with Black, Asian, and Other/Mixed ethnic groups compared to the White British reference. Additional analysis conducted in males and females separately found few differences between the sex of patients (Supplementary

TABLE 2 Unadjusted and fully adjusted logistic regression models for having at least one record of violence victimisation (any type) in 2019.

| | Unadjusted | Fully adjusted |
|---|---|---|
| Predictor | OR(95% CI) | OR(95% CI) |
| Age group | | |
| 0-10 years | __0.50(0.47-0.55)__** | __0.43(0.39-0.47)__** |
| 11-20 years | 0.99(0.47-0.55) | __0.81(0.77-0.86)__** |
| 21-30 years | Reference group | |
| 31-40 years | __1.23(1.16-1.30)__** | 1.13(1.07-1.20)** |
| 41-50 years | __1.47(1.39-1.56)__** | __1.33(1.25-1.41)__** |
| 51-60 years | __1.60(1.51-1.70)__** | __1.36(1.28-1.45)__** |
| 61-70 years | 1.06(1.51-1.70) | __0.88(0.81-0.96)__** |
| 71-80 years | __0.47(0.43-0.52)__** | __0.36(0.33-0.40)__** |
| 81-90 years | __0.28(0.25-0.31)__** | __0.20(0.18-0.23)__** |
| Sex | | |
| Female | Reference group | |
| Male | __0.68(0.66-0.70)__** | __0.64(0.62-0.65)__** |
| Ethnic group | | |
| White British (%) | Reference group | |
| White Other (%) | 1.05(0.99-1.12) | 1.02(0.95-1.08) |
| Black (%) | __1.72(0.99-1.12)__** | __1.73(1.65-1.82)__** |
| Asian (%) | __1.21(1.12-1.30)__** | __1.22(1.12-1.32)__** |
| Other/Mixed (%) | __1.44(1.35-1.53)__** | __1.41-1.32-1.50)__** |
| Not Stated (%) | __0.33(0.31-0.35)__** | __0.30(0.28-0.31)__** |
| Diagnostic group | | |
| F0-F09 | __0.25(0.23-0.27)__** | __0.32(0.29-0.35)__** |
| F10-F19 | 0.92(0.87-0.97)** | __0.63-0.60-0.67)__** |
| F20-F29 | __4.08(3.86-4.32)__** | __3.19(3.00-3.40)__** |
| F30-F39 | __1.59(1.51-1.67)__** | __1.42(1.35-1.50)__** |
| F40-F49 | 1.04(0.99-1.09) | 0.95(0.90-1.00)* |
| PTSD | __4.84(4.19-5.62)__** | __4.21(3.62-4.91)__** |
| F50-F59 | __0.69(0.62-0.75)__** | __0.50(0.45-0.55)__** |
| F60-F69 | __5.69(5.12-6.35)__** | __4.66(4.17-5.22)__** |
| F70-F79 | 1.23(1.07-1.41)** | 0.87(0.75-1.00) |
| F80-F89 | __0.65(0.6-0.70)__** | __0.80-0.63-0.86)__** |
| F90-F98 | __0.59(0.56-0.61)__** | __0.72(0.69-0.75)__** |
| Unspecified | 0.79(0.24-2.49) | 0.51(0.15-1.67) |
| No axis 1 | __0.56(0.52-0.60)__** | __0.68(0.63-0.74)__** |

OR, odds ratio; CI, Confidence intervals. *p<.05, **p<.01. Bold: significant results after controlling for multiple comparisons (0.05/28 tests conducted, new p level=0.00179). Diagnostic group effect sizes: odds of victimisation among those with the diagnoses compared to all other patients.
Adjusted models controlled for age, ethnicity, and sex.

Table 1). For diagnoses, when analysed individually against all other diagnostic groups, significantly higher odds of recorded violence victimisation were observed in patients with schizophrenia and related disorders (F2x), affective disorders (F3x), PTSD, and personality disorders (F6x). In secondary analyses of specific violence types, findings were similar for physical and domestic violence (Supplementary Tables 2, 3, respectively). Findings for sexual violence differed in that no association was found with Asian ethnic groups compared to the White British reference; they were similar in all other respects (Supplementary Table 4).

# Discussion

To our knowledge, this is the first application of NLP algorithms to characterise recorded violence in a large corpus of mental health electronic health records. Considering distribution, violence was most commonly recorded in mid-life age groups, in women compared to men, in patients from minority ethnic groups compared to White groups, and among people diagnosed with schizophrenia and related disorders, affective disorders, PTSD and personality disorders, compared to those with other diagnoses.

The reported prevalence of violence in individuals with a severe mental illness has varied between 4% to 35% (5),, with prevalence of violence in patients with a general mental disorder being 15.2% (compared to 6.9% in those without) (29). Physical, domestic, and sexual violence were recorded in 46%, 40% and 24% of our sample of individuals with a diagnosis of a F0-F9 disorder. These absolute levels should be viewed cautiously in light of the performance levels of the algorithms, which we intend to develop further to improve characterisation accuracy. In particular, it should be borne in mind that status combinations (i.e., between 'victimisation' and each violence type) could only be applied at document level. It was therefore conceivable that the victimisation status applied to a different experience of violence in the same document (e.g., sexual violence might have been recorded as a perpetration event in the same document as physical violence received as victimisation, resulting in a false positive ascertainment for recorded sexual violence victimisation). Sub-optimal precision (positive predictive value) will have resulted in an over-estimation of exposure due to false positive instances, while sub-optimal recall (sensitivity) will have resulted in an under-estimation of exposure due to missed instances. Under-estimation will also clearly result from failure to ascertain or record experiences of violence in the source clinical record. Despite this, the associations with demographic and clinical factors, in the directions anticipated, support the applicability of these algorithms, at least as proxy markers of exposure, for analysis over large datasets, even if the performance levels achieved to date do not yet support their use for individual clinical decision support. Importantly, to our knowledge, there are currently no adequate means for quantifying recorded violence victimisation in mental healthcare records (or clinical records for any specialty), so we feel that the approach here at least represents a step towards more inclusive data capture. Relatively high prevalence of recorded violence is consistent with the 17% prevalence for any victimisation ascertained in case notes from a shorter (3-month) period early in the COVID-19 pandemic, a feature that was found to be prospectively associated with increased risk of acute care, emergency referrals, and mortality (30).

Recorded violence was ascertained most frequently in people with diagnoses of schizophrenia and related disorders, affective disorders, PTSD, and personality disorders. The vulnerability of patients within these diagnostic groups to experiences of interpersonal violence has been strongly supported in previous literature

(31–34). Therefore, our results support the notion of having increased screening (for all victimisation types) and victimisation support for these vulnerable groups. Unexpectedly, patients diagnosed with an organic disorder, substance misuse disorder, stress disorder (excluding PTSD), developmental disorder or a disorder with physiological disturbances had significantly less victimisation mentions than other disorders. Previous research has found at least some of these disorders to be risk factors for victimisation, such as research reporting higher rates of victimisation with a substance use disorder compared to those without (32). However, the observed low effect sizes may suggest that disorders such as schizophrenia should be considered a stronger risk factor. In interpreting these findings, it is important to bear in mind the purpose of the algorithm – namely to ascertain violence that has been clinically recorded. It is possible that the nature of some diagnoses encourages the ascertainment and recording of violence; for example, the diagnosis of PTSD would require identification and recording of an index traumatic event, and diagnoses of affective or personality disorders may prompt (and/or result from) a detailed enquiry as to relevant aetiology. In addition, it is important to bear in mind that longer and/or more intensive clinical contact, accompanied by more extensive health records, will increase the likelihood of events being recorded, something which was not adjusted for in these analyses. Patients with briefer contacts with mental healthcare are likely to have less detailed records, which might account for the lack of association with substance use disorder diagnoses. Of note, it is important to bear in mind that the diagnostic categories used in this analysis are very broad ones. There may well be within-category heterogeneity in associations, particularly within the larger groupings of patients with schizophrenia and related disorders, and mood disorders. Evaluation of more specific diagnostic sub-groups was not attempted in this study, aside from PTSD, and we feel that this would demand more specific investigation within broadly defined clinical groups (e.g., mood disorders) rather than across all mental health service users. However, more granular clinical phenotypes might be better ascertained via recorded symptom profiles than specific diagnostic codes, given the potential variability with which coding is likely to be applied in routine practice.

In relation to sociodemographic factors, patients from most minority ethnic groups had significantly higher risk of recorded violence victimisation compared to White British patients. While patients from minority ethnic groups face more barriers that reduce instances of disclosing victimisation in healthcare settings (35), the findings of higher recorded victimisation in these groups has been consistently highlighted in previous literature (36). Also supporting previous research (37), male patients were at a lower risk of victimisation mentions compared to females; this was consistent within all victimisation types. Future research could helpfully investigate whether incidents of victimisations differ between men and women within different diagnostic groups, to ascertain vulnerability and target further support.

## Strengths and limitations

The study described here has important strengths. Firstly, it provides novel findings on how sociodemographic factors and mental health diagnosis associate with the distribution of recorded violence victimisation within clinical record data. The 12-month time period for assessment allowed victimisation to be assessed across a representative sample of patients receiving secondary mental healthcare services, circumventing seasonal variation of victimisation (38, 39). In addition, 2019 was chosen as a recent time period, but one which preceded the COVID-19 pandemic and consequent disruption to services and, potentially, healthcare records. The large sample size increased the precision for the estimate of prevalence of violence mentions and allowed distributions to be investigated across a wide range of disorders. The development of the NLP victimisation application demonstrated the application of machine learning to unlock a complex but clinically important construct, utilising rich and diverse free-text data from a wide array of clinical professionals and groups (15). This approach helps to automate the measurement of victimisation, increasing the number of cases that can be investigated and providing a method that could be used more routinely to monitor victimisation in patients.

One of the important limitations of the NLP algorithm at its current stage of development is the requirement to combine features at document rather than instance level. This means that the algorithm could be raising documents with mixed experiences, e.g. a document raised by the algorithm as having a positive mention of violence victimisation may also include recorded instances of perpetrated violence. Therefore, prevalence of recorded victimisation should be considered with caution and further development of the NLP algorithm is needed to increase precision and recall. In addition, NLP can only be used to ascertain violence which, if it is recorded at all, is done so using terminology that can be reliably ascertained. This will inevitably underestimate true exposure where this is not enquired about and/or not reported by the patient and/or not recorded by the reviewing clinician (9, 40). Finally, the analyses presented here focused on relatively few characteristics as exposures, and only considered the primary diagnosis of the patient (and, as mentioned, within relatively broad diagnostic groupings), not including the additive effects of comorbid disorders that may strengthen or weaken the risk of victimisation.

Considering future directions, clearly further development is required to construct accurate NLP algorithms to allow combinations of features at instance level, and to differentiate more efficiently between witnessed, perpetrated, and experienced violence, as well as encompassing broader experiences (e.g., including emotional abuse). This would aid in our understanding of the complex relationship between violence and mental health diagnoses. Future research into the clinical benefits of synthesizing previous interpersonal violence experienced by patients could aid in the real time decision making of clinicians, although ethical challenges of using NLP methods in practice need to be considered (15).

## Data availability statement

The datasets presented in this article are not readily available due to the terms of Ethics and Information Governance approvals and clinical source of the data, CRIS datasets must remain within the SLaM firewall. All data used from this study can remade accessible on request from ris.administrator@slam.nhs.uk with an appropriate research passport or appropriate SLaM honorary contract. Requests to access the datasets should be directed to ris.administrator@slam.nhs.uk.

## Ethics statement

The studies involving humans were approved by CRIS has a robust, patient-led governance and data security model and has approval as a data resource for secondary analysis (Oxford Research Ethics Committee C, reference 18/SC/0372). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

## Author contributions

AM: Conceptualisation, Formal analysis, Writing - original draft, Writing - review & editing. VB: Writing - original draft, Writing - review & editing. RB: Resources, Methodology, Writing - original draft. DC: Methodology, Resources, Writing - original draft. LL: Methodology, Resources, Writing - original draft, Writing - review & editing. AM: Methodology, Resources, Writing - original draft. JS: Resources, Writing - original draft, GKS: Writing - original draft, Writing - review & editing. AR: Methodology, Resources, Writing - original draft. MW: Writing - original draft, Writing - review & editing. RS: Funding acquisition, Methodology, Resources, Supervision, Writing - original draft, Writing - review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. AM, JS, DC, and RS are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at the South London and Maudsley NHS Foundation Trust and King's College London. RS is additionally funded by i) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust; ii) the DATAMIND HDR UK Mental Health Data Hub (MRC grant MR/W014386); iii) the UK Prevention Research Partnership (Violence, Health and Society; MR-VO49879/1), an initiative funded by UK Research and Innovation Councils, the Department of Health and Social Care (England) and the UK devolved administrations, and leading health research charities. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. VB is funded by an NIHR Advanced Fellowship NIHR302243 administered by KCL. GK-S has received salary support from the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London; and has also received funding from a researcher-initiated grant from the Violence and Abuse Mental Health Network (VAMHN) UKRI.

## Conflict of interest

RS declares research support received in the last 3 years from Janssen, GSK, and Takeda. GK-S has received research funding from Janssen and H Lundbeck.

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.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2024.1181739/full#supplementary-material

## References

1. Krug GE, Mercy JA, Dahlberg LL, Zwi AB. The world report on violence and health. *Lancet*. (2002) 360:1083–88. doi: 10.1016/S0140-6736(02)11133-0

2. Reza A, Mercy JA, Krug E. Epidemiology of violent deaths in the world. *Injury Prev*. (2001) 7:104–11. doi: 10.1136/ip.7.2.104

3. Olofsson N, Lindqvist K, Shaw BA, Danielsson I. Long-term health consequences of violence exposure in adolescence: A 26–year prospective study. *BMC Public Health*. (2012) 12:1–11. doi: 10.1186/1471-2458-12-411

4. López-Martínez AE, Serrano-Ibáñez ER, Ruiz-Párraga GT, Gómez-Pérez L, Ramírez-Maestre C, Esteve R. Physical health consequences of interpersonal trauma: A systematic review of the role of psychological variables. *Trauma Violence Abuse*. (2018) 19:305–22. doi: 10.1177/1524838016659488

5. Maniglio R. Severe mental illness and criminal victimization: a systematic review. *Acta Psychiatrica Scandinavica*. (2009) 119:180–91. doi: 10.1111/j.1600-0447.2008.01300.x

6. Khalifeh H, Moran P, Borschmann R, Dean K, Hart C, Hogg J, et al. Domestic and sexual violence against patients with severe mental illness. *psychol Med*. (2015) 45:875–86. doi: 10.1017/S0033291714001962

7. Khalifeh H, Johnson S, Howard LM, Borschmann R, Osborn D, Dean K, et al. Violent and non-violent crime against adults with severe mental illness. *Br J Psychiatry*. (2015) 206:275–82. doi: 10.1192/bjp.bp.114.147843

8. Mullen PE. Schizophrenia and violence: from correlations to preventive strategies. *Adv Psychiatr Treat*. (2006) 12:239–48. doi: 10.1192/apt.12.4.239

9. Howard LM, Trevillion K, Agnew-Davies R. Domestic violence and mental health. *Int Rev Psychiatry*. (2010) 22:525–34. doi: 10.3109/09540261.2010.512283

10. Hildersley R, Easter A, Bakolis I, Carson L, Howard LM. Changes in the identification and management of mental health and domestic abuse among pregnant women during the COVID-19 lockdown: regression discontinuity study. *BJPsych Open*. (2022) 8:e96. doi: 10.1192/bjo.2022.66

11. Robson D, Spaducci G, McNeill A, Stewart D, Craig TJ, Yates M, et al. Effect of implementation of a smoke-free policy on physical violence in a psychiatric inpatient setting: an interrupted time series analysis. *Lancet Psychiatry*. (2017) 4:540–6. doi: 10.1016/S2215-0366(17)30209-2

12. Bhavsar V, Sanyal J, Patel R, Shetty H, Velupillai S, Stewart R, et al. The association between neighbourhood characteristics and physical victimisation in men and women with mental disorders. *BJPsych Open*. (2020) 6:e73. doi: 10.1192/bjo.2020.52

13. Cullen AE, Bowers L, Khondoker M, Pettit S, Achilla E, Koeser L, et al. Factors associated with use of psychiatric intensive care and seclusion in adult inpatient mental health services. *Epidemiol Psychiatr Sci*. (2018) 27:51–61. doi: 10.1017/S2045796016000731

14. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *J Biomed Inf*. (2018) 77:34–49. doi: 10.1016/j.jbi.2017.11.011

15. Botelle R, Bhavsar V, Kadra-Scalzo G, Mascio A, Williams MV, Roberts A, et al. Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study. *BMJ Open*. (2022) 12:e052911. doi: 10.1136/bmjopen-2021-052911

16. Florence C, Shepherd J, Brennan I, Simon T. Effectiveness of anonymised information sharing and use in health service, police, and local government partnership for preventing violence related injury: experimental study and time series analysis. *BMJ*. (2011) 342:1–9. doi: 10.1136/bmj.d3313

17. Kadra-Scalzo G, Kornblum D, Stewart R, Howard LM. Adverse outcomes associated with recorded victimization in mental health electronic records during the first UK COVID-19 lockdown. *Soc Psychiatry Psychiatr Epidemiol*. (2023) 58:431–40. doi: 10.1007/s00127-022-02393-w

18. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS foundation trust biomedical research centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry*. (2009) 9:1–12. doi: 10.1186/1471-244X-9-51

19. Perera G, Broadbent M, Callard F, Chang C-K, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) case register: current status and recent enhancement of an electronic mental health record-derived data resource. *BMJ Open*. (2016) 6:e008721. doi: 10.1136/bmjopen-2015-008721

20. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. (2020) 36:1234–40. doi: 10.1093/bioinformatics/btz682

21. Hugging_face, Read Rectangular Text Data . Available online at: https://huggingface.co/docs/transformers/en/model_doc/bert (Accessed 26 April 2024).

22. Maudsley, CRIS NLP service . Available online at: https://www.maudsleybrc.nihr.ac.uk/media/325736/applications-library-v13.pdf (Accessed 26 April 2024).

23. GOV.UK, List of ethnic groups: GOV.UK (2022). Available online at: https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups#2021-census (Accessed 26 April 2024).

24. WHO. *The ICD-10 classification of mental and behavioural disorders*. Geneva, Switzerland: World Health Organization (1993).

25. Wickham H, et al. *Package 'readr'*. Read Rectangular Text Data . Available online at: https://cran.r-project.org/web/packages/readr/readr.pdf (Accessed 23 August 2023).

26. Wickham H, François R, Henry L, Müller K. *dplyr: A Grammar of Data Manipulation*. R package version 0.7. 6. (2018). https://github.com/tidyverse/dplyr.

27. Wickham H, Wickham H. *Data analysis*. New York, United States: Springer (2016).

28. Hair JF, Risher JJ, Sarstedt M, Ringle CM. When to use and how to report the results of PLS-SEM. *Eur Business Rev*. (2019) 31:2–24. doi: 10.1108/EBR-11-2018-0203

29. Silver E, Arseneault L, Langley J, Caspi A, Moffitt TE. Mental disorder and violent victimization in a total birth cohort. *Am J Public Health*. (2005) 95:2015–21. doi: 10.2105/AJPH.2003.021436

30. Kadra G, Dean K, Hotopf M, Hatch SL. Investigating exposure to violence and mental health in a diverse urban community sample: data from the South East London Community Health (SELCoH) survey. *PloS One*. (2014) 9(4):e93660. doi: 10.1371/journal.pone.0093660

31. Kooyman I, Dean K, Harvey S, Walsh E. Outcomes of public concern in schizophrenia. *Br J Psychiatry*. (2007) 191:s29–36. doi: 10.1192/bjp.191.50.s29

32. Fazel S, Gulati G, Linsell L, Geddes JR, Grann M. Schizophrenia and violence: systematic review and meta-analysis. *PloS Med*. (2009) 6:e1000120. doi: 10.1371/journal.pmed.1000120

33. Labrum T, Solomon P, Marcus S. Victimization and perpetration of violence involving persons with mood and other psychiatric disorders and their relatives. *Psychiatr Serv*. (2020) 71:498–501. doi: 10.1176/appi.ps.201900384

34. Mazza M, Marano G, Del Castillo AG, Chieffo D, Monti L, Janiri D, et al. Intimate partner violence: A loop of abuse, depression and victimization. *World J Psychiatry*. (2021) 11:215. doi: 10.5498/wjp.v11.i6.215

35. Heron RL, Eisma MC, Browne K. barriers and facilitators of disclosing domestic violence to the UK health service. *J Family Violence*. (2022) 37:533–43. doi: 10.1007/s10896-020-00236-3

36. Policastro C, Teasdale B, Daigle LE. The recurring victimization of individuals with mental illness: a comparison of trajectories for two racial groups. *J Quantitative Criminology*. (2016) 32:675–93. doi: 10.1007/s10940-015-9271-8

37. Yapp E, Booth T, Davis K, Coleman J, Howard LM, Breen G, et al. Sex differences in experiences of multiple traumas and mental health problems in the UK Biobank cohort. *Soc Psychiatry Psychiatr Epidemiol*. (2021) 58(12):1819–31. doi: 10.1007/s00127-021-02092-y

38. Koutaniemi EM, Einiö E. Seasonal variation in seeking help for domestic violence based on Google search data and Finnish police calls in 2017. *Scandinavian J Public Health*. (2021) 49:254–9. doi: 10.1177/1403494819834098

39. Farrell G, Pease P. CRIM SEASONALITY: *domestic disputes and residential burglary in Merseyside 1988–90. Br J Criminology*. (1994) 34:487–98. doi: 10.1093/oxfordjournals.bjc.a048449

40. Gutmanis I, Beynon C, Tutty L, Wathen CN, MacMillan HL. Factors influencing identification of and response to intimate partner violence: a survey of physicians and nurses. *BMC Public Health*. (2007) 7:1–11. doi: 10.1186/1471-2458-7-12

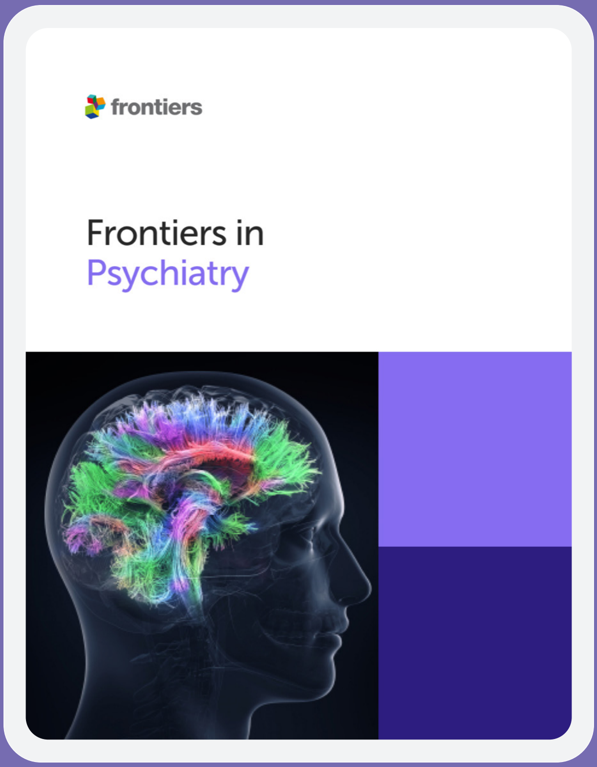
# Frontiers in Psychiatry

**Explores and communicates innovation in the field of psychiatry to improve patient outcomes**

The third most-cited journal in its field, using translational approaches to improve therapeutic options for mental illness, communicate progress to clinicians and researchers, and consequently to improve patient treatment outcomes.

## Discover the latest Research Topics

See more →



frontiers

Frontiers in
Psychiatry

**Frontiers**

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

**Contact us**

+41 (0)21 510 17 00
frontiersin.org/about/contact