frontiers | Research Topics

# Artificial intelligence and mental health care

**Edited by**
Jorge Piano Simoes, Peter ten Klooster, Jannis Kraiss,
Patrick K. A. Neff and Uli Niemann

**Published in**
Frontiers in Public Health
Frontiers in Psychiatry

ISSN 1664-8714
ISBN 978-2-8325-5303-9
DOI 10.3389/978-2-8325-5303-9

## 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

# Artificial intelligence and mental health care

**Topic editors**

Jorge Piano Simoes — University of Twente, Netherlands
Peter ten Klooster — University of Twente, Netherlands
Jannis Kraiss — University of Twente, Netherlands
Patrick K. A. Neff — University of Zurich, Switzerland
Uli Niemann — Otto von Guericke University Magdeburg, Germany

**Citation**

Simoes, J. P., ten Klooster, P., Kraiss, J., Neff, P. K. A., Niemann, U., eds. (2024). *Artificial intelligence and mental health care*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-5303-9

# Table of contents

frontiers | Frontiers in Public Health

TYPE Editorial
PUBLISHED 30 July 2024
DOI 10.3389/fpubh.2024.1461446



OPEN ACCESS

EDITED AND REVIEWED BY
Wulf Rössler,
Charité University Medicine Berlin, Germany

*CORRESPONDENCE
Jorge P. Simões
✉ j.pianosimoes@utwente.nl

RECEIVED 08 July 2024
ACCEPTED 15 July 2024
PUBLISHED 30 July 2024

CITATION
Simões JP, ten Klooster P, Neff PK, Niemann U and Kraiss J (2024) Editorial: Artificial intelligence and mental health care. *Front. Public Health* 12:1461446. doi: 10.3389/fpubh.2024.1461446



# Editorial: Artificial intelligence and mental health care

Jorge P. Simões[1]*, Peter ten Klooster[1], Patrick K. Neff[2,3], Uli Niemann[4] and Jannis Kraiss[1]

[1]Department of Health Psychology and Technology, University of Twente, Enschede, Netherlands, [2]Department of Otorhinolaryngology, Head, and Neck Surgery, University Hospital Zurich, University of Zurich, Zurich, Switzerland, [3]Department of Psychiatry and Psychotherapy, University of Regensburg, Regensburg, Germany, [4]University Library, Otto-von-Guericke-University Magdeburg, Magdeburg, Germany

KEYWORDS

**mental health, healthcare, public health, machine learning, artificial intelligence**

Editorial on the Research Topic
Artificial intelligence and mental health care

## Introduction

Advancements in machine learning (ML) and artificial intelligence (AI) offer significant potential to transform mental health care. These technologies have been utilized for various purposes, such as early detection of mental disorders, optimizing personalized treatments tailored to individual patient characteristics, improving the characterization of disorders that negatively impact mental wellbeing and quality of life, better predicting their progression over time, and developing new treatments and diagnostic tools for mental health care. Despite their considerable potential and occasional breakthroughs, ML and AI have not yet fully realized these objectives in mental health care.

## Aim of this Research Topic

This Research Topic aimed to provide innovative examples of how ML and AI applications can be practically implemented in standard mental health care. The particular focus of this Research Topic was to provide examples of how to use ML and AI to enhance public health by lessening the impact of chronic disorders that adversely affect wellbeing and improving quality of life.

## Research Topic impact

This Research Topic was open between November 10th, 2022, and November 1st, 2023. There were 14 submissions, 12 of which were accepted after peer review, from 64 different authors. While open, the topic had 26,973 views, 19,768 article views, 5,845 article downloads, and 1,360 topic views.

Alsaqqa and Alwawi conducted a scoping review on the characteristics of studies, related concepts, and recommendations for implementing digital interventions in public health. It highlighted the importance of addressing structural inequalities, ensuring personal agency, and social connectedness. The study also emphasized the importance of iterative optimization during study design, involving stakeholders, and using contextual indicators to enhance the effectiveness of digital interventions. An important aspect of the review was the call for more patient and public involvement and the suggestion to adopt standardized metrics to improve research quality and application of digital health interventions.

Morita et al. explored the application of large language models like ChatGPT in public health through SWOT and PESTLE analyses. The identified strengths include personalized health support and data analysis capabilities, weaknesses such as potential miscommunication and data privacy issues, opportunities in improving healthcare access and disease surveillance, and threats including misinformation and bias. The PESTLE analysis identified factors like government policies impacting investment and data governance, cost-effectiveness and job impact considerations, public trust and cultural attitudes toward AI, integration with health systems and algorithmic transparency, privacy laws and ethical guidelines, and the environmental impact of AI infrastructure's energy consumption and carbon footprint.

Wen et al. used 2D gait videos for automatic anxiety assessment among graduate students. By analyzing gait features from time-series data, the authors created anxiety assessment models via machine learning. The study found that dynamic time-frequency features significantly enhance model performance, particularly for women. The models demonstrated reliability and validity, suggesting that 2D gait analysis could be a practical, non-invasive method for real-time anxiety assessment and should be further investigated and evaluated in clinical samples.

Huisman et al. examined the validity of automated sentiment analysis in interpreting emotional content from therapy session notes of patients with eating disorders, comparing it to human raters. The study analyzed 460 records and found moderate agreement between automated analysis and human raters. The findings suggest the potential for automated sentiment analysis in clinical settings but emphasize the need for further refinement before applying the algorithm in clinical settings, particularly by incorporating ED-specific terminology and establishing more relevant benchmarks for validation.

Franken et al. investigated the ability of ML to predict improvement in patients using real-world longitudinal data from specialized outpatient mental health treatment. Different ML models were trained and compared with traditional logistic regression. The models showed moderate predictive ability in an independent test set, with slightly better performance when early change scores were included as predictors. Machine learning algorithms did not outperform simpler logistic regression models. Early change during treatment was a crucial predictor for longer-term outcomes.

Li et al. also aimed to leverage the advantages of an ML approach over traditional statistical methods to predict the risk of depression in people with obstructive sleep apnea hypopnea syndrome using data readily available from the NHANES database. Several features predictive of depression were identified, including demographic, health and lifestyle-related, and socio-economic factors. Interestingly, like in the study by Franken et al., the simple logistic regression model was not inferior—and even superior—to more complex ML models.

Kim et al. used ML methods to examine the performance of classifying states of stress and non-stress using biosignal data measured by a smartwatch. In contrast to the previous studies, this study used an experimental setup where participants were instructed to perform stress-inducing and relaxation tasks. The top 9 features extracted from the heart rate and photoplethysmography data were able to classify stress with an accuracy of >80% with, again, the logistic regression classifier showing the best performance.

Delgadillo et al. performed a study during the COVID-19 pandemic using Bayesian network analyses and modeling interactions between risk and protective factors for suicidal ideation in Austria and the UK. The models achieved high predictive accuracy (AUC $\geq$ 0.84 within-sample and AUC $\geq$ 0.79 out-of-sample), explaining nearly 50% of suicidal ideation variability. Seven consistent factors, including depressive symptoms, loneliness, and anxiety, were identified in both countries. This study shows the potential to predict suicidal risk accurately using these factors.

Jović et al. addressed the challenge of comparing ADHD scores across different scales used by various research consortia. They harmonized scores from the Child Behavior Checklist (CBCL) and Strengths and Difficulties Questionnaire (SDQ) using various test equating and machine learning methods on 1,551 parent reports of children aged 10–11.5 years. The study found that methods utilizing item-level information and treating outcomes as interval measurements, such as regression, were most effective for harmonizing scores.

Pavicic et al. used iterative Random Forests to identify geographic, environmental, and sociodemographic predictors of suicide attempts among U.S. veterans. Analyzing data from 405,540 patients, the model incorporated 1,784 features, including climatic factors, population demographics, and the density of firearms and alcohol vendors. Key findings indicated that areas with higher concentrations of married males have lower suicide attempt rates, while areas with renting and males living alone have higher rates.

Bremer-Hoeve et al. investigated predictors of treatment dropout in patients with post-traumatic stress disorder (PTSD) due to childhood abuse, using elastic net regression. Analyzing data from 121 patients undergoing two different Eye Movement Desensitization and Reprocessing (EMDR) therapy protocols, they identified key dropout predictors: male gender, low education, suicidal thoughts, emotion regulation issues, high general psychopathology, and lack of benzodiazepine use.

Guo et al. explored causal factors of non-suicidal self-injury (NSSI) in children using computational causal analysis. They identified nine key factors: life satisfaction, depression, family dysfunction, sugary beverage consumption, positive youth development (PYD), internet addiction, COVID-19 PTSD, academic anxiety, and sleep duration. The research highlighted four main causal pathways and emphasized the

roles of pandemic-induced lifestyle changes, screen time, adolescent development, and family dynamics in NSSI risk, advocating for targeted interventions addressing these diverse factors.

## Author contributions

JS: Writing – review & editing, Writing – original draft. PK: Writing – review & editing, Writing – original draft. PN: Writing – review & editing, Writing – original draft. UN: Writing – review & editing, Writing – original draft. JK: Writing – review & editing, Writing – original draft.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 01 March 2023
DOI 10.3389/fpubh.2023.1010264



OPEN ACCESS

EDITED BY
Jannis Kraiss,
University of Twente, Netherlands

REVIEWED BY
Sara Nooraeen,
Mayo Clinic, United States
Jorge Piano Simoes,
University Medical Center
Regensburg, Germany

*CORRESPONDENCE
Jaime Delgadillo
✉ jaime.delgadillo@nhs.net

†These authors share senior authorship

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

RECEIVED 02 August 2022
ACCEPTED 06 February 2023
PUBLISHED 01 March 2023

CITATION
Delgadillo J, Budimir S, Barkham M, Humer E, Pieh C and Probst T (2023) A Bayesian network analysis of psychosocial risk and protective factors for suicidal ideation.
*Front. Public Health* 11:1010264.
doi: 10.3389/fpubh.2023.1010264



# A Bayesian network analysis of psychosocial risk and protective factors for suicidal ideation

Jaime Delgadillo[1]*, Sanja Budimir[2,3], Michael Barkham[1], Elke Humer[2], Christoph Pieh[2†] and Thomas Probst[2†]

[1]Clinical and Applied Psychology Unit, Department of Psychology, University of Sheffield, Sheffield, United Kingdom, [2]Department for Psychosomatic Medicine and Psychotherapy, Danube University Krems, Krems an der Donau, Austria, [3]Department of Work, Organization and Society, Ghent University, Ghent, Belgium

**Background:** The aim of this study was to investigate and model the interactions between a range of risk and protective factors for suicidal ideation using general population data collected during the critical phase of the COVID-19 pandemic.

**Methods:** Bayesian network analyses were applied to cross-sectional data collected 1 month after the COVID-19 lockdown measures were implemented in Austria and the United Kingdom. In nationally representative samples ($n = 1{,}005$ Austria; $n = 1{,}006$ UK), sociodemographic features and a multi-domain battery of health, wellbeing and quality of life (QOL) measures were completed. Predictive accuracy was examined using the area under the curve (AUC) within-sample (country) and out-of-sample.

**Results:** The AUC of the Bayesian network models were ≥ 0.84 within-sample and ≥0.79 out-of-sample, explaining close to 50% of variability in suicidal ideation. In total, 15 interrelated risk and protective factors were identified. Seven of these factors were replicated in both countries: depressive symptoms, loneliness, anxiety symptoms, self-efficacy, resilience, QOL physical health, and QOL living environment.

**Conclusions:** Bayesian network models had high predictive accuracy. Several psychosocial risk and protective factors have complex interrelationships that influence suicidal ideation. It is possible to predict suicidal risk with high accuracy using this information.

KEYWORDS

**depression, risk factors, Bayesian network analysis, suicide, COVID-19**

## Background

Suicide is a serious global health problem, with an age-adjusted annual global incidence rate of 11.4 per 100,000 (1). Suicide represents the leading cause of death worldwide among young people, disproportionately affecting males living in environments with high economic inequalities (2). There are indications that for every suicide there are over 20 times more people who attempt suicide (1). This already alarming situation may have been further aggravated by the outbreak of the novel coronavirus disease (COVID-19). The COVID-19 outbreak has dramatically impacted health, economics and social connections around the world (3), thereby exacerbating known risk factors for suicidal ideation and suicide attempts (4, 5). These risk factors include forced isolation, quarantine, reduction of social contacts, health-related anxiety, economic problems, risk of domestic violence, risk of addictive behavior and reduction of access to mental health care (6). The COVID-19 pandemic might

lead to an increase in rates of self-injury or suicide, especially in individuals with pre-existing mental health problems (i.e., depression or anxiety), but also in people under increased stress such health care professionals (6–8). Therefore, this public health emergency calls for advances in suicide research and prevention (7).

Understanding suicide risk is crucial in order to advance the implementation of effective prevention strategies. Traditional attempts at understanding the antecedents of suicide have focused on single risk factors, or a specific domain of risk (i.e., socio-demographics), and thus have been of limited value to the design of effective prevention measures (9). Literature in this field has identified some risk factors such as genetic and biological factors, mental disorders, and stressors such as financial problems or violence (1, 10, 11). However, risk prediction accuracy is still limited due to the low explained variance afforded by these variables (9). Low base-rate events such as suicide are also notoriously difficult to predict, which limits the reliability of risk factor research. Furthermore, the complexity of factors leading to suicidal behavior cannot be adequately addressed by conventional statistical techniques, such as regression analysis or analysis of variance, as they provide limited insight into the interrelationships between the risk factors themselves (12).

Compared to actual suicide attempts, suicidal ideation, which refers to thoughts of engaging in behavior intended to end one's life, is more than three times more prevalent in the general population (13). In this regard, studying suicidal ideation as a proximal antecedent to suicidal behavior could offer a way forward in understanding key risk and protective factors, and enable the development of *just in time adaptive interventions* (14, 15). However, the real-time monitoring of risk factors involves considerable participant burden. Accordingly, deploying these assessments at a population-scale, and even in clinical samples, seems unfeasible (14). A more realistic strategy could be to deploy such interventions in a targeted way, focusing on people at *high risk* of suicide. As the ability to predict suicide risk has not improved in the past 50 years (9), it is necessary to investigate the combined effects of multiple factors to characterize this *high risk* phenotype with greater precision. However, so far the study of factors contributing to suicidal thoughts have rarely examined the combined effect of multiple risk factors and protective factors. Also, large data sets including multiple potential risk and protective factors are required to enable reliable prediction research (9, 16).

Methodological developments such as machine learning and network analysis represent a novel way to predict health-related outcomes and to model complex interrelationships between variables in a causal network (17, 18). Unlike conventional hypothesis-testing studies that specify expected relationships *a priori*; machine learning offers an exploratory and data-driven framework to discover patterns of associations in large datasets. Conventional approaches to model risk factors for suicidal ideation tend to focus on the statistical significance and explained variance attributable to specific variables that are selected based on prior theory or research (i.e., main effects for hypothesized predictors). Machine learning analyses are not necessarily constrained to the modeling of main effects, and can "discover" complex (i.e., non-linear, interactive) relationships between variables, which were not previously known or expected. Rather than prioritizing goodness-of-fit in a single dataset as in conventional regression analysis, machine learning frameworks use cross-validation methods to determine if discovered relationships in the data have adequate predictive accuracy, and are therefore potentially generalizable to new samples. As such, network analysis of suicidal ideation and its risk and protective factors could potentially help to derive new insights in the field of suicide prevention (12). The present paper aims to contribute to the identification of reliable risk and protective factors for suicidal ideation from a data-driven perspective, without prior specification of hypotheses, but using variables that have been selected based on prior evidence described above. To this end, we developed Bayesian network models using data from a cross-sectional survey conducted during the peak of the first COVID-19 lockdown in two European countries, Austria and the United Kingdom.

# Methods

## Design and setting

The objectives of the present study were (1) to identify predictors of suicidal ideation (2), to model complex interactions between these predictors, and (3) to examine their generalizability across two countries. We approached this from a machine learning perspective, using a cross-country cross-validation design to enable us to understand which predictors replicate in samples from two different countries. A cross-sectional online survey was designed to recruit representative samples covering all geographical regions of Austria and the United Kingdom (UK), and reflecting population norms in relation to demographic features. The Qualtrics® population survey platform was used; implementing age, gender, educational, and regional quotas based on available population census data from both countries. The survey measured sociodemographic features and several health, wellbeing and quality of life indicators that were informed by prior evidence. Data collection started 4 weeks after COVID-19 lockdown measures were implemented in Austria and the UK (April 2020), until the point where a representative sample was obtained with a minimum sample size of $n = 1{,}000$ participants from each country, which was specified *a priori*. Participants were recruited from existing pools of research panel participants and received financial incentives. Participants who did not respond to all questions or who failed quality checks, including attention filters and survey timings, were excluded. The goal of the sampling procedure was to obtain large enough samples from each country in order to conduct machine learning analyses, which were nationally representative (covering all regions of each country in a proportionate way, reflective of local demographics), and balanced between both countries (same sample size, to minimize imbalance due to differences in overall population density across countries). Overall, the target sample was attained within 10 days, after which the survey closed.

## Measures

The primary outcome of interest was suicidal ideation, derived from the Patient Health Questionnaire (PHQ-9), which has been shown to be a robust and age-independent predictor of suicide

10.3389/fpubh.2023.1010264

attempts and deaths (19). The PHQ-9 is a measure of depression symptoms, where response options for each of 9 questions are "not at all" (0 points), "several days" (1 point), "more than half of the days" (2 points) or "nearly every day" (3 points), yielding an overall severity score between 0 and 27 (20). A cut-off score of ≥10 has been recommended to screen for clinically significant depression symptoms, with adequate sensitivity (88%) and specificity (88%). Item 9 of the PHQ-9 measure asks, "Over the last 2 weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself in some way?" Response to this question was coded in a binary way to identify the presence of any recent suicidal ideas within the last 2 weeks (1 = item endorsed if response ranged from 1 to 3; 0 = item not endorsed if response was 0). The remaining items (PHQ-8) were used to control for depression severity (21).

## Health and wellbeing indicators

The GAD-7 is a 7-item case-finding measure for anxiety disorders; each item is rated between 0 and 3, with a total severity score between 0 and 21 (22). Stress-severity was measured with the PSS-10, which measures two related domains (perceived helplessness, perceived self-efficacy) using 10 items on a five-point scale ranging from 0-4 (23). The Insomnia Severity Index (ISI) (24) is a measure of sleep quality and insomnia, based on 7 items rated on a five-point scale (from 0 to 4). The WHOQOL-BREF is a 26-item questionnaire that measures four domains of quality-of-life; physical health, psychological health, social relationships, and environment, during the past 2 weeks (25). Social loneliness was measured using the 11-item De Jong-Gierveld scale (26). Resilience was assessed using the 10-item version of the Connor-Davidson resilience scale (CD-RISC-10) (27), where items are rated using a Likert scale from 0 to 4. Single-item questions were used to assess self-reported days of exercise per week and physical illness status.

## Demographics

Participants completed single-item questions to gather the following demographic features: age, gender, highest level of education, marital status, having children requiring care, employment status, net household income, housing type, household number of occupants additional to the respondent.

# Statistical analysis

A cross-country, cross-validation design was used to identify suicidal risk factors that may be country-specific and those that are common across samples. This design, depicted in Figure 1, involved training a prediction model for each country and then testing its generalizability using data from the other country to classify cases as belonging to the suicidal or non-suicidal class.

Each country-specific prediction model was trained using a Tree-Augmented Naïve Bayes (TAN) algorithm (18). Unlike conventional multivariable logistic regression models which only model main effects, or require pre-specification of expected interactions, the TAN method offers a data-driven way to model a network of relationships (called *attribute dependencies*) between

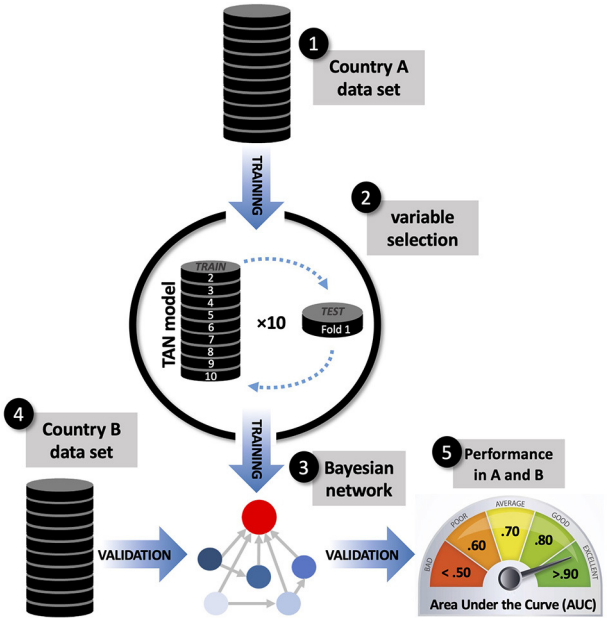



FIGURE 1
Schematic representation of the cross-country, cross-validation design used to identify suicidal ideation risk factors that are country-specific and those that are common across samples. Model training used a tree-augmented naïve Bayes algorithm, applying 10-fold internal cross-validation for variable selection. Classification accuracy was assessed within-sample and out-of-sample using the Area Under the Curve (AUC).

predictors and their joint influence over a target outcome. TAN produces a simple and parsimonious network model where each predictor is allowed to depend on one additional predictor, thus modeling multiple two-way interactions. The risk of suicidal ideation is thus estimated based on the combined weight (e.g., joint modeling) of the conditional probabilities attributed to each predictor in a Bayesian network model.

Like other machine learning approaches, the performance of the TAN algorithm depends largely on the adequacy of variable selection. In order to build Bayesian networks composed only with reliable predictors, we entered all available variables listed above in the Measures section and performed variable selection using a ten-fold cross-validation (CV10) approach (28). Two noise variables (continuous and categorical) were modeled on the distribution of PHQ-9 (mean, standard deviation) and gender (base rate of males and females). Noise variables were introduced as predictors in the TAN analysis, along with all other candidate predictors listed in Table 1. The CV10 approach produced ten Bayesian network models (within each country-specific sample) with their respective variable importance plots, which ranked candidate variables according to their predictive value. We performed variable selection by only retaining the variables that were consistently ranked as more important than both noise variables in more than half (>5) of the trained models. The selected variables were entered into a final country-specific Bayesian network model, which was visualized using a directed acyclic graph (29). The CV10 procedure was strictly used for variable selection and not for hyperparameter

10.3389/fpubh.2023.1010264

tuning. TAN was applied using pre-specified hyperparameters (using likelihood ratio as the independence test; significance level of 0.01; maximum conditional set size = 5; using Bayes adjustment for small cell counts).

Once the country-specific Bayesian network models were trained, we calculated their explained variance based on Nagelkerke $R^2$. Classification accuracy was assessed within-sample and out-of-sample using the area under the curve (AUC), positive and negative predictive values (PPV, NPV).

# Results

## Sample characteristics

Sample characteristics for the Austrian ($n$ = 1,005) and UK ($n$ = 1,006) samples are displayed in Table 1. The prevalence of suicidal ideation was higher in the UK sample (31.7%) compared to the Austrian sample (17.3%). As expected, participants with clinically significant depression symptoms (PHQ-9 ≥ 10) tended to have a high prevalence of suicidal ideas (Austria = 55.0%; UK = 64.5%). However, around 8% of non-depressed participants also endorsed suicidal ideas, indicating that other risk factors may also be relevant.

Country-specific Bayesian network models are presented in Figures 2 and 3, along with normalized (0–100%) variable importance indices that quantify each variable's contribution to explained variance. The upward red arrows denote factors that increase risk of suicidal ideation, and the downward green arrows denote protective factors. The model also shows inter-relationships between the variables.

The Austrian model included ten variables, of which the five most important ones were loneliness, depression, anxiety, perceived self-efficacy, and quality of life related to physical health. The model also showed multiple inter-relationships between the factors. The effect of wellbeing was moderated by self-efficacy, depression, resilience, and quality of relationships. The effect of depression was moderated by anxiety and insomnia. The effect of relationship quality was moderated by loneliness. The effect of physical health was moderated by quality of the environment. Overall, this network model explained 47.1% of variability in suicidal ideation in the Austrian sample. The model's classification accuracy was similar within-sample (AUC = 0.84; PPV = 0.69; NPV = 0.94) and out-of-sample (AUC = 0.80; PPV = 0.61; NPV = 0.83), with minimal prediction shrinkage (AUC = 0.04).

The UK model included 12 variables, of which the five most important ones were depression, age, perceived helplessness, loneliness and anxiety. All variables interacted with other variables in the network. The effect of physical health was moderated by exercise, housing space, and quality of the environment. The effect of helplessness was moderated by anxiety and self-efficacy. The effect of resilience was moderated by loneliness. The effect of anxiety was moderated by depression, which in turn was moderated by physical health. The effect of age was moderated by helplessness and having children requiring care. Younger parents were at increased risk of suicidal ideation relative to younger people without children requiring care; but older parents (≥45) were at reduced risk compared to older people without children requiring

TABLE 1 Sample characteristics.

| | Austria ($N$ = 1,005) | United Kingdom ($N$ = 1,006) |
|---|---|---|
| **Demographics** | | |
| Age group, % ($n$) | | |
| 18–24 | 11.7 (118) | 9.7 (98) |
| 25–34 | 16.5 (166) | 20.2 (203) |
| 35–44 | 18.4 (185) | 18.9 (190) |
| 45–54 | 22.1 (222) | 19.3 (194) |
| 55–64 | 18.0 (181) | 17.2 (173) |
| 65+ | 13.2 (133) | 14.7 (148) |
| Females, % ($n$) | 52.7(530) | 54.1 (544) |
| Education, % ($n$) | | |
| None at all | 0.0 (0) | 1.6 (16) |
| Elementary school | 0.10 (1) | 3.5 (35) |
| High school | 2.6 (26) | 40.3 (405) |
| Vocational training | 31.9 (321) | 14.2 (143) |
| College degree | 28.7 (288) | 12.8 (129) |
| University degree | 36.7 (369) | 27.6 (278) |
| Children, % ($n$) | 23.7 (238) | 30.6 (308) |
| Employment status, % ($n$) | | |
| Unemployed | 26.8 (269) | 47.5 (478) |
| Employed | 55.8 (561) | 38.5 (387) |
| Retired | 17.4 (175) | 14.0 (141) |
| Household income, % ($n$) | | |
| Band 1 | 7.1 (71) | 13.7 (138) |
| Band 2 | 23.4 (235) | 34.1 (343) |
| Band 3 | 30.2 (304) | 25.4 (256) |
| Band 4 | 19.5 (196) | 14.6 (147) |
| Band 5 | 19.8 (199) | 12.1 (122) |
| Housing type, % ($n$) | | |
| Flat | 23.2 (233) | 20.1 (202) |
| Apartment with terrace | 34.4 (346) | 5.6 (56) |
| House | 42.4 (426) | 74.4 (748) |
| Household occupants, mean (SD) | 1.74 (1.34) | 1.89 (1.43) |
| **Health and wellbeing** | | |
| Illness reported, % ($n$) | 6.9 (69) | 10.3 (104) |
| Days exercise per week, mean (SD) | 2.70 (1.44) | 2.29 (1.59) |
| Suicidal ideas, % ($n$) | 17.3 (174) | 31.7 (319) |
| Suicidal ideas with depression*, % ($n$) | 55.0 (553) | 64.5 (649) |
| Suicidal ideas without depression*, % ($n$) | 7.3 (73) | 8.8 (89) |
| PHQ-8, mean (SD) | 5.93 (5.00) | 8.38 (6.99) |

*(Continued)*

10.3389/fpubh.2023.1010264

TABLE 1 (Continued)

| | Austria ($N = 1,005$) | United Kingdom ($N = 1,006$) |
|---|---|---|
| GAD-7, mean (SD) | 5.84 (4.70) | 8.03 (6.52) |
| PSS10 helplessness, mean (SD) | 9.37 (5.19) | 10.34 (6.05) |
| PSS10 self-efficacy, mean (SD) | 9.40 (3.13) | 8.63 (3.42) |
| Insomnia severity index, mean (SD) | 8.31 (5.70) | 10.43 (7.05) |
| Loneliness scale, mean (SD) | 4.58 (3.67) | 6.41 (3.21) |
| CD-RISC-10, mean (SD) | 27.27 (7.20) | 24.56 (8.12) |
| WHOQOL physical, mean (SD) | 15.57 (2.77) | 14.58 (3.31) |
| WHOQOL psychological, mean (SD) | 15.17 (2.99) | 13.38 (3.42) |
| WHOQOL relationships, mean (SD) | 14.41 (3.47) | 13.67 (3.83) |
| WHOQOL environment, mean (SD) | 15.96 (2.43) | 14.35 (2.97) |

*n*, frequencies; SD, standard deviation; Monthly household income bands = Austria (<€1k, €1k to €2k, €2k to €3k, €3k to €4k, >€4k), UK (<£900, £900–1,800, £1,800–2,700, £2,700–3,600, >£3,600); PHQ-8, depression severity measure excluding suicidal ideation; GAD-7, anxiety severity; CD-RISC-10, resilience; WHOQOL, quality of life across four domains. *depression status is based on PHQ-9 ≥ 10.

care. Overall, this network model explained 49.5% of variability in suicidal ideation in the UK sample. The model's classification accuracy was better within-sample (AUC = 0.93; PPV = 0.75; NPV = 0.90) than out-of-sample (AUC = 0.79; PPV = 0.51; NPV = 0.91), with a prediction shrinkage of AUC = 0.14.

## Discussion

Using large and representative samples from two European countries, this study identified psychosocial risk and protective factors for suicidal ideation during the acute phase of the COVID-19 lockdown. Fifteen relevant factors were identified, of which seven were replicated in both countries: depression, loneliness, anxiety, self-efficacy, resilience, and quality of life related to physical health and the living environment. These results are consistent with evidence from prior meta-analyses and systematic reviews focusing on mood disorders (13, 30, 31), loneliness (32), and poor physical health (33), which are well-known risk factors for suicidal thoughts and behavior. Similarly, self-efficacy (34) and resilience (12) have been found to be inversely related to suicide ideation as supported by the present findings.

The COVID-19 pandemic might have exacerbated the impact of some of these risk factors. For example, the significant negative consequences of isolation and social distancing might increase loneliness (35), which was found to range among the most important risk factors in both countries, explaining 16.3% of variability in suicidal ideation in the Austrian sample and 10.9% in the UK sample. Depression and anxiety, which also ranged among the five most important factors for suicidal ideation in both countries, were also found to significantly increase during the COVID-19 lockdown as compared to previous epidemiological data (36, 37).

The substantial effects of the COVID-19 pandemic on the global economy have been predicted to cause an increase in suicides related to an increase in the unemployment rate of about 2,135 (low scenario) to 9,570 (high scenario) per year (5). Also a narrative historical paper examining how previous disasters (natural disasters, violence, war, epidemics/pandemics, and economic recession) affected suicidal behavior, found that among all the types of disasters, economic recession had the most significant impact on suicide rates (38). Contrary to these studies, the current analysis revealed no association between employment status or net household income and risk of suicidal ideation in Austria as well as UK. However, the downsizing of the economy might lead to unintended long-term problems if unemployment rates rise. Therefore, results might differ from the time during the COVID-19 lockdown or some weeks/months later, as unemployment rates might increase with time, which might also cause a change in the relationship of employment status and income with suicidal ideation.

A direct comparison of the prevalence of suicidal thoughts (17.3% in Austria, 31.7% in UK) with pre-pandemic values is not possible due to a lack of comparable data. However, in the UK face-to-face interviews conducted in 2014 revealed that 5.4% of 16–74 year old participants experienced suicidal thoughts in the past year (39). Even a recent study conducted in outpatients treated for mental disorders did not report suicidal thoughts over the last 2 weeks in the majority (80%) of the patients using the same measure of suicidal thoughts as we did (19). Therefore, it can be assumed that the situation around the COVID-19 pandemic considerably increased suicidal thoughts in the general population, with more than a 1.8-fold higher prevalence in the UK compared to Austria. One explanation for the higher prevalence in the UK might be that the UK was more badly affected by the pandemic than Austria. According to available information from the World Health Organization (WHO), the UK was among the most affected countries in Europe with the highest death rates at the time of the COVID-19 lockdown, while Austria was among the less affected countries. At the time of the start of the online survey, the cumulative number of confirmed deaths related to the COVID-19 pandemic was 28.6 per 100,000 population in UK compared to 3.3 deaths per 100,000 population in Austria (40, 41). However, further studies are required to reveal the underlying causes in the different prevalence rates of suicidal ideation. A number of culture-specific differences between both countries exist. For instance, the mental healthcare system is organized differently in both countries. While in the UK mental health care is widely available through the National Health Service (NHS), providing free of charge mental health services for individuals who are eligible for it (42), in Austria no general agreement covering psychotherapeutic care by national health services or social insurance institutions exists, with only a small fraction of all patients receiving a full refund of treatment costs, while the majority receives a small subsidization and funds their psychotherapeutic treatment themselves (43).

Furthermore, distinctive risk factors were identified in each country, providing evidence that suicidality is also influenced by culturally specific factors. For example, in the Austrian sample, insomnia increased risk, whereas psychological wellbeing and quality of relationships were protective factors. In the UK sample, suicidality was influenced by age, housing space, children

10.3389/fpubh.2023.1010264

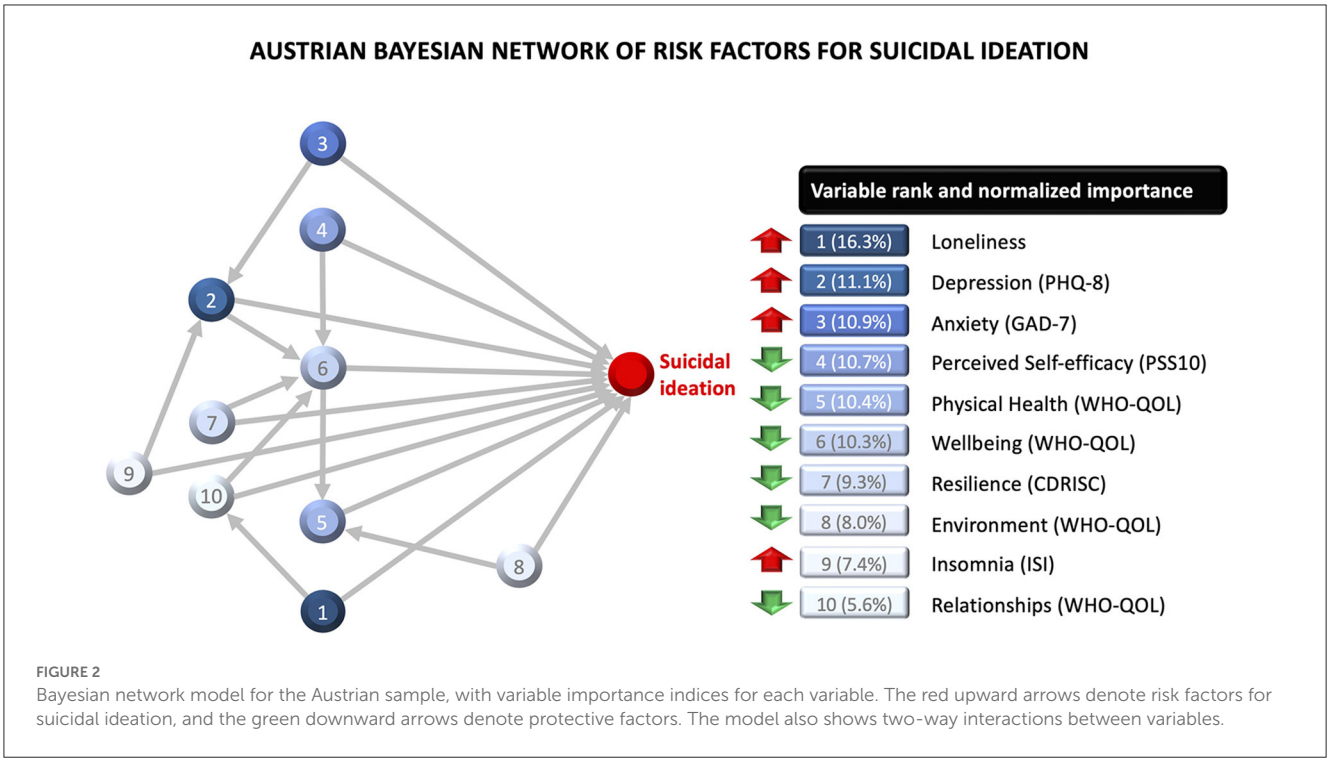



FIGURE 2
Bayesian network model for the Austrian sample, with variable importance indices for each variable. The red upward arrows denote risk factors for suicidal ideation, and the green downward arrows denote protective factors. The model also shows two-way interactions between variables.

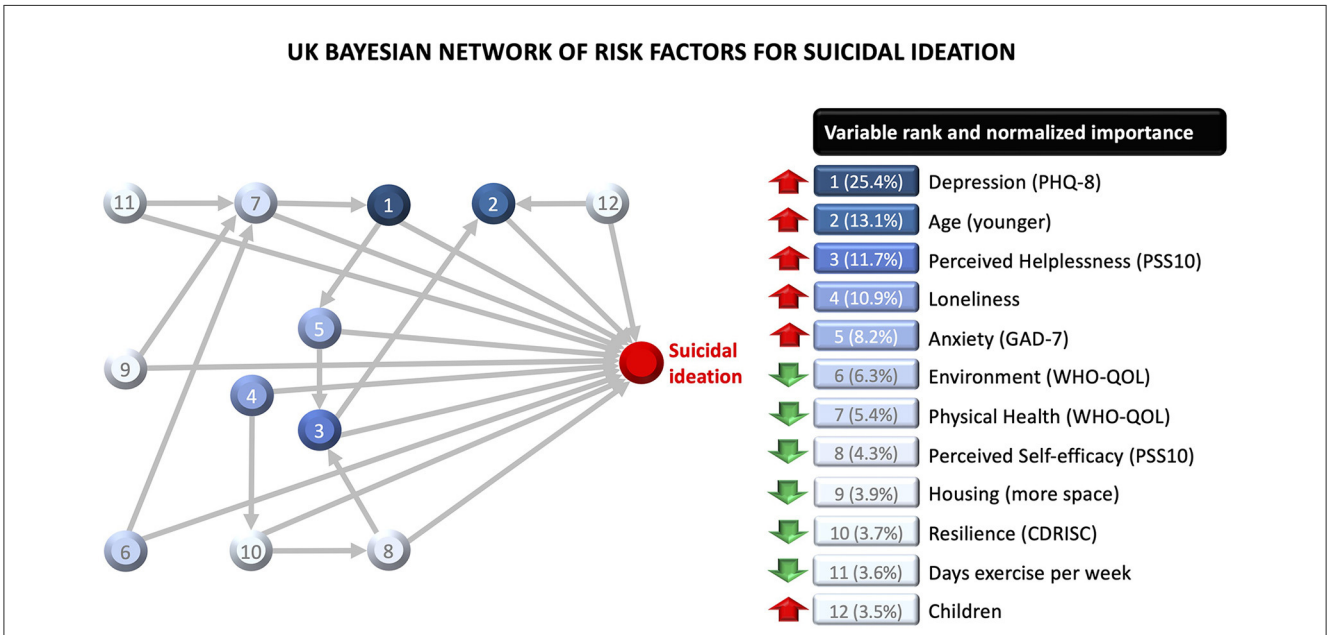



FIGURE 3
Bayesian network model for the British sample, with variable importance indices for each variable. The red upward arrows denote risk factors for suicidal ideation, and the green downward arrows denote protective factors. The model also shows two-way interactions between variables.

requiring care, exercise and perceived helplessness. The application of Bayesian network models enabled the discovery of complex interrelationships between protective and risk factors. Observed interactions indicate that suicidality is influenced by an interplay of relational (parenthood, loneliness, quality of relationships), health indicators (physical health, depression, anxiety, exercise) and living conditions (housing space, quality of the environment). Of note, the effect of physical health was moderated by quality of the living environment in both countries. This fits with wider evidence that people living in socioeconomically deprived neighborhoods tend to have poorer overall physical and mental health (44–46), and with the notion that adverse life circumstances can lead to a sense of defeat and entrapment—as posited by the *integrated motivational–volitional model* of suicide (47). Furthermore, the

important role of loneliness in the networks modeled in both countries is also consistent with contemporary theories such as the *interpersonal theory* (48) and the *three-step theory* of suicide (49). These results support the notion that the pathways to suicide ideation are complex, resulting from an interplay between several risk and protective factors (47, 50). A more precise understanding of the interrelations between key risk and protective factors can advance our efforts to rapidly identify people "at risk" of suicide, and to intervene early enough to prevent a transition from ideation to action, which is a central goal of most theories related to suicide prevention (51).

Aside from enabling the discovery of complex relationships among variables, the Bayesian network models had high predictive accuracy, explaining close to 50% of variability in suicidal ideation, which is a major improvement in terms of prognostic assessment and the identification of "at risk" cases. Furthermore, the variable importance indices displayed in Figures 1 and 2 demonstrate that this predictive value is not mainly driven by well-known risk factors such as depression and anxiety severity. In fact, depression and anxiety accounted for 22.0% (Austria) to 33.6% (UK) of the predictive value of the full network model. Classification accuracy was good (AUC .84) to excellent (AUC .93) within-sample, according to conventional standards in clinical medicine (52). The Austrian network model generalized impressively well to the UK sample, with minimal prediction shrinkage, since it was less complex and the majority of its predictors were common across countries. Higher out-of-sample prediction shrinkage was observed for the UK model, since it had a greater number of predictors that were country-specific. Overall, this cross-country prediction analysis indicates that the features contained in the more parsimonious of the two network models (Austria) has impressive generalizability to cases from a different sample and geographical region.

## Strengths and limitations

The major strengths of the study are the large, representative sample sizes and the cross-country cross-validation design. The conduct of the study in two countries, which were affected differently by the COVID-19 pandemic, allowed the investigation of the generalizability of predictors of suicidal ideation across countries. A further strength is the extensive battery of psychosocial variables and the application of machine learning approaches, enabling the modeling of interrelationships between several factors in a data-driven way. However, whether these high accuracies can be maintained or not in a non-pandemic context with lower base rates of suicidal ideation needs to be evaluated in further studies.

One major limitation of the study is its cross-sectional design, which does not allow a clear elucidation of the direction of the identified relationships, as suicidal ideation and behavior likely follow a cyclical nature (47). As no longitudinal assessments of the different risk and protective factors were conducted, this study was also not able to capture potential dynamics of changes in risk and protection states (14). A further limitation is that the network analysis applied in this study is only able to reveal two-way interactions between variables, whereas associations between three or more variables were not modeled. Furthermore, only self-ratings were used in the current study and clinician assessments were not applied, which might overestimate prevalence as people are often biased when they report their own experiences (53).

## Conclusions

Suicidal ideation can be accurately predicted using data from multiple risk and protective factors. Some of these factors were replicated across different countries, which is indicative of generalizability. The adverse consequences of the COVID-19 pandemic on increased depressive and anxiety symptoms, loneliness, and their strong connection to risk of suicidal ideation highlight the need to take urgent steps to prevent increased suicide rates during as well as in the aftermath of the COVID-19 pandemic.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation. Data requests should be addressed in writing to TP (Thomas.Probst@donau-uni.ac.at).

## Ethics statement

This study involving human participants was reviewed and approved by an Independent Research Ethics Committee at Danube University Krems. Informed consent was obtained from all study participants.

## Author contributions

CP and TP were joint principal investigators and responsible for the design and conduct of the study. SB and EH supported survey design, data collection, and preparation. JD conducted data analysis. All authors contributed to the interpretation of results, writing, editing, and approval of the manuscript.

## 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

# References


1. Saxena S, Krug EG, Chestnov O. *Preventing Suicide: A Global Imperative*. World Health Organization (2014). Available online at: https://www.who.int/publications/i/item/9789241564779 (accessed July 27, 2020).

2. Glenn CR, Kleiman EM, Kellerman J, Pollak O, Cha CB, Esposito EC, et al. Annual research review: a meta-analytic review of worldwide suicide rates in adolescents. *J Child Psychol Psychiatry*. (2020) 61:294–308. doi: 10.1111/jcpp.13106

3. Hasson-Ohayon I, Lysaker PH. Special challenges in psychotherapy continuation and adaption for persons with schizophrenia in the age of coronavirus (COVID-19). *Couns Psychol Q*. (2020) 17:1–9. doi: 10.1080/09515070.2020.1781595

4. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *Lancet*. (2020) 395:912–20. doi: 10.1016/S0140-6736(20)30460-8

5. Kawohl W, Nordt C. COVID-19, unemployment, and suicide. *Lancet Psychiatry*. (2020) 7:389–90. doi: 10.1016/S2215-0366(20)30141-3

6. Aquila I, Sacco MA, Ricci C, Gratteri S, Montebianco Abenavoli L, Oliva A, et al. The role of the COVID-19 pandemic as a risk factor for suicide: what is its impact on the public mental health state today? *Psychol Trauma*. (2020) 12:S120–2. doi: 10.1037/tra0000616

7. Klomek AB. Suicide prevention during the COVID-19 outbreak. *Lancet Psychiatry*. (2020) 7:390. doi: 10.1016/S2215-0366(20)30142-5

8. Zhai Y, Du X. Mental health care for international Chinese students affected by the COVID-19 outbreak. *Lancet Psychiatry*. (2020) 7:e22. doi: 10.1016/S2215-0366(20)30089-4

9. 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

10. Gili M, Castellví P, Vives M, de la Torre-Luque A, Almenara J, Blasco MJ, et al. Mental disorders as risk factors for suicidal behavior in young people: a meta-analysis and systematic review of longitudinal studies. *J Affect Disord*. (2019) 245:152–62. doi: 10.1016/j.jad.2018.10.115

11. Yoshimasu K, Kiyohara C, Miyashita K, Stress Research Group of the Japanese Society for Hygiene. Suicidal risk factors and completed suicide: meta-analyses based on psychological autopsy studies. *Environ Health Prev Med*. (2008) 13:243–56. doi: 10.1007/s12199-008-0037-x

12. De Beurs D, Fried EI, Wetherall K, Cleare S, O' Connor DB, Ferguson E, et al. Exploring the psychology of suicidal ideation: a theory driven network analysis. *Behav Res Ther*. (2019) 120:103419. doi: 10.1016/j.brat.2019.103419

13. Nock MK, Borges G, Bromet EJ, Alonso J, Angermeyer M, Beautrais A, et al. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *Br J Psychiatry*. (2008) 192:98–105. doi: 10.1192/bjp.bp.107.040113

14. Allen NB, Nelson BW, Brent D, Auerbach RP. Short-term prediction of suicidal thoughts and behaviors in adolescents: can recent developments in technology and computational science provide a breakthrough? *J Affect Disord*. (2019) 250:163–9. doi: 10.1016/j.jad.2019.03.044

15. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med*. (2018) 52:446–62. doi: 10.1007/s12160-016-9830-8

16. Junqué de Fortuny E, Martens D, Provost F. Predictive modeling with big data: is bigger really better? *Big Data*. (2013) 1:215–26. doi: 10.1089/big.2013.0037

17. Contreras A, Nieto I, Valiente C, Espinosa R, Vazquez C. The study of psychopathology from the network analysis perspective: a systematic review. *Psychother Psychosom*. (2019) 88:71–83. doi: 10.1159/000497425

18. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Mach Learn*. (1997) 29:131–63. doi: 10.1023/A:1007465528199

19. Rossom RC, Coleman KJ, Ahmedani BK, Beck A, Johnson E, Oliver M, et al. Suicidal ideation reported on the PHQ9 and risk of suicidal behavior across age groups. *J Affect Disord*. (2017) 215:77–84. doi: 10.1016/j.jad.2017.03.037

20. 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

21. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord*. (2009) 114:163–73. doi: 10.1016/j.jad.2008.06.026

22. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Ann Intern Med*. (2007) 146:317–25. doi: 10.7326/0003-4819-146-5-200703060-00004

23. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. *J Health Soc Behav*. (1983) 24:385–96. doi: 10.2307/2136404

24. Morin CM, Belleville G, Bélanger L, Ivers H. The Insomnia Severity Index: psychometric indicators to detect insomnia cases and evaluate treatment response. *Sleep*. (2011) 34:601–8. doi: 10.1093/sleep/34.5.601

25. Skevington SM, Lotfy M, O'Connell KA. The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. *Qual Life Res*. (2004) 13:299–310. doi: 10.1023/B:QURE.0000018486.91360.00

26. De Jong-Gierveld J, Kamphuls F. The development of a Rasch-type loneliness scale. *Appl Psychol Meas*. (1985) 9:289–99. doi: 10.1177/014662168500900307

27. Campbell-Sills L, Stein MB. Psychometric analysis and refinement of the Connor–Davidson Resilience Scale (CD-RISC): validation of a 10-item measure of resilience. *J Trauma Stress*. (2007) 20:1019–28. doi: 10.1002/jts.20271

28. Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans Pattern Anal Mach Intell*. (2009) 32:569–75. doi: 10.1109/TPAMI.2009.187

29. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol*. (2008) 8:70. doi: 10.1186/1471-2288-8-70

30. Mars B, Heron J, Klonsky ED, Moran P, O'Connor RC, Tilling K, et al. Predictors of future suicide attempt among adolescents with suicidal thoughts or non-suicidal self-harm: a population-based birth cohort study. *Lancet Psychiatry*. (2019) 6:327–37. doi: 10.1016/S2215-0366(19)30030-6

31. Solano P, Aguglia A, Caprino M, Conigliaro C, Giacomini G, Serafini G, et al. The personal experience of severe suicidal behaviour leads to negative attitudes towards self-and other's suicidal thoughts and behaviours: a study of temperaments, coping strategies, and attitudes towards suicide among medical students. *Psychiatry Res*. (2019) 272:669–75. doi: 10.1016/j.psychres.2018.12.116

32. McClelland H, Evans JJ, Nowland R, Ferguson E, O'Connor RC. Loneliness as a predictor of suicidal ideation and behaviour: a systematic review and meta-analysis of prospective studies. *J Affect Disord*. (2020) 274:880–96. doi: 10.1016/j.jad.2020.05.004

33. Russell D, Turner RJ, Joiner TE. Physical disability and suicidal ideation: a community-based study of risk/protective factors for suicidal thoughts. *Suicide Life Threat Behav*. (2009) 39:440–51. doi: 10.1521/suli.2009.39.4.440

34. Kobayashi Y, Fujita K, Kaneko Y, Motohashi Y. Self-efficacy as a suicidal ideation predictor: a population cohort study in rural Japan. *Open J Prev Med*. (2015) 5:61–71. doi: 10.4236/ojpm.2015.52007

35. Levi-Belz Y, Aisenberg D. Together we stand: suicide risk and suicide prevention among Israeli older adults during and after the COVID-19 world crisis. *Psychol Trauma*. (2020) 12:S123–5. doi: 10.1037/tra0000667

36. Pieh C, Budimir S, Probst T. The effect of age, gender, income, work, and physical activity on mental health during coronavirus disease (COVID-19) lockdown in Austria. *J Psychosom Res*. (2020) 136:110186. doi: 10.1016/j.jpsychores.2020.110186

37. Pieh C, Budimir S, Delgadillo J, Barkham M, Fontaine JRJ, Probst T. Mental health during COVID-19 lockdown in the United Kingdom. *Psychosom Med*. (2021) 84:328–37. doi: 10.1097/PSY.0000000000000871

38. Devitt P. Can we expect an increased suicide rate due to Covid-19? *Ir J Psychol Med*. (2020) 37:264–8. doi: 10.1017/ipm.2020.46

39. McManus S, Bebbington P, Jenkins R, Brugha T. *Mental Health and Wellbeing in England: Adult Psychiatric Morbidity Survey*. (2014). Leeds: NHS Digital.

40. *EUROSTAT Population on 1 January by Age and Sex*. Available online at: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_pjan&lang=en (accessed July 15, 2020).

41. World Health Organization (WHO). *WHO Coronavirus Disease (COVID-19) Dashboard*. Available online at: https://covid19.who.int/ (accessed July 15, 2020).

42. National Health Service (NHS). *How You Can Access NHS Mental Health Services*. Available online at: https://www.nhs.uk/mental-health/social-care-and-your-rights/how-to-access-mental-health-services/ (accessed January 20, 2023).

43. Heidegger KE. *The Situation of Psychotherapy in Austria*. Österreichischer Bundesverband für Psychotherapie (2017). Available online at: https://www.europsyche.org/app/uploads/2019/05/Situation-Psychotherapy-in-Austria-2017-10-20.pdf (accessed January 20, 2023).

44. Fryers T, Melzer D, Jenkins R. Social inequalities and the common mental disorders. *Soc Psychiatry Psychiatr Epidemiol*. (2003) 38:229–37. doi: 10.1007/s00127-003-0627-2

45. Silva M, Loureiro A, Cardoso G. Social determinants of mental health: a review of the evidence. *Eur J Psychiatry*. (2016) 30:259–92.

46. Wilkinson RG, Pickett KE. The problems of relative deprivation: why some societies do better than others. *Soc Sci Med.* (2007) 65:1965–78. doi: 10.1016/j.socscimed.2007.05.041

47. O'Connor RC, Kirtley OJ. The integrated motivational–volitional model of suicidal behaviour. *Philos Trans R Soc Lond B Biol Sci.* (2018) 373:20170268. doi: 10.1098/rstb.2017.0268

48. Joiner T. *Why People Die by Suicide.* Cambridge: Harvard University Press (2007).

49. Klonsky ED, May AM. The three-step theory (3ST): a new theory of suicide rooted in the "ideation-to-action" framework. *Int J Cogn Ther.* (2015) 8:114–29. doi: 10.1521/ijct.2015.8.2.114

50. Klonsky ED, May AM, Saffer BY. Suicide, suicide attempts, and suicidal ideation. *Annu Rev Clin Psychol.* (2016) 12:307–30. doi: 10.1146/annurev-clinpsy-021815-093204

51. Klonsky ED, Saffer BY, Bryan CJ. Ideation-to-action theories of suicide: a conceptual and empirical update. *Curr Opin Psychol.* (2018) 22:38–43. doi: 10.1016/j.copsyc.2017.07.020

52. Swets JA. Measuring the accuracy of diagnostic systems. *Science.* (1988) 240:1285–93. doi: 10.1126/science.3287615

53. Devaux M, Sassi F. Social disparities in hazardous alcohol use: self-report bias may lead to incorrect estimates. *Eur J Public Health.* (2016) 26:129–34. doi: 10.1093/eurpub/ckv190

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 17 March 2023
DOI 10.3389/fpubh.2023.1082139



OPEN ACCESS

EDITED BY
Patrick K. A. Neff,
University of Zurich, Switzerland

REVIEWED BY
Venkata Ramana Murthy Oruganti,
Amrita Vishwa Vidyapeetham University, India
Abdul Rehman Javed,
Air University, Pakistan

*CORRESPONDENCE
Tingshao Zhu
✉ tszhu@psych.ac.cn
Shaoshuai Gao
✉ ssgao@ucas.ac.cn

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

RECEIVED 27 October 2022
ACCEPTED 27 February 2023
PUBLISHED 17 March 2023

CITATION
Wen Y, Li B, Liu X, Chen D, Gao S and Zhu T (2023) Using gait videos to automatically assess anxiety. *Front. Public Health* 11:1082139. doi: 10.3389/fpubh.2023.1082139



# Using gait videos to automatically assess anxiety

Yeye Wen[1,2], Baobin Li[3], Xiaoqian Liu[2], Deyuan Chen[1], Shaoshuai Gao[1]* and Tingshao Zhu[2,4]*

[1]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China, [2]Institute of Psychology, Chinese Academy of Sciences, Beijing, China, [3]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China, [4]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

**Background:** In recent years, the number of people with anxiety disorders has increased worldwide. Methods for identifying anxiety through objective clues are not yet mature, and the reliability and validity of existing modeling methods have not been tested. The objective of this paper is to propose an automatic anxiety assessment model with good reliability and validity.

**Methods:** This study collected 2D gait videos and Generalized Anxiety Disorder (GAD-7) scale data from 150 participants. We extracted static and dynamic time-domain features and frequency-domain features from the gait videos and used various machine learning approaches to build anxiety assessment models. We evaluated the reliability and validity of the models by comparing the influence of factors such as the frequency-domain feature construction method, training data size, time-frequency features, gender, and odd and even frame data on the model.

**Results:** The results show that the number of wavelet decomposition layers has a significant impact on the frequency-domain feature modeling, while the size of the gait training data has little impact on the modeling effect. In this study, the time-frequency features contributed to the modeling, with the dynamic features contributing more than the static features. Our model predicts anxiety significantly better in women than in men ($r_{Male} = 0.666$, $r_{Female} = 0.763$, $p < 0.001$). The best correlation coefficient between the model prediction scores and scale scores for all participants is 0.725 ($p < 0.001$). The correlation coefficient between the model prediction scores for odd and even frame data is 0.801~0.883 ($p < 0.001$).

**Conclusion:** This study shows that anxiety assessment based on 2D gait video modeling is reliable and effective. Moreover, we provide a basis for the development of a real-time, convenient and non-invasive automatic anxiety assessment method.

KEYWORDS

**anxiety assessment, mental health, gait video, machine learning, reliability and validity**

## 1. Introduction

The increasing pressure of modern life has led to a decline in global mental health and an increase in anxiety and depression (1). Anxiety disorders are the most common mental health problems worldwide and may cause physiological reactions such as irritability, fatigue, and increased heart rate. A long-term intense anxious state not only affects an individual's social, life, and work responsibilities but also has a serious impact on their physical health (2). Therefore, to improve the mental health of different groups, the demand for mental health services has increased worldwide (3, 4). Fortunately, in recent years, researchers have made new progress in the treatment of mental diseases such as anxiety and depression (5, 6). At the same time, we urgently need to develop a convenient and timely method for assessing anxiety states.

In psychology, the anxiety scale has been carefully designed, revised and tested, and various scale-based assessment methods have been developed (7). Self-reports rely on individuals reporting their symptoms, behaviors, and attitudes (8). At present, self-reports remain the most commonly used and most effective anxiety assessment method (9). However, scale-based assessments have some limitations and are not applicable in some scenarios (10). For example, in scenarios that require multiple measurements, participants completing the same questionnaire multiple times can lead to practice effects (11). In scenarios such as job interviews, scale results may be inaccurate due to social desirability (12). In addition, the self-report method is not suitable for certain populations, such as illiterate or dyslexic individuals. Therefore, we hope to develop more objective indicators to assess anxiety.

Anxiety can affect an individual's physiological responses. Anxious individuals may experience shortness of breath and accelerated heartbeat (2). In addition, fear is a typical symptom of anxiety disorders, and patients may experience muscle tension (13), sweating, trembling (14), and skin conductance and heart rate changes (15). Anxiety-induced fear can also be reflected through facial expressions (16). Giannakakis et al. showed that some specific facial cues, such as eye and mouth movements, are suitable as discriminative indicators of anxiety (17). Anxiety may also be reflected in voice changes. In anxious states, individuals tend to speak quickly at a loud volume (18), showing fewer voice changes and more pauses (19). Gait and anxiety are also related. Gait posture and movement characteristics can indicate a variety of emotions (20, 21). For example, individuals with anxiety tend to pace back and forth (22). Feldman et al. found that compared with healthy people, anxious patients have shorter stride distances and take fewer steps per minute, displaying movement disorders to some extent (23). Other researchers have noted similar characteristics, such as slow gait (24, 25) and balance dysfunction (26, 27). In addition, arm swings, vertical head movements, and lateral upper body swings have also been associated with anxiety (28). Among the various physiological and behavioral characteristics related to anxiety, gait has several advantages, including large variations, non-invasiveness and ease of observation. Thus, gait can serve as an objective indicator for assessing anxiety.

To acquire gait data, some researchers have used body-worn sensors (29), human motion capture systems (30, 31), Kinects (Xbox One Kinect Sensor) (32) and other devices. However, these devices are expensive and complex to operate, which is not conducive to improving the applicability of anxiety assessment methods. In this study, we recorded 2D gait videos using a common camera that is simple to operate, increasing the ease of obtaining data.

In recent years, with the development of machine learning technology, various researchers have used gait to assess anxiety. Jing et al. found that a prediction model based on gait features performed better than a prediction model based on speech features (33). Miao et al. and Zhao et al. established anxiety assessment models, and the correlation coefficients between the anxiety prediction score and the scale score reached 0.4 (34) and 0.51 (35), respectively. Both studies considered the basic statistics of the gait time series data and the amplitude in the frequency domain after a Fourier transform as features. These features are relatively simple, which may increase the make it difficult to express the rich movement characteristics of gait. In addition, these features lack biological or kinematic interpretations. Stark et al. considered five main gait parameters to identify anxiety, namely, the turning angle, neck variance, lumbar rotation, lumbar movement in the sagittal plane, and arm movement (36). Although the above studies established different anxiety assessment models, they did not comprehensively evaluate the model reliability and validity, and did not adequately validate the performance of their models.

In this study, we used 2D gait videos to construct static and dynamic time-domain features and frequency-domain features and established anxiety prediction models through machine learning algorithms. To validate the proposed models, we examined the effects of different frequency-domain feature construction methods, training data sizes and gender on model performance and compared the contributions of different time-frequency features to the modeling results. In addition, we tested the odd-even split-half reliability of the proposed anxiety assessment model. The goal of this study is to provide a convenient auxiliary anxiety assessment method.

The contributions of this study are as follows:

- Build anxiety assessment models using easily accessible 2D gait videos, reducing cost and increasing convenience of anxiety assessment. It was verified that a good anxiety assessment model can be built without using longer gait videos.
- We constructed static and dynamic time-domain features and frequency-domain features with biological kinematic significance, and proved the rationality and necessity of constructing features.
- This study carefully evaluated the performance (validity and reliability) of the anxiety assessment model through experiments. We validated differences in anxiety assessment between men and women, and verified the robustness of our model in a video odd-even split-half test.

The rest of this paper is organized as follows. First, we introduce the research methods and experiments in Section Methods, including the collection and preprocessing of gait data, feature engineering and modeling, and experimental procedures. Then, the results of several comparative experiments are reported in Section Results. A general discussion of the results is given in Section Discussion, explaining the findings of the study and illustrating further work. Finally, concluding remarks is presented in Section Conclusion.

## 2. Methods

In this study, we used a camera to capture participant gait videos (walking back and forth) indoors. The specific gait video collection method is similar to the method described in Wen et al. (37).

After the gait videos were collected, the participants immediately completed a 7-item Generalized Anxiety Disorder (GAD-7) scale assessment. The GAD-7 assessment is a valid and efficient tool for identifying GAD and assessing its severity in clinical practice and research (9). It evaluates anxiety states in the previous 2 weeks and divides anxiety into four levels according

10.3389/fpubh.2023.1082139

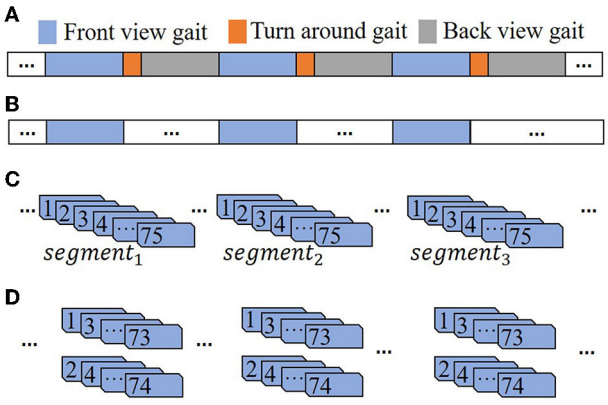



FIGURE 1
Gait video data segmentation process. **(A)** Full gait video. **(B)** Only the front-view gait segments of the video are kept. **(C)** Keep 75 frames. **(D)** Split odd and even frame segments.

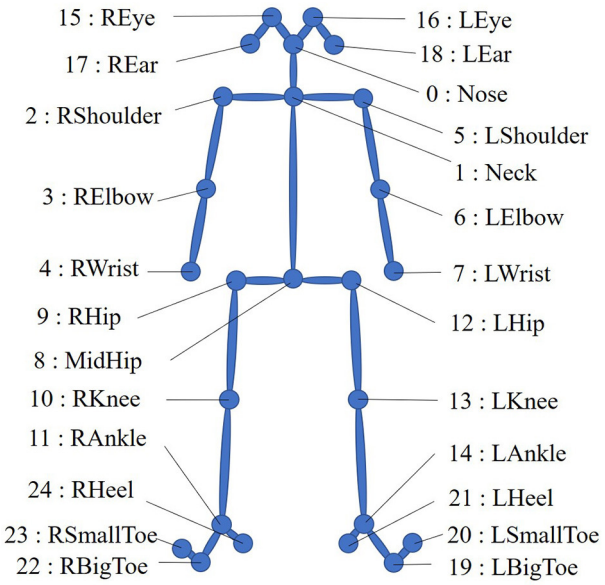



FIGURE 2
Twenty-five human body key points in OpenPose.

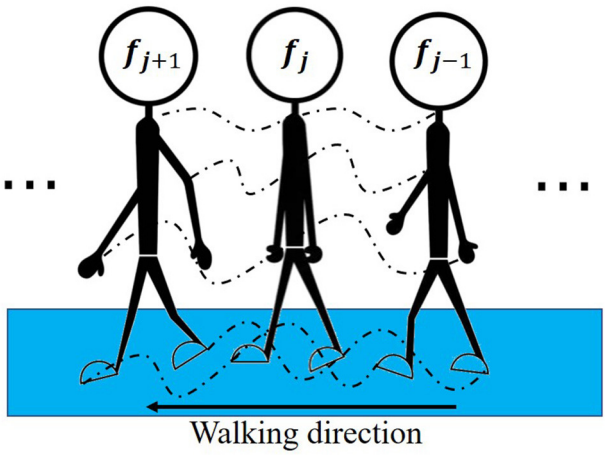



FIGURE 3
Diagram of the interframe difference. $f_{j-1}$, $f_j$ and $f_{j+1}$ represent three adjacent gait images in the gait video. The dotted line represents the movement trajectory of the key point.

to the scale scores, namely, minimal anxiety (0–4), mild anxiety (5–9), moderate anxiety (10–14), and severe anxiety (15–21). The GAD-7 assessment shows good internal consistency (Cronbach $\alpha = 0.92$) and test-retest reliability (intraclass correlation = 0.83) (9).

Permission for the above protocol was obtained from the Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences (Approval number: H15010).

We obtained ~2-min gait videos for each participant, including front and back gaits. Since the front-view gait skeleton evaluation is more accurate than that the back-view evaluation (38), we analyzed skeletons only from the front view to obtain more precise features. Previous studies have shown that good models can be built using a small number of gait frames (35). We kept three consecutive front-view gait segments for each participant, and each segment included 75 frames. To assess the odd-even split-half reliability of the model, we divided the first 74 frames in the gait data into two sets by considering odd and even frames. The gait data segmentation process is shown in Figure 1.

The preprocessing method is similar to the approach proposed in Wen et al. (37). We used OpenPose (39) (a multiperson 2D pose recognition system) to extract the 2D coordinates of 25 body key points from the gait videos and performed coordinate translation (with the MidHip key point as the coordinate origin) and smoothing on the coordinate sequence. Figure 2 shows the 25 human body key points in OpenPose.

The gait coordinate sequence obtained after preprocessing includes only isolated coordinate points and thus does not reflect changes between frames and variations between different key points. We call the features obtained from such data static time-domain features. To reflect the changing gait characteristics (40), we calculate the interframe difference and construct the distances between joints (see Supplementary Table A) and angles between joints (see Supplementary Table B) to express dynamic information. We term these features dynamic time-domain features. The method for obtaining the static and dynamic time-domain features is similar to Wen et al. (37). Figure 3 shows a diagram of the interframe difference between $f_{j-1}$, $f_j$, and $f_{j+1}$ in a gait video. The motion track of the key points between each frame contains the interframe difference information.

In gait, some movement patterns are more easily reflected in the frequency domain (41). Relevant studies have extracted frequency-domain gait features through Fourier transforms (34, 35). However, Fourier transforms (42) cannot be applied in multiresolution analyses in the frequency domain. Thus, we use wavelet transforms (43) to analyze the frequency variation characteristics of the joint distances in the frequency domain.

We use the *db*1 wavelet base to decompose the distance between joints into an approximation coefficient array $A_3$ representing low-frequency signals and detail coefficient arrays $D_1$, $D_2$, and $D_3$

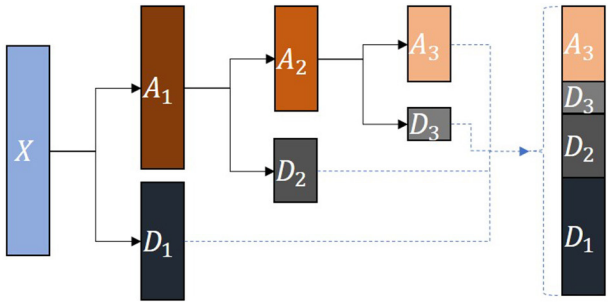



FIGURE 4
Wavelet decomposition process. $X$ represents the source signal. $A_1$, $A_2$ and $A_3$ are the approximation coefficient arrays obtained by decomposing each layer. $D_1$, $D_2$ and $D_3$ are the detail coefficient arrays obtained by decomposing each layer.

**Algorithm:** Sequential Forward Selection.
**Input:**
$X$: The whole feature set
$J$: The model evaluation rules (Using RMSE)
**Output:**
S: The best subset of features
**Method:**
(1) Create an empty subset $Z = \{\varnothing\}$
(2) **repeat**
(3) Select best remaining feature:
$x = \arg\min_{x' \epsilon Z_k}[J(Z + x)]$
(4) Update $Z = Z + x$
(5) $S = Z$
(6) **until** not decreased in $J$ OR $Z = X$

Algorithm 1. Pseudocode for the Sequential Forward Selection algorithm.

representing high-frequency signals. Figure 4 shows the three-layer wavelet decomposition process.

We used 10 feature extraction functions to extract the above time-domain and frequency-domain features. These functions include the maximum, minimum, mean, median, variance, root mean square, skewness, kurtosis, absolute energy, and coefficient of variation in the sequence data. The specific feature extraction functions are shown in Supplementary Table C.

We used z-score standardization (44) to eliminate differences in the values and dimensions of features. The z-score standardization is defined as:

$$x' = \frac{x - \bar{x}}{\sigma_x}$$

Where $x$ is the sample mean and $\sigma_x$ is the sample standard deviation. Then, we used principal component analysis (PCA) (45) to remove redundant features and sequential forward selection (SFS) (46) to automatically identify feature combinations that resulted in optimal model performance. SFS is a greedy search algorithm. At each stage, according to the evaluation rules, the SFS algorithm continuously selects the optimal feature from the remaining features to determine the optimal feature subset. The SFS pseudocode is shown in Algorithm 1.

We selected 3 typical machine learning regression algorithms for modeling, namely, Gaussian process regression (GPR), linear regression (LR), and support vector regression (SVR), where the SVR models included the "linear," "poly," "rbf," and "sigmoid" kernel functions. We trained and tested the models with 10 rounds of 10-fold cross validation. The complete modeling process is shown in Figure 5.

In computer science, the root mean square error (RMSE) is often used to evaluate regression model performance (47) and is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(Model_n - Scale_n\right)^2}$$

Where $Model_n$ and $Scale_n$ represent the anxiety model prediction score and anxiety scale score of the nth participant, respectively.

To comprehensively evaluate the performance of the proposed anxiety assessment models, we considered reliability and validity assessment methods used in psychology. We used the Pearson correlation between the anxiety assessment model prediction scores and the anxiety scale scores as the model criterion validity. In addition, we fed different data segments into the model to obtain prediction scores and used the Pearson correlation between these different model prediction scores to evaluate model reliability.

To explore the influence of the number of wavelet decomposition layers during the construction of the frequency-domain features on the prediction results, we set the wavelet decomposition *level* parameter from 1 to 4 (the *level* parameter controls the number of wavelet decomposition layers). Figure 6 shows the effect of decomposing the original time series signal according to different numbers of wavelet layers. The signals in each column can be restored to the original signal $X$ after they are superimposed on each other.

To explore the influence of the gait video training data size on the model, we used gait segments with different numbers of frames to build various models and compared the model performance. In gait data segmentation, each participant has three segments of gait data, as shown in Figure 1. First, we used $segment_1$, $segment_2$ and $segment_3$ to establish three single-segment models. Then, two of the three segments were combined to establish three double-segment fusion models. Finally, the three segments were combined to establish a three-segment fusion model. The gait segments were combined as follows:

$$segment_{12} = segment_1 + segment_2$$
$$segment_{13} = segment_1 + segment_3$$
$$segment_{23} = segment_2 + segment_3$$
$$segment_{123} = segment_1 + segment_2 + segment_3$$

The Pearson correlation coefficients between the model prediction scores and the anxiety scale scores were calculated to evaluate the influence of the number of gait segment frames on the performance of the models.

In machine learning, some neural network components can be removed to understand their impact on the network (48). In this study, we explored the impact of different features on model

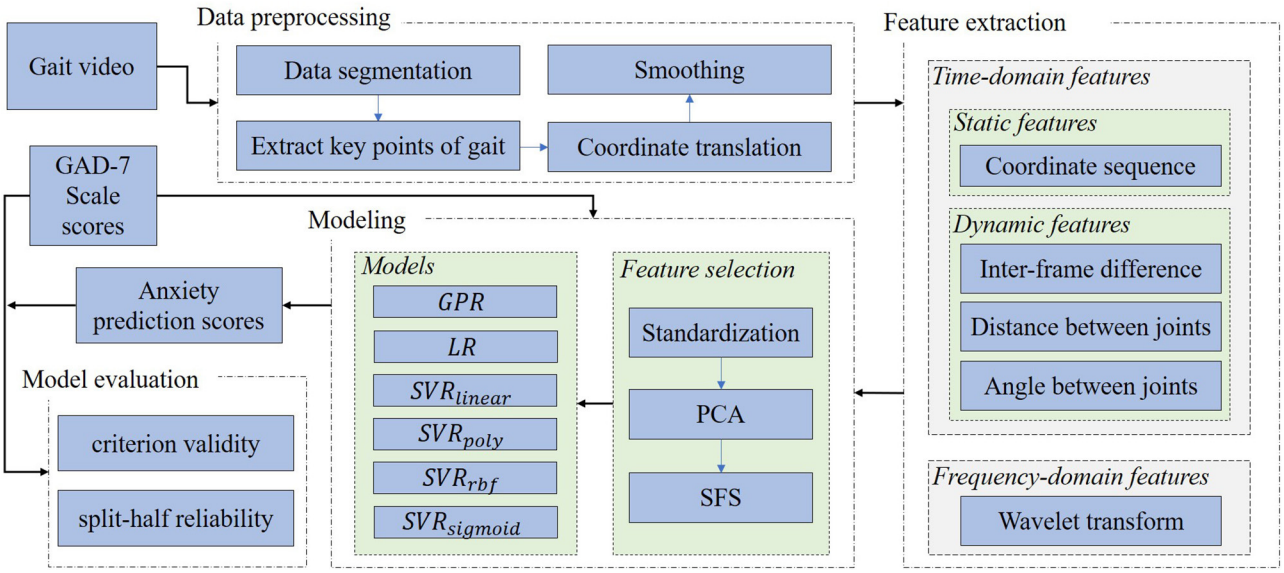



FIGURE 5
Modeling process. PCA, principal component analysis; SFS, sequential forward selection; GPR, Gaussian process regression; LR, linear regression. $SVR_{linear}$, $SVR_{poly}$, $SVR_{rbf}$, and $SVR_{sigmoid}$ represent support vector regression using linear, poly, rbf, and sigmoid kernel functions, respectively.

performance through feature ablation studies to determine whether the constructed features are effective. We used the static time-domain features, dynamic time-domain features, all time-domain features (including dynamic and static features), frequency-domain features, and all features (including all time-domain and frequency-domain features) to build 5 anxiety assessment models. The Pearson correlation coefficients between the model prediction scores and the scale scores were used to evaluate the contribution of different features to the model.

We also explored whether gender has an effect on anxiety prediction models. To accomplish this, we input the male and female gait data into the anxiety assessment model. Then, we calculated the Pearson correlation coefficients between the anxiety prediction scores of males and females and the corresponding scale scores to evaluate whether gender impacts the anxiety prediction model.

In psychology, odd-even split-half reliability is often used to characterize the degree of internal consistency of scales (49). We input the odd and even frame gait data into the anxiety assessment model to obtain the corresponding model prediction scores and used the Pearson correlation coefficient between the two prediction scores to evaluate the robustness and reliability of the model.

## 3. Results

We recruited 152 participants. According to the experimental processing requirements, 150 valid data remained after screening, including 79 males (52.67%) and 71 females (47.33%). The proportion of males and females was essentially balanced. The ages of the participants ranged from 21 to 28 years (mean = 22.99, SD = 1.07). The mean and standard deviation of the participant GAD-7 scores were 4.31 and 4.45, respectively. As shown in Table 1, the participants mainly showed minimal and mild anxiety, with 132 participants at this anxiety level (88%). There were 5 participants with severe anxiety, and all were women.

Table 2 show that in terms of the different algorithms, the *GPR* and *LR* models had the best effect, regardless of the number of wavelet decomposition layers. In terms of the number of wavelet decomposition layers, except for the $SVR_{poly}$ model (the $SVR_{poly}$ model had the best effect when $level = 2$), the performance of the other models continuously improved as the number of layers increased from $level = 1$ to $level = 3$ (the mean values of $r_{L_1}$, $r_{L_2}$ and $r_{L_3}$ were 0.401, 0.504, and 0.565, respectively). When $level = 4$, the model performance declined (the mean value of $r_{L_4}$ was 0.464). In summary, the *GPR* and *LR* models showed optimal performance when $level = 3$ ($r_{L_3\_GPR} = 0.677$, $r_{L_3\_LR} = 0.677$, $p < 0.001$, and their *RMSE* values were less than those of the other algorithms). We determined the optimal number of wavelet decomposition layers by iteratively searching parameters.

As shown in Table 3, among the 7 data combinations, the *GPR* and *LR* models had the best results. In the *GPR* and *LR* models, the modeling effects of the $segment_1$, $segment_{12}$,$segment_{13}$ and $segment_{123}$ gait segments (which all contained $segment_1$ and had mean $r_{s_1}$, $r_{s_{12}}$, $r_{s_{13}}$ and $r_{s_{123}}$ values of 0.559, 0.495, 0.495, and 0.516, respectively) were better than those of the other segments (the mean values of $r_{s_2}$, $r_{s_3}$ and $r_{s_{23}}$ were 0.425, 0.435, and 0.447, respectively). Similar trends were found for the $SVR_{rbf}$ and $SVR_{sigmoid}$ models. In conclusion, the *GPR* and *LR* models had the best performance when modeled on $segment_1$ ($r_{s_1\_GPR} = 0.731$, $r_{s_1\_LR} = 0.702$, $p < 0.001$). We found that there are some differences in the modeling effect of gait segments in different periods. Moreover, the increase in the number of gait segments did not significantly improve the model effect.

As shown in Table 4, the modeling effects of the *GPR* and *LR* models on different features were significantly better than those of

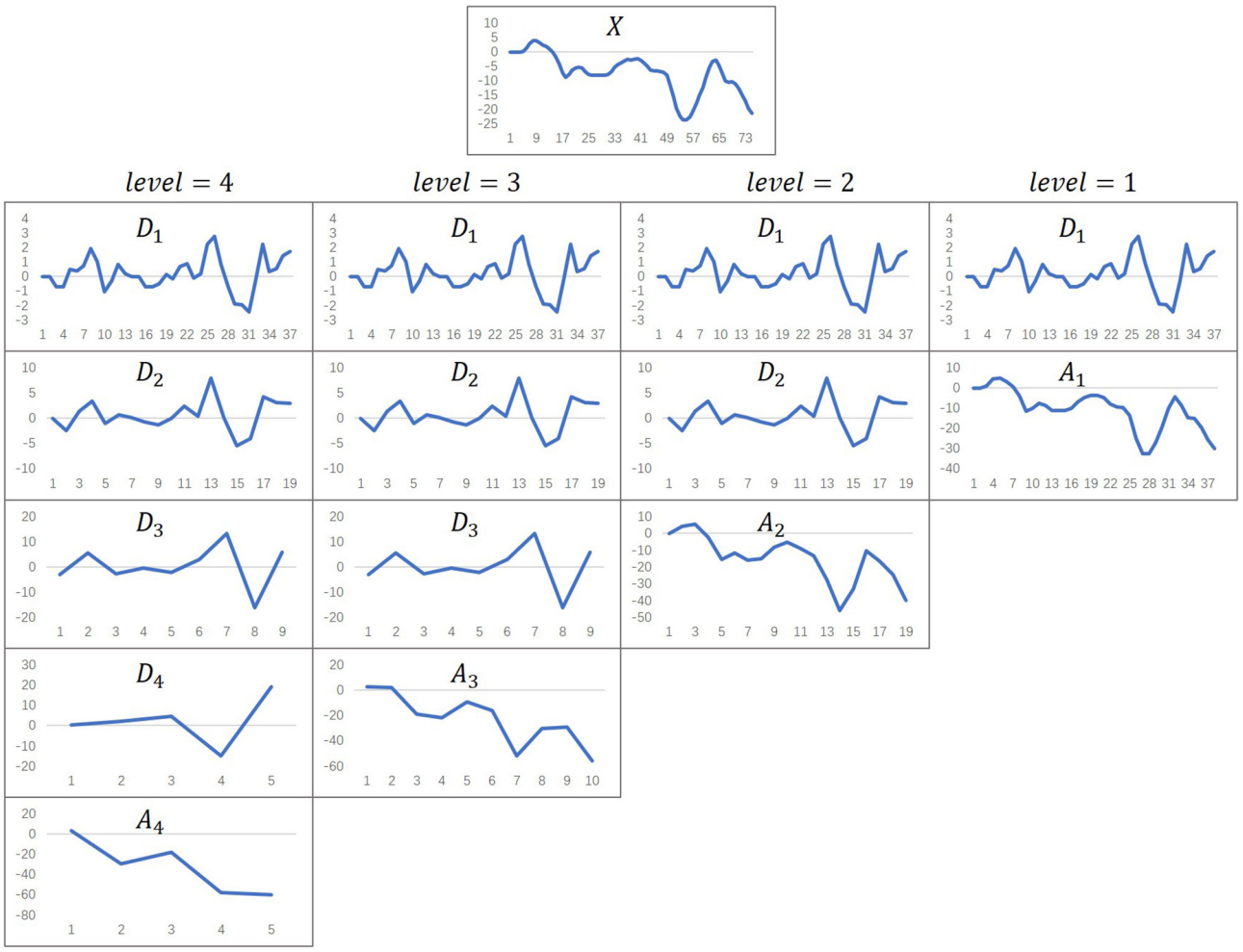



FIGURE 6
The effect of wavelet decomposition. The level parameter represents the number of wavelet decomposition layers. *X* represents the original time series data. $D_i$ $(i \in \{1, 2, 3, 4\})$ represent detail coefficient arrays. $A_j$ $(j \in \{1, 2, 3, 4\})$ represent approximation coefficient arrays.

**TABLE 1** Population distribution of GAD-7 scale scores.

| | GAD-7 scale score range | | | | Total |
|---|---|---|---|---|---|
| | 0∼4 | 5∼9 | 10∼14 | 15∼21 | |
| Male | 55 | 18 | 6 | 0 | 79 |
| Female | 41 | 18 | 7 | 5 | 71 |
| Total | 96 | 36 | 13 | 5 | 150 |

GAD-7, the 7-item Generalized Anxiety Disorder scale; 0∼4, minimal anxiety; 5∼9, mild anxiety; 10∼14, moderate anxiety; 15∼21, severe anxiety.

the other models. The *GPR* model achieved the best modeling effect on all features, including the time-domain and frequency-domain features ($r_{5\_GPR} = 0.725, p < 0.001$). The mean values of $r_1$, $r_2$, $r_3$, $r_4$, and $r_5$ were 0.399, 0.446, 0.536, 0.565, and 0.560, respectively, showing a slow increasing trend. These trends were particularly noticeable in the *GPR* and *LR* models, with $r_{5\_GPR} > r_{4\_GPR}$ and $r_{5\_LR} > r_{4\_LR}$ ($p < 0.001$). We found that the anxiety assessment models are sensitive to different gait features. And gait features with kinematic characteristics can significantly improve the performance of the model.

As shown in Table 5, the *GPR* model performed significantly better than the other models ($r_{All\_GPR} = 0.725$, $r_{Male\_GPR} = 0.666$, $r_{Female\_GPR} = 0.763$, $p < 0.001$, and its *RMSE* value was lower than those of the other algorithms). The anxiety prediction effect was better for women than for men (the mean values of $r_{Male}$ and $r_{Female}$ were 0.547 and 0.566, respectively). Except for the $SVR_{linear}$ and $SVR_{poly}$ models, all other models reflected this characteristic. We found that the prediction performance of anxiety assessment model for different groups is different.

As shown in Table 6, except for $SVR_{poly}$, all models showed good reliability, and their odd-even split-half reliability was > 0.8. This proved the stability of the model to a certain extent. In conclusion, the *GPR* model obtained the best criterion validity and split-half reliability performance.

Gait-based anxiety assessment methods have not been fully established. Here we migrated our method to a similar dataset (34). The results showed that the GPR model had the best effect. The Pearson correlation coefficient between the predicted scores of the

TABLE 2 Criterion validity of frequency-domain feature modeling using different numbers of wavelet decomposition layers.

| | $RMSE_{L_1}$ | $r_{L_1}$ | $RMSE_{L_2}$ | $r_{L_2}$ | $RMSE_{L_3}$ | $r_{L_3}$ | $RMSE_{L_4}$ | $r_{L_4}$ |
|---|---|---|---|---|---|---|---|---|
| **GPR** | 4.027 | 0.475 | 3.568 | 0.594 | 3.273 | 0.677 | 3.830 | 0.564 |
| **LR** | 4.092 | 0.471 | 3.593 | 0.594 | 3.291 | 0.677 | 3.859 | 0.565 |
| **$SVR_{linear}$** | 4.024 | 0.408 | 3.967 | 0.430 | 3.619 | 0.562 | 3.946 | 0.441 |
| **$SVR_{poly}$** | 4.223 | 0.269 | 3.772 | 0.520 | 3.915 | 0.437 | 4.085 | 0.409 |
| **$SVR_{rbf}$** | 4.105 | 0.405 | 4.008 | 0.434 | 3.967 | 0.496 | 4.071 | 0.390 |
| **$SVR_{sigmoid}$** | 4.045 | 0.375 | 3.952 | 0.451 | 3.773 | 0.542 | 3.988 | 0.416 |

The subscripts $L_1$, $L_2$, $L_3$ *and* $L_4$ indicate that the numbers of wavelet decomposition layers are 1, 2, 3, and 4 (the *level* parameter ranges from 1 to 4), respectively, when constructing the frequency-domain features. *RMSE and r* represent the root mean square error and criterion validity of the model established using the frequency-domain features, respectively. All correlation coefficients are highly significant ($p < 0.001$).

TABLE 3 Criterion validity of modeling with different training data sizes.

| | $r_{s_1}$ | $r_{s_2}$ | $r_{s_3}$ | $r_{s_{12}}$ | $r_{s_{13}}$ | $r_{s_{23}}$ | $r_{s_{123}}$ |
|---|---|---|---|---|---|---|---|
| **GPR** | 0.731 | 0.543 | 0.578 | 0.633 | 0.592 | 0.545 | 0.634 |
| **LR** | 0.702 | 0.547 | 0.578 | 0.630 | 0.583 | 0.545 | 0.637 |
| **$SVR_{linear}$** | 0.542 | 0.276 | 0.320 | 0.362 | 0.540 | 0.426 | 0.494 |
| **$SVR_{poly}$** | 0.403 | 0.386 | 0.372 | 0.392 | 0.314 | 0.425 | 0.354 |
| **$SVR_{rbf}$** | 0.460 | 0.454 | 0.403 | 0.526 | 0.487 | 0.425 | 0.490 |
| **$SVR_{sigmoid}$** | 0.518 | 0.346 | 0.359 | 0.424 | 0.454 | 0.314 | 0.488 |

$r_{s_1}$, $r_{s_2}$ *and* $r_{s_3}$ represent the criterion validity of the models established using gait segments $segment_1$, $segment_2$ and $segment_3$, respectively. $r_{s_{12}}$, $r_{s_{13}}$ *and* $r_{s_{23}}$ represent the criterion validity of the models established after combining any two of the three gait segments. $r_{s_{123}}$ represents the criterion validity of the model established after combining all three gait segments. All correlation coefficients are highly significant ($p < 0.001$).

TABLE 4 Ablation studies with different modeling features.

| | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ |
|---|---|---|---|---|---|
| **GPR** | 0.462 | 0.602 | 0.681 | 0.677 | 0.725 |
| **LR** | 0.461 | 0.595 | 0.680 | 0.677 | 0.704 |
| **$SVR_{linear}$** | 0.349 | 0.274 | 0.498 | 0.562 | 0.540 |
| **$SVR_{poly}$** | 0.410 | 0.368 | 0.467 | 0.437 | 0.404 |
| **$SVR_{rbf}$** | 0.378 | 0.428 | 0.459 | 0.496 | 0.457 |
| **$SVR_{sigmoid}$** | 0.336 | 0.407 | 0.432 | 0.542 | 0.528 |

$r_1$, $r_2$, $r_3$, $r_4$ *and* $r_5$ represent the criterion validity of the models developing using static time-domain features, dynamic time-domain features, all time-domain features (including dynamic and static features), frequency-domain features, and all features (including all time-domain and frequency-domain features), respectively. All correlation coefficients are highly significant ($p < 0.001$).

TABLE 5 Criterion validity of the anxiety assessment model for males and females.

| | *RMSE* | $r_{All}$ | $r_{Male}$ | $r_{Female}$ |
|---|---|---|---|---|
| **GPR** | 3.185 | 0.725 | 0.666 | 0.763 |
| **LR** | 3.430 | 0.704 | 0.639 | 0.722 |
| **$SVR_{linear}$** | 3.698 | 0.540 | 0.632 | 0.446 |
| **$SVR_{poly}$** | 4.018 | 0.404 | 0.404 | 0.361 |
| **$SVR_{rbf}$** | 3.948 | 0.457 | 0.469 | 0.512 |
| **$SVR_{sigmoid}$** | 3.823 | 0.528 | 0.474 | 0.590 |

*RMSE*, root mean square error. $r_{All}$, $r_{Male}$ *and* $r_{Female}$ represent the criterion validity of the model for all participants, male participants, and female participants, respectively. All correlation coefficients are highly significant ($p < 0.001$).

TABLE 6 The odd-even split-half reliability of anxiety assessment models.

| | *GPR* | *LR* | $SVR_{linear}$ | $SVR_{poly}$ | $SVR_{rbf}$ | $SVR_{sigmoid}$ |
|---|---|---|---|---|---|---|
| **$r_{split-half}$** | 0.803 | 0.801 | 0.808 | −0.696 | 0.876 | 0.883 |

$r_{split-half}$ represents the odd-even split-half reliability. All correlation coefficients are highly significant ($p < 0.001$).

anxiety assessment model and the scale scores reached 0.6, which was higher than the 0.4 reported by Miao et al. (34). In addition, we also tested the odd-even split-half reliability of the model on this dataset to 0.8. This shows that our anxiety assessment model has good robustness.

## 4. Discussion

We demonstrated that automated anxiety assessment using 2D gait videos is feasible. Based on 2D gait videos, we constructed and fused static and dynamic time-domain features and frequency-domain features and used machine learning methods to establish anxiety assessment models. Moreover, we evaluated the criterion validity and split-half reliability of the proposed anxiety prediction models. We also assessed the effects of different frequency-domain feature construction methods, gait training data sizes, and gender differences on the modeling results, verifying the contributions of various time-domain and frequency-domain features. Our results showed that the proposed gait video-based anxiety assessment method had good reliability and validity.

People with anxiety disorders tend to be between 15 and 35 years old (50). Higher education levels appear to have a protective effect on anxiety and depression (51). In our study, the participants ranged from 21 to 28 years old, their educational backgrounds were mainly involved postgraduate education, and their anxiety levels were concentrated between minimal and mild anxiety. This showed that our sample had a certain representativeness in the higher education student groups.

We used the *RMSE* to evaluate the relative performance of different models. Smaller *RMSE* and larger *r* values indicate better

model performance. In Tables 2, 4, the *RMSE* and *r* values showed inverse trends. This result showed that it was reasonable to use the criterion validity to evaluate the performance of the models.

As the number of wavelet decomposition layers increases, we can obtain more detail coefficient arrays representing high-frequency information and more approximate coefficient arrays representing low-frequency information. Since our sequence length was 75, the coefficient arrays that cannot be divided into half are filled with zeros in each wavelet decomposition. When the wavelet decomposition level was too high, the length of the coefficient array was too short, and the zero-padding operation introduced more errors, which led to inaccurate frequency-domain features. This was why the mean value of $r_{L_4}$ was smaller than that of $r_{L_3}$. Therefore, in wavelet decomposition, as the number of decomposition layers increases, we can more easily distinguish between low-frequency and high-frequency signals. However, the interference errors caused by the continuous subdivision also increase.

In general, in machine learning, more training data leads to better model effects (52). In our experiments, the model performance did not improve and even decreased as the number of gait training segment frames increased. For example, as shown in Table 3, the modeling effect after fusing two or three gait segments was worse than that of single gait segment modeling. On the one hand, gait is a periodic process (53). More gait segments lead to redundant information that does not contribute to modeling. Therefore, it is sufficient to model with fewer gait frames, which is similar to previous research results (34, 35, 37). On the other hand, different gait segments are discontinuous, and directly merging these sequences may cause mutations that reduce model performance to some extent. We also observed that the modeling effect of gait data including $segment_1$ was better than that of data including other segments, which may be due to the fatigue of participants walking back and forth in a narrow space, which led to inaccuracies in the subsequent gait videos.

Feature ablation studies were performed to examine how different features contribute to modeling. Taking the GPR model with good reliability and validity as an example, $r_{3\_GPR} > r_{2\_GPR} > r_{1\_GPR}$ verified that gait contains both dynamic and static information and that dynamic information expresses gait characteristics better than static information. Moreover, $r_{5\_GPR} > r_{4\_GPR}$ and $r_{5\_GPR} > r_{3\_GPR}$ verified that time-domain and frequency-domain information both contribute to modeling. The results of the feature ablation studies showed that the various constructed features were effective and necessary.

Previous studies have shown that the muscular strength of anxious women is significantly lower than that of healthy women and that these two groups show differences in gait, while these differences are not obvious among males (23). In addition, anxiety differs between the genders, and females are more likely to be anxious than males (54). This may be the reason why the anxiety prediction results are better for women than for men. This fact also supports the finding that participants with severe anxiety in Table 1 were all women.

Cronbach's alpha for the GAD-7 scale was 0.92 (9). In general, an alpha value >0.7 is considered to indicate acceptable reliability. In this study, except for the $SVR_{poly}$ model, the split-half reliability of the models was > 0.8. This result indicates that the odd-even split-half reliability can be applied to evaluate model performance.

This study is a continuation and extension of our previous work (37). We have optimized the methods of data segmentation, frequency-domain feature construction, and feature selection in experiments. Compared with previous studies, we explored in detail the impact of various factors (different features, gait dataset size, gender) on the model through comparative experiments with various parameters. In this study, the modeling method is more objective and reasonable, and the robustness and predictive performance of the anxiety assessment model are improved. Our research has some limitations. During data collection, a single camera was used to capture gait videos of the participants walking back and forth. Thus, the data contained some gait segments (such as turning and back gaits) that were not suitable for modeling. During preprocessing, the segmentation and recombination of different gait segments might introduce data breakpoints that can impact the model effects. In the future, we set the gait data collection scene as participants walking normally on the treadmill, ensuring that only the participants' front-view gait videos are recorded. We will try to avoid damaging the continuity of gait videos in preprocessing. In addition, although we verified the feasibility of assessing anxiety state based on gait videos, the participants were mainly college graduate students. Since this model was trained on only one social group, the generalizability may be insufficient. Thus, we will recruit participants from different groups according to the differences in age, gender, region, culture and economic background to increase the diversity of training data.

Due to the convenience, real-time, and non-invasive properties of our model, our approach can be applied in various scenarios. For example, the model can be applied for personal daily anxiety assessment. Moreover, companies can learn the employee anxiety levels through video data to provide psychological counseling in a timely manner and improve work efficiency. Using this method to assess the anxiety level of social groups in a timely manner can help to improve community mental health and public health. In future work, our proposed method still has some room for improvement. First, our current research uses traditional machine learning models and artificially constructed features. Although we have demonstrated the rationality and effectiveness of the constructed features in experiments, we still rely on a lot of subjective experience in the early stage. In recent years, many studies have made breakthroughs using deep learning (55). So next we will apply deep neural network to automatically extract gait features and train anxiety assessment models with better predictive performance. Second, our current research needs to convert gait video frame by frame into human body key point coordinates, and then calculate and analyze based on these 2D coordinates. In the process of extracting key points, some gait information will be lost, which will affect the model's learning of gait information. In the future work, we will use image streams for modeling directly based on gait video, so that the neural network can capture more detailed information in the gait.

## 5. Conclusion

In this study, we developed a convenient and timely anxiety assessment method that may contribute to improving mental health services. Our experiments show that gait can be used

as an objective cue to measure anxiety, the gait video-based anxiety assessment model has good criterion validity and split-half reliability, and the model has a better prediction effect on females than males. In addition, due to the periodicity of gait, increasing the number of gait training segment frames has little effect on the performance of the anxiety assessment model. The results of comparative experiments showed that the static and dynamic time-domain features and frequency-domain features improved model performance. Our preliminary study provides ideas for developing a convenient real-time anxiety assessment method.

## Data availability statement

To protect the privacy of the participants, the original datasets in the article cannot be made public. If necessary, feature datasets of gait are available from the corresponding author on reasonable request. Requests to access the datasets should be directed to TZ, tszhu@psych.ac.cn.

## Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

YW, BL, XL, and TZ proposed the idea of the research and designed the research method. DC and SG put forward constructive suggestions. TZ and XL provided research data. YW completed the data analysis and modeling and completed the first draft of the manuscript. TZ and SG guided the research process. All authors participated in the editing and reviewing of manuscripts and contributed to the article and approved the submitted version.

## Funding

This research was funded by the Scientific Foundation of Institute of Psychology, Chinese Academy of Sciences, No. E2CX4735YZ.

## 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/fpubh.2023.1082139/full#supplementary-material

## References

1. Vindegaard N, Benros ME. Covid-19 pandemic and mental health consequences: systematic review of the current evidence. *Brain Behav Immun.* (2020) 89:531–42. doi: 10.1016/j.bbi.2020.05.048

2. Association AP. *Diagnostic and Statistical Manual of Mental Disorders : Dsm-5.* 5th Ed Arlington VA Washington DC: American Psychiatric Association (2013).

3. Salari N, Hosseinian-Far A, Jalali R, Vaisi-Raygani A, Rasoulpoor S, Mohammadi M, et al. Prevalence of stress, anxiety, depression among the general population during the covid-19 pandemic: a systematic review and meta-analysis. *Global Health.* (2020) 16:11. doi: 10.1186/s12992-020-00589-w

4. World Health Organization. *World Health Statistics 2022: Monitoring Health for the Sdgs, Sustainable Development Goals* Geneva: World Health Organization (2022).

5. Chalah MA, Ayache SS. Noninvasive brain stimulation and psychotherapy in anxiety and depressive disorders: a viewpoint. *Brain Sci.* (2019) 9:82. doi: 10.3390/brainsci9040082

6. Chalah MA, Ayache SS. Disentangling the neural basis of cognitive behavioral therapy in psychiatric disorders: a focus on depression. *Brain Sci.* (2018) 8:150. doi: 10.3390/brainsci8080150

7. Rossi PH, Wright JD, Anderson AB. *Handbook of Survey Research.* London: Academic press (2013).

8. Jupp V. *The Sage Dictionary of Social Research Methods.* London: SAGE Publications, Ltd (2006).

9. Spitzer RL, Kroenke K, Williams JB, 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

10. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. *J Multidiscip Healthc.* (2016) 9:211. doi: 10.2147/JMDH.S104807

11. Duff K, Beglinger LJ, Schultz SK, Moser DJ, McCaffrey RJ, Haase RF, et al. Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index. *Arch Clin Neuropsychol.* (2007) 22:15–24. doi: 10.1016/j.acn.2006.08.013

12. Podsakoff PM, MacKenzie SB, Podsakoff NP. Sources of method bias in social science research and recommendations on how to control it. *Annu Rev Psychol.* (2012) 63:539–69. doi: 10.1146/annurev-psych-120710-100452

13. Hemmings C, Bouras N. *Psychiatric and Behavioural Disorders in Intellectual and Developmental Disabilities*, 3rd Edn. Cambridge: Cambridge University Press (2016).

14. Stein MB, Sareen J. Generalized anxiety disorder. *N Engl J Med.* (2015) 373:2059–68. doi: 10.1056/NEJMcp1502514

15. Marks I, Marset P, Boulougouris J, Huson J. Physiological accompaniments of neutral and phobic imagery. *Psychol Med.* (1971) 1:299–307. doi: 10.1017/S0033291700042264

16. Surcinelli P, Codispoti M, Montebarocci O, Rossi N, Baldaro B. Facial emotion recognition in trait anxiety. *J Anxiety Disord.* (2006) 20:110–7. doi: 10.1016/j.janxdis.2004.11.010

17. Giannakakis G, Pediaditis M, Manousos D, Kazantzaki E, Chiarugi F, Simos PG, et al. Stress and anxiety detection using facial cues from videos. *Biomed Signal Process Control.* (2017) 31:89–101. doi: 10.1016/j.bspc.2016.06.020

18. Siegman AW, Boyle S. Voices of fear and anxiety and sadness and depression: the effects of speech rate and loudness on fear and anxiety and sadness and depression. *J Abnorm Psychol.* (1993) 102:430–7. doi: 10.1037/0021-843X.102.3.430

19. Wortwein T, Morency LP, Scherer S, Ieee, editors. Automatic assessment and analysis of public speaking anxiety: a virtual audience case study. In: *6th AAAC Affective Computing and Intelligent Interaction International Conference (ACII); 2015 Sep 21-24; Xian, Peoples Republic China*. New York: IEEE (2015). doi: 10.1109/ACII.2015.7344570

20. Roether CL, Omlor L, Christensen A, Giese MA. Critical features for the perception of emotion from gait. *J Vis.* (2009) 9:15. doi: 10.1167/9.6.15

21. Montepare JM, Goldstein SB, Clausen A. The identification of emotions from gait information. *J Nonverbal Behav.* (1987) 11:33–42. doi: 10.1007/BF00999605

22. Seligman ME, Walker EF, Rosenhan DL. *Abnormal Psychology.* Norton (2001).

23. Feldman R, Schreiber S, Pick CG, Been E. Gait, balance, mobility and muscle strength in people with anxiety compared to healthy individuals. *Hum Mov Sci.* (2019) 67:10. doi: 10.1016/j.humov.2019.102513

24. Reelick MF, van Iersel MB, Kessels RP, Rikkert MGO. The influence of fear of falling on gait and balance in older people. *Age Ageing.* (2009) 38:435–40. doi: 10.1093/ageing/afp066

25. Staab JP, Balaban CD, Furman JM, editors. Threat assessment and locomotion: clinical applications of an integrated model of anxiety and postural control. *Semin Neurol.* (2013) 33:297–306. doi: 10.1055/s-0033-1356462

26. Hainaut J-P, Caillet G, Lestienne FG, Bolmont B. The role of trait anxiety on static balance performance in control and anxiogenic situations. *Gait Posture.* (2011) 33:604–8. doi: 10.1016/j.gaitpost.2011.01.017

27. Bolmont BT, Gangloff P, Vouriot A, Perrin PP. Mood states and anxiety influence abilities to maintain balance control in healthy human subjects. *Neurosci Lett.* (2002) 329:96–100. doi: 10.1016/S0304-3940(02)00578-5

28. Michalak J, Troje NF, Fischer J, Vollmar P, Heidenreich T, Schulte D. Embodiment of sadness and depression—gait patterns associated with dysphoric mood. *Psychosom Med.* (2009) 71:580–7. doi: 10.1097/PSY.0b013e3181a2515c

29. Greene BR, Foran TG, McGrath D, Doheny EP, Burns A, Caulfield B, et al. Comparison of algorithms for body-worn sensor-based spatiotemporal gait parameters to the gaitrite electronic walkway. *J Appl Biomech.* (2012) 28:349–55. doi: 10.1123/jab.28.3.349

30. Moeslund TB, Hilton A, Kruger V. A survey of advances in vision-based human motion capture and analysis. *Comput Vis Image Underst.* (2006) 104:90–126. doi: 10.1016/j.cviu.2006.08.002

31. Cloete T, Scheffer C, Ieee, editors. Benchmarking of a full-body inertial motion capture system for clinical gait analysis. In: *30th Annual International Conference of the IEEE-Engineering-in-Medicine-and-Biology-Society; 2008 Aug 20–24; Vancouver, CANADA*. New York: IEEE (2008).

32. Li QN, Wang YF, Sharf A, Cao Y, Tu CH, Chen BQ, et al. Classification of gait anomalies from kinect. *Visual Comput.* (2018) 34:229–41. doi: 10.1007/s00371-016-1330-0

33. Jing C, Liu X, Zhao N, Zhu T, editors. Different performances of speech and natural gait in identifying anxiety and depression. In: *International Conference on Human Centered Computing.* Cham: Springer (2019). doi: 10.1007/978-3-030-37429-7_20

34. Miao B, Liu X, Zhu T. Automatic mental health identification method based on natural gait pattern. *PsyCh J.* (2021) 10:453–64. doi: 10.1002/pchj.434

35. Zhao N, Zhang Z, Wang Y, Wang J, Li B, Zhu T, et al. See your mental state from your walk: recognizing anxiety and depression through kinect-recorded gait data. *PLoS ONE.* (2019) 14:e0216591. doi: 10.1371/journal.pone.0216591

36. Stark M, Huang H, Yu L-F, Martin R, McCarthy R, Locke E, et al. Identifying individuals who currently report feelings of anxiety using walking gait and quiet balance: an exploratory study using machine learning. *Sensors.* (2022) 22:3163. doi: 10.3390/s22093163

37. Wen Y, Li B, Chen D, Zhu T. Reliability and validity analysis of personality assessment model based on gait video. *Front Behav Neurosci.* (2022) 16:901568. doi: 10.3389/fnbeh.2022.901568

38. Fang J, Wang T, Li C, Hu X, Ngai E, Seet B-C, et al. Depression prevalence in postgraduate students and its association with gait abnormality. *IEEE Access.* (2019) 7:174425–37. doi: 10.1109/ACCESS.2019.2957179

39. Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans Pattern Anal Mach Intell.* (2021) 43:172–86. doi: 10.1109/TPAMI.2019.2929257

40. Murray MP. Gait as a total pattern of movement: including a bibliography on gait. *Am J Phys Med Rehabil.* (1967) 46:290–333.

41. Orović I, Stanković S, Amin M. A new approach for classification of human gait based on time-frequency feature representations. *Signal Process.* (2011) 91:1448–56. doi: 10.1016/j.sigpro.2010.08.013

42. Nussbaumer HJ. The fast fourier transform. In: Nussbaumer HJ, editor. *Fast Fourier Transform and Convolution Algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg (1981). p. 80–111. doi: 10.1007/978-3-662-00551-4_4

43. Daubechies I. *The Wavelet Transform, Time-Frequency Localization and Signal Analysis.* Princeton, NJ: Princeton University Press (2009). doi: 10.1515/9781400827268.442

44. Zill DG. *Advanced Engineering Mathematics.* Burlington, MA: Jones & Bartlett Publishers (2020).

45. Bishop CM. *Pattern Recognition and Machine Learning.* New York, NY: Springer (2006).

46. Reeves SJ, Zhe Z. Sequential algorithms for observation selection. *IEEE Trans Signal Process.* (1999) 47:123–32. doi: 10.1109/78.738245

47. Zhou Z-H. *Machine Learning.* Singapore: Springer (2021). doi: 10.1007/978-981-15-1967-3

48. Meyes R, Lu M, de Puiseau CW, Meisen T. Ablation studies in artificial neural networks. *arXiv [Preprint].* (2019). *arXiv*: 1901.08644. doi: 10.48550/arXiv.1901.08644

49. Bartko JJ, Carpenter WT. On the methods and theory of reliability. *J Nerv Ment Dis.* (1976) 163:307–17. doi: 10.1097/00005053-197611000-00003

50. Craske MG, Stein MB. Anxiety. *Lancet.* (2016) 388:3048–59. doi: 10.1016/S0140-6736(16)30381-6

51. Bjelland I, Krokstad S, Mykletun A, Dahl AA, Tell GS, Tambs K. Does a higher educational level protect against anxiety and depression? The Hunt study. *Soc Sci Med.* (2008) 66:1334–45. doi: 10.1016/j.socscimed.2007.12.019

52. Luyckx K, Daelemans W. The effect of author set size and data size in authorship attribution. *Lit Linguist Comput.* (2011) 26:35–55. doi: 10.1093/llc/fqq013

53. Baker R, Hart HM. *Measuring Walking: A Handbook of Clinical Gait Analysis.* London: Mac Keith Press (2013).

54. Lewinsohn PM, Gotlib IH, Lewinsohn M, Seeley JR, Allen NB. Gender differences in anxiety disorders and anxiety symptoms in adolescents. *J Abnorm Psychol.* (1998) 107:109. doi: 10.1037/0021-843X.107.1.109

55. Amanat A, Rizwan M, Javed AR, Abdelhaq M, Alsaqour R, Pandya S, et al. Deep learning for depression detection from textual data. *Electronics.* (2022) 11:676. doi: 10.3390/electronics11050676

frontiers | Frontiers in Public Health

TYPE Opinion
PUBLISHED 03 July 2023
DOI 10.3389/fpubh.2023.1225861



OPEN ACCESS

EDITED BY
Peter Kokol,
University of Maribor, Slovenia

REVIEWED BY
Jernej Zavrsnik,
Health Center dr. Adolf Drolc, Slovenia

*CORRESPONDENCE
Plinio P. Morita
✉ plinio.morita@uwaterloo.ca

RECEIVED 20 May 2023
ACCEPTED 16 June 2023
PUBLISHED 03 July 2023

CITATION
Morita PP, Abhari S, Kaur J, Lotto M, Miranda PADSES and Oetomo A (2023) Applying ChatGPT in public health: a SWOT and PESTLE analysis. *Front. Public Health* 11:1225861. doi: 10.3389/fpubh.2023.1225861



# Applying ChatGPT in public health: a SWOT and PESTLE analysis

Plinio P. Morita[1,2,3,4,5]*, Shahabeddin Abhari[1], Jasleen Kaur[1], Matheus Lotto[1,6], Pedro Augusto Da Silva E. Souza Miranda[1] and Arlene Oetomo[1]

[1]School of Public Health Sciences, University of Waterloo, Waterloo, ON, Canada, [2]Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada, [3]Research Institute for Aging, University of Waterloo, Waterloo, ON, Canada, [4]Centre for Digital Therapeutics, Techna Institute, University Health Network, Toronto, ON, Canada, [5]Dalla Lana School of Public Health, Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, ON, Canada, [6]Department of Pediatric Dentistry, Orthodontics, and Public Health, Bauru School of Dentistry, University of São Paulo, Bauru, Brazil

KEYWORDS

**public health, artificial intelligence, ChatGPT, SWOT analysis, healthcare**

## 1. Introduction

Public health is a multidisciplinary field that aims to promote and protect the health of communities through various interventions, such as disease prevention, health promotion, and policy development. It involves analyzing data and applying evidence-based approaches to improve the health outcomes of populations (1–3). One of the main challenges in public health is the emergence and re-emergence of infectious diseases such as COVID-19, which can pose a significant threat to public health (4). Other challenges include the rise of non-communicable diseases (NCDs) such as heart disease and cancer, which are often linked to lifestyle factors such as poor diet and physical inactivity (5). Health inequities also pose a challenge to public health, as some groups, such as marginalized and vulnerable populations, may experience poorer health outcomes due to commercial determinants of health (1, 3, 5). Furthermore, there are challenges associated with the collection, management, and analysis of public health data. Ensuring data privacy and security, addressing biases in data collection and analysis, and making data accessible to all stakeholders are all critical issues in the field of public health. Finally, the need for effective communication and collaboration among stakeholders is essential to address these challenges and improve public health outcomes (2, 6, 7).

The growth of Artificial Intelligence (AI) in healthcare has been exponential in recent years, with advancements in machine learning, natural language processing (NLP), and image analysis. AI is increasingly being used to improve disease surveillance, drug discovery, and personalized medicine, among other applications, with the potential to transform healthcare delivery (8–10). Also, AI methods have shown great promise in addressing various public health challenges. Machine learning algorithms, NLP, and other AI techniques can be used to analyze large datasets, identify patterns and trends, and generate insights that can inform public health interventions (11, 12). ChatGPT is a state-of-the-art NLP model developed by OpenAI that has shown impressive performance in a variety of tasks, including language translation, text completion, and sentiment analysis. The model's ability to generate coherent and contextually appropriate responses to text inputs has made it a promising tool for a wide range of applications, including public health (13). For example, ChatGPT can be used to help patients manage chronic conditions by providing reminders for medication, diet, and exercise, and answering questions about symptoms and treatment options. ChatGPT can also be used to help patients find healthcare providers, schedule

appointments, and access healthcare information. Moreover, ChatGPT can be used to improve patient engagement and education. Patients can interact with ChatGPT in natural language and receive tailored responses based on their medical history, preferences, and needs. ChatGPT can also provide patients with reliable and up-to-date health information, such as disease prevention tips, symptom management advice, and resources for mental health support (13–15).

In this paper, we conducted a SWOT analysis and PESTLE analysis (16, 17) to evaluate the applying of ChatGPT in public health. SWOT stands for Strengths, Weaknesses, Opportunities, and Threats, and is a widely used strategic planning tool that helps in identifying the internal and external factors that can impact the success of a project or initiative (17). As well as PESTLE analysis is a strategic planning tool used to assess and analyze external factors that can impact an organization, project, or industry. It examines the Political, Economic, Sociocultural, Technological, Legal, and Environmental factors that can influence the environment in which an entity operates (16). By conducting a SWOT analysis of ChatGPT, we aim to identify the strengths, weaknesses, opportunities, and threats associated with the application of this technology in public health. Also, we want to determine main external factors that can have an effect on applying this technology in public health.

The methods used in this paper involve a comprehensive literature review of previous studies and publications related to the applications of ChatGPT in public health (18). To identify pertinent studies for our research, a comprehensive search was conducted in reputable databases, employing relevant keywords. The databases used for this purpose were Pubmed, Scopus, and Google Scholar. The selection of relevant articles was based on a combination of crucial keywords that proved effective in narrowing down the search results. The selected keywords for this study were "ChatGPT" AND ("public health" OR "healthcare"). To ensure consistency and focus, the search was limited to English-language papers published before April 15, 2023. Through this systematic approach, a total of 106 articles were initially identified across the three aforementioned databases. After evaluation of these articles based on their alignment with the research topic, 16 papers were ultimately deemed relevant and included for further analysis. After identifying relevant articles, a qualitative content analysis was conducted to extract and classify relevant information related to the SWOT analysis and PESTLE analysis of ChatGPT in public health.

## 2. SWOT analysis for applying ChatGPT in public health

The extracted data were organized into four categories: strengths, weaknesses, opportunities, and threats.

### 2.1. Strengths

The use of ChatGPT in public health has several strengths. One of the main strengths of ChatGPT is its ability to provide personalized health information and support to individuals. Chatbots powered by ChatGPT can be available 24/7, which can improve access to health information and support for people who may not be able to seek care during regular business hours. Additionally, ChatGPT can process and analyze large amounts of data quickly and accurately, which can support disease surveillance and outbreak detection. Chatbots can monitor social media and other online platforms for signs of emerging health threats, such as outbreaks of infectious diseases. They can also provide real-time information to individuals and healthcare providers about outbreaks in their area, which can help to prevent the spread of disease (13–15, 19–23).

### 2.2. Weaknesses

The use of ChatGPT in public health also has several weaknesses. One of the main weaknesses is the potential for misinterpretation or miscommunication, as language models may not always accurately understand the nuances of human language and context. This could result in chatbots providing incorrect or misleading health information (22, 24). Additionally, privacy is a concern, as chatbots may be vulnerable to hacking or data breaches, which can compromise sensitive health information. There is also the potential for ChatGPT to perpetuate biases in health data if the underlying data used to train the model is biased. For example, if the data used to train ChatGPT is biased toward certain demographics, the chatbot may not provide accurate information to all populations (15, 20–28).

### 2.3. Opportunities

There are several opportunities associated with the use of ChatGPT in public health. One of the main opportunities is the ability of chatbots to provide personalized health information and support to individuals. This is particularly useful for individuals who may not have access to healthcare services or who may be hesitant to seek care due to stigma or other barriers (11, 13, 21, 27). ChatGPT can also assist with disease surveillance and epidemic identification, which helps stop the spread of disease. Chatbots can keep an eye on social media and other online platforms for indications of emerging health risks, such infectious disease epidemics. They can also notify people and healthcare professionals in real time about epidemics in their region. Another opportunity is that ChatGPT can facilitate communication and collaboration between healthcare providers and patients, which can improve the quality of care and health outcomes (11, 13, 15, 21–25).

### 2.4. Threats

The use of ChatGPT in public health also has several threats. The possibility of chatbots distributing inaccurate or deceptive health information is one of the key dangers. This can be as a result of biased training data or mistakes in the language model's comprehension of human language and context (11, 14, 21, 24). Additionally, as was already said, there is a chance that chatbots would maintain the biases that now exist in health data. Another

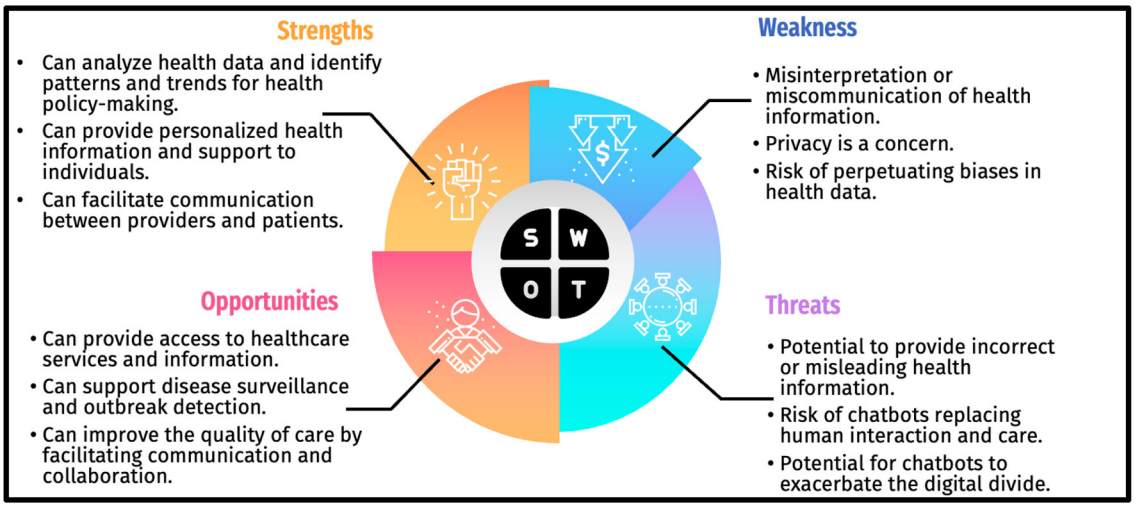



FIGURE 1
SWOT analysis for applying chatbot in public health.

danger is that chatbots may take the role of human connection and care, which would reduce empathy and prevent personalized care from being provided. Finally, there is a chance that chatbots will widen the digital divide by excluding those without access to technology or who are uncomfortable utilizing it from ChatGPT's public health advantages (9–11, 21–24, 26–29). The summary of SWOT analysis showed in Figure 1.

## 3. PESTLE analysis for applying ChatGPT in public health

The application of ChatGPT in public health can be analyzed using the PESTLE framework, which examines the Political, Economic, Sociocultural, Technological, Legal, and Environmental factors that influence its implementation and impact. This analysis aims to provide a comprehensive understanding of the external factors that shape the context of ChatGPT's application in public health.

(1) The Political factor explores the political environment surrounding the use of ChatGPT in public health. This includes government policies, regulations, and political support or resistance to AI technologies. Political factors play a crucial role in determining the level of investment, data governance, and ethical considerations in deploying ChatGPT in public health (24, 28, 30).

(2) The Economic factor examines the economic implications of using ChatGPT in public health. This involves evaluating the cost-effectiveness, affordability, and sustainability of implementing the technology. Economic factors also consider the potential for job displacement or creation, economic disparities in access to AI technologies, and the overall financial implications for healthcare systems (23, 24, 27, 29, 30).

(3) Sociocultural factors play a significant role in the application of ChatGPT in public health. These factors involve understanding public acceptance, trust, and perception of AI technologies. Sociocultural considerations also encompass issues of privacy, data security, and the potential impact on the patient-provider relationship. Moreover, cultural norms, beliefs, and attitudes toward AI-driven healthcare interventions need to be taken into account (15, 23, 27, 30).

(4) The Technological factor analyzes the technological landscape for ChatGPT in public health. This includes advancements in natural language processing, machine learning algorithms, and the integration of ChatGPT with existing health information systems. Technological factors also encompass the potential for bias, algorithmic transparency, and the need for continuous updates and maintenance of the technology (22, 27, 29, 30).

(5) Legal factors examine the legal and regulatory framework governing the use of ChatGPT in public health. This includes privacy regulations, data protection laws, intellectual property rights, and ethical guidelines for AI applications in healthcare. Compliance with legal requirements and adherence to ethical principles are critical for the responsible deployment of ChatGPT in public health (14, 27, 30).

(6) The Environmental factor focuses on the environmental implications of implementing ChatGPT in public health. This involves considering the energy consumption and carbon footprint associated with AI infrastructure and data centers. It also includes assessing the environmental impact of data collection, storage, and disposal practices (23, 31, 32).

## 4. Discussion

The findings highlight the potential benefits and limitations of using chatbots powered by ChatGPT in the public health context. Consequently, they can provide personalized health information and support to individuals, disease surveillance, and outbreak detection, besides facilitating individual and shared decision-making. Nevertheless, there are also limitations associated

with ChatGPT that should be considered, such as the potential misinterpretation or miscommunication, privacy concerns, and the risk of perpetuating biases in health data. By conducting a PESTLE analysis, policymakers, healthcare organizations, and researchers can gain insights into the broader contextual factors that can influence the application of ChatGPT in public health. This analysis can inform decision-making, help anticipate challenges, and guide the development of ethical guidelines and regulatory frameworks to maximize the benefits and mitigate potential risks associated with ChatGPT's implementation in the public health domain.

Indeed, future studies must clarify whether the advantages of utilizing chatbots outweigh their risks in assisting public health actions. Although artificial intelligence-based approaches can improve healthcare outcomes, the threats should be carefully considered to avoid inappropriate decision-making and the deepening of health inequalities especially the digital divide which continues to grow especially in the Global South. Hence, the widespread adoption of the disruptive technology of AI chatbots in public health will require careful oversight and time as authorities must first understand the optimal scenarios for the ethics and legalities of its implementation and application. However, these actions cannot be delayed because hundreds of AI tools are being released and are being used for learning about public health whether by design or not.

## Author contributions

PPM and SA: conceptualization and writing original draft. JK, SA, ML, PM, and AO: writing review and editing. PPM, SA, JK, ML, PM, and AO: conceptualization, supervision, and writing review and editing. 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



## References

1. Edemekong PF, Tenny S. *Public Health. StatPearls*. Treasure Island (FL): StatPearls Publishing Copyright© 2023, StatPearls Publishing LLC (2023).

2. Berg J. Data in public health. *Science.* (2017) 355:669. doi: 10.1126/science.aam9455

3. Schneider MJ. *Introduction to Public Health*. Burlington: Jones & Bartlett Learning (2020).

4. Cruz MP, Santos E, Cervantes MV, Juárez MLJRCE. COVID-19, a worldwide public health Emergency. *Revista Clínica Española.* (2021) 221:55–61. doi: 10.1016/j.rceng.2020.03.001

5. Times YJTC-RTVRotPHP. The future of public health. *New Eng J Med.* (2022) 4:143.

6. Zhao Y, Liu L, Qi Y, Lou F, Zhang J, Ma WJJoi, et al. Evaluation and design of public health information management system for primary health care units based on medical and health information. *J Infect Public Health.* (2020) 13:491–6. doi: 10.1016/j.jiph.2019.11.004

7. Thiébaut R, Thiessard F. Public health and epidemiology informatics. *Yearb Med Inform.* (2017) 26:248–51. doi: 10.15265/IY-2017-036

8. Chen M, Decary M. Artificial intelligence in healthcare: an essential guide for health leaders. *Healthc Manage Forum.* (2020) 33:10–8. doi: 10.1177/0840470419873123

9. Yu K-H, Beam AL. Kohane ISJNbe. *Arti Intell Healthcare.* (2018) 2:719–31. doi: 10.1038/s41551-018-0305-z

10. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Med Inform Dec Making.* (2021) 21:1–23. doi: 10.1186/s12911-021-01488-9

11. Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. *Lancet Dig Health.* (2019) 1:e13–e4. doi: 10.1016/S2589-7500(19)30002-0

12. Gunasekeran DV, Tseng RM, Tham YC, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. *NPJ Dig Med.* (2021) 4:40. doi: 10.1038/s41746-021-00412-9

13. Jungwirth D, Haluza D. Artificial intelligence and public health: an exploratory study. *Int J Environ Res ublic Health.* (2023) 20:4541. doi: 10.3390/ijerph20054541

14. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel).* (2023) 11:6. doi: 10.3390/healthcare11060887

15. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst.* (2023) 47:33. doi: 10.1007/s10916-023-01925-4

16. Basu R. Implementing Quality: A Practical Guide to Tools and Techniques: Enabling the Power of Operational Excellence. Cengage Learning EMEA. Toronto (2004).

17. Namugenyi C, Nimmagadda SL, Reiners T. Design of a SWOT analysis model and its evaluation in diverse digital business ecosystem contexts. *Procedia Comput Sci.* (2019) 159:1145–54.

18. Kokol P, BlaŽun Vošner H, Završnik J. Application of bibliometrics in medicine: a historical bibliometrics analysis. *Health Inform Lib J.* (2021) 38:125–38. doi: 10.1111/hir.12295

19. Xue VW, Lei P, Cho WC. The potential impact of ChatGPT in clinical and translational medicine. *Clin Transl Med.* (2023) 13:e1216. doi: 10.1002/ctm2.1216

20. Sallam M, Salim NA, Al-Tammemi AB, Barakat M, Fayyad D, Hallit S, et al. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. *Cureus.* (2023) 15:e35029. doi: 10.7759/cureus.35029

21. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H, Schonfeld J. Artificial intelligence in public health: Challenges and opportunities for public health made possible by advances in natural language processing. *Cana Commun Dis Report.* (2020) 46:161. doi: 10.14745/ccdr.v46i06a02

22. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test. *Observ Study Demonstr Opp Limit Primary Care.* (2023) 9:e46599. doi: 10.2196/46599

23. Biswas SS. Role of Chat GPT in Public health. *Ann Biomed Eng.* (2023). doi: 10.1007/s10439-023-03172-7

24. Komorowski M. del Pilar Arias López M, Chang ACJICM. Komorowski M, del Pilar Arias López M, Chang AC. How could ChatGPT impact my practice as an intensivist? An overview of potential applications, risks and limitations. *Inten Care Med.* (2023) 4:1–4. doi: 10.1007/s00134-023-07096-7

25. Ferres JML, Weeks WB, Chu LC, Rowe SP, Fishman EK. Beyond chatting: the opportunities and challenges of ChatGPT in medicine and radiology. *Diagn Interv Imaging*. (2023) 3:6. doi: 10.1016/j.diii.2023.02.006

26. Snoswell CL, Snoswell AJ, Kelly JT, Caffery LJ, Smith AC. Artificial intelligence: augmenting telehealth with large language models. *J Telemed Telecare*. (2023) 3:1357633X231169055. doi: 10.1177/1357633X231169055

27. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCounc *Transact Benchmarks Stand Evaluat*. (2023) 3:100105. doi: 10.1016/j.tbench.2023.100105

28. Parray AA, Inam ZM, Ramonfaur D, Haider SS, Mistry SK, Pandya AK. ChatGPT and global public health: applications, challenges, ethical considerations and mitigation strategies. *Elsevier*. (2023). doi: 10.1016/j.glt.2023.05.001

29. Paul J, Ueno A, Dennis C. ChatGPT and consumers: Benefits, pitfalls and future research agenda. *Int J Cons Stud*. Wiley Online Library. (2023) 25:928. doi: 10.1111/ijcs.12928

30. Gill SS, Kaur R. ChatGPT: vision and challenges. *Internet Things Cyber-Phy Sys*. (2023). doi: 10.1016/j.iotcps.2023.05.004

31. Galaz V, Centeno MA, Callahan PW, Causevic A, Patterson T, Brass I, et al. Artificial intelligence, systemic risks, and sustainability. *Technol Soc*. (2021) 67:101741. doi: 10.1016/j.techsoc.2021.101741

32. Van Wynsberghe A. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics*. (2021) 1:213–8. doi: 10.1007/s43681-021-00043-6

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 01 August 2023
DOI 10.3389/fpsyt.2023.1178633



OPEN ACCESS

EDITED BY
Jorge Piano Simoes,
University of Twente, Netherlands

REVIEWED BY
Clara Puga,
Otto von Guericke University Magdeburg, Germany
Laura Basso,
Charité University Medicine Berlin, Germany

*CORRESPONDENCE
Nathan A. Kimbrel
✉ nathan.kimbrel@va.gov
Daniel A. Jacobson
✉ jacobsonda@ornl.gov

†These authors have contributed equally to this work

RECEIVED 03 March 2023
ACCEPTED 21 June 2023
PUBLISHED 01 August 2023

CITATION
Pavicic M, Walker AM, Sullivan KA, Lagergren J, Cliff A, Romero J, Streich J, Garvin MR, Pestian J, McMahon B, Oslin DW, Beckham JC, Kimbrel NA and Jacobson DA (2023) Using iterative random forest to find geospatial environmental and Sociodemographic predictors of suicide attempts.
*Front. Psychiatry* 14:1178633.
doi: 10.3389/fpsyt.2023.1178633



# Using iterative random forest to find geospatial environmental and Sociodemographic predictors of suicide attempts

Mirko Pavicic[1†], Angelica M. Walker[2†], Kyle A. Sullivan[1], John Lagergren[1], Ashley Cliff[2], Jonathon Romero[2], Jared Streich[1], Michael R. Garvin[1] on behalf of MVP Suicide Exemplar Workgroup, the Million Veteran Program, John Pestian[1,3], Benjamin McMahon[4], David W. Oslin[5,6], Jean C. Beckham[7,8,9], Nathan A. Kimbrel[7,8,10,11]*and Daniel A. Jacobson[1]*

[1]Oak Ridge National Laboratory, Computational and Predictive Biology, Oak Ridge, TN, United States, [2]The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee Knoxville, Knoxville, TN, United States, [3]Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH, United States, [4]Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, United States, [5]VISN 4 Mental Illness Research, Education, and Clinical Center, Center of Excellence, Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, United States, [6]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [7]Durham Veterans Affairs Health Care System, Durham, NC, United States, [8]VA Mid-Atlantic Mental Illness, Research, Education, and Clinical Center, Seattle, WA, United States, [9]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, NC, United States, [10]Duke University School of Medicine, Duke University, Durham, NC, United States, [11]VA Health Services Research and Development Center of Innovation to Accelerate Discovery and Practice Transformation, Durham, NC, United States

**Introduction:** Despite a recent global decrease in suicide rates, death by suicide has increased in the United States. It is therefore imperative to identify the risk factors associated with suicide attempts to combat this growing epidemic. In this study, we aim to identify potential risk factors of suicide attempt using geospatial features in an Artificial intelligence framework.

**Methods:** We use iterative Random Forest, an explainable artificial intelligence method, to predict suicide attempts using data from the Million Veteran Program. This cohort incorporated 405,540 patients with 391,409 controls and 14,131 attempts. Our predictive model incorporates multiple climatic features at ZIP-code-level geospatial resolution. We additionally consider demographic features from the American Community Survey as well as the number of firearms and alcohol vendors per 10,000 people to assess the contributions of proximal environment, access to means, and restraint decrease to suicide attempts. In total 1,784 features were included in the predictive model.

**Results:** Our results show that geographic areas with higher concentrations of married males living with spouses are predictive of lower rates of suicide attempts, whereas geographic areas where males are more likely to live alone and to rent housing are predictive of higher rates of suicide attempts. We also identified climatic features that were associated with suicide attempt risk by age group. Additionally, we observed that firearms and alcohol vendors were associated with increased risk for suicide attempts irrespective of the age group examined, but that their effects were small in comparison to the top features.

**Discussion:** Taken together, our findings highlight the importance of social determinants and environmental factors in understanding suicide risk among veterans.

KEYWORDS

suicide prevention, explainable artificial intelligence, geospatial analysis, public health, veterans' health, firearms, alcohol misuse

# 1. Introduction

Suicide rates in the United States (U.S.) have increased in recent years despite these rates declining globally (1). According to the biopsychosocial model of suicide risks, there are distal, developmental, and proximal factors that affect the probability of suicide attempt (2). Distal factors are related to familial and genetic predisposition and early-life adversity. Developmental factors include personality traits associated with suicidal behavior, cognitive deficits, and chronic substance misuse. Proximal factors include but are not limited to psychiatric, psychological, socioeconomic, and environmental factors. Several studies have found associations between demographic factors and suicide such as age, ethnicity, socioeconomic status, marital status, religion, etc. (3). The impact of climate on suicidal behavior is significant, although the relationship between climate and suicide is complex and not yet fully understood. One possible explanation for how climate could affect suicidal behavior is through seasonal changes. Research has shown a pattern of increased suicide attempts and deaths during the spring and early summer months, indicating a seasonality in such events (4, 5). Sunlight and temperature are among the most relevant climatic features associated with this seasonality, as they may directly influence various mood disorders related to suicide risk (6–9). While several studies have attempted to link other climatic features to suicide risk, their findings have been contradictory or inconclusive (10–15). Interestingly, the majority of these studies have been conducted using extensive geospatial regions. An investigation carried out in Taipei, Taiwan examined suicide mortality at high geospatial resolution using neighborhoods known as "li" as geospatial units (16). The findings of this study revealed a significant geospatial variation in suicide mortality across neighborhoods, indicating that the analysis of aggregated data in broader geographic areas may attenuate predictive signals (16).

Other relevant proximal factors include access to means and substance misuse (17, 18). For example, occupations with access to lethal means are associated with increased risk of death by suicide (19, 20). Moreover, controlling access to lethal means is an effective strategy for decreasing suicide risk (21). In the U.S., death by suicide is the leading cause of violent deaths, and firearms are responsible for approximately half of these deaths (22). Substance misuse also plays an important role in suicide prevention because acute substance intoxication can increase an individual's disinhibition. For example, a study showed that suicide decedents have an increased risk of alcohol ingestion and intoxication before their death relative to controls (23).

The objective of the present research was to conduct an analysis of climatic and socio-demographic factors that are associated with increased risk for suicide attempts among U.S. veterans using an explainable artificial intelligence (X-AI) model. We were also interested in the relationship of the number of firearms and alcohol vendors per 10,000 people as proxies for access to means and decreased restraint, which were also included in the final model. Together, we identify several novel factors at zip code-level resolution that impact individual-level risk for attempting suicide.

# 2. Materials and methods

## 2.1. Data and data pre-processing

### 2.1.1. Patient data

The cohort and suicide attempt phenotype used in this study were initially described in Kimbrel et al. (24). A total of 405,540 participating patients in this study were enrolled in the U.S. Department of Veterans Affairs' (VA) Million Veteran Program (MVP). All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human subjects/patients were approved by VA Central Institutional Review Board (cIRB# 18-11) after all subjects provided written and signed informed consent. Race, ethnicity and gender were self-reported. Suicide attempt phenotype was created from electronic health records (EHR) from the VA corporate data warehouse (CDW), using International Classification of Diseases (ICD) diagnostic codes, survey data from the mental health domain, and the Suicide Prevention Application Network (SPAN) data set (25). Veteran participants were considered controls if there was no recorded evidence of them ever attempting suicide or experiencing suicidal thoughts throughout their lives. This determination was based on qualifying ICD codes, reports of suicide behavior, or responses from mental health surveys previously mentioned. It is worth noting that veterans who had a history of having suicidal thoughts but had not attempted suicide were specifically excluded from the current analysis. This was done to guarantee that the control group consisted of individuals who had no prior instances of engaging in or contemplating suicidal thoughts or behaviors. Thus, cases were defined as having a history of one or more suicide attempts (including both fatal and non-fatal), whereas controls were defined has having no history of suicide attempts or ideation. In total, this resulted in 391,409 controls and 14,131 cases. The mean age was 62.4 years for the whole cohort, 63.4 for males and 50.8 years for females. As reflective of the veteran population, this cohort is predominantly male but does include females as well as racial and ethnic minorities. The sex distribution was 8% female and 92% male. This cohort was 73% white, 18% black or African American, 1% American Indian or Alaska Native, 1% Asian, and 7% mixed race, other, missing, or unknown. This is generally close to the proportions of racial groups in the United States in 2021 (78.6% white, 12.2% black/African American, 0.7% American Indian or Alaska Native, 5.6% Asian, 2.8% mixed race) with an overrepresentation of black/African American and an underrepresentation of Asian people. This includes Latinx people spread across those racial categories. The

Supplementary Table S1 provides a breakdown of cohort demographics by cases and controls. Most of the suicide attempts in this cohort were concentrated around 60 years of age, likely since this age group is overrepresented among these patients (Figures 1A,B). However, the proportion of attempts grouped by age decreased abruptly after age 60 (Figure 1C). Due to this rapid decrease in attempts after age 60, we analyzed patients greater than 60 years of age separately from those under 60 years of age. The split with patients above or equal to 60 contained a total of 267,447 individuals with 4,231 attempts and 263,216 controls. The split with patients below age 60 was composed of 138,093 individuals with 9,900 suicide attempts and 128,193 controls. Each attempt and control were associated with climatic and socio-demographic features by patient ZIP code. For patients with multiple ZIP codes (less than ~0.6% of the cohort), we used the most recent ZIP code since not all suicide attempts had a corresponding date. The proportion of attempts grouped by age decreased steadily after age 60 (Figure 1). Therefore, we explored the socio-demographic and climatic features that were associated with suicide attempts in patients greater than 60 years of age separately from those under 60 years of age.

### 2.1.2. Climatic features

The climatic features included two groups: static measurements and monthly measurements. Monthly measurements included longitudinal features such as monthly average precipitation and maximum temperature. There were 12 distinct measurement types recorded each month, totaling 144 features. The 30 static features included features such as elevation and percent urban cover. This led to a total of 174 climatic and weather-related features, mapped to 33,144 ZIP codes across the U.S. (26–32) (Supplementary Table S2).

### 2.1.3. Socio-demographic features

The socio-demographic features were collected from the 2019 American Community Survey, produced by the United States Census Bureau (33). These 1,606 features were captured using the tidycensus software package in R using 5 years estimate for 2019 (34). These features were normalized to represent a percentage of the total population or age bracket within each ZIP code. We also included two additional features: population density (people per square mile) and the ratio of water to land area, which led to a total of 1,608 demographic features that were mapped to 33,120 ZIP codes across the U.S (Supplementary Table S3).

### 2.1.4. Alcohol and firearms features

From the Historical Business Database (35) we extracted the number of firearms vendors per 10,000 residents in each ZIP code, averaged across the years 2010 and 2019, and the similarly calculated number of alcohol vendors. This information was included for 31,378 ZIP codes across the U.S (Supplementary Table S4).

## 2.2. Explainable artificial intelligence analysis

In this study, we used iterative Random Forest (iRF) (36, 37), an X-AI algorithm that ranks input features by importance through iterative feature weighting, to associate proximal environmental features with suicide attempts. There are two motivations for our choice of methods: (i) identifying which predictor features explain changes in the output and (ii) scaling to high-performance computing (HPC) systems. iRF has been implemented in C++ to use regression trees and leverages massive parallelism to scale to very large datasets. The model was trained to predict cases and controls, where the predicted values at each leaf node, is the average value of all samples that reached that specific leaf node. Thus, with binary values encoded as case (1) and control (0), iRF identifies the proportional change in outcome as a function of each input feature, thereby ranking features by feature importance.

High pairwise correlations among input features can have a negative impact on the explainability of iRF models. This occurs because when one feature is strongly correlated with others, its importance is divided across the correlated features. As a result, the overall importance of these features in predicting suicide attempts decreases. To identify highly correlated features, we calculated Pearson's correlation coefficients between all 1,784 features. Through this analysis, we identified feature groups with correlation coefficients equal to or greater than 0.90 in absolute value, and therefore considered as highly correlated. From each feature group of strongly correlated

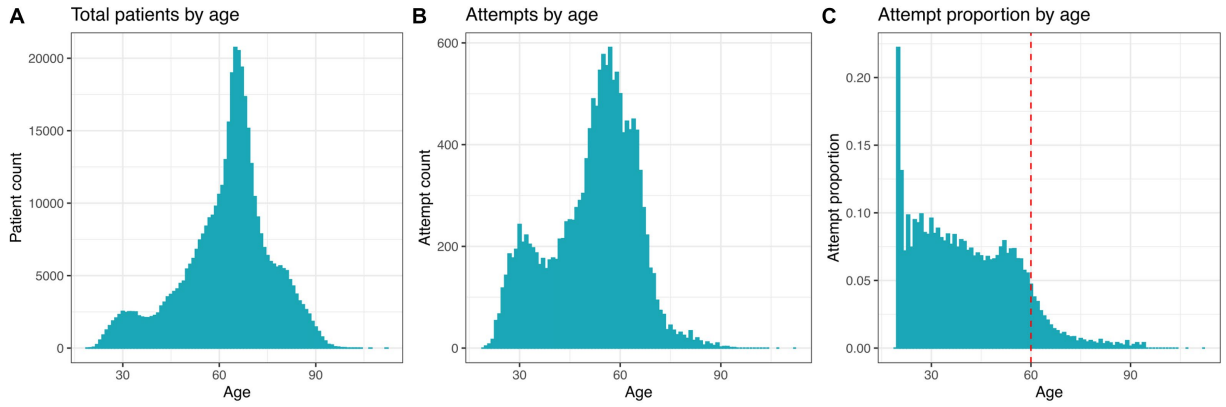



FIGURE 1
Distribution of suicide attempts by age of 405,540 participating patients. **(A)** Total patient count by age. **(B)** Total suicide attempt count by age. **(C)** Age-specific suicide rate. The red dashed line shows a cutoff to highlight the decreased proportion of suicide attempts in individuals equal to/above 60 years of age.

features we selected a representative feature to be included in the iRF model. Thus, this led to the removal of 248 features, reducing the total number of features to 1,536 (Supplementary Table S5). Please refer to Supplementary Table S6 for a comprehensive description of all the features utilized in this study.

## 2.3. iRF k-fold cross validation and accuracy calculation

To obtain accuracy scores for the iRF model, 5-fold cross-validation was used. The cross-validation technique employed in this study was group shuffle split. In this methodology, the dataset was randomly divided into 80% for training and 20% for testing, while incorporating a grouping factor to ensure that samples from the same group were not utilized for both training and testing in the same run. By employing patient zip codes as a grouping factor, we aimed to mitigate geospatial bias in our analysis. To further enhance the robustness of our analysis, we replicated this process five times, resulting in a total of five accuracy estimations for the models. Prediction accuracy was calculated using the average Area Under the Precision Recall curve (AUPRC), where precision was defined as *true positives/true positives + false positives* and recall as *true positives/true positives + false negatives*. Each random forest in the iRF model includes 1,000 trees with a leaf node size of 1,000 patients. We set the number of iRF iterations to five to rank the importance of each feature in predicting suicide attempt. In addition to ranking input features by importance, we next identified the feature-level explainability of our model by determining whether each feature was predictive for or against suicide attempt. To estimate if a feature predicts suicide attempt or controls, the result of each split was averaged using the given feature and mapping those to a linear effect. This provided both the feature effect in the slope of the line and an $R^2$ with how closely that related to each of the set of splits. If the slope of the line was positive, then the feature effect size direction was positive, i.e., the value and the feature were positively correlated. If the slope of the line was negative, then the feature effect size direction was negative.

## 2.4. Model selection

Model selection was based on research interest and accuracy gain. Initially, we aimed to address three questions: predicting suicide attempts based on climatic features, demographic features, and access to firearms and alcohol. We used iRF models with k-fold cross-validation for each feature group, and later employed all 1,536 features together to identify the most important predictors for suicide attempts across all feature groups in patients above or equal to 60 years and below 60 years. We then combined the top 20 most important features from the model with all features with alcohol and firearm vendor data per 10,000 residents. The reduced models showed better accuracy than using all features, as indicated by the area under the precision-recall curves (AUPRC) across 5 data splits (Figures 2, 3). Using all features introduced noise and reduced model accuracy and interpretability. Most models demonstrated predictability, with AUPRC values above random chance, especially in the age group below 60 years (red dashed line in Figure 3). Based on these results, we selected the model with the top 20 features, alcohol, and firearm vendors for model explanation.

## 2.5. iRF-LOOP

To show how features or groups of features interconnect each other, we applied iRF-Leave One Out Prediction (iRF-LOOP), which is an extension of the iRF model (36). In this framework, iRF was used to compute all-against-all predictions of each vector of data from all other vectors. The results of this analysis were captured as networks, in which nodes (i.e., features) were connected by an edge if the pair of features were predictive of each other, thereby revealing functional relationships and subgroups within and across data layers. We performed iRF-LOOP using the pre-processed input matrix which consisted of climate, census, and alcohol and firearm business data, with a total of 1,536 features and 31,378 samples or ZIP codes. This analysis created an all-to-all directed feature association network that captured the relationships between data layers. The resulting network was then filtered to the top 1% of edges to capture the most important connections between features.

# 3. Results

## 3.1. Features that were most strongly associated with suicide attempts

iRF predictive models can compute the importance of each feature predicting suicide attempt and if they predict either cases or controls. We examined iRF models trained with 1,536 features to identify the 20 most important predictive features in patients below, and above or equal to age 60, plus *firearms and alcohol vendors per 10,000 residents* (Supplementary Figures S1, S2). The resulting features were further analyzed to determine directionality (i.e., if they predict suicide attempts or controls) and can be aggregated in groups related to emotional support, housing, ancestry, commuting and mobility, access to healthcare, cognitive difficulties, access to means, decrease restraint and climate (Figures 4, 5).

In the model that includes patients aged 60 years or older, specific features were found to be associated with either a decreased or increased risk of attempting suicide within the emotional support category. The features *population of 18 years and older that live with a spouse*, *married males* and *household with spouse present between 35–64, and 65 years of age and above* were associated with decreased risk for attempting suicide, whereas *males never married of 75–84 years*, and *married males of 85 years of with spouse absent* were associated with increased risk for attempting suicide (Figure 4). Ancestry also appeared among the top features explaining suicide attempts. *French Canadian*, *Northern European*, and *Pennsylvania German ancestries* were associated with increased risk for suicide attempts, as well as *Native Hawaiian and other Pacific Islander females of ages 45–54* (Figure 4). *Dutch ancestry* and *speaking French (Haitian or Cajun dialects) at home* were associated with reduced risk (Figure 4). In the housing category, *occupied housing units that are occupied by a renter* was predictive of suicide attempts (Figure 4). For commuting and mobility, we observed that *moving abroad last year, 75 years and older*, and *females with a work commute lasting 10–14 min*, were also associated with increased risk (Figure 5). In access to means and decrease restraint groups, we observed that number of firearms and alcohol vendors per 10,000 residents were predictive of suicide attempt (Figure 5). Several climate features were also predictive: *precipitation*

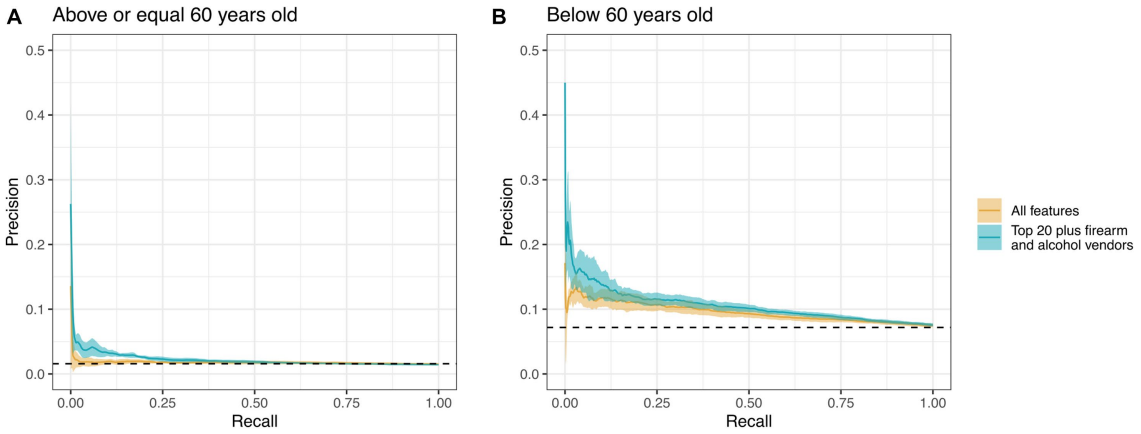



FIGURE 2
Precision-recall curves for an iRF model using all features (yellow line) vs. using only the top 20 features plus firearm and alcohol vendors per 10,000 residents (teal line). Each line represents the average of 5 runs along their respective 95% confidence intervals. **(A)** Model comprised patients who were 60 years of age or older ($n$ = 267,447). **(B)** The model focused exclusively on patients below the age of 60 ($n$ = 138,093). The dashed line represents the random chance of correct classification without iRF.

*in April*, *solar radiation in summer* (which represents solar radiation in June, through September), and *wind speed late spring through summer* (which represents wind speed in May through September) were associated with decreased risk, whereas *wind speed in autumn through spring* (which represents wind speed for the rest of the year) and *temperature in late autumn through winter* (which represents average temperature from October through April, minimum temperature from November through March, and maximum temperature from October through March) were associated with increased risk (Figure 5). For representative relationships among features (Pearson >0.9) see Supplementary Table S5.

For the model using only individuals under 60 years of age, the two most important features associated with decreased risk were *married males*, and *females with no cognitive difficulties between 35 and 64 years of age* [which represents a lack of disabilities, including no hearing, vision, ambulatory, or self-care difficulties in females 35–64 years of age (Figures 4, 5)]. Conversely, the features *living alone* and *living with unmarried partner* were associated with increased risk (Figure 4). Similarly, *commuting between 10 and 14 min,* regardless of gender, *monthly housing costs between $700 and $800, cash rent between $600 and $649*, *house heating using gas*, *moved within county,* and *having two types of public insurance* were all associated with increased risk (Figures 4, 5). On the other hand, only *Pennsylvania German ancestry* was associated with increased risk in the ancestry category (Figure 4). Similarly, to the model from patients 60 years of age or above, *number of firearm* and *alcohol vendors per 10,000 inhabitants* were associated with increased risk (Figure 5). Regarding climatic features, *precipitation in winter* (which represents precipitation December through February) and *Water vapor in mid spring through mid autumn* (which represents water vapor from April through October and minimum temperature in May) were associated with reduced risk (Figure 5). Contrary to the first model, in patients under 60 years of age the climatic features *solar radiation in summer*, and both *wind speed in late spring through summer* and *wind speed in autumn through spring* were predictive of individuals with a history of suicide attempt (Figure 5). Additionally, *terrain elevation* was also associated with suicide attempts (Figure 5).

## 3.2. Feature network by iRF-LOOP

It is well known that alcohol abuse increases suicide risk (2). Thus, we included the number of alcohol vendors per 10,000 people in our final model. Interestingly, when used in combination with the top 20 features, alcohol vendors per 10,000 people ranked 12th and 20th in feature importance in patients above or equal to 60 and below 60 years of age respectively, even though it ranked 360th (above age of 60) and 74th (below age of 60) in the models using all features (Supplementary Figures S1, S2). This could mean that other features are competing with number of alcohol vendors per 10,000 people within the iRF model to classify suicide attempt and controls. Therefore, we created a feature network using iRF-LOOP to explore how features relate to each other. Figure 6 shows the immediate neighbors of the feature measuring the number of *alcohol vendors per 10,000 people*, where each node in the subnetwork represents a feature and the arrows represent the edges (connections) between features. Edges are weighted by their normalized feature importance value and the arrow direction shows what feature is predicting another one. Several features related to *geographic mobility*, *population density*, *low gross rent with cash*, and *high home value* were observed. Further, the subnetwork shows associations among features including *number of widowed females*, and several *European ancestries*. Regarding climate features, we observed *temperature in late autumn through winter* and *temperature in late spring through summer*, which represents average temperature from May to August, maximum temperature in June, and minimum temperature from June to September.

# 4. Discussion

Most prior suicide prevention studies have focused on a relatively small number of features. Moreover, most have typically relied on individual-level information only (e.g., clinical features) or aggregated data for a given geographical area with low geospatial resolution. In the present study we showed that ZIP code-level data can improve prediction in individuals with a history of suicide attempt greater than

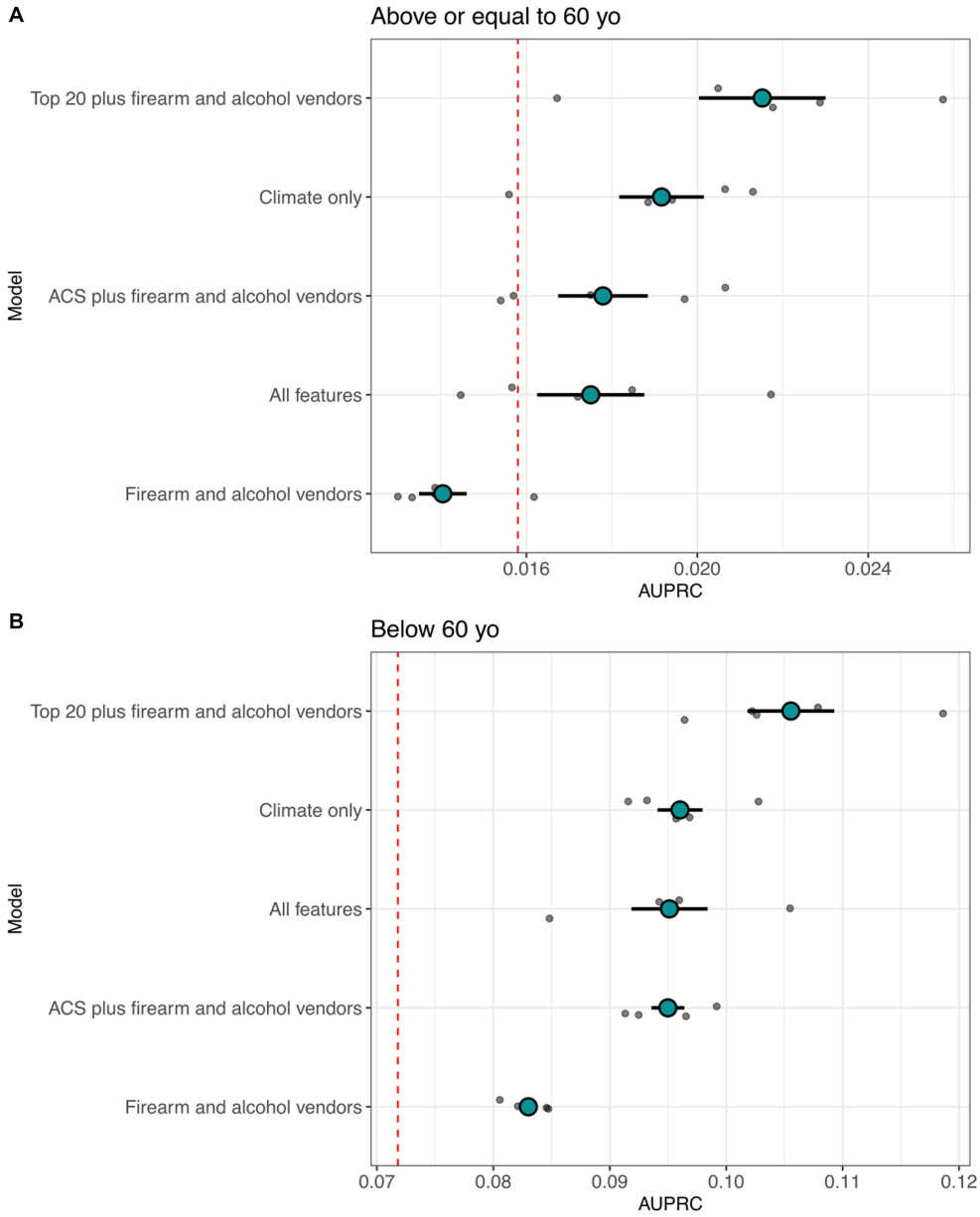



FIGURE 3
Distribution of the area under the precision recall curves (AUPRC) for suicide attempt predictive capacity using environmental features. **(A)** AUPRC value distributions for equal to/above 60 years of age for iRF cross validation models. **(B)** AUPRC value distributions for below 60 years of age. The red dashed line is the area under the curve of a base model (random chance of correct classification without iRF). The total patient count for the models was 267,447 for individuals aged 60 years or older and 138,093 for those below the age of 60.

random chance, supporting the role of the surrounding environment as proximal/precipitating factors that influence the propensity for an individual to attempt suicide.

Our selection of the iRF model was based on its suitability for our research objectives. In particular, its capacity to efficiently handle large, high-dimensional datasets was crucial for the present study. Furthermore, the model's feature weighting technique helps to address data overfitting by prioritizing the most informative features while removing the non-relevant ones. This method also enhances the interpretability of the results by generating a ranked list of the most influential features. Lastly, the use of decision trees allowed for the estimation of the directionality of the prediction, enabling a deeper interpretation of the environmental factors contributing to suicidal attempts. Thus, by using this X-AI we were able to screen more than a 1,700 demographic and climatic features to obtain several that appear to potentially protect or increase veterans' propensity for attempting suicide.

An individual's environment can have profound effects on their psychological state and subsequent risk of suicide. In fact, it has been well documented that socioeconomic and demographic factors such as poverty, education and population density are related to suicide risk (38–40). In our study we used the electronic health records, demographic and climate data to build a predictive model of suicide attempts in the U.S. veteran population. It is worth highlighting that

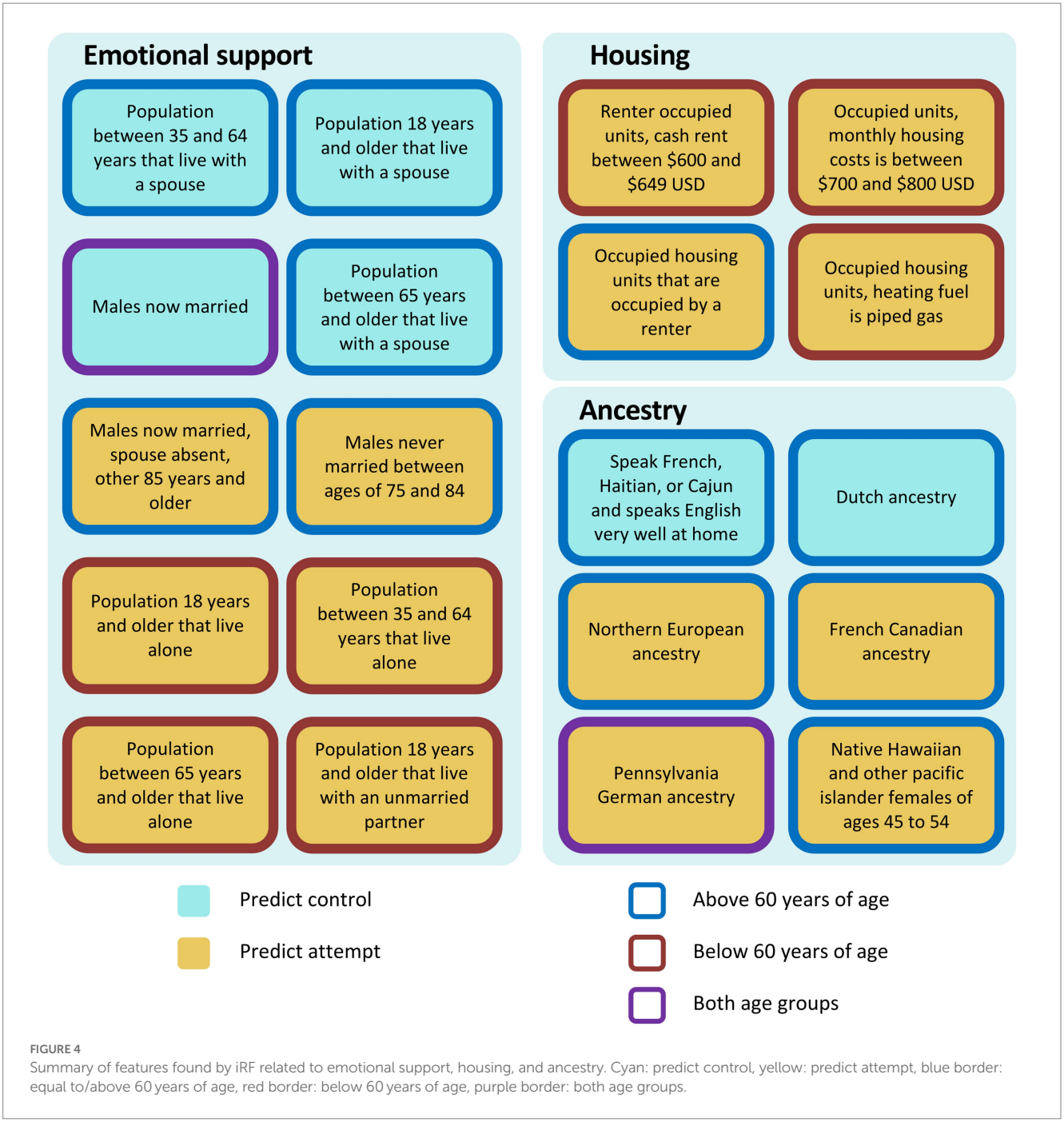



FIGURE 4
Summary of features found by iRF related to emotional support, housing, and ancestry. Cyan: predict control, yellow: predict attempt, blue border: equal to/above 60 years of age, red border: below 60 years of age, purple border: both age groups.

demographic and climate data are not direct information from the individuals under study, but rather a representation of the environment in which they currently reside. Moreover, in the cohort used here, there was a rapid decline in the number of suicide attempts after the age of 60 (Figure 1), which is in agreement with a surveillance summary from the Center for Disease Control and Prevention, where the highest suicide rates were observed in age group between 35 and 64 years (41). Thus, we divided the population into cohorts below or equal to/above 60 years of age, since age groups may be affected differently by risk factors (42).

Interestingly, the top features can be classified into nine main groups namely: emotional support, housing, ancestry, commuting and mobility, climate, decreased restraint, access to means, cognitive difficulties and access to healthcare (Figures 4, 5). In the emotional support group, the protective effect of marital status on suicide risk is well-documented (43). Studies have shown that married people showed the lowest rate of suicide rates (43–45), whereas divorced or separated persons are twice as likely to commit suicide than married persons, and this effect is stronger in divorced males (46). In our study, for all age groups we observed that living in areas with high proportions of married males and individuals living with a spouse were protective factors against suicide attempts for those 60 years of age or above (similar to the individual-level protective factor of marital status). Conversely, areas with high proportions of individuals living alone and unmarried males were associated with higher rates of suicide attempts (47).

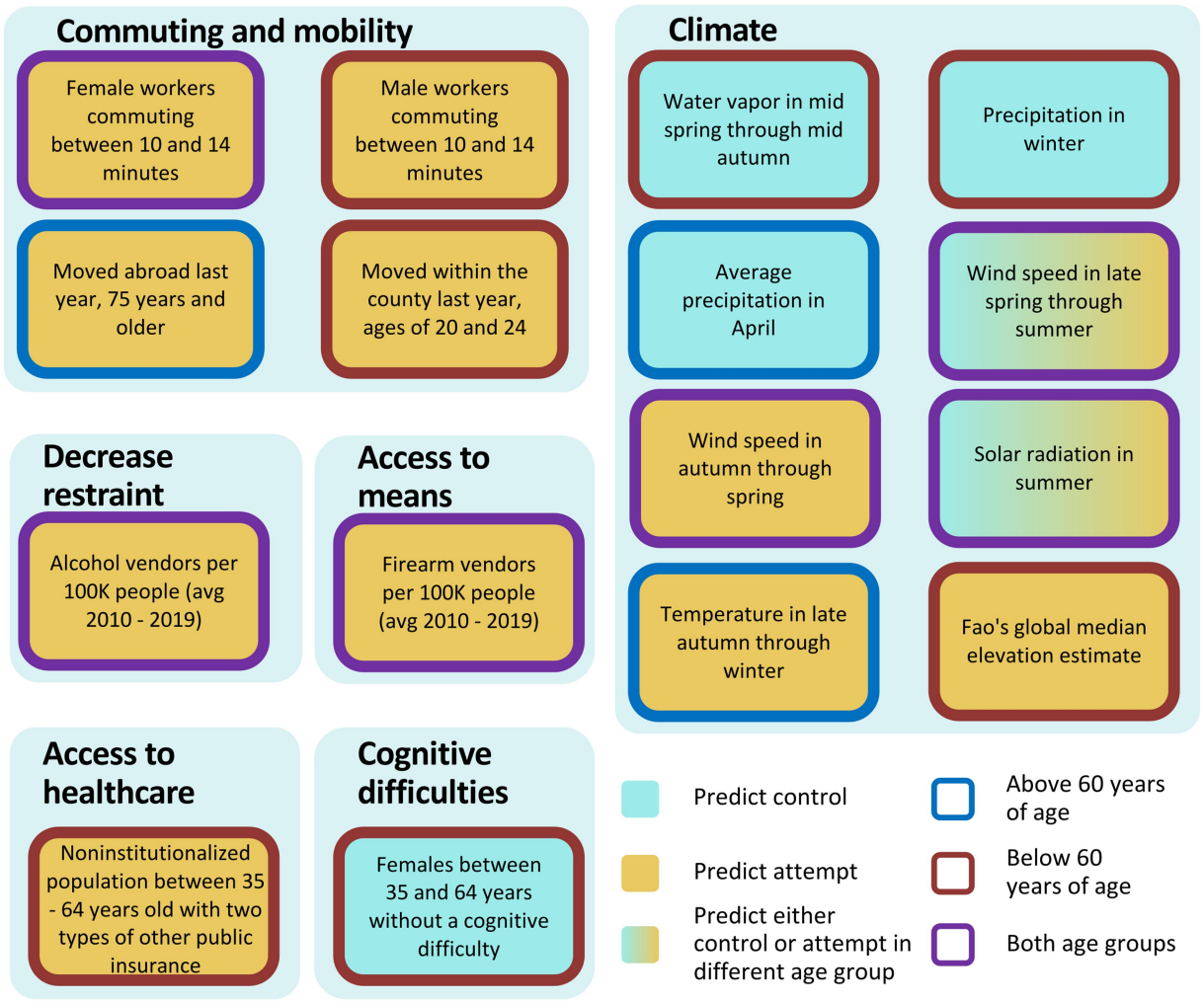



FIGURE 5
Summary of features found by iRF related to commuting and mobility, climate, decreased restraint, access to means, access to healthcare and cognitive difficulties (proxy for disabilities). Cyan: predict control, yellow: predict attempt, blue border: equal to/above 60 years of age, red border: below 60 years of age, purple border: both age groups.

Regarding the commuting and mobility group, commuting time has been associated with depression in a dose responsive manner (48). Our results showed that living in areas with high proportions of individuals with commutes between 10 and 14 min has been associated with increased risk for suicide attempts, irrespective of age. The iRF-LOOP network showed that commuting features are predictive of each other regardless of the commuting time (Supplementary Figure S3). Thus, commuting between 10 and 14 min may be acting as a proxy for commuting in general. We also observed housing-related features for both age groups in relation to suicide attempts. These results are consistent with Lorant et al. (49), who showed that higher education and home ownership status decreases the risk of suicide. Other studies have found that living in rented units increased the risk for suicide in middle-aged males and females (40). This effect might be especially important for females who live in large urban areas (50).

In the ancestry group, we observed that living in areas with a higher proportion of Northern European ancestry was associated with suicide attempts. Although suicide attempts vary widely in Europe, countries of Finno-Ugrian origin show disproportionally higher suicide rates than the rest of Europe, suggesting a genetic cause (51). Voracek et al. (52), tested this hypothesis in the U.S.A. using state-level self-reported ancestry from census data. The study found support for this hypothesis using historical data from 1913–1924 and 1928–1932, but not from 1990–1994. Taken together, these findings encourage an analysis with higher geospatial resolution of the sample (e.g., ZIP or county level) and a genetic characterization of the individuals ancestries, since self-reporting may be inaccurate. We also found that living in a ZIP code with higher proportion of *Native Hawaiian and other pacific islander females between ages 45 to 54* was associated with suicide attempts. These results are consistent with Ji et al. (53) finding, who showed that Asian or Pacific Islander ancestry was a risk factor for suicide in healthcare professionals.

Climate can have a significant impact on suicidal behavior, although the relationship between climate and suicide is complex and not yet fully understood. An explanation of how climate could impact suicidal behavior could be due to seasonal changes. The

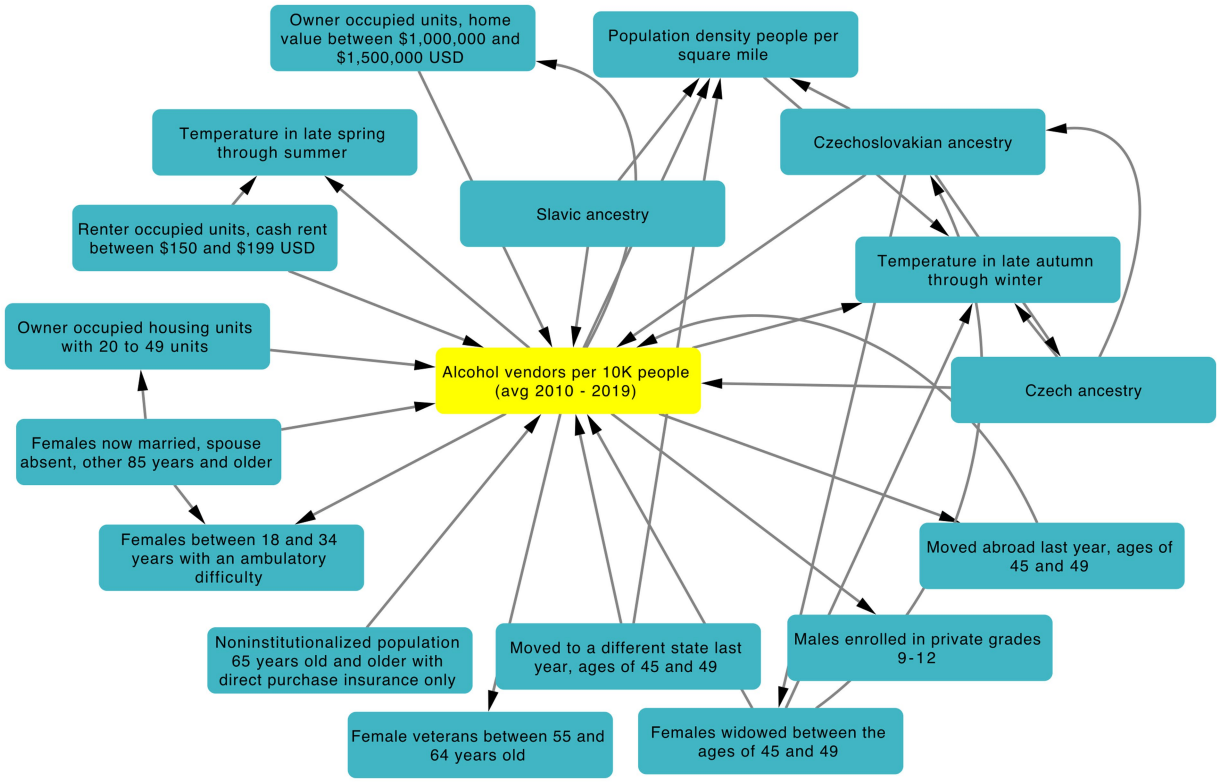



FIGURE 6
iRF-LOOP Subnetwork showing first neighbors of alcohol vendors per 10,000 people. Only the top 1% of edges weighed by normalized importance are shown. Arrow direction corresponds to one feature predicting another feature.

phenomenon of seasonality in death by suicide has been consistently observed across various studies. A comprehensive literature review encompassing a time span of three decades (1979–2009) revealed a prevailing pattern of increased suicide deaths occurring during the spring and early summer months (4). A subsequent systematic review conducted in 2016 revealed comparable seasonal patterns for suicide attempts, indicating a peak occurrence during the spring and summer months (5). While not all studies in this review identified this seasonal pattern, the majority of them appear to agree with this observation.

Several mood disorders related to suicidal behavior such as depression, bipolar and seasonal affective disorder (SAD) also show seasonality patterns (8). A plausible explanation of this seasonality is the availability of the neurotransmitter serotonin and its receptor, which may be dysregulated in patients with these psychiatric disorders (54–58). However, the exact role of serotonin in major depression is currently under discussion (59, 60). Seasonal variations in serotonin levels have been observed, with winter displaying the lowest levels and summer exhibiting the highest (61). The same study found a direct relationship between serotonin levels and the amount of sunlight on the day of assessment, without significant influence from preceding days. Another study found that individuals exposed to lower solar radiation in the days preceding measurement exhibit reduced postsynaptic serotonin receptor levels (62). Throughout the day, serotonin receptor levels increase while its transporter, which modulates serotonin availability, decreases (63). Intriguingly, longer days coincide with enhanced availability of the serotonin receptor (63). These findings underscore the intricate interplay between sunlight, serotonin, and seasonal fluctuations.

In the present study, we found that solar radiation in summer predicted suicide attempt in patients younger than 60 years. Consistent with this result, a study found that intentional drug overdose deaths in the United States were higher with longer day lengths, which correlates with months with higher solar radiation (64). Conversely, we also found that solar radiation in summer was protective against suicide attempts in patients of 60 years and older. A possible explanation for this variation among different age groups is that older individuals appear to exhibit greater resilience to the seasonal fluctuations of mood changes, as indicated by a self-reported study (65). Nevertheless, the underlying factors contributing to this resilience have yet to be thoroughly explored.

High temperatures are linked to suicide rates (6-8), and they also contribute to an increase in emergency department visits for mental health issues (66). A recent systematic review and meta-analysis revealed that heat negatively affects mental health, potentially by disrupting neurotransmitter balance, causing neuro-inflammation, and disrupting sleep patterns, with the elderly being particularly vulnerable (67). While heat waves typically occur in summer, our study identified that temperatures in late autumn through winter were associated with a higher risk of suicide among patients aged 60 and above. Ajdacic-Gross et al. (68) propose a possible explanation for this inconsistency. Their study, which used moving frames to analyze time series data, found a minor peak of association between suicide risk and summer frames, and a major peak during winter frames. However,

it is important to interpret these findings cautiously, as the study was conducted in Switzerland and may not be generalizable to other regions. The biological mechanism underlying these findings has yet to be explored.

Numerous studies have explored the relationship between water vapor (humidity), rainfall, and suicide risk (14, 15). However, the findings either lack evidence of association or yield contradictory results, making it challenging to determine the impact of these features on suicide risk. Our study indicates a potential link between precipitation, water vapor, and reduced suicide risk, but the protective mechanisms behind these features remain inconclusive.

Limited research also exists on the relationship between wind speed and suicide risk. Some studies yield contradictory or inconclusive results (10-14). In our study, we found that wind speed during autumn through spring predicted suicide attempts in both age groups, while wind speed in late spring through summer protected against suicide attempts in patients aged 60 and above. One possible explanation for the predictive nature of wind speed in autumn through spring is its alignment with the tornado season, particularly May, which historically experiences the highest tornado activity. A systematic review revealed that exposure to tornadoes was associated with adverse mental health outcomes (69). While tornadoes have been linked to mental health issues, a separate study found associations between hurricanes, rather than tornadoes, and suicide rates (70).

Although terrain elevation is not inherently a climatic feature, we classified it as such due to its association with various climatic factors that undergo changes alongside it. Our study revealed a positive association between altitude and suicide attempts, aligning with a systematic review that showed a similar trend in 17 out of 19 analyzed studies (70, 71). Hypobaric hypoxia-induced dysregulation of serotonin levels has been proposed as a potential mechanism for increased suicide risk at high altitudes, but further empirical testing is required (72).

Another goal of the present work was to explore decreased restraint and access to means for suicide attempts using the number of alcohol and firearms vendors per 10,000 residents as proxies. Importantly, regardless of age, we observed that living in areas with more alcohol and firearm vendors was associated with increased risk for attempting suicide (albeit less so than several of the climatic and socio-demographic features identified as important). These results provide additional evidence regarding the importance of access to means as one of several risk factors for suicide, especially in the U.S., where a larger proportion of suicides are committed using firearms in comparison to other high-income countries where only 5% use them (22, 73, 74). Our findings also confirm the importance of including alcohol abuse in models of suicide risk (75, 76).

In this study we demonstrated the use of X-AI to explore the impact of more than 1,700 demographic and climatic features on suicide attempt risk with high geospatial resolution. This research provides additional evidence for the role of several demographic and climatic features in suicide attempts and demonstrates the utility of using geospatial features from the area (e.g., neighborhoods/communities) in which patients live within an X-AI framework to improve suicide risk prediction. By focusing solely on sociodemographic and environmental factors, we aimed to determine their distinct influence on suicide risk, regardless of clinical or psychiatric factors. We also aimed to investigate their influence from a public health standpoint. Analyzing these factors can provide valuable insights for developing effective interventions and policies to decrease suicide risk. By identifying the key factors that contribute to suicide risk, researchers and policymakers can prioritize interventions that specifically target these factors. For example, understanding the localized variations in risk factors can help target interventions to specific regions or communities. Suicide prevention measures can be tailored based on the unique sociodemographic and environmental characteristics of different areas, allowing for more effective and efficient allocation of resources. Interventions could include promoting social support networks, providing tailored mental health services, and implementing community-based programs that address the unique needs and challenges of vulnerable populations. Overall, the findings contribute valuable insights to suicide prevention measures by highlighting the role of sociodemographic and environmental factors and emphasizing the need for a holistic, geospatially-informed approach. By integrating these findings into policies and interventions, it is possible to develop more effective strategies for preventing suicide and promoting mental well-being in at-risk populations.

## 4.1. Limitations

The present study had a number of limitations that should be considered when interpreting these findings. First, we utilized a cross-sectional design, and a lifetime suicide attempts variable. Thus, additional work is still needed to examine the degree to which the features identified in the present study might be predictive of future suicide attempts. Second, our findings regarding alcohol and firearms vendors should be interpreted cautiously, as there are several other features (Supplementary Figure S4) that could also potentially explain their association with suicide attempts which should be considered in future work. Third, it is important to note that the cohort examined in this study primarily consisted of males of Caucasian descent, with an average age of about 60. Consequently, caution must be exercised when generalizing these findings to individuals of different racial backgrounds, age groups, and genders. Fourth, while the present study provides clear evidence that zip code-level geospatial features can predict suicide attempt risk better than random chance, the degree to which these features might predict suicide attempts above and beyond well-validated individual-level risk factors (e.g., psychiatric diagnosis) remains unknown. While such work is beyond the scope of the present study, future work is still needed to ascertain if and how geospatial features might interact with individual-level data to predict suicide risk.

# Data availability statement

Suicide attempt data from individual patients is not readily available because this data is considered Protected Health Information by the U.S. Department of Veterans Affairs. All climate data and socio-demographics can be made available upon request.

# Ethics statement

The studies involving human participants were reviewed and approved by VA Central Institutional Review Board (cIRB# 18-11). The patients/participants provided their written informed consent to participate in this study.

## Author contributions

MP, JL, AW, KS, MG, NK, and DJ: conceptualization. MP, JL, AC, JR, and DJ: methodology. MP: validation, investigation, and visualization. MP and AW: formal analysis. MP, AW, and JS: data curation. MP, KS, JS, MG, JP, and NK: data interpretation. MP, JL, and AW: writing of the original draft. MP, JL, KS, JS, MG, BM, DO, JB, NK, and DJ: writing—review and editing. DJ: funding. NK and DJ: supervision. All authors contributed to the article and approved the submitted version.

## Funding

This research is based on data from the Million Veteran Program, Office of Research and Development (ORD), Veterans Health Administration (VHA), and was supported by award #I01CX001729 from the Clinical Science Research and Development (CSR&D) Service of VHA ORD. This work was also supported in part by the joint U.S. Department of Veterans Affairs and US Department of Energy MVP CHAMPION program.

## Acknowledgments

The authors thank and acknowledge MVP and the MVP Suicide Exemplar Workgroup for their contributions to this manuscript.

## 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



## Author disclaimer



## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1178633/full#supplementary-material

## References

1. WHO. World Health Statistics data visualizations dashboard. Noncommunicable diseases and mental health. *World Health Organization* (2019). Available at: https://www.who.int/data/gho/data/themes/mental-health/suicide-rates (Accessed December 6, 2021).

2. Turecki G, Brent DA, Gunnell D, O'Connor RC, Oquendo MA, Pirkis J, et al. Suicide and suicide risk. *Nat Rev Dis Prim*. (2019) 5:74. doi: 10.1038/s41572-019-0121-0

3. Huang X, Ribeiro JD, Musacchio KM, Franklin JC. Demographics as predictors of suicidal thoughts and behaviors: A meta-analysis. *PLoS One*. (2017) 12:e0180793. doi: 10.1371/journal.pone.0189461

4. Christodoulou C, Douzenis A, Papadopoulos FC, Papadopoulou A, Bouras G, Gournellis R, et al. Suicide and seasonality. *Acta Psychiatr Scand*. (2012) 125:127–46. doi: 10.1111/j.1600-0447.2011.01750.x

5. Coimbra DG, PEAC S, de Sousa-Rodrigues CF, Barbosa FT, de Siqueira Figueredo D, Araújo Santos JL, et al. Do suicide attempts occur more frequently in the spring too? A systematic review and rhythmic analysis. *J Affect Disord*. (2016) 196:125–37. doi: 10.1016/j.jad.2016.02.036

6. Burke M, González F, Baylis P, Heft-Neal S, Baysan C, Basu S, et al. Higher temperatures increase suicide rates in the United States and Mexico. *Nat Clim Chang*. (2018) 8:723–9. doi: 10.1038/s41558-018-0222-x

7. Dixon PG, Kalkstein AJ. Where are weather-suicide associations valid? An examination of nine US counties with varying seasonality. *Int J Biometeorol*. (2018) 62:685–97. doi: 10.1007/s00484-016-1265-1

8. Heo S, Lee W, Bell ML. Suicide and associations with air pollution and ambient temperature: a systematic review and meta-analysis. *Int J Environ Res Public Health*. (2021) 18:7699. doi: 10.3390/ijerph18147699

9. Zhang R, Volkow ND. Seasonality of brain function: role in psychiatric disorders. *Transl Psychiatry*. (2023) 13:65. doi: 10.1038/s41398-023-02365-x

10. Perwira S, Yudianto A. Analysis of climate factors on suicide cases in East Java Province Indonesia in 2015-2018. (2023). Preprint (Version 1) available at Research Square. doi: 10.21203/rs.3.rs-2692709/v1

11. Grjibovski AM, Kozhakhmetova G, Kosbayeva A, Menne B. Associations between air temperature and daily suicide counts in Astana, Kazakhstan. *Medicina (Kaunas)*. 49:379–85. doi: 10.3390/medicina49080059

12. Lester D. A hazardous environment and city suicide rates. *Percept Mot Skills*. (1996) 82:1330. doi: 10.2466/pms.1996.82.3c.1330

13. Maes M, De Meyer F, Thompson P, Peeters D, Cosyns P. Synchronized annual rhythms in violent suicide rate, ambient temperature and the light-dark span. *Acta Psychiatr Scand*. (1994) 90:391–6. doi: 10.1111/j.1600-0447.1994.tb01612.x

14. Deisenhammer EA. Weather and suicide: the present state of knowledge on the association of meteorological factors with suicidal behaviour. *Acta Psychiatr Scand*. (2003) 108:402–9. doi: 10.1046/j.0001-690X.2003.00209.x

15. Woo J-M, Okusaga O, Postolache TT. Seasonality of suicidal behavior. *Int J Environ Res Public Health*. (2012) 9:531–47. doi: 10.3390/ijerph9020531

16. Lin C-Y, Hsu C-Y, Gunnell D, Chen Y-Y, Chang S-S. Spatial patterning, correlates, and inequality in suicide across 432 neighborhoods in Taipei City, Taiwan. *Soc Sci Med*. (2019) 222:20–34. doi: 10.1016/j.socscimed.2018.12.011

17. Pompili M, Serafini G, Innamorati M, Dominici G, Ferracuti S, Kotzalidis GD, et al. Suicidal behavior and alcohol abuse. *Int J Environ Res Public Health*. (2010) 7:1392–431. doi: 10.3390/ijerph7041392

18. Nock MK, Green JG, Hwang I, McLaughlin KA, Sampson NA, Zaslavsky AM, et al. Prevalence, correlates, and treatment of lifetime suicidal behavior among adolescents: results from the National Comorbidity Survey Replication Adolescent Supplement. *JAMA Psychiatry*. (2013) 70:300–10. doi: 10.1001/2013.jamapsychiatry.55

19. Milner A, Witt K, Maheen H, LaMontagne AD. Access to means of suicide, occupation and the risk of suicide: a national study over 12 years of coronial data. *BMC Psychiatry*. (2017) 17:125. doi: 10.1186/s12888-017-1288-0

20. Skegg K, Firth H, Gray A, Cox B. Suicide by occupation: does access to means increase the risk? *Aust N Z J Psychiatry*. (2010) 44:429–34. doi: 10.3109/00048670903487191

21. Sarchiapone M, Mandelli L, Iosue M, Andrisano C, Roy A. Controlling access to suicide means. *Int J Environ Res Public Health*. (2011) 8:4550–62. doi: 10.3390/ijerph8124550

22. Ertl A, Sheats KJ, Petrosky E, Betz CJ, Yuan K, Fowler KA. Surveillance for Violent Deaths - National Violent Death Reporting System, 32 States, 2016. *MMWR Surveill Summ*. (2019) 68:1–36. doi: 10.15585/mmwr.ss.6809a1

23. Kaplan MS, Huguet N, McFarland BH, Caetano R, Conner KR, Giesbrecht N, et al. Use of alcohol before suicide in the United States. *Ann Epidemiol*. (2014) 24:588–592.e2. doi: 10.1016/j.annepidem.2014.05.008

24. Kimbrel NA, Ashley-Koch AE, Qin XJ, Lindquist JH, Garrett ME, Dennis MF, et al. A genome-wide association study of suicide attempts in the million veterans program identifies evidence of pan-ancestry and ancestry-specific risk. *Mol Psychiatry*. 27:2264–72. doi: 10.1038/s41380-022-01472-3

25. Hoffmire C, Stephens B, Morley S, Thompson C, Kemp J, Bossarte RM. VA suicide prevention applications network: a national health care system-based suicide event tracking system. *Public Health Rep*. (2016) 131:816–21. doi: 10.1177/0033354916670133

26. Fick SE, Hijmans RJ. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol*. (2017) 37:4302–15. doi: 10.1002/joc.5086

27. Fischer G, Nachtergaele F, Prieler S. *Global agro-ecological zones assessment for agriculture (GAEZ 2008)*. Laxenburg: IIASA (2008).

28. Fischer G, Nachtergaele FO, Prieler S. *Global Agro-Ecological Zones (GAEZ v3. 0): Model Documentation*. Laxenburg: International Institute for Applied systems Analysis (IIASA) (2012).

29. Igbp-Dis S. *A program for creating global soil-property databases*. France: IGBP Global Soils Data Task (1998).

30. Trabucco A, Zomer RJ. High-resolution global soil-water balance explicit for climate–standard vegetation and soil conditions In:. *CGIAR Consortium for Spatial Information* (2010)

31. Zomer JR, Bossio AD, Trabucco A, Yuanjie L, Gupta CD, et al. *Trees and water: smallholder agroforestry on irrigated lands in Northern India*. IWMI (2007). 122:41.

32. Zomer RJ, Trabucco A, Bossio DA, Verchot LV. Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agricult Ecosyst Environ*. (2008) 126:67–80. doi: 10.1016/j.agee.2008.01.014

33. United States Census Bureau. *American Community Survey: Data Profiles*. (2019). Available at: https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2019/

34. Walker K, Herman M. tidycensus: Load US Census Boundary and Attribute Data as "tidyverse" and "sf"-Ready Data Frames. (2021). Available at: https://CRAN.R-project.org/package=tidycensus.

35. Data Axle. *2010-2019 Historical Business Data*. (2021).

36. Cliff A, Romero J, Kainer D, Walker A, Furches A, Jacobson D. A high-performance computing implementation of iterative random forest for the creation of predictive expression networks. *Genes (Basel)*. (2019) 10:996. doi: 10.3390/genes10120996

37. Basu S, Kumbier K, Brown JB, Yu B. Iterative random forests to discover predictive and stable high-order interactions. *Proc Natl Acad Sci USA*. (2018) 115:1943–8. doi: 10.1073/pnas.1711236115

38. Bantjes J, Iemmi V, Coast E, Channer K, Leone T, McDaid D, et al. Poverty and suicide research in low- and middle-income countries: systematic mapping of literature published in English and a proposed research agenda. Global Mental. *Health*. (2016) 3. doi: 10.1017/gmh.2016.27

39. Vichi M, Vitiello B, Ghirini S, Pompili M. Does population density moderate suicide risk? An Italian population study over the last 30 years. *Eur Psychiatry*. (2020) 63:e70. doi: 10.1192/j.eurpsy.2020.69

40. Johansson LM, Sundquist J, Johansson S-E, Qvist J, Bergman B. The influence of ethnicity and social and demographic factors on Swedish suicide rates. *Soc Psychiatry Psychiatr Epidemiol*. (1997) 32:165–70. doi: 10.1007/BF00794616

41. Ivey-Stephenson AZ, Crosby AE, Jack SPD, Haileyesus T, Kresnow-Sedacca M-J. Suicide trends among and within urbanization levels by sex, race/ethnicity, age group, and mechanism of death — United States, 2001–2015. *MMWR Surveillance Summaries*. (2017) 66:1–16. doi: 10.15585/mmwr.ss6618a1

42. Suresh Kumar PN, Anish PK, George B. Risk factors for suicide in elderly in comparison to younger age groups. *Indian J Psychiatry*. (2015) 57:249–54. doi: 10.4103/0019-5545.166614

43. Rendall MS, Weden MM, Favreault MM, Waldron H. The protective effect of marriage for survival: a review and update. *Demography*. (2011) 48:481–506. doi: 10.1007/s13524-011-0032-5

44. Smith JC, Mercy JA, Conn JM. Marital status and the risk of suicide. *Am J Public Health*. (1988) 78:78–80. doi: 10.2105/AJPH.78.1.78

45. Kyung-Sook W, SangSoo S, Sangjin S, Young-Jeon S. Marital status integration and suicide: A meta-analysis and meta-regression. *Soc Sci Med*. (2018) 197:116–26. doi: 10.1016/j.socscimed.2017.11.053

46. Kposowa AJ. Marital status and suicide in the National Longitudinal Mortality Study. *J Epidemiol Community Health*. (2000) 54:254–61. doi: 10.1136/jech.54.4.254

47. Calati R, Ferrari C, Brittner M, Oasi O, Olié E, Carvalho AF, et al. Suicidal thoughts and behaviors and social isolation: A narrative review of the literature. *J Affect Disord*. (2019) 245:653–67. doi: 10.1016/j.jad.2018.11.022

48. Wang X, Rodríguez DA, Sarmiento OL, Guaje O. Commute patterns and depression: Evidence from eleven Latin American cities. *J Transp Health*. (2019) 14:100607. doi: 10.1016/j.jth.2019.100607

49. Lorant V, Kunst AE, Huisman M, Costa G, Mackenbach J. EU Working Group on Socio-Economic Inequalities in Health. Socio-economic inequalities in suicide: a European comparative study. *Br J Psychiatry*. (2005) 187:49–54. doi: 10.1192/bjp.187.1.49

50. Johansson LM, Sundquist J, Johansson SE, Bergman B. Ethnicity, social factors, illness and suicide: a follow-up study of a random sample of the Swedish population. *Acta Psychiatr Scand*. (1997) 95:125–31. doi: 10.1111/j.1600-0447.1997.tb00385.x

51. Marusic A, Farmer A. Genetic risk factors as possible causes of the variation in European suicide rates. *Br J Psychiatry*. (2001) 179:194–6. doi: 10.1192/bjp.179.3.194

52. Voracek M. Ancestry, genes, and suicide: a test of the Finno-Ugrian Suicide Hypothesis in the United States. *Percept Mot Skills*. (2006) 103:543–50. doi: 10.2466/pms.103.2.543-550

53. Ji YD, Robertson FC, Patel NA, Peacock ZS, Resnick CM. Assessment of Risk Factors for Suicide Among US Health Care Professionals. *JAMA Surg*. (2020) 155:713–21. doi: 10.1001/jamasurg.2020.1338

54. Mc Mahon B, Andersen SB, Madsen MK, Hjordt LV, Hageman I, Dam H, et al. Seasonal difference in brain serotonin transporter binding predicts symptom severity in patients with seasonal affective disorder. *Brain*. (2016) 139:1605–14. doi: 10.1093/brain/aww043

55. Tyrer AE, Levitan RD, Houle S, Wilson AA, Nobrega JN, Meyer JH. Increased Seasonal Variation in Serotonin Transporter Binding in Seasonal Affective Disorder. *Neuropsychopharmacology*. (2016) 41:2447–54. doi: 10.1038/npp.2016.54

56. Mc Mahon B, Nørgaard M, Svarer C, Andersen SB, Madsen MK, Baaré WFC, et al. Seasonality-resilient individuals downregulate their cerebral 5-HT transporter binding in winter - A longitudinal combined C-DASB and C-SB207145 PET study. *Eur Neuropsychopharmacol*. (2018) 28:1151–60. doi: 10.1016/j.euroneuro.2018.06.004

57. Mahmood T, Silverstone T. Serotonin and bipolar disorder. *J Affect Disord*. (2001) 66:1–11. doi: 10.1016/S0165-0327(00)00226-3

58. Oquendo MA, Hastings RS, Huang Y-Y, Simpson N, Ogden RT, Hu X-Z, et al. Brain serotonin transporter binding in depressed patients with bipolar disorder using positron emission tomography. *Arch Gen Psychiatry*. (2007) 64:201–8. doi: 10.1001/archpsyc.64.2.201

59. Moncrieff J, Cooper RE, Stockmann T, Amendola S, Hengartner MP, Horowitz MA. The serotonin theory of depression: a systematic umbrella review of the evidence. *Mol Psychiatry*. (2022). doi: 10.1038/s41380-022-01661-0

60. Bartova L, Lanzenberger R, Rujescu D, Kasper S. Reply to: "The serotonin theory of depression: a systematic umbrella review of the evidence" published by Moncrieff J, Cooper RE, Stockmann T, Amendola S, Hengartner MP, Horowitz MA in Molecular Psychiatry (2022 Jul 20. doi: 10.1038/s41380-022-01661-0). *Mol Psychiatry*. (2023). doi: 10.1038/s41380-023-02093-0

61. Lambert GW, Reid C, Kaye DM, Jennings GL, Esler MD. Effect of sunlight and season on serotonin turnover in the brain. *Lancet*. (2002) 360:1840–2. doi: 10.1016/S0140-6736(02)11737-5

62. Spindelegger C, Stein P, Wadsak W, Fink M, Mitterhauser M, Moser U, et al. Light-dependent alteration of serotonin-1A receptor binding in cortical and subcortical limbic regions in the human brain. *World J Biol Psychiatry*. (2012) 13:413–22. doi: 10.3109/15622975.2011.630405

63. Matheson GJ, Schain M, Almeida R, Lundberg J, Cselényi Z, Borg J, et al. Diurnal and seasonal variation of the brain serotonin system in healthy male subjects. *Neuroimage*. (2015) 112:225–31. doi: 10.1016/j.neuroimage.2015.03.007

64. Han B, Compton WM, Einstein EB, Cotto J, Hobin JA, Stein JB, et al. Intentional Drug Overdose Deaths in the United States. *Am J Psychiatry*. (2022) 179:163–5. doi: 10.1176/appi.ajp.2021.21060604

65. Höller Y, Gudjónsdottir BE, Valgeirsdóttir SK, Heimisson GT. The effect of age and chronotype on seasonality, sleep problems, and mood. *Psychiatry Res*. (2021) 297:113722. doi: 10.1016/j.psychres.2021.113722

66. Nori-Sarma A, Sun S, Sun Y, Spangler KR, Oblath R, Galea S, et al. Association between ambient heat and risk of emergency department visits for mental health among US adults, 2010 to 2019. *JAMA Psychiatry*. (2022) 79:341–9. doi: 10.1001/jamapsychiatry.2021.4369

67. Liu J, Varghese BM, Hansen A, Xiang J, Zhang Y, Dear K, et al. Is there an association between hot weather and poor mental health outcomes? A systematic review and meta-analysis. *Environ Int*. (2021) 153:106533. doi: 10.1016/j.envint.2021.106533

68. Ajdacic-Gross V, Lauber C, Sansossio R, Bopp M, Eich D, Gostynski M, et al. Seasonal associations between weather conditions and suicide--evidence against a classic hypothesis. *Am J Epidemiol*. (2007) 165:561–9. doi: 10.1093/aje/kwk034

69. Lee S, First JM. Mental health impacts of tornadoes: a systematic review. *Int J Environ Res Public Health*. (2022) 19:13747. doi: 10.3390/ijerph192113747

70. Krug EG, Kresnow M, Peddicord JP, Dahlberg LL, Powell KE, Crosby AE, et al. Suicide after natural disasters. *N Engl J Med*. (1998) 338:373–8. doi: 10.1056/NEJM199802053380607

71. Brown A, Hellem T, Schreiber J, Buerhaus P, Colbert A. Suicide and altitude: a systematic review of global literature. *Public Health Nurs*. (2022) 39:1167–79. doi: 10.1111/phn.13090

72. Kious BM, Kondo DG, Renshaw PF. Living high and feeling low: altitude, suicide, and depression. *Harv Rev Psychiatry*. (2018) 26:43–56. doi: 10.1097/HRP.0000000000000158

73. Yip PSF, Caine E, Yousuf S, Chang S-S, Wu KC-C, Chen Y-Y. Means restriction for suicide prevention. *Lancet*. (2012) 379:2393–9. doi: 10.1016/S0140-6736(12)60521-2

74. Grinshteyn E, Hemenway D. Violent death rates in the US compared to those of the other high-income countries, 2015. *Prev Med*. (2019) 123:20–6. doi: 10.1016/j.ypmed.2019.02.026

75. Arsenault-Lapierre G, Kim C, Turecki G. Psychiatric diagnoses in 3275 suicides: a meta-analysis. *BMC Psychiatry*. (2004) 4:37. doi: 10.1186/1471-244X-4-37

76. Dumais A, Lesage AD, Alda M, Rouleau G, Dumont M, Chawky N, et al. Risk factors for suicide completion in major depression: a case-control study of impulsive and aggressive behaviors in men. *Am J Psychiatry*. (2005) 162:2116–24. doi: 10.1176/appi.ajp.162.11.2116

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 04 August 2023
DOI 10.3389/fpsyt.2023.1194669



OPEN ACCESS

EDITED BY
Jannis Kraiss,
University of Twente, Netherlands

REVIEWED BY
Albert Eduard Boon,
Parnassia Psychiatric Institute, Netherlands
Sander De Vos,
University of Twente, Netherlands

*CORRESPONDENCE
Noortje I. van Vliet
✉ n.vanvliet@dimence.nl

†These authors share first authorship

RECEIVED 27 March 2023
ACCEPTED 19 July 2023
PUBLISHED 04 August 2023

CITATION
Bremer-Hoeve S, van Vliet NI, van Bronswijk SC, Huntjens RJC, de Jongh A and van Dijk MK (2023) Predictors of treatment dropout in patients with posttraumatic stress disorder due to childhood abuse[1]. *Front. Psychiatry* 14:1194669. doi: 10.3389/fpsyt.2023.1194669



# Predictors of treatment dropout in patients with posttraumatic stress disorder due to childhood abuse[1]

Susanne Bremer-Hoeve[1†], Noortje I. van Vliet[1*†], Suzanne C. van Bronswijk[2,3], Rafaele J.C. Huntjens[4], Ad de Jongh[5,6,7] and Maarten K. van Dijk[1]

[1]Dimence Mental Health Group, Deventer, Netherlands, [2]Department of Psychiatry and Neuropsychology, School for Mental health and Neuroscience, Faculty of Health Medicine and Life Sciences, Maastricht University, Maastricht, Netherlands, [3]Department of Psychiatry and Psychology, Maastricht University Medical Center, Maastricht, Netherlands, [4]Department of Experimental Psychotherapy and Psychopathology, University of Groningen, Groningen, Netherlands, [5]Department of Social Dentistry and Behavioral Sciences, University of Amsterdam and Vrije Universiteit, Amsterdam, Netherlands, [6]School of Health Sciences, Salford University, Manchester, United Kingdom, [7]Institute of Health and Society, University of Worcester, Worcester, United Kingdom

**Background:** Knowledge about patient characteristics predicting treatment dropout for post-traumatic stress disorder (PTSD) is scarce, whereas more understanding about this topic may give direction to address this important issue.

**Method:** Data were obtained from a randomized controlled trial in which a phase-based treatment condition (Eye Movement Desensitization and Reprocessing [EMDR] therapy preceded by Skills Training in Affect and Interpersonal Regulation [STAIR]; $n = 57$) was compared with a direct trauma-focused treatment (EMDR therapy only; $n = 64$) in people with a PTSD due to childhood abuse. All pre-treatment variables included in the trial were examined as possible predictors for dropout using machine learning techniques.

**Results:** For the dropout prediction, a model was developed using Elastic Net Regularization. The ENR model correctly predicted dropout in 81.6% of all individuals. Males, with a low education level, suicidal thoughts, problems in emotion regulation, high levels of general psychopathology and not using benzodiazepine medication at screening proved to have higher scores on dropout.

**Conclusion:** Our results provide directions for the development of future programs in addition to PTSD treatment or for the adaptation of current treatments, aiming to reduce treatment dropout among patients with PTSD due to childhood abuse.

KEYWORDS

**PTSD, childhood abuse, EMDR, dropout, predictors, machine learning**

1 The study protocol was approved by the Medical Ethics Committee, reference number P16–03. The study design was registered in a national trial register (https://www.trialregister.nl/trialreg/admin/rctview.asp?TC=5,991) NTR5991.

# 1. Introduction

Posttraumatic stress disorder (PTSD) has a major impact on social and occupational functioning and individuals' quality of life (1). In addition, individuals with PTSD are at great risk of attempting suicide (2–4). Although many treatments targeting PTSD have proven to be effective [e.g., (5)], a recent meta-analysis of randomized controlled trials (RCTs) found an average dropout rate of 21% for guideline-recommended PTSD treatments (6). This is problematic as untreated PTSD compared to treated PTSD may lead to a worse prognosis, and many societal consequences (7). Considering the societal impact of PTSD, it is important to ensure that individuals suffering from PTSD complete their treatment because evidence-based PTSD treatments significantly improve their prognosis (7). Completing treatment may be of particular importance for those with a history of childhood abuse since they are at risk of displaying more severe symptoms of PTSD or developing symptoms of Complex PTSD [ISTSS, 2012; (8)].

In order to establish a Complex PTSD diagnosis individuals need to experience symptoms of "Disturbances in Self-Organization" [DSO; i.e., problems with affect regulation, negative self-concept, and interpersonal problems; (9)], in addition to all diagnostic criteria of PTSD. It has been suggested that existing evidence-based trauma-focused therapies may lead to less favorable outcomes and more dropout in patients with a history of childhood abuse (10–12), and that these patients are at risk of developing Complex PTSD (8). Previous attempts to decrease the dropout rate from PTSD treatments in this target group by adding treatment programs that specifically target Complex PTSD symptoms in addition to actual trauma-focused treatment (11, 13) did not lead to less dropout compared to only applying direct trauma-focused treatment (14–16). The identification of patient characteristics that predict early treatment termination is essential for developing new target strategies to prevent dropout.

Most meta-analyses studying potential predictors of dropout regarding PTSD treatments focus on the kind of therapy as a predictive factor [i.e., (17–21)]. As far as we know, only Varker et al. (6) performed a meta-analysis in which they also considered patient characteristics in the comparison of several PTSD treatments among patients with military and civilian trauma. PTSD chronicity, PTSD severity, medication use, age, employment status, relationship status, sex, baseline depression scores, or baseline anxiety scores were included separately as possible predictors, but none of the included patient characteristics were found to be predictive of dropout. To this end, given the scarcity of studies and variables considered so far and the lack of meta-analytic support for specific comorbid symptoms as potential predictors for treatment dropout, more research on patient characteristics predicting dropout is needed. As, until now, not one specific characteristic has been found to be predictive of dropout, it may be even more important to determine whether a combination of characteristics may be predictive.

The purpose of the present study was to identify patients who are at risk of dropping out, and to identify patient characteristics that predict the dropout of PTSD treatments in patients with PTSD related to childhood abuse. To achieve this aim, we used machine learning techniques. One advantage of machine learning techniques is that all available pretreatment variables and variable combinations are examined (22), thus not limiting the number of possible predictors. Another advantage is that it is particularly appropriate for identifying small effects in predicting outcomes (22). For the current study, the dataset of a multi-center randomized controlled trial was used, that compared a phase-based treatment with a direct-trauma-focused treatment in patients with PTSD related to a history of childhood abuse [Van Vliet et al., 2017; (16)].

# 2. Method

## 2.1. Design and participants

A total of 121 patients were recruited in two mental health organizations; Dimence GGZ and GGZ Oost-Brabant. Included patients were aged between 18 and 65 years, and diagnosed with PTSD based on the Clinician-Administered PTSD scale for DSM-5 [CAPS-5; (23)]. Participants had to be a victim of repeated sexual and/or physical abuse before the age of 18 by a caretaker or a person in a position of authority, which was identified by the Life Events Checklist for the DSM-5 [LEC-5; (23)]. Patients were excluded when they did not master the Dutch language sufficiently, or in case of acute suicidality for which direct crisis intervention was needed, as assessed by an item of the Beck Depression Inventory-II [BDI; (24)]. In addition, patients were excluded if they had received at least eight sessions of any well-evaluated treatment for PTSD in the past year, reported being a victim of ongoing physical and/or sexual abuse, reported severe use of alcohol or drugs, or had an intellectual disability. The study design was registered in a national trial registry[2] and approved by the medical ethics committee Twente NL.[3] Further details on the trial methodology and patient sample can be found in the study protocol (25) and main outcome paper of the study (16).

## 2.2. Interventions

After patients were found to be eligible for participation in the study, they were randomized to two treatment conditions: both contained 16 sessions of EMDR (Eye Movement Desensitization and Reprocessing) as the trauma-focused treatment, and in one condition these sixteen sessions of EMDR were preceded by eight sessions of the first phase treatment STAIR (in total 24 sessions). Both the STAIR and the EMDR therapy were delivered twice a week for 90 min. STAIR was conducted according to the program described by Cloitre and her colleagues (11). EMDR therapy was conducted according to the protocol by Shapiro using the Dutch translation of the treatment protocol (26). Before actual treatment, patients first received one 90 min session consisting of psycho-education in which a hierarchy of relevant traumatic experiences was determined. For a full description of the two treatments, see Van Vliet et al. (25).

## 2.3. Measures

### 2.3.1. Drop out

Patients were considered to have dropped out if they discontinued treatment prematurely after the first session, which included psychoeducation and case conceptualization just before the actual

2 NTR5991.

3 56641.044.16 CCMO.

treatment began. This applies to cases where the patient did not complete the total number of treatment sessions as per the study's requirements [see the study protocol; (25)], and failed to complete treatment for all the traumas that were selected for treatment during the first treatment session. The outcome variable was a dichotomous variable: dropout versus completer status.

## 2.3.2. Pre-treatment variables

Tables 1 and 2 show an overview of the pre-treatment variables and the differences between the two groups (dropout versus completer status). The reasons given by the patients for dropout are shown in Supplementary Table S1.

### 2.3.2.1. Demographic characteristics

The following demographic characteristics were determined in the study: Gender, level of education, employment status, marital status, and age.

### 2.3.2.2. PTSD variables

PTSD symptom severity at the start of the treatment was measured with the CAPS-5 (23). The CAPS-5 is a clinical interview that includes 20 items on a 5-point Likert scale, resulting in a total score of between 0 and 80. The CAPS-5 has good psychometric properties [(27); see for the Dutch version (28)]. The inter-rater reliability was assessed by calculating the interclass correlation coefficient, which was 0.999, which is an excellent score.

### 2.3.2.3. Suicidality

Item 9 of the BDI-II (24) was used to measure suicidality. The scale of this item ranges from 0 to 3, with 0 for the absence of suicidal thoughts, 1 for indicating the presence of suicidal thoughts, but no intention to carry them out, and, 2 for indicating suicidal thoughts accompanied by a clear intention to commit suicide. Patients were excluded when they scored a 3 on this scale, which said that they would commit suicide whenever they had the chance. In that case they were assigned to a direct crisis intervention. In the current analysis, suicidality was used as a dichotomous value: absence (score of 0) or presence of suicidality (score of 1 or 2).

### 2.3.2.4. Borderline personality disorder

The Structured Clinical Interview for DSM-IV Axis II personality disorders [SCID-II interview; (29, 30)] was used to determine the presence of a borderline personality disorder. The psychometric properties are fair to good for this instrument.

Self-Injury: The severity of self-injury was determined by items 97 and 98 of the Dutch version of the SCID-II interview (30), with 1 for absent, 2 for doubtful, and 3 for present.

### 2.3.2.5. Dissociative symptoms

The severity of dissociative symptoms was indexed using the Dissociative Experiences Scale (DES-II; (31); Cronbach's $\alpha = 0.93$ in the present study at baseline). This is a 28-item self-report questionnaire with scores ranging from 0 to 100 (32). The presence of a dissociative subtype of PTSD was determined using the CAPS-5 (23).

### 2.3.2.6. Complex PTSD

The presence and severity of Complex PTSD was measured by the Structured Interview for Disorders of Extreme Stress [SIDES; (33)], the 38-item version developed by Ford et al. (34). The SIDES has good psychometric properties as a dichotomous measure in determining whether Complex PTSD is present or not [SIDES Manual by (35)].

TABLE 1 Continuous baseline variables pre-treatment for completers and dropouts and comparison of the means.

| Variables pre-treatment | Completer ($n = 90$) Mean (*Sd*) | Dropout ($n = 21$) Mean (*Sd*) | *t* | df | *p* |
|---|---|---|---|---|---|
| Age | 37.91 (12.72) | 34.59 (10.57) | 1.11 | 109 | 0.27 |
| | 38.32 (9.27) | 39.57 (8.27) | | | |
| CAPS-5 | 31.28 (11.73) | 34.19 (14.31) | −0.57 | 109 | 0.57 |
| | 29.03 (9.02) | 31.19 (7.09) | | | |
| SIDES | 23.55 (14.82) | 26.06 (14.64) | −0.98 | 109 | 0.33 |
| | 112.79 (24.87) | 120.52 (22.15) | | | |
| PSS-SR | 3.98 (1.40) | 4.29 (1.07) | −1.03 | 109 | 0.31 |
| | 1.63 (0.57) | 1.71 (0.53) | | | |
| DES | 1.83 (0.72) | 2.09 (0.75) | −0.7 | 109 | 0.49 |
| | 37.91 (12.72) | 34.59 (10.57) | | | |
| DERS | 38.32 (9.27) | 39.57 (8.27) | −1.31 | 109 | 0.19 |
| | 31.28 (11.73) | 34.19 (14.31) | | | |
| PTCI Self-Esteem | 29.03 (9.02) | 31.19 (7.09) | −0.96 | 109 | 0.34 |
| | 23.55 (14.82) | 26.06 (14.64) | | | |
| IIP | 112.79 (24.87) | 120.52 (22.15) | −0.59 | 109 | 0.56 |
| | 3.98 (1.40) | 4.29 (1.07) | | | |
| BSI | 1.63 (0.57) | 1.71 (0.53) | −1.47 | 109 | 0.14 |
| | 1.83 (0.72) | 2.09 (0.75) | | | |

CAPS-5, Clinician Administered PTSD Scale for DSM-5, SIDES = Structured Interview for Disorders of Extreme Stress-Revised, PSS-SR, PTSD Symptoms Scale-self report, DES, Dissociative Experiences Scale, DERS, Difficulties in Emotion Regulation Scale, PTCI, Posttraumatic Cognitions Inventory, IIP, Inventory of Interpersonal Problems, BSI, Brief Symptom Inventory.

### 2.3.2.7. Interpersonal difficulties

Interpersonal difficulties were indexed using the Inventory of Interpersonal Problems [IIP; (36)]. The psychometric properties of the IIP are satisfying (37). The IIP contains 32 items that can be scored on a 5-point scale from 0 (not at all) to 4 (very strongly). The reliability at baseline in this study was high (Cronbach's $\alpha = 0.85$).

TABLE 2 Categorical baseline variables for completers and droupouts and comparison of the amounts.

| Variable | Completer ($n = 90$) | Dropout ($n = 21$) | $\chi^2$ / Fisher's exact | *df* | *p* |
|---|---|---|---|---|---|
| **Gender** | | | | | |
| Woman | 65 | 11 | 2.25 | 1 | 0.13 |
| Man | 25 | 10 | | | |
| **Education** | | | | | |
| low | 42 | 12 | 4.06 | 2 | 0.13 |
| middle | 33 | 9 | | | |
| high | 15 | 0 | | | |
| **Employment** | | | | | |
| unemployed | 54 | 14 | 0.33 | 2 | 0.85 |
| employed | 25 | 5 | | | |
| student | 11 | 2 | | | |
| Living together | 53 | 12 | 0.00 | 1 | 1.00 |
| Married | 40 | 8 | 0.08 | 1 | 0.78 |
| Sexual Abuse | 69 | 14 | 0.45 | 1 | 0.50 |
| Physical Abuse | 69 | 16 | 0.00 | 1 | 1.00 |
| Dissociative subtype | 31 | 6 | 0.66 | 1 | 0.80 |
| Complex PTSD | 22 | 10 | 3.40 | 1 | 0.07 |
| Borderline personality disorder | 16 | 5 | 0.11 | 1 | 0.74 |
| **Self-injury** | | | | | |
| no self-injury | 53 | 12 | 0.09 | 2 | 0.96 |
| doubtful | 14 | 3 | | | |
| self-injury | 23 | 6 | | | |
| Suicidality | | | 5.75 | 1 | 0.02* |
| no suicidality | 27 | 1 | | | |
| suicidality | 63 | 20 | | | |
| Psychiatric medication use | 48 | 13 | 0.22 | 1 | 0.64 |
| Benzodiazepine medication use at screening | 22 | 2 | 2.24 | 1 | 0.24 |

*$p < 0.05$.

#### 2.3.2.8. Emotion regulation

Difficulties in emotion regulation were measured with the Difficulties in Emotion Regulation Scale [DERS; (38)], a questionnaire that has been validated in clinical populations (38, 39) and nonclinical populations (38, 40). Each item of the DERS is rated on a 5-point scale. The reliability in this study at baseline was high (Cronbach's $\alpha = 0.92$).

#### 2.3.2.9. Problems In self-esteem

To index problems in self-esteem the self-esteem subscale of the Posttraumatic Cognitions Inventory [PTCI; (41)] was used. Items are scored on a Likert scale from 1 ("I totally disagree") to 7 ("I totally agree"), and psychometric properties for the Dutch version (42) are good. The PTCI score for self-esteem showed a high reliability at baseline in this study (Cronbach's $\alpha = 0.94$).

#### 2.3.2.10. General psychopathology

The Brief Symptom Inventory (43, 44) was used to measure the severity of general psychopathology symptoms. The severity of each item can be rated on a 5-point scale from 0 (not at all) to 4 (a lot). The Dutch version has good psychometric properties [(45); Cronbach's $\alpha = 0.95$ at baseline in the present study].

## 3. Statistical analysis

All analyses were carried out using RStudio (46). The R code used is available in the Supplementary Material.

### 3.1. Data pre-processing

Following Cohen et al. (47) individuals with less than 50% missing pre-treatment values were included in this study. As a result, ten individuals were excluded. For the remaining 111 participants, variables with missing data were imputed using a random forest imputation algorithm [R package 'MissForest', (48)]. The benefits of this approach are that no pre-processing is required and that it is robust for noisy data and multicollinearity, so that it can be applied to mixed data types (48, 49). The imputation method was verified by

TABLE 3 Variable transformation.

| Variable | Included | Reason excluded | Transformation |
|---|---|---|---|
| Sex | yes | | Centered (male: −0.5, female: 0.5) |
| Education | yes | | Centered (low: −0.5, middle: 0; high: 0.5) |
| Employment | yes | | Centered (unemployed: −0.5, student: 0; employed: 0.5) |
| Age | yes | | Transformed (lambda 0.4) |
| Married | yes | | Centered (not married: −0.5, married: 0.5) |
| Living together | no | Substantial overlap with variable Married | |
| LEC-5 Sexual Abuse | yes | | Centered (no sexual abuse: −0.5, sexual abuse: 0.5) |
| LEC-5 Physical Abuse | yes | | Centered (no physical abuse: −0.5, physical abuse: 0.5) |
| Dissociative subtype | yes | | Centered (no dissociative subtype: −0.5, dissociative subtype: 0.5) |
| BDI-II item 9 Suicidal thoughts | yes | | Centered (no suicidal thoughts: −0.5, suicidal thoughts: 0.5) |
| Complex posttraumatic stress disorder (PTSD)_ | yes | | Centered (no complex PTSD: −0.5, complex PTSD: 0.5) |
| SCID-II item 97 and 98 Borderline personality disorder (BPD) | yes | | Centered (no BPD: −0.5, BPD: 0.5) |
| Kind of Personality Disorder | no | Substantial overlap with variable Presence BPD | |
| Self-injury | yes | | Centered (no self-injury: −0.5, doubtful: 0; self-injury: 0.5) |
| Posttraumatic stress disorder (PTSD)_ | no | Near-zero variance | |
| Psychoactive medication use at screening | yes | | Centered (no Psychoactive medication: −0.5, Psychoactive medication: 0.5) |
| Benzodiazepine medication use at screening | yes | | Centered (no Benzodiazepine medication: −0.5, Benzodiazepine medication: 0.5) |
| CAPS-5 total score | yes | | |
| SIDES total score | yes | | |
| PSS-SR total score | yes | | |
| DES total score | yes | | Winsorized 2 high outliers; Transformed (lambda 0.4) |
| BSI total score | yes | | |
| PTCI total score | no | Subscale included | |
| PTCI Self-Esteem | yes | | Transformed (lambda 1.2) |
| DERS total score | yes | | Transformed (lambda 1.5) |
| IIP total score | yes | | |

Transformations were performed with the BoxCox method.
BDI-II, Beck Depression Inventory; CAPS-5, Clinician-Administered PTSD scale for DSM-5; SIDES, Structured Interview for Disorders of Extreme Stress; PSS-SR, PTSD Symptoms Scale-Self Report; DES-II, Dissociative Experiences Scale; BSI, Brief Symptom Inventory; PTCI, Posttraumatic Cognitions Inventory; DERS, Difficulties in Emotion Regulation Scale; IIP, Inventory of Interpersonal Problems; SCID-II, Structured Clinical Interview for DSM-IV Axis II; LEC-5, Life Events Checklist for DSM-5.

applying this method to the non-missing dataset with completely at random removed values. After imputing the missing values, the performance was assessed using the normalized root mean squared error (NRMSE) and the proportion of falsely classified (PFC), which is defined by comparing the complete values with the imputed values (48).

In case of highly correlated variables (cor. > 0.70), one of the variables was dropped to avoid redundancy and multicollinearity. The decision which variable to remove was made by the research team. Outliers were winsorized and continuous variables with a non-normal distribution were transformed using the Box-Cox method [(50); R package 'Caret'; (51)]. Finally, continuous variables were standardized and categorical variables were centered (see Table 3 for details about transformations for each variable). This data pre-processing procedure resulted in a dataset of 111 participants with 22 pre-treatment variables. Dropouts did not differ significantly between patients who received STAIR-EMDR (21.8%) and EMDR only therapy [16.1%; $\chi 2$ (1) = 6.00, $p = 0.440$]. A prognostic model for dropout was developed independently of treatment conditions because the subsample of dropouts was too small to create a prescriptive model for dropout depending on the treatment conditions.

## 3.2. Imbalanced dataset

After data pre-processing, the ratio between dropout and completers was checked, because it was expected that the dataset would be imbalanced due to a relatively smaller proportion of

dropouts compared to completers. This is of importance as an imbalanced dataset causes classification performance problems in machine learning algorithms (52). To deal with this class imbalance, the proven effective synthetic minority oversampling technique (SMOTE) was applied (53). This method over-samples the minority class (i.e., dropout), by artificially creating new samples using the nearest neighbours of the cases, and under-samples the majority class (53).

### 3.3. Model building

For the dropout prediction, a model was developed using Elastic Net Regularization [ENR, R package 'glmnet', (54)]. ENR is a combination of ridge regression and lasso regression where alpha is the tuning parameter between ridge regression (alpha = 0) and lasso [alpha = 1; (55)]. Another tuning parameter is the lambda which determines the shrinkage or penalty of the coefficients, the larger its value the stronger the shrinkage (55). The alpha and lambda were determined using 10-fold cross-validation, where the selected alpha was based on the highest area under the curve [AUC; (56)]. After determination of the alpha and lambda, the final model was built.

### 3.4. Model evaluation

The model was evaluated on the initial (imbalanced) dataset using three different measures: accuracy, the area under the curve (AUC), and the Brier score. The selected variables in the final model were evaluated for significance and variable importance. The variable importance was calculated using the vip package (57). The accuracy is the percentage of correct dropout predictions ranging from 0% (worst prediction) to 100% (best prediction). The AUC refers to the overall performance of a classifier. When the AUC is 1, predictions are 100% correct, and when the AUC is 0, all predications are incorrect (56). The Brier score can be described as a parameter that measures the accuracy of probabilistic predictions ranging from 0 (best prediction) to 1 [worst prediction; (58)].

## 4. Results

### 4.1. Data pre-processing

The total sample consisted of 111 patients. From that number, 80 individuals (72%) had no missing data and the percentage of missing baseline variables was 2.5 percent. All patients who dropped out prematurely and were assessed after dropping out still fulfilled the diagnostic criteria for PTSD. The number of sessions before dropping out was registered and the median of the number of sessions before dropping out was 4 sessions. The missing values were imputed and the performance of the imputation method was acceptable (NRMSE = 0.15; PFC 0.26).

### 4.2. Imbalanced dataset and dropout rate

The dropout rate in this study was 19%, indicating that the distribution between dropout and completers was unequal, leading to an imbalanced dataset. To balance the dataset, SMOTE was applied. After applying SMOTE, the dropout rate was 43%.

TABLE 4 Variable importance indicators.

| Variable | Importance |
|---|---|
| Suicidality | 1.22 |
| Benzodiazepine medication use at screening | 1.14 |
| Education | 0.97 |
| Complex PTSD | 0.89 |
| Gender | 0.68 |
| SIDES | 0.37 |
| DERS | 0.35 |
| Sexual Abuse | 0.34 |
| PTCI Self-Esteem | 0.32 |
| Borderline personality disorder | 0.26 |
| BSI | 0.24 |
| DES | 0.10 |
| CAPS-5 | 0.07 |

### 4.3. Model building

The final ENR model selected the following variables for dropout risk: suicidality, benzodiazepine medication use at screening, education, gender, sexual abuse, Borderline personality disorder, PTCI Self-Esteem, SIDES, DERS, complex PTSD, BSI, DES, and CAPS-5 scores.

### 4.4. Predictor evaluation

The best ENR model (alpha = 0.25, lambda = 0.18) correctly predicted 81.6% of all individuals (accuracy). For this model, the AUC was 0.85 and the Brier score was 0.31, indicating sufficient predictions. The sensitivity and specificity are 75 and 87%, respectively. Predictors were evaluated based on both significance and variable importance. Based on significance (i.e., value of p lower than 0.05), the main variables for risk of dropout in the model were gender, education, suicidality, score on the DERS, score on the BSI and benzodiazepine medication use at screening. The (i.e., an importance score greater than or close to 1), the main variables of the model were suicidality, education, and benzodiazepine use at screening (for a total overview of the variable importance, see Table 4). More specifically, low education level, presence of suicidal thoughts, and not using benzodiazepine medication during screening were most strongly related to dropout risk according to both significance and variable importance. Given our focus on prediction, we will further focus on these three variables with the highest variable importance score and not on all significant variables.

## 5. Discussion

The purpose of the present study was to identify patients with PTSD related to childhood abuse who are at risk of dropping out of

treatment. Therefore, we applied machine learning techniques to analyze all available pre-treatment variables from a recent RCT. The best model was able to correctly predict dropout in 81.6% of the individuals. We consider our research endeavor to be a promising approach for identifying individuals at risk of dropout and, to this end, could be considered a first step in developing models for predicting patient groups that are at an elevated risk of dropout. This makes it possible to develop interventions that support these patients during treatment and prevent them from dropping out; for example, extra nursing support or treatment in clinical settings.

Our model evaluation indicated that the within-sample predictions were accurate based on three different measures (accuracy, AUC, and Brier score). To this end, the combination of pretreatment suicidality, low education, and non-use of benzodiazepines was found to be a risk factor for dropout. Interestingly, from their meta-analysis Guina et al. (59) concluded that the use of benzodiazepines is contraindicated for the treatment of PTSD. They found that the use of benzodiazepines is associated with poor treatment outcomes. To this end, the lack of experienced treatment benefits and the inability to use benzodiazepines as missed opportunities to ease and regulate feelings of tension may explain why patients discontinued therapy. Also, regarding the other predictors, it is conceivable that this could have prompt the patient to discontinue treatment early. This holds true, for instance, if one conceives suicidality as an indicator of feelings of entrapment or hopelessness, and a low level of education as a possible indicator of higher trait impulsivity or low frustration tolerance. However, before speculating too much on how these factors together might explain the increased risk of treatment dropout, more research is needed to first try and replicate these results and rule out the possibility of chance. Our method is the first step toward developing personalized medicine to prevent dropouts from PTSD treatment. Our findings further suggest that one of the ways to reduce dropout rates may be to intensify trauma treatments, as most patients in this study dropped out in an early stage (between the third and fourth treatment sessions) and other studies found that a more intensive PTSD treatment can lead to a lower dropout rate [for example, (60, 61)]. Clearly, replication studies are needed to provide reliable evidence for out-of-sample prediction accuracy for use in clinical practice.

The first strength of this study is that, to our knowledge, this is the first study to analyze predictors of dropout among patients with PTSD related to childhood abuse using machine learning, which has the advantage that it takes into account all possible pre-treatment variables. A second strength is that we handled an imbalanced dataset by applying the proven effective synthetic minority oversampling technique [SMOTE; (53)], which provides more opportunities to compare dropouts with completers, as in most studies, there is a large difference between the sample sizes of both groups. As in any study, some limitations of this study need to be noted. The first limitation is that the study sample was quite small, thereby limiting the possibility of distinguishing between the two treatment conditions (62). Future studies should focus on examining patient characteristics that might predict dropout by machine learning, using larger patient samples and including more types of trauma-focused therapies. A second limitation is the missing data for which we had to impute the data, leading to less reliable outcomes. However, the performance of our imputation method proved reliable (53), which was verified in this study. Third, we were not able to externally validate the data on an independent dataset to answer the question of whether this model leads to significant effects in new samples (63, 64). Before these outcomes are confirmed by external validation, the results cannot be generalized to clinical practice. Although we attempted to overcome this using 10-fold cross-validation, it is unknown how this model will perform on a truly independent dataset. For example, Isaksson et al. (65) argued that cross-validation was unreliable for classifying small samples.

In conclusion, this study identified a combination of variables predicting the dropout rate of patients with PTSD due to childhood abuse in trauma-focused treatments. A challenging task for future research is to examine whether these results can be replicated in larger patient samples. Another challenge is to examine these potential dropout predictors in a more profound way; although our study may help identify patients who are at risk of dropping out of therapy, the results do not reveal the mechanisms that explain the elevated risk. Experimental studies are required to elucidate the exact mechanisms involved, which could be fundamental for future preventive interventions. Developments in the field of machine learning are moving rapidly, and for follow-up research, it may be interesting to look at other models that use tidy modeling.

## 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 Medical Ethics Committee Twente. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

SBre, NV, SBro, RH, AJ, and MD have made substantial contributions to the design of this study. NV was responsible for the implementation of this study and for the inclusion process and the data collection. AJ supervised the therapists involved in the study. SBre performed the analyses under supervision of SBro. SBre and NV drafted the paper under supervision of SBro, MD, RH, and AJ. All authors contributed to the article and approved the submitted version.

## Funding

Funding for this study was provided by Stichting tot Steun VCVGZ, Dimence Mental Health Care, and the Dutch EMDR association.

## Conflict of interest

AJ receives income from published books on EMDR therapy and training of postdoctoral professionals in this method.

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.2023.1194669/full#supplementary-material

## References


1. Alonso J, Angermeyer MC, Bernert S, Bruffaerts R, Brugha TS, Bryson H, et al. Disability and quality of life impact of mental disorders in Europe: results from the European study of the epidemiology of mental disorders (ESEMeD) project. *Acta Psychiatr Scand*. (2004) 420:38–46. doi: 10.1111/j.1600-0047.2004.00329.x

2. Kessler RC. Posttraumatic stress disorder: the burden to the individual and to society. *J Clin Psychiatry*. (2000) 61:4–12; discussion 13-4.

3. Panagioti M, Gooding PA, Tarrier N. A meta-analysis of the association between posttraumatic stress disorder and suicidality: the role of comorbid depression. *Compr Psychiatry*. (2012) 53:915–30. doi: 10.1016/j.comppsych.2012.02.009

4. Sareen J, Cox BJ, Stein MB, Afifi TO, Fleet C, Asmundson GJ. Physical and mental comorbidity, disability, and suicidal behavior associated with posttraumatic stress disorder in a large community sample. *Psychosom Med*. (2007) 69:242–8. doi: 10.1097/PSY.0b013e31803146d8

5. Mavranezouli I, Megnin-Viggers O, Daly C, Dias S, Welton N, Stockton S, et al. Psychological treatments for post-traumatic stress disorder in adults: a network meta-analysis. *Psychol Med*. (2020) 50:542–55. doi: 10.1017/S0033291720000070

6. Varker T, Jones KA, Arjmand HA, Hinton M, Hiles SA, Freijah I, et al. Dropout from guideline-recommended psychological treatments for posttraumatic stress disorder: a systematic review and meta-analysis. *J Affective Disorders Reports*. (2021) 4:100093–3. doi: 10.1016/j.jadr.2021.100093

7. Usman M, Rehman A, Bakhtawar N, Bhatti ABH. Prognosis of PTSD in treated vs. *Non-Treated Groups J Pakistan Psychiatric Soc*. (2015) 12:39.

8. Rink J, Lipinska G. Evidence of distinct profiles of ICD-11 post-traumatic stress disorder (PTSD) and complex PTSD in a south African sample. *Eur J Psychotraumatol*. (2020) 11:1818965. doi: 10.1080/20008198.2020.1818965

9. World Health Organization. *The ICD-11 for mortality and morbidity statistics. World Health Organization. Guidelines for the management of conditions specifically related to stress*. Geneva: World Health Organization (2018).

10. Cloitre M., Courtois C. A., Ford J. D., Green B. L., Alexander P., Van der Hart O., et al. (2012). The ISTSS expert consensus treatment guidelines for complex PTSD in adults. Available at: https://terrorvictimresponse.ca/wp-content/uploads/ISTSS-Expert-Concesnsus-Guidelines-for-Complex-PTSD-Updated-060315.pdf

11. Cloitre M, Koenen KC, Cohen LR, Han H. Skills training in affective and interpersonal regulation followed by exposure: a phase-based treatment for PTSD related to childhood abuse. *J Consult Clin Psychol*. (2002) 70:1067–74. doi: 10.1037/0022-006X.70.5.1067

12. Karatzias T, Murphy P, Cloitre M, Bisson J, Roberts N, Shevlin M, et al. Psychological interventions for ICD-11 complex PTSD symptoms: systematic review and meta-analysis. *Psychol Med*. (2019) 49:1761–75. doi: 10.1017/S0033291719000436

13. Cloitre M, Stovall-McClough KC, Nooner K, Zorbas P, Cherry S, Jackson CL, et al. Treatment for PTSD related to childhood abuse: a randomized controlled trial. *Am J Psychiatr*. (2010) 167:915–24. Available from: 10.1176/appi.ajp.2010.09081247. doi: 10.1176/appi.ajp.2010.09081247

14. Oprel DAC, Hoeboer CM, Schoorl M, De Kleine RA, Cloitre M, Wigard IG, et al. Effect of prolonged exposure, intensified prolonged exposure and STAIR+prolonged exposure in patients with PTSD related to childhood abuse: a randomized controlled trial. *Eur J Psychotraumatol*. (2021) 12:1851511. doi: 10.1080/20008198.2020.1851511

15. Raabe S, Ehring T, Marquenie L, Arntz A, Kindt M. Imagery Rescripting versus STAIR plus imagery Rescripting for PTSD related to childhood abuse: a randomized controlled trial. *J Behav Ther Exp Psychiatry*. (2022) 77:101769. doi: 10.1016/j.jbtep.2022.101769

16. Van Vliet NI, Huntjens RJC, Van Dijk MK, Bachrach N, Meewisse ML, De Jongh A. Phase-based treatment versus immediate trauma-focused treatment for post-traumatic stress disorder due to childhood abuse: randomised clinical trial. *British J Psychiatry Open*. (2021) 7:E211. doi: 10.1192/bjo.2021.1057

17. Bisson JI, Ehlers A, Matthews R, Pilling S, Richards D, Turner S. Psychological treatments for chronic post-traumatic stress disorder. Systematic review and meta-analysis. *Br J Psychiatry J Ment Sci*. (2007) 190:97–104. doi: 10.1192/bjp.bp.106.021402

18. Bradley R, Greene J, Russ E, Dutra L, Westen D. A multidimensional meta-analysis of psychotherapy for PTSD. *Am J Psychiatry*. (2005) 162:214–27. doi: 10.1176/appi.ajp.162.2.214

19. Hembree EA, Foa EB, Dorfan NM, Street GP, Kowalski J, Tu X. Do patients drop out prematurely from exposure therapy for PTSD? *J Trauma Stress*. (2003) 16:555–62. doi: 10.1023/B:JOTS.0000004078.93012.7d

20. Imel ZE, Laska KM, Jakupcak M, Simpson TL. Meta-analysis of dropout in treatments for posttraumatic stress disorder. *J Consult Clin Psychol*. (2013) 81:394–404. doi: 10.1037/a0031474

21. Lewis C, Roberts NP, Gibson S, Bisson JI. Dropout from psychological therapies for post-traumatic stress disorder (PTSD) in adults: systematic review and meta-analysis. *Eur J Psychotraumatol*. (2020) 11:1709709. doi: 10.1080/20008198.2019.1709709

22. 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: official J World Psychiatric Association (WPA)*. (2021) 20:154–70. doi: 10.1002/wps.20882

23. Weathers F. W., Blake D. D., Schnurr P. P., Kaloupek D. G., Marx B. P., Keane T. M. (2013). *The life events checklist for DSM-5 (LEC-5). National Center for posttraumatic stress disorder*. Available online at: https://www.ptsd.va.gov/

24. Beck AT, Steer RA, Brown GK. *Manual for the Beck depression inventory II*. San Antonio, TX: Psychological corporation (1996).

25. Van Vliet NI, Huntjens RJC, Van Dijk MK, de Jongh A. Phase-based treatment versus immediate trauma-focused treatment in patients with childhood trauma-related posttraumatic stress disorder: study protocol for a randomized controlled trial. *Trials*. (2018) 19:1–10. doi: 10.1186/s13063-018-2508-8

26. De Jongh A, Ten Broeke E. *Handboek EMDR, een geprotocolleerde behandelmethode voor de gevolgen van psychotrauma [EMDR manual: A Protocolized treatment method for the consequences of Psychotrauma]. 6th* ed. Amsterdam: Benelux Pearson Education BV (2013).

27. Weathers FW, Bovin MJ, Lee DJ, Sloan DM, Schnurr PP, Kaloupek DG, et al. The clinician-administered PTSD scale for DSM-5 (CAPS-5): development and initial psychometric evaluation in military veterans. *Psychol Assess*. (2017) 30:383–95. doi: 10.1037/pas0000486

28. Boeschoten MA, Van der Aa N, Bakker A, Ter Heide FJJ, Hoofwijk MC, Jongedijk RA, et al. Development and evaluation of the Dutch clinician-administered PTSD scale for DSM-5 (CAPS-5). *Eur J Psychotraumatol*. (2018) 9:1546085. doi: 10.1080/20008198.2018.1546085

29. First MB, Spitzer RL, Gibbon M, Williams JBW, Benjamin LS. *Structured clinical interview for DSM-IV Axis II personality disorders (SCID-II)*. Washington, DC: American Psychiatric Press (1997).

30. Weertman A, Arntz A, Dreessen L, Van Velzen C, Vertommen S. Short-interval test-retest interrater reliability of the Dutch version of the structured clinical interview for DSM-IV personality disorders (SCID-II). *J Personal Disord*. (2003) 17:562–7. doi: 10.1521/pedi.17.6.562.25359

31. Carlson EB, Putnam FW. An update on the dissociative experiences scale. *Dissociation*. (1993) 6:16–27.

32. Van IJzendoorn MH, Schuengel C. The measurement of dissociation in normal and clinical populations: meta-analytic validation of the dissociative experiences scale (DES). *Clin Psychol Rev*. (1996) 16:365–82. doi: 10.1016/0272-7358(96)00006-2

33. Scoboria A, Ford J, Lin H, Frisman L. Revision of the structured interview for disorders of extreme stress (SIDES): an exploratory and confirmatory factor analytic approach. *Assessment*. (2008) 15:404–25. doi: 10.1177/1073191108319005

34. Ford J, Stockton P, Kaltman S, Green BL. Disorders of extreme stress (DESNOS) symptoms are associated with type and severity of interpersonal trauma exposure in a sample of healthy young women. *J Interpers Violence*. (2006) 21:1399–416. doi: 10.1177/0886260506292992

35. Spinazzola J. (2019). Structured interview for disorders of extreme stress (SIDES) & self-report inventory for disorders of extreme stress (SIDES-SR). Unpublished

manuscript Available at https://complextrauma.org/wp-content/uploads/2019/03/SIDES-Manual-Spinazzola-2019.pdf

36. Horowitz LM, Alden LE, Wiggins JS, Pincus AL. *Inventory of interpersonal problems*. London: Psychological Corporation (2000).

37. Barkham M, Hardy GE, Startup M. The IIP-32: a short version of the inventory of interpersonal problems. *Br J Clin Psychol*. (1996) 35:21–35. doi: 10.1111/j.2044-8260.1996.tb01159.x

38. Gratz KL, Roemer L. Multidimensional assessment of emotion regulation and dysregulation: development, factor structure, and initial validation of the difficulties in emotion regulation scale. *J Psychopathol Behav Assessment*. (2004) 26:41–54. doi: 10.1023/B:JOBA.0000007455.08539.94

39. Fox HC, Axelrod SR, Paliwal P, Sleeper J, Sinha R. Difficulties in emotion regulation and impulse control during cocaine abstinence. *Drug Alcohol Depend*. (2007) 89:298–301.

40. Johnson KA, Zvolensky MJ, Marshall EC, Gonzalez A, Abrams K, Vujanovic AA. Linkages between cigarette smoking outcome expectancies and negative emotional vulnerability. *Addict Behav*. (2008) 33:1416–24. doi: 10.1016/j.addbeh.2008.05.001

41. Foa EB, Ehlers A, Clark DM, Tolin DF, Orsillo SM. The posttraumatic cognitions inventory (PTCI). *Dev validation Psychol Assessment*. (1999) 11:303–14. doi: 10.1037/1040-3590.11.3.303

42. Van Emmerik AAP, Schoorl M, Emmelkamp PMG, Kamphuis JH. Psychometrische kenmerken van de Nederlandstalige Posttraumatische Cognities Inventarisatielijst (PTCI). *Gedragstherapie [Behav Therapy]*. (2007) 40:269–84.

43. De Beurs E. *Brief symptom inventory*. Handleiding. Leiden: PITS (2006).

44. Derogatis LR. Brief symptom inventory. Baltimore, MD: Clinical (1975).

45. De Beurs E, Zitman FG. De brief symptom inventory (BSI): De betrouwbaarheid en validiteit van een handzaam alternatief voor de SCL-90 [reliability and validity of a handy alternative for the SCL-90]. *Maandblad Geestelijke Volksgezondheid [monthly magazine for mental health]*. (2006) 61:120–41.

46. RStudio Team. *RStudio: integrated development for R*. Oxford: Oxford University Press (2020).

47. Cohen ZD, Kim TT, Van HL, Dekker JJ, Driessen E. A demonstration of a multi-method variable selection approach for treatment selection: recommending cognitive–behavioural versus psychodynamic therapy for mild to moderate adult depression. *Psychother Res*. (2019) 1–14:1563312. doi: 10.1080/10503307.2018.1563312

48. Stekhoven DJ, Bühlmann P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. (2012) 28:112–8. doi: 10.1093/bioinformatics/btr597

49. Tang F, Ishwaran H. Random Forest missing data algorithms. *Stat Analysis Data Mining*. (2017) 10:363–77. doi: 10.1002/sam.11348

50. Box G, Cox D. An analysis of transformations. *J R Statistical Soc Series B (Methodological)*. (1964) 26: 211–52. doi: 10.1111/j.2517-6161.1964.tb00553.x

51. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. (2008) 28:1–26. doi: 10.18637/jss.v028.i05

52. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. *Intelligent Data Analysis*. (2002) 6:429–49. doi: 10.3233/IDA-2002-6504

53. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. (2002) 16:321–57. doi: 10.1613/jair.953

54. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. (2010) 33:1.

55. Kuhn M, Johnson K. *Applied predictive modeling*. New York: Springer (2013).

56. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning with applications in R*. New York: Springer (2003).

57. Greenwell BM, Boehmke BC. Variable importance plots—an introduction to the vip package. *The R J*. (2020) 12:343–66. doi: 10.32614/RJ-2020-013

58. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. (1950) 78:1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

59. Guina J, Rossetter SR, DeRhodes BJ, Nahhas RW, Welton RS. Benzodiazepines for PTSD: a systematic review and meta-analysis. *J Psychiatr Pract*. (2015) 21:281–303. doi: 10.1097/PRA.0000000000000091

60. Hoppen TH, Kip A, Morina N. Are psychological interventions for adult PTSD more efficacious and acceptable when treatment is delivered in higher frequency? A meta-analysis of randomized controlled trials. *J Anxiety Disord*. (2023) 95:102684. doi: 10.1016/j.janxdis.2023.102684

61. Van Woudenberg C, Voorendonk EM, Bongaerts H, Zoet HA, Verhagen M, Lee CW, et al. Effectiveness of an intensive treatment programme combining prolonged exposure and eye movement desensitization and reprocessing for severe post-traumatic stress disorder. *Eur J Psychotraumatol*. (2018) 9:1487225. doi: 10.1080/20008198.2018.1487225

62. Luedtke A, Sadikova E, Kessler RC. Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clin Psychol Sci*. (2019) 7:445–61. doi: 10.1177/2167702618815466

63. Bleeker S, Moll H, Steyerberg E, Donders A, Derksen-Lubsen G, Grobbee D, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol*. (2003) 56:826–32. doi: 10.1016/S0895-4356(03)00207-5

64. Van Bronswijk SC, Bruijniks SJ, Lorenzo-Luaces L, Derubeis RJ, Lemmens LH, Peeters FP, et al. Cross-trial prediction in psychotherapy: external validation of the personalized advantage index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression. *Psychother Res*. (2021) 31:78–91. doi: 10.1080/10503307.2020.1823029

65. Isaksson A, Wallman M, Göransson H, Gustafsson MG. Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recogn Lett*. (2008) 29:1960–5. doi: 10.1016/j.patrec.2008.06.018

66. Foa EB, Riggs DS, Dancu CV, Rothbaum BO. Reliability and validity of a brief instrument for assessing posttraumatic stress disorder. *J Trauma Stress*. (1993) 6:459–73. doi: 10.1002/jts.2490060405

frontiers | Frontiers in Public Health

TYPE Systematic Review
PUBLISHED 18 September 2023
DOI 10.3389/fpubh.2023.1142443



OPEN ACCESS

EDITED BY
Ajaya Bhattarai,
Tribhuvan University, Nepal

REVIEWED BY
Victoria Ramos Gonzalez,
Carlos III Health Institute (ISCIII), Spain
Graciela Rojas,
University of Chile, Chile

*CORRESPONDENCE
Hatem H. Alsaqqa
✉ hs-mch@hotmail.com

RECEIVED 11 January 2023
ACCEPTED 04 September 2023
PUBLISHED 18 September 2023

CITATION
Alsaqqa HH and Alwawi A (2023) Digital intervention for public health: searching for implementing characteristics, concepts and recommendations: scoping review. *Front. Public Health* 11:1142443. doi: 10.3389/fpubh.2023.1142443



# Digital intervention for public health: searching for implementing characteristics, concepts and recommendations: scoping review

Hatem H. Alsaqqa[1,2]* and Abdallah Alwawi[3]

[1]Deanship of Scientific Research, Al-Quds University, Jerusalem, Palestine, [2]Ministry of Health, Gaza Strip, Palestine, [3]Anesthesia and Resuscitation Technology, Health Professions Faculty, Al Quds University, Jerusalem, Palestine

Studying the impact of digital interventions on public health can help ensure that the offered services produce the desired results. In order to address these factors, the subsequent study uses a scope review to evaluate the state of the field while concentrating on ideas and suggestions that represent factors that have been crucial in the management of digital intervention for public health. To shed light on the traits, ideas and suggestions related to public health digital intervention, a scoping review was carried out. Five electronic databases were used to locate pertinent research that were published before February 2022. All texts were examined, and study abstracts were scrutinized to determine their eligibility. The last analysis of this study included fifteen publications; five reviews, four qualitative studies, two quantitative studies, one viewpoint study, one mixed-method study, one perspective study, and one interventional study. The key ideas for digital interventions in population management and health studies are presented in this overview. Many concepts, implementation characteristics and recommendations have been raised which highlight the future role of these interventions to enhance public engagement and health equity.

KEYWORDS

**characteristics, concepts, digital interventions, implementation, public health**

## Introduction

The choice of an appropriate theory to administrator the implementation course and the technique selection, assuring that proper consideration is paid to planning implementation, and having a flexible tactic that allows for response to recently evolving obstacles are all of the utmost importance.

Public health professionals and academics have proposed a variety of responses in their search for solutions, including stepping up current initiatives to promote information and health literacy, coming up with plans for widely refuting distortion, and educating clinicians and public health professionals on how to discourse misinformation one-on-one (1).

Better data science ought to lead to healthier behaviors and wiser health decisions. In this way, technologically mediated health data processing might support patient empowerment and individual sovereignty (2, 3). However, because human decision-making is complex and influenced by environment and intellectual biases, mixing emotion and rational, the embracing of healthy habits does not occur linearly as a result of better health knowledge. In addition to being

the area of health care that is focused with promoting healthy behavior, health promotion is viewed as crucial to attempts to avoid diseases (4).

However, methods from all these fields are needed since these studies sit at the nexus of biological, behavioral, computational, and engineering research. Related research answers encompass identifying the issue and the expected assistance of the digital health intervention (DHI), which in flip necessitates determining the intervention's likely reach and uptake, the causal model outlining how the intervention will produce the optimally selected, key elements and how they socialize with one another, and assessing the actual advantage in terms of effectiveness, cost effectiveness, and harms (5).

Distinguishing the implication of digital technology in this expanse and in pandemic preparation planning has become crucial since the future of public health is anticipated to be more and more digital. Technology productions and other noteworthy players in the digital area ought to work together as longstanding allies in readiness rather than only through crises times (6).

DHIs, which are therapies given through digital technology like smartphones or websites, have a huge potential to provide efficient, affordable, safe, and scaled interventions to promote healthcare. DHIs can be used to optimize outcomes for those with longstanding conditions, such as cardiovascular disease, diabetes, and mental health issues, as well as to provide remote access to effective treatments. They are frequently intricate interventions with numerous parts, and many of them have multiple goals, such as empowering users to learn more about their health, connect with others in a similar situation, alter perceptions and beliefs about it, monitor certain health conditions or behaviors, titrate prescription, identify treatment priorities, and enhance patient-provider interactions (5).

This review aims to examine the series of digital inventions used universally to address public health challenges, as well as their restrictions and implementation footraces, such as those associated to the law, ethics, and privacy as well as legislative and personnel issues. The objective of the paper is to identify mechanism-based explanations for how and in what contexts digital intervention for public health achieved its effects.

## Appraising digital health interventions

Integrating assessment from the start of the DHI progress process allows for the development of evaluation thinking, abilities, and tools. The resulting evaluation service gives non-academics and digital developers the ability to use evaluation approaches and thinking when designing, developing, and implementing their DHI. By doing this, it demystifies evaluation, which has historically been the purview of academia, and uses people's motivations to make sure that their DHI is as effective as it can be while enhancing the health and wellbeing of end users (7).

Diagnostic or population health interventions, digital product design, product and service design, as well as communication and health promotion, are all components of the interdisciplinary endeavor known as DHIs. Therefore, interdisciplinary approaches to evaluation are best for understanding the effectiveness of DHIs as well as their usableness and attractiveness, with success criteria that consider the various parties involved in the hiring, design, and growth of a DHI as well as its end users (8). To validate the appraisal enterprise path for a DHI, seven key ideas for evaluating DHIs have been identified; evaluation thought, review image, contract assistant, testing tools, progress history, data hub, and published health results.

## Argue the structural and epistemic aspects

Along with the crucial problems of safety, data privacy, and the value of human caring touch, structural unfairness raises concerns. DHIs for "reporting and evidence building" urge users to actively recount their involvements and join with other survivors' tales, creating a shared epistemic space for people who have come into contact with violence. Users are encouraged to apply their epistemic capacity, are recognized as epistemic subjects, and are able to communicate and possibly advance their knowledge within the user community by allowing consumers to express their opinions, even if only digitally through a digital application.

These DHIs can thus stress the need to respond not only on an interpersonal basis but also on a structural level and may aid in better comprehending patterns and clusters of violence (e.g., societies, regimes). These DHIs might also make people impress less alone in their involvements, teach them coping mechanisms, and help them locate guidance and support. Even if applications are created in a way that considers and reflects systemic factors of violence, their impact would be dubious if not everyone can use them. By endorsing digital intervention tactics that can only be opened by users with specific advantages, this runs the risk of highlighting and strengthening structural and epistemic unfairness.

The main epistemic circumstances and traits are recognized under the categorization: data and information structures are related to psychological effects in four ways: (1) they are caused by psychological properties, (2) they are caused by information features, (3) they are related to psychological properties, and (4) information features (co-) constitute psychological properties (9).

Additionally, as DHIs become more prevalent as an intervention strategy, problems like the loss of private contact in intervention settings (such as social workforces providing resident counseling) as a result of a change to digital technology, the possibilities for mistranslation of the details given owed to the absence of non-verbal cues, as well as matters with language and comprehension and access to technology, may crop up (10).

Meanwhile, social media usage for purposes related to health has the ability to create interpersonal networks that support specific epistemic positions on medical matters, which could have a negative impact on public health. Another illustration is the impact of technology-mediated interaction on the connection between the patient and the healthcare professional (11).

## Personal agency and motivation

As patients and the general public tended to engross with and enroll in DHIs because they wanted to be healthy or have more influence over how they managed their welfare, the first topic that arose was personal agency and motivation. Information technology was believed that using technology could help people stay motivated to engage in physical activity, reduce their weight, and stay healthy (12,

Abbreviations: DHI, Digital Health Intervention; DHIs, Digital Health Interventions; PPI, Patient and Public Involvement.

13). As a result of having the freedom to obtain health information whenever and wherever they pleased, many people joined a DHI, which in some circumstances helped lower anxiety (14, 15). The level of regulating knowledge provided for tracking and comprehending health-related behaviors, such as food and exercise, or for managing chronic diseases on one's own, was also well-liked by users, which prompted registration (12).

### Personal life and values

The recurring topic was how patients' and the general public's capacity to participate in and participate in DHIs was impacted by a busy personal life with many conflicting demands. People tended to sign up for new technology if they felt it was useful, could be customized to meet their needs, and was simple to integrate into their daily lives (15, 16). Additionally, individuals who were digitally literate (14, 16) and had experience with or were already familiar with utilizing technology (14) found it simpler to enroll since they possessed the necessary knowledge and abilities. Some people registered because they valued the privacy that online health services offered, being secure and protected from the discrimination and disgrace they occasionally encountered in the actual life (12, 16).

## Perceived fit perceived

In contrast to a one-size-fits-all approach, perceived fit describes how much users handled the intervention was acceptable, applicable to their culture and values, and/or oriented at others who were similar to them. For instance, the information's applicability to their current circumstance (17, 18) and the ability to adapt or tailor the intervention (19, 20) made it more likely that it would fit. Users' ability to relate to the intervention's presenters, who may be coaches, teachers, or samples of people with comparable situations, was a facilitating element (21). Culturally appropriate material (22), level of literacy (23), and content given with little use of technical terminology (24) are all examples of elements that make the information pertinent and in a vocabulary suited for the user.

## Perceived usefulness

The term "perceived usefulness" describes how a user feels about an intervention and how they judge if it will be helpful to them. Users' ability to comprehend the information presented to them (104,117,170), the clarity of the action they should take (17, 25), and the perception that the intervention offered a distinct advantage over previous or ongoing care received (17, 26) all contributed to this perception. Facilitators were identified as making it simpler for users to get services they would not otherwise have access to (26, 27) and removing the need for them to travel far to a health center (28).

## Level of guidance

The amount of assistance a user receives when using an intervention—for instance, through reminders or a web-based supporter—determines how much accountability they receive to frequently engage with the information. If the intervention raised the level of control, leading to users perceived more responsibility over their own health, it would be a facilitating factor for utilizing DHIs (29, 30). Participants had trouble interacting with interventions that were entirely self-guided, and they occasionally failed to use the intervention (31). The demand for more structured use was voiced by the participants, for examples of this structured use include app alerts or routine human coaching checks (32, 33).

## Social connectedness

User engagement was revealed to be facilitated by an intervention's impact on participants' feelings of social connectivity. Another facilitating element was if the intervention (34, 35) helped to mainstream lived perspectives by giving instances of others who had comparable experiences. Additional beneficial outcomes that could promote engagement included enhanced abilities (36, 37), a greater understanding of users' health (38, 39), and a sense of empowerment over having control over their wellness (40).

### Iterative methods to adjust an intervention

Regular stakeholder involvement, new scientific information, evolving government directives, quick qualitative research (telephone interviews and open-text questionnaires), and usage data analysis all influenced the optimizations. All comments were quickly compiled, and potential improvements were prioritized according to their likelihood of having an influence on behavior change (41).

In order to improve a health intervention and/or its execution to achieve stakeholder-specified public health benefits within supply limits, optimization can be defined as a purposeful, iterative, and data-driven procedure. This study sought to characterize the core ideas, procedures, or processes of selected frameworks to maximize the efficacy of health interventions and/or their administration (42).

The optimization step's goal was to test and perfect the intervention's reasoning and program theory in order to comprehend intervention mechanisms and increase its effectiveness. This frequently happened by doing numerous or repeated "little experiments." If the "experimentation" step was unsuccessful, the intervention may then go back to the preparation or "theoretical/literature base" phase (42).

An intentional, iterative, and data-driven effort to enhance a health intervention is what is meant by optimization (43). In-depth qualitative research pinpoints obstacles to behavior and intervention adoption and iteratively improves the intervention to get around them (44, 45). The theory- and evidence-based behavioral analysis is integrated into this strategy to choose the most suitable set of efficient behavior change strategies (46).

A live intervention's effectiveness was improved quickly and iteratively to keep pace with the terrifying and continuously changing environment of an international crisis. Making sure the intervention's contents is inspiring, credible, and convincing may be more crucial for fostering involvement than making changes to the intervention's design (41).

## Key principles of intervention development

Important guidelines for developing interventions should follow certain key tenets, including being dynamic, iterative, innovative, adaptable, and forward-looking in terms of evaluation and application in the future. When developing an intervention, developers are likely to switch back and forth between redundant tasks including examining the available data, using preexisting theory, and interacting with stakeholders. Iterative cycles will also be used to establish a rendition of the intervention, with feedback from stakeholders used to identify issues, possible alternatives put into action, their appropriateness evaluated, and the cycle repeated until evaluation of subsequent iterations of the intervention shows rare variations.

The need for the intervention, its style, substance, or method of delivery may be strongly held beliefs when the intervention is first being developed. Throughout the design process, keeping open to various alternatives may result in abandoning the project or moving both backward and forward. Being adaptable is a good idea since it could lessen the likelihood that you produce an intervention that fails in a later evaluation or is never used in practice. In order to prepare for this and highlight lessons learned and significant uncertainties that need to be addressed in future evaluations, developers may also gain from anticipating how the intervention will be appraised (47).

Monitoring and iterative evaluation should be prioritized to the greatest extent possible, and results should be regularly discussed and understood in collaboration with stakeholders, as well as thoughtfully and continually implemented in any system redesign or anticipated adaptations/modifications (48). The iterative method of data collection, assessment, evaluation, review, and change responds to the dynamic nature of evidence and the requirement for learning from and with stakeholders, such as populations and field workers (49).

# Patient and public involvement

By raising disease understanding and recognizing patients as active participants in their own conditions, patient and public involvement (PPI) can support patient empowerment (50). However, PPI varies significantly between nations and research organizations, and even today, many patients and the general public do not participate in or have access to study protocols (51). Cohort studies are increasingly including digital resources like websites, social media, and connected gadgets, which could be used to boost PPI (52). Digital tools can also help PPI by facilitating feedback and communication between study collaborators and patients (53). PPI is an effective strategy for raising the relevance of research efforts. We have demonstrated that PPI must be designed from the early stages of the construction of a original epidemiological study and then deliberated as the research project progresses. The most successful technique for raising the caliber of research appears to be combining various PPI approaches (54).

# Contextual indicators

The contextual indicators basis was created to offer direction on the interdependent context, implementation, and setting characteristics that may have an impact on the efficient delivery of complex interventions (55). What is crucial is that the framework explains how ambitious contextual factors outside of the administrative environment may affect how a complicated intervention with community-facing components is implemented. The seven dimensions that contextual indicators consider are; geographical, epidemiological, sociocultural, socioeconomic, ethical, legal, and political setting. Here, researchers look into the interactions between the political (healthcare infrastructure), epidemiological (blood pressure, body mass index, and older population) and geographical (region, urbanicity) domains of the contextual indicators (56).

## Share viewpoints and knowledge from public health experts

The deeper understandings into participants' involvements, opinions, views, and tips helped the researchers produce more detailed data and also helped others. Focus group involvement provided community members with new perspectives on issues they were discussing as well as a sense of insertion and community development, according to their reports (57).

Future studies should concentrate on a few unanswered problems about the use of digital forms in community-based health promotion interventions. If digital forms potentially take the place of outdated setups for health promotion and prevention actions, notably in vulnerable groups, this should be seriously evaluated. One of the most pressing unknowns, in our opinion, is whether the use of digital health promotion interventions results in an extra enlargement of a selection bias or whether such interventions combat this bias and are utilized and recognized by vulnerable groups and environments where inclusion struggle (58).

The researchers came to the conclusion that emphasizing participation in DHIs and utilizing standardized metrics to describe DHIs will aid future research and potentially open up more possibilities for meta-analyses of DHI results. This is further confirmed by Zanaboni et al. (59), who state that more emphasis should be given to clinical research in the form of high-quality randomized controlled trials in order to run a credible evidence base about the use of digital health and health results. According to Blandford et al. (8), established health research methodologies need to be flexible and modified in order to evaluate DHIs in study.

# Methodology

By using scoping review approach, this study investigates DHIs for public health. A review of the literature known as a "scoping study" or "scoping review" has the goal of "fastly mapping the major concepts driving a research topic and the main sources and forms of evidence available, notably where a subject is intricate or has not been studied thoroughly before" (60).

This kind of scoping review may not go into individual study findings, but rather maps and visualizes the body of knowledge that exists within the confines of the research field (61). Data were gathered and evaluated throughout five stages, as per the scoping review process delineated by Arksey and O'Malley (61), which is described below.

According to the stated rules for writing systematic reviews, peer reviews, and research articles, a systematic review was planned and

carried out. The literature on digital interventions for public health has undergone a thorough assessment. The articles' quality was not evaluated because it is not a part of the typical scoping review technique.

The main review interrogation was; "what are the implementing characteristics, concepts and recommendations of the digital interventions for public health considered." In addition to updates in five databases, an electronic search of digital interventions for public health was conducted. We searched databases from EBSCO, PubMed, ScienceDirect, Scopus and the Cochrane Library. We looked for the words "digital intervention" and "public health" in the article titles. In relation to the objective of the study, it was determined that these terms should be used the most. Duplicate articles were removed, and articles had to have been released in English before February 2022. All scholarly investigations underwent a thorough search of the peer-reviewed literature.

Four main inclusion criteria were defined (Figure 1):

- Published papers as peer-reviewed.
- Original research articles.
- Papers with full access possibility.
- Not targeted mental, sexual or productive health research.
- Papers written in the English language.
- Published before February/2022.

Studies that did not match the aforementioned requirements were excluded, while those that did were listed and subjected to further evaluation. Studies were assessed and given a critical review. Extraction of the key conclusions from each repossessed study and literature screening (a three-stage technique involving exclusion by reading the title, the abstract, and the full text). The following details were taken from each of the studies that were included (Table 1): title, authors, country, study design, research objective, and key findings.

# Results

The following research question was developed:

What implementing characteristics, concepts and recommendations that encourage digital intervention in public health? The terms "digital health intervention" were recognized as the use of digital, mobile, and wireless technologies to support the achievement of health objectives (71), encompassing both mHealth and eHealth. Arksey and O'Malley (61) advise using a broad definitional approach and propose that search words can be modified and reduced later to manage bibliographic references after the entire breadth of information within a given field has been attained. Given that it applies a uniform analytical framework to all studies, which is considered as a standard practice in scoping reviews, this methodology reflects a "descriptive-analytical" approach to charting.

From the publications, this study obtained both qualitative and quantitative data. This study's major objective was to conceptually clarify the characteristics of digital intervention for public health.

The results of this study may have also been exaggerated by other search parameters, such as restricting results to English-language articles. The current study's goal was to regulate the existing status of digital intervention for public health and make recommendations. The methodology was suitable for a policy analysis topic like this one. The limitations found in the literature highlight the need for public health practice information and more rigorous study approaches.

## Concepts of digital intervention for public health in the context of the reviewed articles

The articles focused on diverse concepts for the digital intervention for public health and also on different methods on the topic. Article

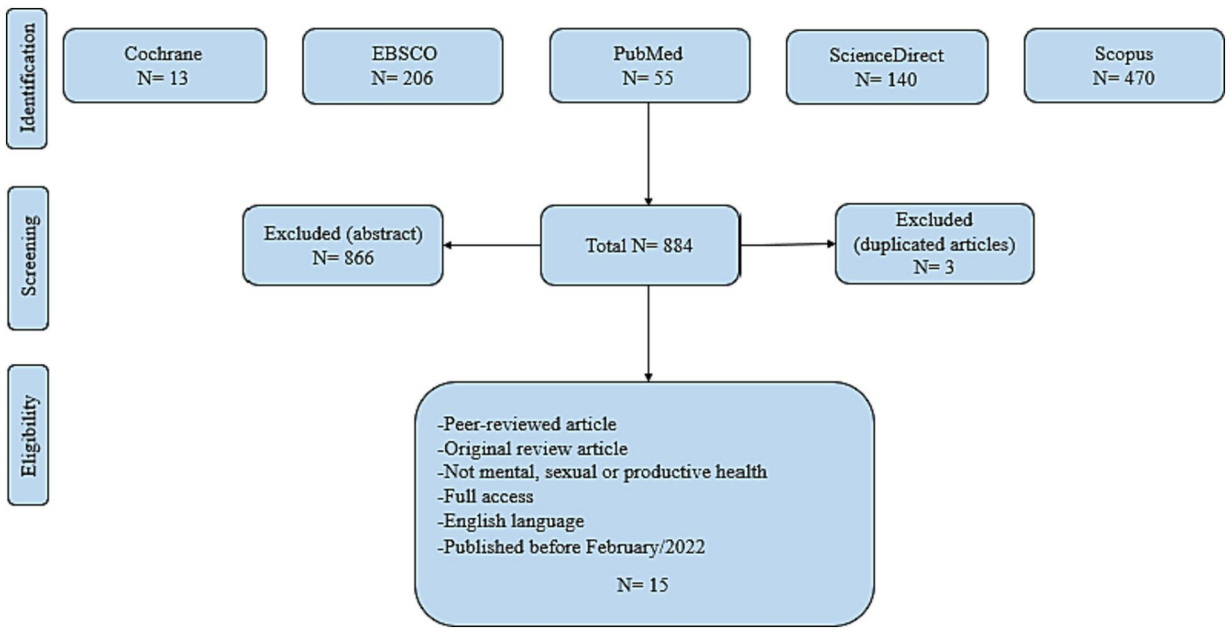



FIGURE 1
PRISMA flow chart of the literature review search.

TABLE 1 A summary of reviewed studies.

| Author | Setting/design | Aim | Study focus and findings | Recommendation |
|---|---|---|---|---|
| Burke and Bloss (62) | United States viewpoint | To hire commercial corporations to adopt cutting-edge technologies that monitor and manipulate students' social media activities | There is a need for creative answers to the problems of student health and safety.<br>The assertions made by social media monitoring firms and the schools that employ them that these technologies can solve the wide range of public health issues affecting pupils are unsupported by any evidence | The issues that young individuals already confront, particularly those from historically oppressed groups, may only be made worse by these digital surveillance tools |
| Susser (63) | United States perspective | Various strategies have been proposed by officials and public health experts, including; stepping up current initiatives to advance information and health literacy, coming up with plans for publicly refuting falsehoods, and training clinicians and public health authorities to deal with falsehoods one-on-one | Dealing with ethical challenges successfully will require balancing difficult tradeoffs.<br>The vast amounts of personal information that have been gathered about each of us are tremendously illuminating, and the instruments for using that information to target digital communications are strong and easily accessible.<br>It is simple to comprehend why academics and professionals in public health are keen to investigate the potential good they may do with them.<br>The ethical costs of targeted digital public health initiatives may be high | Targeting technologies frequently infringe the security of personal information by using data that has been collected in this way.<br>With these technology, disadvantaged populations could be targeted with discriminating messages.<br>Targeted digital public health interventions pose a risk of interfering with our autonomy by influencing our decision-making. Each situation should be evaluated individually to determine whether the advantages of these interventions exceed the disadvantages.<br>Practitioners should weigh the seriousness of the health problems they are addressing as well as their ability to reduce potential effects when making these decisions.<br>It proposes a starting point for conversations on the morality of targeted digital public health interventions |
| Karpathakis et al. (7) | London multidisciplinary project team user-centered qualitative research | A framework for evaluation that integrates biological and digital methods was intended to be operationalized as part of the Public Health England (PHE) effort. Shows how effective, affordable, and beneficial DHIs are for improving public health | Seven key ideas for evaluating DHIs have been identified: evaluation thinking, evaluation canvas, contract assistant, testing tools, development history, data hub, and publish health outcomes.<br>The planned PHE evaluation service for public health DHIs was developed after additional testing and refinement of three concepts that were given priority | PHE was able to integrate the skills of academic and biomedical fields with the knowledge of non-academic and digital developers through the use of an iterative, user-centered design methodology.<br>Using design-led techniques in public health settings can be beneficial. The following service is now offered by health organizations in the UK and is called evaluating digital health products |
| Sauerborn et al. (10) | Germany evaluation content and functions apps | Constructed on an awareness of structural, societal, and individual dimensions of violence against women and girls as a multidimensional, global public health concern, and positioning it within the theoretical framework of structural injustice | Make the case that while technical tools like apps may be helpful in the battle against violence against women and girls, they must be positioned within the larger context of public health that takes into account the structural aspects of such violence.<br>Along with major considerations for safety, data privacy, the value of human supportive touch, and other issues, structural injustice concerns are significant features in the ethical evaluation of such apps | Research on the function and applicability of apps as tactics to deal with the structural and epistemic aspects of violence is still lacking |

*(Continued)*

TABLE 1 (Continued)

| Author | Setting/design | Aim | Study focus and findings | Recommendation |
|---|---|---|---|---|
| O'Connor et al. (64) | UK systematic review | In order to guide future implementation efforts, it is important to identify and integrate the qualitative research on the factors that influence recruitment and involvement in DHIs | Four major topics that influence the participation of patients and the general public in DHIs emerged:<br>(1) Individual agency and motivation; (2) the individual's life and values; (3) the recruitment and engagement strategy; and (4) the DHI's level of quality.<br>Outlines the recruitment and engagement techniques used.<br>To highlight the crucial steps, a draft digital health engagement model (DIEGO) was created.<br>Future research recommendations are created after identifying existing knowledge gaps | Summarizes and elucidates the complexities of the recruitment and participation processes in digital health, as well as the problems that must be resolved before patients and the general public commit to digital health.<br>It will take more effort to develop personalized, higher-quality digital solutions that are clinically accredited and endorsed when necessary.<br>Additionally, more money is required to boost computer literacy and make sure that technologies are available and cheap for individuals who want to subscribe to them |
| Holst et al. (65) | Norway mixed methods (nonrandomized controlled trial and qualitative interviews) | To evaluate the DHI's impact on rural communities' long-term acquisition and retention of health knowledge | (1) Compare the intervention group's knowledge ratings at baseline and immediately following the intervention.<br>(2) The baseline knowledge score disparity between the intervention and control groups | Analyzing a DHI's results in light of pertinent health messages |
| Budd et al. (66) | United Kingdom review | To document the range of technological developments for the global public health response to COVID-19 and its shortcomings | To identify obstacles to its execution, such as those posed by the law, morality, and privacy concerns, as well as those posed by organizations and the workforce | Examine the necessity of coordinating global strategies for the control, assessment, and application of digital technology to improve pandemic readiness and future COVID-19 and other infectious disease preparedness |
| Morton et al.(41) | United Kingdom review | Offers a collection of iterative techniques for quickly modifying and improving an intervention as it is being implemented | The intervention was clinically correct thanks to tight collaboration with clinical stakeholders.<br>Contributors to patient and public involvement (PPI) recognized critical clarifications to the intervention's content and made sure that data concerning challenging behaviors (such self-isolation) was encouraging and practical | According to calls for more expeditious, practical health research techniques, quick optimization techniques of this kind may be utilized in the future to enhance the speed and efficiency of adaption, refinement, and implementation of interventions |
| Ross et al. (67) | United Kingdom implemented intervention (HeLP-Diabetes) | Give an illustration of how to create a theoretically based implementation strategy and how to openly disclose it | For the purpose of integrating HeLP-Diabetes into everyday practice, a new implementation strategy was created. The normalization process theory served as a guide for the selection and development of specific component techniques. These tactics included involving local opinion leaders, distributing instructional materials, hosting educational visits and meetings, conducting audits, receiving feedback, and reminding people. Barriers that surfaced during deployment were iteratively addressed with additional solutions. Having trouble allocating funds to put the intervention into practice within ordinary treatment was a major barrier | Others who are working on planning and carrying out implementation activities in regular healthcare can benefit from the knowledge gained from this study.<br>The choice of an acceptable theory to direct the process of implementation and the choice of tactics; making sure that adequate attention is paid to planned implementation and a flexible approach that permits responsiveness to developing hurdles |

*(Continued)*

TABLE 1 (Continued)

| Author | Setting/design | Aim | Study focus and findings | Recommendation |
|---|---|---|---|---|
| Bevens et al. (68) | Australia a practical overview | Aims to share information and thoughts from public health academics who have taken part in the process of digitally transforming a face-to-face lifestyle management training program | Information on the digital transformation of lifestyle education programs is scarce, and this is especially true for initiatives focused on chronic conditions.<br>Higher education has produced a significant body of work that has experienced fast digital transition. Much can be learned from this area of study.<br>Additionally, academics looking to design, develop, and implement DHIs have access to a well-established area of design approaches and frameworks | Gives a detailed explanation of how the processes of higher education's digital transformation can be combined with the use of a current development model for DHIs |
| Patel et al.(56) | USA cross-sectional analysis | It assessed the current healthcare system's ability to support digital health treatments and looked at the correlates of the system's epidemiological, socioeconomic, and geographic contexts | The availability of critical personnel was lower than the availability of IT infrastructure for all locations except subcenters.<br>Higher blood pressure, body mass index, and urban residents were associated with better infrastructure for all hospitals except district hospitals | When compared to apex facilities in India, lower and mid-tier healthcare facilities more commonly lack the IT infrastructure needed to facilitate digital health initiatives.<br>Physical infrastructure gaps were typically higher than staffing ones, indicating that, in addition to IT infrastructure, shortages of key personnel place serious restrictions on the adoption of digital health solutions |
| Schroeer et al. (58) | Germany a scoping review | Seeks to map the body of research on digital platforms that encourage community meeting in the field of health promotion and prevention | There were two studies on interaction with peers, five studies that used qualitative participatory research, one study on empowerment, and five studies that used crowdsourcing.<br>The digital tools employed ranged greatly and included social networking sites, message boards, websites for online forums, and specialized web hosts and applications.<br>The majority of research cited convenience, flexibility, and anonymity as advantages of digital interventions.<br>Some articles noted drawbacks, such as issues with interpreting data that can only be read in writing or the potential for selection bias brought on by the digital divide | There is a study gap on this subject, as the review only found a few studies that were pertinent to our goal.<br>It was discovered that digital formats are especially well suited for activities where confidentiality and adaptability are advantageous, like online peer-to-peer assistance programs |
| Harte et al. (46) | USA exploratory | Explains the purpose and plan of a trial that examines the combined impact of community health worker and digital health support on hemoglobin and glucose self-monitoring | The population of interest was low-income people, the study purpose was explicitly to advance knowledge beneficial for increasing health equity, and the study protocols were developed in partnership with frontline community health professionals | It enhances understanding of whether integrating community health worker interventions with digital health can enhance glucose self-monitoring and outcomes related to diabetes in a high-risk group |

*(Continued)*

TABLE 1 (Continued)

| Author | Setting/design | Aim | Study focus and findings | Recommendation |
|---|---|---|---|---|
| Chen et al. (69) | China public surveys | Analyze the circumstances and important players in China's quick deployment of digital health solutions in response to COVID-19, and record and disseminate the lessons collected | The wide adoption of digital health technology revealed contextual elements and important enabling mechanisms in case studies that were identified under each category | The prosperous digital health expanse before COVID-19, the public sector's flexibility in introducing regulatory flexibilities, and incentives to energize the private sector are among the contextual factors and key permitting mechanisms through the practice of policy instruments to encourage DHIs for COVID-19 in China. These factors also include the route of policy advices affecting the private sector using a regionalized approach |
| Batta and Iwokwagh (70) | Nigeria inductive content analysis | It examines how Nigerian teaching hospitals make use of social media and new media.<br>It examines whether new and social media are used as public relations tools (to increase their visibility, promote their services, and enhance their corporate image), educational tools (to provide health information, revelation, and education in order to prevent disease and promote health), and social tools (to facilitate communication between people) (to deepen interactions and exchanges between healthcare providers and healthcare recipients) | Nigerian teaching hospitals mostly use new and social media to solicit customer input (100%), provide their vision and mission statements (65%), post details about their administrative and staff structures (65%), and provide contract information (60%). For financial transactions (10%) and the promotion of health (25%), these media are seldom ever used | Teaching hospitals should make more use of social media and new media to give patients and family members a platform to share their stories and to give informed advice on medical and health issues |

focus on commercial corporations to adopt cutting-edge technologies (62), advance information and health literacy (63), framework for evaluation that integrates biological and digital methods (7), multidimensional, global public based on an awareness of structural, societal, and individual extents of violence against women and girls (10), factors that influence recruitment and involvement in DHIs (64), DHI's impact on rural communities' (66), range of technological developments for the global public health (65), collection of iterative techniques (41), theoretically based implementation strategy (67), share information and thoughts from public health academics, healthcare system's ability to support digital health treatments (68), encourage community engagement (56), digital health support on hemoglobin (46), circumstances and important players in China's quick deployment of digital health solutions (69) and teaching hospitals that make use of social media (70).

## Discussion

This scoping review only discovered uncommon studies that used a digital platform to empower substantial community involvement in health promotion and prevention, highlighting a research gap in this area. Digital formats were discovered to be appropriate for situations where obscurity is advantageous. This was evident in the included studies' qualitative participatory research investigations, notably in the virtual focus groups where contributors had to discuss difficult topics. Additionally, it indicated that anonymity and ease of access were helpful in assisting marginalized and disadvantaged communities, such as through interaction with peers and social exchange programs (58).

With the help of this scoping study, we were able to map the body of research on digital platforms that encourage community involvement in the field of health promotion and prevention. In addition, we obtained a deeper awareness of the fundamental ideas in this field in terms of the sorts of involvement that can be facilitated, the ways to use digital forms, and the advantages and drawbacks associated with them (58).

DHIs are provided through digital channels, such as websites and mobile applications, with the goal of providing care or promoting health (5). Such DHIs are anticipated to combine the effectiveness of individualized therapies with the influence of large-scale population campaigns. DHIs are also meant to expand access and capacity for

public health efforts by offering services in places where face-to-face choices are absent or inadequate to satisfy demand (7).

However, it is also important to analyze how institutional inequality, particularly epistemic injustice, affects the content and purposes of the DHIs utilized in public health treatments. They outline and emphasize the significance of violence against women and girls as a global public health concern and briefly evaluate its multifaceted character on structural, societal, and personal levels (10). According to the author, technical solutions like DHIs may be a useful tool in the battle against violence against women or gender inequality, but they must be placed within the larger context of public health that recognizes and the structural components of such fierceness.

The vast amounts of personal information gathered are highly illuminating, and the means for using that information to target digital communications are strong and easily accessible. It is simple to comprehend why academics and professionals in public health are keen to investigate the potential good they may do with them. Such technologies run the risk of discriminatory message targeting against disadvantaged groups. Targeted digital public health interventions pose a risk of interfering with our autonomy by influencing our decision-making. Each situation should be evaluated individually to determine whether the advantages of these interventions exceed the disadvantages. Practitioners should weigh the seriousness of the health risks they are targeting (e.g., promoting a healthy diet as opposed to intervening in suicide cases or eradicating health misinformation during a pandemic) as well as their ability to lessen probable harms (e.g., whether messaging can be clear and collected data respect entities' privacy) (63).

Collecting the quantitative and qualitative results will produce a strong set of data that can be used to adapt the intervention's execution with access to the digital health platform as well as to evaluate the DHI. The study illustrates a participatory and community-based component that has the opportunity to have an improved, context-specific influence on local communities' digital health education by leveraging upon conclusions from both research techniques to enhance the intervention (65).

In addition to a list of the hurdles and implementors that patients and the general public encounter while appealing with and enrolling in DHIs, this review gives an overview of reported engagement and recruitment tactics. In line with the findings of our review, literacy abilities (72) and financial resources (73) do have an impact on people's capacity to interact with and use DHIs.

Digital technologies must be integrated into the current public healthcare systems since they cannot function alone (74). For instance, as one of many approaches, South Korea and Singapore effectively implemented contact-tracing DHIs to support massive teams of manual contact tracers (66). The digital infrastructure and public health systems' readiness, which include secondary, primary, and social care systems, will be key factors in the analysis and utilization of these data. With multiple symptom-reporting sites in a single nation, coordination of therapies is especially difficult and runs the danger of fragmentation (66).

The intervention, however, was clinically correct since tight collaboration with clinical stakeholders guaranteed that the information concerning transmission and exposure was compatible with the available data, for instance. Contributors to patient and public involvement (PPI) identified crucial justifications to the intervention's content, such as whether epidemics can spread through the air as well as surfaces and made sure that evidence about challenging behaviors (like self-isolation) was encouraging and practical (41).

Furthermore, the author has created knowledge about some of the enablers and barriers to putting DHIs into reality. In a system with limited resources, we discovered that requiring personnel to assist patients in registering to use a DHI was a barrier (67). A live intervention's effectiveness was improved quickly and iteratively to keep pace with the terrifying and continuously changing environment of an international crisis. A rich approach for swift stakeholder assignation was crucial for apprising decisions about how to discourse these obstacles, and the variety of methods assisted in developing a thorough grasp of the potential hurdles to the target behaviors (41).

## Conclusion

Understanding the variables connected to digital interventions for public health begins with this scoping review of the literature. The review has given ideas about the factors that contribute to success and insight into some of the techniques used to identify high achievers, but it has also highlighted the need for new approaches to understanding what counts as high impact and how to enhance elements that are crucial to population health. As, the public health is likely to become more and more digital in the future, the author examines the requirement for the synchronization of global approaches for the regulation, assessment and use of digital technologies in order to improve population health supervision and imminent alertness for diseases.

The author contends that elements that go beyond the inter-individual level must be considered for any intervention technique to be successful and long-lasting. There is little research on the function and importance of DHIs as tactics for addressing the structural and epistemological components. The participants and those around them will gain more awareness about health issues by receiving health messages in a digital format, which may change how they seek out health care. More work is required to develop effective engagement tactics, significantly greater, individualized digital solutions, and to obtain clinical accreditation and support where necessary.

The choice of an appropriate theory to direct the course of implementation and strategy selection is essential. The reporting of implementation strategies using terms that are clear and defined, and using a flexible approach are all important considerations. In addition, physical infrastructure gaps were typically indicating that beyond information technology infrastructure, shortages of indispensable staff enforce significant barriers to the adoption of DHIs.

To sum up, the author's work outlines an iterative, cross-disciplinary, participatory progression for creating, implementing, and appraising DHI, emphasizing the adjacent collaboration between behavior scientists, designers, data engineers, software developers, and data scientists as well as on a constant reaction circle from end users. A defined approach for swift stakeholder involvement was crucial for guiding decisions about how to discourse these obstacles, and the variety of ways contributed to the development of a deep consideration of the potential barriers to the target behaviors. Making sure the intervention's content is inspiring, reliable, and convincing may be more crucial for fostering engagement than making changes to the intervention's design (66).

DHI offers a viewpoint that emphasizes a considerable larger series of issues related to the sociotechnical system involved by a specific digital health technology and the health of the numerous communities. This study could be used in other areas of public health policy and practice and will attend as a source for enduring discussion in this area.

## Author contributions

HA held the main parts of the research, writing, collecting the data, and results and discussion. AA helped HA in reviewing the paper and gave notes. 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. Chou WYS, Oh A, Klein WM. Addressing health-related misinformation on social media. *JAMA*. (2018) 320:2417–8. doi: 10.1001/jama.2018.16865

2. Human Rights Watch (2020). Myanmar: end world's longest internet shutdown. Available at: https://www.hrw.org/news/2020/06/19/myanmar-end-worlds-longestinternetshutdown.

3. Rimmer A. COVID-19: disproportionate impact on ethnic minority healthcare workers will be explored by government. *Br Med J*. (2020) 369:m1562. doi: 10.1136/bmj.m1562

4. World Health Organization. *Education for health: a manual on health education in primary health care* World Health Organization (1988) Available at: https://apps.who.int/iris/handle/10665/77769.

5. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. *Am J Prev Med*. (2016) 51:843–51. doi: 10.1016/j.amepre.2016.06.008

6. Tognotti E. Lessons from the history of quarantine, from plague to influenza A. *Emerg Infect Dis*. (2013) 19:254. doi: 10.3201/eid1902.120312

7. Karpathakis K, Libow G, Potts HW, Dixon S, Greaves F, Murray E. An evaluation service for digital public health interventions: user-centered design approach. *J Med Internet Res*. (2021) 23:e28356. doi: 10.2196/28356

8. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. *Digit Health*. (2018) 4:2055207618770325. doi: 10.1177/2055207618770325

9. Coghlan S, D'Alfonso S. Digital phenotyping: an epistemic and methodological analysis. *Philos Technol*. (2021) 34:1905–28. doi: 10.1007/s13347-021-00492-1

10. Sauerborn E, Eisenhut K, Ganguli-Mitra A, Wild V. Digitally supported public health interventions through the lens of structural injustice: the case of mobile apps responding to violence against women and girls. *Bioethics*. (2022) 36:71–6. doi: 10.1111/bioe.12965

11. Oudshoorn N. *Telecare technologies and the transformation of healthcare*. Houndmills, UK: Palgrave Macmillan. (2011).

12. Greenhalgh T, Wood GW, Bratan T, Stramer K, Hinder S. Patients' attitudes to the summary care record and HealthSpace: qualitative study. *BMJ*. (2008) 336:1290–5. doi: 10.1136/bmj.a114

13. Bardus M, Blake H, Lloyd S, Suzanne Suggs L. Reasons for participating and not participating in a e-health workplace physical activity intervention: A qualitative analysis. *Int. J. Workplace Health Manag*. (2014) 7:229–46.

14. Lorimer K, Martin S, McDaid LM. The views of general practitioners and practice nurses towards the barriers and facilitators of proactive, internet-based chlamydia screening for reaching young heterosexual men. *BMC Fam Pract*. (2014) 15:1–10. doi: 10.1186/1471-2296-15-127

15. Trujillo Gómez JM, Díaz-Gete L, Martín-Cantera C, Fábregas Escurriola M, Lozano Moreno M, Burón Leandro R, et al. Intervention for smokers through new communication technologies: what perceptions do patients and healthcare professionals have? A qualitative study. *PLoS One*. (2015) 10:e0137415. doi: 10.1371/journal.pone.0137415

16. Winkelman WJ, Leonard KJ, Rossos PG. Patient-perceived usefulness of online electronic medical records: employing grounded theory in the development of information and communication technologies for use by patients living with chronic illness. *J Am Med Inform Assoc*. (2005) 12:306–14. doi: 10.1197/jamia.M1712

17. Lundgren J, Johansson P, Jaarsma T, Andersson G, Köhler AK. Patient experiences of web-based cognitive behavioral therapy for heart failure and depression: qualitative study. *J Med Internet Res*. (2018) 20:e10302. doi: 10.2196/10302

18. Wilhelmsen M, Lillevoll K, Risør MB, Høifødt R, Johansen ML, Eisemann M, et al. Motivation to persist with internet-based cognitive behavioural treatment using blended care: a qualitative study. *BMC Psychiatry*. (2013) 13:1–9. doi: 10.1186/1471-244X-13-296

19. Wachtler C, Coe A, Davidson S, Fletcher S, Mendoza A, Sterling L, et al. Development of a mobile clinical prediction tool to estimate future depression severity and guide treatment in primary care: user-centered design. *JMIR Mhealth Uhealth*. (2018) 6:e9502. doi: 10.2196/mhealth.9502

20. Pagliari C, Burton C, McKinstry B, Szentatotai A, David D, Ferrini L, et al. Psychosocial implications of avatar use in supporting therapy for depression. *Annu Rev Cyberther Telemed*. (2012) 2012:329–333. doi: 10.1111/nyas.13336

21. Gonsalves PP, Hodgson ES, Kumar A, Aurora T, Chandak Y, Sharma R, et al. Design and development of the "POD adventures" smartphone game: a blended problem-solving intervention for adolescent mental health in India. *Front Public Health*. (2019) 7:238. doi: 10.3389/fpubh.2019.00238

22. Povey J, Mills PPJR, Dingwall KM, Lowell A, Singer J, Rotumah D, et al. Acceptability of mental health apps for aboriginal and Torres Strait islander Australians: a qualitative study. *J Med Internet Res*. (2016) 18:e5314. doi: 10.2196/jmir.5314

23. Hunter-Jones JJ, Gilliam SM, Carswell AL, Hansen NB. Assessing the acceptability of a mindfulness-based cognitive therapy intervention for African-American women living with HIV/AIDS. *J Racial Ethn Health Disparities*. (2019) 6:1157–66. doi: 10.1007/s40615-019-00617-5

24. Henshall C, Marzano L, Smith K, Attenburrow MJ, Puntis S, Zlodre J, et al. A web-based clinical decision tool to support treatment decision-making in psychiatry: a pilot focus group study with clinicians, patients and carers. *BMC Psychiatry*. (2017) 17:1–10. doi: 10.1186/s12888-017-1406-z

25. Walsh DM, Moran K, Cornelissen V, Buys R, Cornelis N, Woods C. Electronic health physical activity behavior change intervention to self-manage cardiovascular disease: qualitative exploration of patient and health professional requirements. *JMIR*. (2018) 20:e163.

26. Feijt MA, de Kort YA, Bongers IM, WA IJ. Perceived drivers and barriers to the adoption of eMental health by psychologists: the construction of the levels of adoption of eMental health model. *J Med Internet Res*. (2018) 20:e9485. doi: 10.2196/jmir.9485

27. Apolinário-Hagen J, Vehreschild V, Alkoudmani RM. Current views and perspectives on e-mental health: an exploratory survey study for understanding public attitudes toward internet-based psychotherapy in Germany. *JMIR Ment health*. (2017) 4:e6375. doi: 10.2196/mental.6375

28. Jordan SE, Shearer EM. An exploration of supervision delivered via clinical video telehealth (CVT). *Train Educ Prof Psychol*. (2019) 13:323. doi: 10.1037/tep0000245

29. Kumarasiri J, Jubb C. Framing of climate change impacts and use of management accounting practices. *Asian Acad Manag J Account Finance*. (2017) 13. doi: 10.21315/aamjaf2017.13.2.3

30. Huerta-Ramos E, Escobar-Villegas MS, Rubinstein K, Unoka ZS, Grasa E, Hospedales M, et al. Measuring users' receptivity toward an integral intervention model based on mHealth solutions for patients with treatment-resistant schizophrenia (m-RESIST): A qualitative study. *JMIR Mhealth Uhealth*. (2016) 4:e5716. doi: 10.2196/mhealth.5716

31. Donkin L, Glozier N. Motivators and motivations to persist with online psychological interventions: a qualitative study of treatment completers. *J Med Internet Res*. (2012) 14:e2100. doi: 10.2196/jmir.2100

32. Wallin E, Norlund F, Olsson EMG, Burell G, Held C, Carlsson T. Treatment activity, user satisfaction, and experienced usability of internet-based cognitive behavioral therapy for adults with depression and anxiety after a myocardial infarction: mixed-methods study. *J Med Internet Res*. (2018) 20:e9690. doi: 10.2196/jmir.9690

33. Anderson AP, Fellows AM, Binsted KA, Hegel MT, Buckey JC. Autonomous, computer-based behavioral health countermeasure evaluation at HI-SEAS Mars analog. *Aerosp Med Hum Perform*. (2016) 87:912–20. doi: 10.3357/AMHP.4676.2016

34. Jonathan G, Carpenter-Song EA, Brian RM, Ben-Zeev D. Life with FOCUS: a qualitative evaluation of the impact of a smartphone intervention on people with serious mental illness. *Psychiatr Rehabil J*. (2019) 42:182. doi: 10.1037/prj0000337

35. Williams A, Fossey E, Farhall J, Foley F, Thomas N. Recovery after psychosis: qualitative study of service user experiences of lived experience videos on a recovery-oriented website. *JMIR Ment Health*. (2018) 5:e9934. doi: 10.2196/mental.9934

36. Görges F, Oehler C, von Hirschhausen E, Hegerl U, Rummel-Kluge C. GET. HAPPY2—user perspectives on an internet-based self-management positive psychology intervention among persons with and without depression: results from a retrospective survey. *J Clin Psychol*. (2020) 76:1030–46. doi: 10.1002/jclp.22886

37. Lawal AM, Idemudia ES, Senyatsi T. Emotional intelligence and mental health: an exploratory study with south African university students. *J Psychol Afr*. (2018) 28:492–7. doi: 10.1080/14330237.2018.1540229

38. Eisner E, Drake RJ, Berry N, Barrowclough C, Emsley R, Machin M, et al. Development and long-term acceptability of ExPRESS, a mobile phone app to monitor basic symptoms and early signs of psychosis relapse. *JMIR Mhealth Uhealth*. (2019) 7:e11568. doi: 10.2196/11568

39. Edbrooke-Childs J, Edridge C, Averill P, Delane L, Hollis C, Craven MP, et al. A feasibility trial of power up: smartphone app to support patient activation and shared decision making for mental health in young people. *JMIR Mhealth Uhealth*. (2019) 7:e11677. doi: 10.2196/11677

40. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. *J Med Internet Res*. (2021) 23:e24387. doi: 10.2196/24387

41. Morton K, Ainsworth B, Miller S, Rice C, Bostock J, Denison-Day J, et al. Adapting behavioral interventions for a changing public health context: a worked example of implementing a digital intervention during a global pandemic using rapid optimisation methods. *Front Public Health*. (2021) 9:668197. doi: 10.3389/fpubh.2021.668197

42. McCrabb S, Mooney K, Elton B, Grady A, Yoong SL, Wolfenden L. How to optimise public health interventions: a scoping review of guidance from optimisation process frameworks. *BMC Public Health*. (2020) 20:1–12. doi: 10.1186/s12889-020-09950-5

43. Long JA, Jahnle EC, Richardson DM, Loewenstein G, Volpp KG. Peer mentoring and financial incentives to improve glucose control in African American veterans: a randomized trial. *Ann Intern Med*. (2012) 156:416–24. doi: 10.7326/0003-4819-156-6-201203200-00004

44. Palmas W, March D, Darakjy S, Findley SE, Teresi J, Carrasquillo O, et al. Community health worker interventions to improve glycemic control in people with diabetes: a systematic review and meta-analysis. *J Gen Intern Med*. (2015) 30:1004–12. doi: 10.1007/s11606-015-3247-0

45. Heisler M, Vijan S, Makki F, Piette JD. Diabetes control with reciprocal peer support versus nurse care management: a randomized trial. *Ann Intern Med*. (2010) 153:507–15. doi: 10.7326/0003-4819-153-8-201010190-00007

46. Harte R, Norton L, Whitehouse C, Lorincz I, Jones D, Gerald N, et al. Design of a randomized controlled trial of digital health and community health worker support for diabetes management among low-income patients. *Contemp Clin Trials Commun*. (2022) 25:100878. doi: 10.1016/j.conctc.2021.100878

47. O'Cathain A, Croot L, Duncan E, Rousseau N, Sworn K, Turner KM, et al. Guidance on how to develop complex interventions to improve health and healthcare. *BMJ Open*. (2019) 9:e029954. doi: 10.1136/bmjopen-2019-029954

48. Chambers DA. Considering the intersection between implementation science and COVID-19. *Implement Res Pract*. (2020) 1:0020764020925994. doi: 10.1177/0020764020925994

49. Yousefi Nooraie R, Shelton RC, Fiscella K, Kwan BM, McMahon JM. The pragmatic, rapid, and iterative dissemination and implementation (PRIDI) cycle: adapting to the dynamic nature of public health emergencies (and beyond). *Health Res Policy Syst*. (2021) 19:1–10. doi: 10.1186/s12961-021-00764-4

50. Bereczky T. Patient advocacy-with a feeling patient citizenship-the description of an affective model of patient advocacy In: *Doctoral dissertation*. Hungary: ELTE University (2019) Available at: https://www.academia.edu/42773922/Patient_Advocacy_With_a_Feeling_Patient_citizenship_The_description_of_an_affective_model_of_patient_advocacy

51. Biddle MS, Gibson A, Evans D. Attitudes and approaches to patient and public involvement across Europe: a systematic review. *Health Soc Care Community*. (2021) 29:18–27. doi: 10.1111/hsc.13111

52. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digi Med*. (2019) 2:1–11. doi: 10.1038/s41746-019-0166-1

53. Algeo N, Hunter D, Cahill A, Dickson C, Adams J. Usability of a digital self-management website for people with osteoarthritis: a UK patient and public involvement study. *Int J Ther Rehabil*. (2017) 24:78–82. doi: 10.12968/ijtr.2017.24.2.78

54. Aguayo GA, Goetzinger C, Scibilia R, Fischer A, Seuring T, Tran VT, et al. Methods to generate innovative research ideas and improve patient and public involvement in modern epidemiological research: review, patient viewpoint, and guidelines for implementation of a digital cohort study. *J Med Internet Res*. (2021) 23:e25743. doi: 10.2196/25743

55. Pfadenhauer LM, Gerhardus A, Mozygemba K, Lysdahl KB, Booth A, Hofmann B, et al. Making sense of complexity in context and implementation: the context and implementation of complex interventions (CICI) framework. *Implement Sci*. (2017) 12:1–17. doi: 10.1186/s13012-017-0552-5

56. Patel SA, Vashist K, Jarhyan P, Sharma H, Gupta P, Jindal D, et al. A model for national assessment of barriers for implementing digital technology interventions to improve hypertension management in the public health care system in India. *BMC Health Serv Res*. (2021) 21:1–11. doi: 10.1186/s12913-021-06999-9

57. Nared J, Bole D. *Participatory research and planning in practice*. Berlin, Germany: Springer. (2020). 227 p.

58. Schroeer C, Voss S, Jung-Sievers C, Coenen M. Digital formats for community participation in health promotion and prevention activities: a scoping review. *Front Public Health*. (2021) 9:713159. doi: 10.3389/fpubh.2021.713159

59. Zanaboni P, Ngangue P, Mbemba GIC, Schopf TR, Bergmo TS, Gagnon MP. Methods to evaluate the effects of internet-based digital health interventions for citizens: systematic review of reviews. *J Med Internet Res*. (2018) 20:e10202. doi: 10.2196/10202

60. Mays N, Roberts E, Popay J. Synthesising research evidence In: N Fulop, P Allen, A Clarke and N Black, editors. *Studying the organization and delivery of health services: research methods*. London: Routledge. (2001)

61. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. (2005) 8:19–32. doi: 10.1080/1364557032000119616

62. Burke C, Bloss C. Social media surveillance in schools: rethinking public health interventions in the digital age. *J Med Internet Res*. (2020) 22:e22612. doi: 10.2196/22612

63. Susser D. Ethical considerations for digitally targeted public health interventions. *Am J Public Health*. (2020) 110:S290–1. doi: 10.2105/AJPH.2020.305758

64. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. *BMC Med Inform Decis Mak*. (2016) 16:1–15. doi: 10.1186/s12911-016-0359-3

65. Holst C, Sukums F, Ngowi B, Diep LM, Kebede TA, Noll J, et al. Digital health intervention to increase health knowledge related to diseases of high public health concern in Iringa, Tanzania: protocol for a mixed methods study. *JMIR Res Protoc*. (2021) 10:e25128. doi: 10.2196/25128

66. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. *Nat Med*. (2020) 26:1183–92. doi: 10.1038/s41591-020-1011-4

67. Ross J, Stevenson F, Dack C, Pal K, May C, Michie S, et al. Developing an implementation strategy for a digital health intervention: an example in routine healthcare. *BMC Health Serv Res*. (2018) 18:1–13. doi: 10.1186/s12913-018-3615-7

68. Merolli M. Insights from public health researchers into the digital transformation of an educational lifestyle course. *Healthier Lives, Digitally Enabled: Selected Papers from the Digital Health Institute Summit 2020*, vol. *276*. (2021) p. 14.

69. Chen M, Xu S, Husain L, Galea G. Digital health interventions for COVID-19 in China: a retrospective analysis. *Intell Med*. (2021) 1:29–36. doi: 10.1016/j.imed.2021.03.001

70. Batta HE, Iwokwagh NS. Optimising the digital age health-wise: utilisation of new/social media by Nigerian teaching hospitals. *Procedia Soc Behav Sci*. (2015) 176:175–85. doi: 10.1016/j.sbspro.2015.01.459

71. World Health Organization. *Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment*. Geneva, Switzerland: World Health Organization (2016).

72. Kontos E, Blake KD, Chou WYS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the health information National Trends Survey 2012. *J Med Internet Res*. (2014) 16:e3117. doi: 10.2196/jmir.3117

73. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. *J Med Internet Res*. (2012) 14:e1619. doi: 10.2196/jmir.1619

74. Gong M, Liu L, Sun X, Yang Y, Wang S, Zhu H. Cloud-based system for effective surveillance and control of COVID-19: useful experiences from Hubei, China. *J Med Int Res*. (2020) 22:e18948. doi: 10.2196/18948

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 25 September 2023
DOI 10.3389/fpsyt.2023.1236551



OPEN ACCESS

EDITED BY
Matthias Jaeger,
Psychiatrie Baselland, Switzerland

REVIEWED BY
Yannik Terhorst,
University of Ulm, Germany
Seyed-Ali Sadegh-Zadeh,
Staffordshire University, United Kingdom

*CORRESPONDENCE
Katinka Franken
✉ c.p.m.franken@utwente.nl

RECEIVED 11 June 2023
ACCEPTED 11 September 2023
PUBLISHED 25 September 2023

CITATION
Franken K, ten Klooster P, Bohlmeijer E, Westerhof G and Kraiss J (2023) Predicting non-improvement of symptoms in daily mental healthcare practice using routinely collected patient-level data: a machine learning approach.
*Front. Psychiatry* 14:1236551.
doi: 10.3389/fpsyt.2023.1236551



# Predicting non-improvement of symptoms in daily mental healthcare practice using routinely collected patient-level data: a machine learning approach

Katinka Franken*, Peter ten Klooster, Ernst Bohlmeijer, Gerben Westerhof and Jannis Kraiss

Department of Psychology, Health and Technology, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, Netherlands

**Objectives:** Anxiety and mood disorders greatly affect the quality of life for individuals worldwide. A substantial proportion of patients do not sufficiently improve during evidence-based treatments in mental healthcare. It remains challenging to predict which patients will or will not benefit. Moreover, the limited research available on predictors of treatment outcomes comes from efficacy RCTs with strict selection criteria which may limit generalizability to a real-world context. The current study evaluates the performance of different machine learning (ML) models in predicting non-improvement in an observational sample of patients treated in routine specialized mental healthcare.

**Methods:** In the current longitudinal exploratory prediction study diagnosis-related, sociodemographic, clinical and routinely collected patient-reported quantitative outcome measures were acquired during treatment as usual of 755 patients with a primary anxiety, depressive, obsessive compulsive or trauma-related disorder in a specialized outpatient mental healthcare center. ML algorithms were trained to predict non-response (< 0.5 standard deviation improvement) in symptomatic distress 6 months after baseline. Different models were trained, including models with and without early change scores in psychopathology and well-being and models with a trimmed set of predictor variables. Performance of trained models was evaluated in a hold-out sample (30%) as a proxy for unseen data.

**Results:** ML models without early change scores performed poorly in predicting six-month non-response in the hold-out sample with Area Under the Curves (AUCs) < 0.63. Including early change scores slightly improved the models' performance (AUC range: 0.68–0.73). Computationally-intensive ML models did not significantly outperform logistic regression (AUC: 0.69). Reduced prediction models performed similar to the full prediction models in both the models without (AUC: 0.58–0.62 vs. 0.58–0.63) and models with early change scores (AUC: 0.69–0.73 vs. 0.68–0.71). Across different ML algorithms, early change scores in psychopathology and well-being consistently emerged as important predictors for non-improvement.

**Conclusion:** Accurately predicting treatment outcomes in a mental healthcare context remains challenging. While advanced ML algorithms offer flexibility, they showed limited additional value compared to traditional logistic regression in this study. The current study confirmed the importance of taking early change scores in both psychopathology and well-being into account for predicting longer-term outcomes in symptomatic distress.

KEYWORDS

machine learning, mental disorder, well-being, psychopathology, prediction, non-improvement

# 1. Introduction

## 1.1. Prevalence and impact of psychiatric disorders

Worldwide, around one in eight people has one or more mental disorders (1). Mental disorders are the leading cause of years lived with disability (YLDs), accounting for one in every six YLDs globally (1). They contribute significantly to a lack of quality of life (2) and the direct and indirect economic and societal costs are substantial (1). Depression and anxiety alone result in the loss of nearly US$ 1 trillion and 12 billion working days every year (3). The increasing demand for care in combination with limited treatment effects puts pressure on waiting lists in mental healthcare (4). Insights in predicting who is less likely to improve early in treatment would be helpful to make treatment more efficient, reduce waste of financial and human resources and tailor treatment to the individual (5–7).

## 1.2. Predicting treatment effects

Studies show that 60% of patients with a mental disorder do not benefit from evidence-based treatments (8–12). At present, no convincing evidence has been found for a difference in treatment effect for any specific treatment, neither for mood disorders (13, 14) nor anxiety disorders (14, 15). Norcross and Lambert (16) argue that fitting psychotherapy to patient characteristics is necessary for treatment success. Clinical practice, however, shows that a DSM-classification alone does not give sufficient direction to appropriate treatment (17–19). This underlines the relevance of adopting a more transdiagnostic approach in clinical practice and searching for predictors across the main diagnoses. Early identification of non-responders can increase treatment effectiveness as it may support personalized treatment recommendations (20).

Mental disorders are complex and trajectories of treatment can depend on many factors, making the prediction of treatment outcomes challenging. Previous studies have incidentally found several predictors for treatment outcomes in various populations, including sociodemographic features (age, gender, employment status), symptom severity, emotion regulation abilities, problem duration, level of functioning, interpersonal problems, prior treatments, comorbidity of personality disorders or medical conditions, treatment non-adherence and alliance [e.g., (6, 7, 21–33)]. However, no consistent pre-treatment characteristics have been identified that reliably predict treatment outcomes (34, 35).

## 1.3. Importance of analyzing longitudinal data from the real-world psychiatric context

Findings about predictors for treatment outcomes often stem from data from randomized controlled trials (RCTs), which may be problematic for several reasons. First, only a selective and limited number of potential predictors are usually included in randomized controlled trials (RCTs), only allowing for limited conclusions about what predictors are relevant for treatment outcomes. Second, RCTs often do not meet the required sample size needed for detecting significant predictors (36–41). Third, many RCTs, especially efficacy trials, tend to have rather strict in- and exclusion criteria and controlled study procedures. While the use of such criteria leads to relatively high internal validity, it may decrease external validity and limit generalizability to patient populations treated in daily clinical practice (42–44).

Considering that most patients are not treated in RCTs, but in naturalistic clinical institutions, using real-world clinical data to identify predictors of treatment outcomes is likely to be more externally valid (45). Large observational studies using data collected in the real-world context may be a valuable alternative to develop more generalizable prediction models (28). Longitudinal routinely collected patient-reported outcome data of psychopathology and well-being are increasingly available that provide information about treatment outcomes [e.g., (46, 47)]. For instance, electronic health records (EHRs) of psychiatric patients contain large amounts of potentially useful clinical information. However, despite increased external validity, such routinely collected data presents challenges as well. Predictive features are heterogeneous and may interact with each other in ways that traditional statistical models may not be able to capture. By including a larger number of features there is also a risk for overfitting. In addition, using real-world data is often challenging, especially due to high attrition and missing data rates.

## 1.4. The potential of machine learning

Recent improvements in computational power and the refinement of the applications of machine learning (ML) technologies have been suggested to offer possibilities to develop robust and generalizable prediction models for treatment response using real-world data (18, 48, 49). ML has shown promise within clinical psychology in helping to understand large-scale health data (50–55). ML is a subfield of artificial intelligence that involves the development of algorithms and statistical models that enable computers to learn and make predictions or decisions based on data without being explicitly programmed to do so (56).

ML can predict treatment effects using high-quality data such as patient characteristics and questionnaire scores over time [e.g., (55, 57, 58)]. The techniques used in building ML models depend on the type of data and can be based on supervised learning, unsupervised learning, and reinforcement learning. Supervised learning, as applied in current study, involves training a model on labeled data, where the desired output is already known. The ultimate goal is to build a model that can accurately predict future outcomes (59). Aafjes-van Doorn, Kamsteeg, Bate, and Aafjes (60) systematically reviewed 51 studies of ML in psychotherapy and concluded that most model development

studies used supervised learning techniques to classify or predict labeled treatment process or outcome data, whereas some used unsupervised techniques to identify clusters in the unlabeled patient or treatment data.

In ML models, the main statistic of interest is the prediction accuracy of the algorithm in a hold-out sample. The hold-out sample is a random subset of the original dataset that is held back and not used during training. For categorical outcomes the accuracy is usually reported as the accuracy, sensitivity (or recall) and specificity, and area under the curve (AUC) computed from the confusion matrix of the predicted against the observed labels of the observations.

Application of ML has various potential advantages above traditional statistical methods. First, by employing robust statistical and probabilistic techniques, ML has the ability to make predictions regarding treatment effects, enabling the comprehension of complex, integrated datasets consisting of heterogeneous features (57, 60). Second, ML methods require less restrictive assumptions regarding the non-linear relationship of high-dimensional data and the skewed distribution of features (61). The potential of ML has been demonstrated by improved accuracy compared to regular methods such as regression (62, 63). Third, the application of cross-validation techniques, which are common in ML methods but usually not applied in traditional prediction analyzes such as significance-based regression, reduces the risk for overfitting (64). Fourth, ML increases the generalizability of the predictions since some ML algorithms might perform better than traditional analysis techniques in complex datasets involving many features (65).

### 1.5. Predicting non-improvement by ML using outcome data of psychopathology and well-being

Real-world mental health data have been used in various ML applications, such as modeling disease progression (66), predicting disease deterioration (67), predicting risk factors for adverse outcomes, such as mortality, readmission or prolonged length of stay (68) as well as predicting treatment outcomes (69). However, research predicting outcomes using real-world clinical data is still scarce. Some studies have shown that compared to traditional research methods ML can increase prediction accuracy using sociodemographic, clinical and biological data (19, 63, 64, 70–74). However, ML has not often been applied to the routine collection of patient-level outcome data in combination with sociodemographic and clinical data.

Hence, the objective of the current study is to evaluate and compare the performance of different ML models in predicting treatment outcomes in an observational sample of patients treated in routine specialized mental healthcare. This will be done by predicting non-improvement in psychopathology 6 months after start of treatment in a group of patients with anxiety and mood disorders. A range of routinely available clinical, demographic and self-reported outcome features will be used to predict treatment outcomes. Several models are explored, such as those involving the incorporation of change scores early in treatment as supplementary predictors, and models that are trained on a reduced set of features using feature reduction techniques.

## 2. Methods

### 2.1. Study design and data collection

The present study concerned an exploratory machine learning based prediction analysis of routinely collected observational longitudinal quantitative data. The recommendations for reporting machine learning analyzes in clinical research (75) were followed. We used data collected in the context of routinely collected patient-level outcome data of psychopathology and well-being, a standardized service to measure treatment effects. Patients in a mental healthcare center in the Netherlands completed online questionnaires every 3 months from the initial interview to end of treatment. Data were collected before start of treatment (T0), and three (T1), six (T2), nine (T3), and 12 (T4) months after treatment commenced. Invitations to complete the questionnaires were sent automatically and data from the completed questionnaires were stored anonymously by an independent data controller in a database generated for this longitudinal study. The data were gathered between March 2015 and November 2019. About 19% ($n = 145$) were lost to 3-month follow-up, 34% ($n = 254$) did not complete the six-month follow-up assessment, and about 58% ($n = 439$) did not complete the 12-month follow-up.

Patients provided passive informed consent for their anonymized data to be used for scientific research. As data were collected in the context of regular care and only anonymized data were analyzed, the study did not require medical ethical approval according to Dutch law. Inclusion criteria were: (1) aged between 18 to 65 years, (2) full completion of the questionnaires on the same day, and (3) diagnosed by depressive, bipolar, anxiety, trauma related or obsessive-compulsive disorder. The diagnosis was based on an extensive interview by a licensed clinical psychologist or psychiatrist. The diagnosis and related (evidence- and practice-based) treatment options were discussed and confirmed in a multidisciplinary team.

### 2.2. Baseline features

An overview of all available baseline features that were included in the models can be found in Table 1. These include sociodemographic (e.g., gender, age), diagnostic (e.g., main diagnosis, comorbidity), and clinical characteristics of patients (e.g., number of treatments in the past, social problems). One additional clinical feature was created that was labeled as treatment intensity. This feature represents the ratio of number of treatments in the past and total duration of past treatments. Routinely collected self-reported psychological features included the total and subscales scores of the Outcome Questionnaire [OQ-45; (76)] and the Mental Health Continuum-Short Form [MHC-SF; (77, 78)]. The OQ-45 is a 45-item self-report measure of psychopathology and includes four subscales, namely symptomatic distress (e.g., "I'm anxious"), interpersonal relations (e.g., "Often I have fights"), somatic complaints (e.g., "I tire quickly"), and social roles performance (e.g., "I feel like I'm not doing well with my work"). Items are answered on a five-point Likert scale ranging from 0 (*never*) to 4 (*almost always*). Previous studies have shown that the OQ-45 is a reliable and valid instrument across different cultural contexts (76, 79, 80). The 14-item MHC-SF measures the presence of different well-being dimensions during the past month on three subscales: emotional (e.g., "Feeling satisfied with

10.3389/fpsyt.2023.1236551

TABLE 1 Overview of baseline features.

| Sociodemographic | Psychological |
|---|---|
| Age | OQ-45 Total score |
| Gender (male/female) | OQ-45 Symptomatic distress |
| Education (low/moderate/high) | OQ-45 Anxiety and somatic distress |
| Marital status (no partner/partner/other) | OQ-45 Interpersonal relationships |
| | OQ-45 Social role adjustment |
| | MHC-SF Total score |
| | MHC-SF Emotional well-being |
| | MHC-SF Social well-being |
| | MHC-SF Psychological well-being |
| | GAF score |
| **Diagnostic** | **Clinical** |
| Main diagnosis (depressive disorder, anxiety disorder, bipolar disorder, OCD, traumatic disorder) | Axis II problem (no/yes) |
| | Axis IV financial problem (no/yes) |
| | Axis IV relationship problems (no/yes) |
| First comorbidity (no/yes) | Axis IV social problems (no/yes) |
| Second comorbidity (no/yes) | Axis IV work problems (no/yes) |
| Somatic comorbidity (no/yes) | Number of treatments in the past (0–4/5–10/10+) |
| | Years since first time enrolled (0–3/3–10/10+) |
| | Sum of previous enrollments in years (0–2/2–5/5+) |
| | Log-transformed treatment intensity[a] |

[a]Treatment intensity was calculated as the ratio of number of treatments in the past and total duration of past treatments.

life"), social (e.g., "Feeling that you belong to a community"), and psychological well-being (e.g., "Feeling that your life as a sense of direction or meaning to it"). Items are answered on a six-point Likert scale ranging from 0 (*never*) to 5 (*every day*). The MHC-SF has shown good psychometric properties in the general population [e.g., (77, 78)] and in clinical groups (81). In total, 41 baseline features were included in the models.

## 2.3. Response variable

Non-improvement on the OQ-45 total scores at six-month follow-up was used as binary response variable. Cases were labeled as 'not improved' if the change from baseline in the symptomatic distress scale of the OQ-45 6 months after baseline was smaller than half a standard deviation (0.5 SD). The choice of this cut-off is motivated by a previous systematic review of 38 studies, suggesting that half a standard deviation consistently reflected a minimally important difference for health-related quality of life instruments across studies (82). Half a standard deviation also corresponds with a medium effect size according to Cohen's conventional rule of thumb (83). The reason to use improvement at six-month follow-up as response variable, was that missing data become too high at later follow-up points and because 6 months was considered a time period long enough to be clinically relevant. Besides, hardly any additional average treatment effects were observed after that time in the dataset.

## 2.4. Preprocessing

Descriptive analyzes were done in the statistical package for social sciences (SPSS) version 27 (84). All other ML analyzes were conducted in R (85) using the caret R-package (86). Data, syntax and output files can be found on the Open Science Framework website (https:osf.io/xwme4/).

All categorical features were dummy coded and continuous features were visually checked for approximate normal distribution. The feature 'treatment intensity' was log-transformed, since it was not normally distributed and right-skewed. Cases that did not complete the OQ-45 at 6 months after baseline were removed. Only complete cases were used, since imputing the response variable might overestimate the performance of the ML algorithms, as common imputation techniques (e.g., random forest) would be similar to what ML algorithms would use to predict non-improvement at follow-up. After data preprocessing and cleaning, the remaining data was randomly split into a training (70%) and hold-out sample (30%). Next, missing baseline data (0.8%) was globally imputed (before conducting k-fold cross-validations) and separately for training and hold-out data, using random forest imputation (87).

## 2.5. Machine learning models and model performance

The goal of ML is to identify patterns in observed high complex data in high dimensional settings, make accurate predictions or classifications, and improve their performance over time by learning from new data [e.g., (57, 58, 63, 88–90)]. ML algorithms involve three main components, which are (1) a model, (2) data for training, testing and validation, and (3) an optimization algorithm. The model represents the data and relationships between features. The training data is used to optimize model weights using cross-validation (CV) to minimize error or loss, while the optimization algorithm finds the optimal values of the model weights. ML algorithms are conducted in two steps: training and testing. During training, the objective is to find a balance between identifying specific patterns in the patient data and preventing overfitting (training data so well that it negatively affects its performance on new data, which occurs when the algorithm fits too closely to the random noise in the data). In the test phase, the accuracy of the predictions made by the algorithm is computed by comparing the predictions made for new data with the actual values observed in the new sample. CV optimizes the ML model by assessing skills of the ML model and testing its performance (or accuracy) in new data later.

Different ML algorithms were compared to predict non-improvement at six-month follow-up. The following algorithms were used: Logistic regression (LR), random forest (RF), support vector machine (SVM) with linear, radial and polynomial kernels, and gradient boosting machine (GBM). These algorithms differ in their underlying principles and modeling techniques. LR focuses on estimating probabilities based on linear relationships, RF combines decision trees for predictions, SVM find optimal hyperplanes for classification, and GBM sequentially build models to minimize prediction error. The rationale for choosing these algorithms was to be able to compare this study with previous studies that used similar algorithms [e.g., (19, 74)]. Furthermore, we not only wanted to

include flexible and less interpretable algorithms (e.g., GBM or SVM), but also techniques that are easier to interpret, while being less flexible (91).

All models were trained on the training set using repeated k-fold cross-validation with 10 folds and 10 repetitions (90). As the response variable was imbalanced, up-sampling was used for training purposes, which randomly replicates instances of the minority class. We explored the effect of class imbalance before applying up-sampling. If no up-sampling was used models performed comparably well in terms of overall accuracy, but were not useful because the sensitivity was extremely high (often higher than 90%), while the specificity was often extremely low (often about 10–20%). We therefore decided to use up-sampling techniques for training the model, in order to create models that are more balanced in terms of sensitivity and specificity.

Using class weights (i.e., imposing a heavier cost for errors made in the minority class) was tested as an alternative to up-sampling, but did not lead to a substantially different performance.

Depending on the model, different hyperparameters were tuned for training the models. For RF models, the number of features used at each split was tuned. For linear SVM, the C hyperparameter was tuned, for SVM with radial basis function kernel the C and sigma parameters were tuned, for SVM with polynomial basis function the C, degree, and scale parameters were tuned, and for GBM number of iterations and complexity of the tree were tuned, while shrinkage and minimum number of training set samples in a node to commence splitting was held constant at 0.1 and 10, respectively. Model training was done in different settings. First, models were fitted that only included baseline features (T0). Second, models were fitted that additionally included three-month change scores in OQ-45 (psychopathology) and MHC-SF (well-being) subscales and total scores. Change scores were included in the second setting, because early improvements in treatment have been shown to be a strong and unique indicator for ongoing improvement at a later moment across a range of psychiatric disorders (92–94). If such a model would perform substantially better, it would be of added value for practice to (additionally) use this model some months after the treatment started to make more accurate predictions.

Third, additional feature reduction was used in both settings, because this might avoid overfitting and lead to better generalizability and increased performance on the test set. The practical usefulness of a model would increase if a reduced set of features yields comparable or even superior performance in predicting non-improvement. Least absolute shrinkage and selection operator regression (LASSO) was used to reduce the number of features. LASSO has the advantage of shrinking less relevant weights to zero, allowing to use it to reduce the number of features (90, 95, 96). In total, this resulted in four settings used for training the models: (1) no change scores and not reduced, (2) no change scores and reduced, (3) change scores and not reduced, and (4) change scores and reduced.

The trained models were then validated in the hold-out sample using a default probability cut-off of 0.5 (82). This means that every case that had a probability higher than 50% of not being improved, was classified as 'not improved'. Performance of all models was evaluated using balanced accuracy, sensitivity, specificity, and area under the curve (AUC). Sensitivity, also known as True Positive Rate (TPR) or recall, focuses on the model's ability to correctly detect positive instances whereas specificity, also known as True Negative Rate (TNR), assesses the model's ability to correctly identify negative instances. Both sensitivity and specificity refer to a specific prediction threshold of the outcome. The AUC, on the other hand, provides a global evaluation, capturing the model's performance across the entire range of threshold choices. AUC thus provides a holistic view of performance, independent of thresholds, making it a valuable measure to assess the overall discriminatory power of our binary classification model (improvement versus non-improvement). Therefore, the AUC was used as the primary evaluation measure in this study. Guidelines for interpreting AUC scores suggest that scores from 0.5 to 0.59 can be seen as extremely poor, from 0.60 to 0.69 as poor, 0.70 to 0.79 as fair, 0.80 to 0.89 as good and > 0.90 as excellent (97).

To be better able to interpret the models and for reasons of conciseness, we additionally determined the top five most important features in the hold-out sample of each model in the four different settings. Feature importance was determined using the varImp evaluation function from the caret package, a generic calculation method and analysis technique for statistical modeling. It evaluates the impact of each predictor feature by assessing how much the model's performance deteriorates when a particular feature is removed. By measuring the relative contribution of the features, it helps in understanding the ranking of influence on the prediction of non-improvement, ensuring further model optimization. Depending on the type of model, different metrics are used to determine feature importance [for an overview, see Kuhn, (86)].

# 3. Results

## 3.1. Sample

At baseline, 755 patients receiving outpatient treatments within multidisciplinary teams consisting of psychologists, psychiatrists, nurses and art therapists, were included in the dataset. Most patients were female, followed lower (43%), intermediate (37.1%) or higher (19.9%) vocational education, and lived with a partner and children (see Table 2). Almost one third had social, relation and/or work problems. The respondents were classified into five common psychopathological groups based on their primary diagnosis: depressive disorder ($n = 417$; 55.2%), bipolar disorder ($n = 79$; 10.5%), anxiety disorder ($n = 114$; 15.1%), trauma related disorder ($n = 115$; 15.2%) or obsessive-compulsive disorder ($n = 30$; 4.0%). Most patients had comorbid disorders ranging from attention deficit hyperactivity disorder (ADHD), depression, anxiety, trauma or addiction, and/or had personality problems respectively: depressive disorder (5.0%; 31.3%), bipolar disorder (6.3%; 1.3%), anxiety disorder (8.8%; 29.8%), trauma related disorder (14.8%; 32.2%) or obsessive-compulsive disorder (OCD; 10.0%; 26.7%).

## 3.2. Psychopathology and well-being per diagnosis over time

For descriptive purposes, Figure 1 shows the average OQ-45 symptomatic distress scale scores over the 12-month time span for

10.3389/fpsyt.2023.1236551

TABLE 2 Major characteristics of respondents ($N = 755$).

| | Depression ($n = 417$) (55.2%) | | Bipolar ($n = 79$) (10.5%) | | Anxiety ($n = 114$) (15.1%) | | Trauma ($n = 115$) (15.2%) | | OCD ($n = 30$) (4.0%) | | Total ($N = 755$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender *n* (%) | | | | | | | | | | | | |
| Male | 190 | (45.6) | 32 | (40.5) | 45 | (39.5) | 35 | (30.4) | 8 | (26.7) | 310 | (41.1) |
| Female | 227 | (54.4) | 47 | (59.5) | 69 | (60.5) | 80 | (69.6) | 22 | (73.3) | 445 | (58.9) |
| Age | | | | | | | | | | | | |
| Mean | 46.0 | | 45.6 | | 39.3 | | 41.0 | | 36.8 | | 43.8 | |
| Range | 20–65 | | 25–64 | | 21–62 | | 19–63 | | 21–65 | | 19–65 | |
| SD | 10.8 | | 10.0 | | 10.4 | | 10.6 | | 12.0 | | 11.1 | |
| Level of education *n* (%)[a] | | | | | | | | | | | | |
| Low | 182 | (46.8) | 13 | (19.1) | 43 | (39.4) | 60 | (53.6) | 6 | (20.7) | 304 | (43.0) |
| Moderate | 143 | (36.8) | 29 | (42.6) | 45 | (41.3) | 33 | (29.5) | 12 | (41.4) | 262 | (37.1) |
| High | 64 | (16.5) | 26 | (38.2) | 21 | (19.3) | 19 | (17.0) | 11 | (37.9) | 141 | (19.9) |
| Marital status *n* (%) | | | | | | | | | | | | |
| Single without children | 77 | (18.9) | 14 | (19.7) | 17 | (14.9) | 30 | (26.5) | 2 | (6.7) | 140 | (19.0) |
| Single with children | 30 | (7.4) | 6 | (8.5) | 13 | (11.4) | 16 | (14.2) | 1 | (3.3) | 66 | (9.0) |
| Married without children | 93 | (22.9) | 12 | (15.2) | 22 | (19.3) | 17 | (15.0) | 9 | (30.0) | 153 | (20.8) |
| Married with children | 161 | (39.6) | 33 | (46.5) | 36 | (31.6) | 33 | (29.2) | 11 | (36.7) | 274 | (37.3) |
| Other | 46 | (11.3) | 6 | (8.5) | 26 | (22.8) | 17 | (15.0) | 7 | (23.3) | 102 | (13.9) |
| Comorbid society problems *n* (%) | | | | | | | | | | | | |
| House problem | 18 | (4.3) | 0 | (0) | 2 | (1.8) | 2 | (1.7) | 2 | (6.7) | 24 | (3.2) |
| Work problem | 112 | (27.0) | 14 | (17.7) | 31 | (27.4) | 29 | (25.2) | 6 | (20.0) | 192 | (25.3) |
| Relation problem | 109 | (26.3) | 3 | (3.8) | 21 | (18.6) | 28 | (24.3) | 3 | (10.0) | 164 | (21.8) |
| Social problem | 126 | (30.4) | 10 | (12.7) | 30 | (26.5) | 33 | (28.7) | 6 | (20.0) | 205 | (27.3) |
| Financial problem | 59 | (14.2) | 1 | (1.3) | 11 | (9.7) | 13 | (11.3) | 3 | (10.0) | 87 | (11.6) |
| Somatic problem | 61 | (14.6) | 3 | (3.8) | 21 | (18.4) | 8 | (7.0) | 2 | (6.7) | 95 | (12.6) |
| Comorbid diagnosis *n* (%) | | | | | | | | | | | | |
| None | 273 | (65.5) | 58 | (73.4) | 67 | (58.8) | 42 | (36.5) | 19 | (63.3) | 459 | (60.8) |
| Two or more | 21 | (5.0) | 5 | (6.3) | 10 | (8.8) | 17 | (14.8) | 3 | (10.0) | 56 | (7.4) |
| Personality problems | 130 | (31.3) | 1 | (1.3) | 34 | (29.8) | 37 | (32.2) | 8 | (26.7) | 225 | (29.8) |
| Nature all comorbid diagnoses *n* (%) | | | | | | | | | | | | |
| ADHD | 24 | (5.8) | 12 | (15.2) | 6 | (5.3) | 16 | (13.9) | 1 | (3.3) | 59 | (7.8) |
| Depression | – | – | – | – | 25 | (21.9) | 27 | (23.5) | 6 | (20.0) | 58 | (7.7) |
| Anxiety | 32 | (7.7) | 0 | (0) | – | – | 4 | (3.5) | 1 | (3.3) | 37 | (4.9) |
| Trauma | 34 | (8.2) | 5 | (6.3) | 6 | (5.3) | – | – | 0 | (0) | 45 | (6.0) |
| Addiction | 19 | (4.6) | 2 | (2.5) | 2 | (1.8) | 6 | (5.2) | 1 | (3.3) | 30 | (4.0) |
| Other | 35 | (8.4) | 2 | (2.5) | 8 | (7.0) | 20 | (17.4) | 2 | (6.7) | 67 | (8.9) |

[a]Low = primary school, lower vocational education; moderate = secondary school, intermediate vocational education; high = higher vocational education, university.

the different diagnostic categories, as well as the percentages of patients who did or did not improve by more than half an SD compared to baseline. For patients with depressive disorder, a continuous improvement from baseline to 12-month follow-up seemed to be present in the total OQ-45 scores. For patients with anxiety disorder, it seemed that on average no improvement was present after six-month follow-up. The binary improvement data suggests that the largest proportion of improvement happened within the first 3 months. The increase in percentage improved after this point seemed very small for all diagnostic groups. The percentage of improved patients in the trauma-related disorder group seemed especially small.

Figure 2 summarizes the course of total well-being scores over the period of 12 months and the proportion of patients that improved

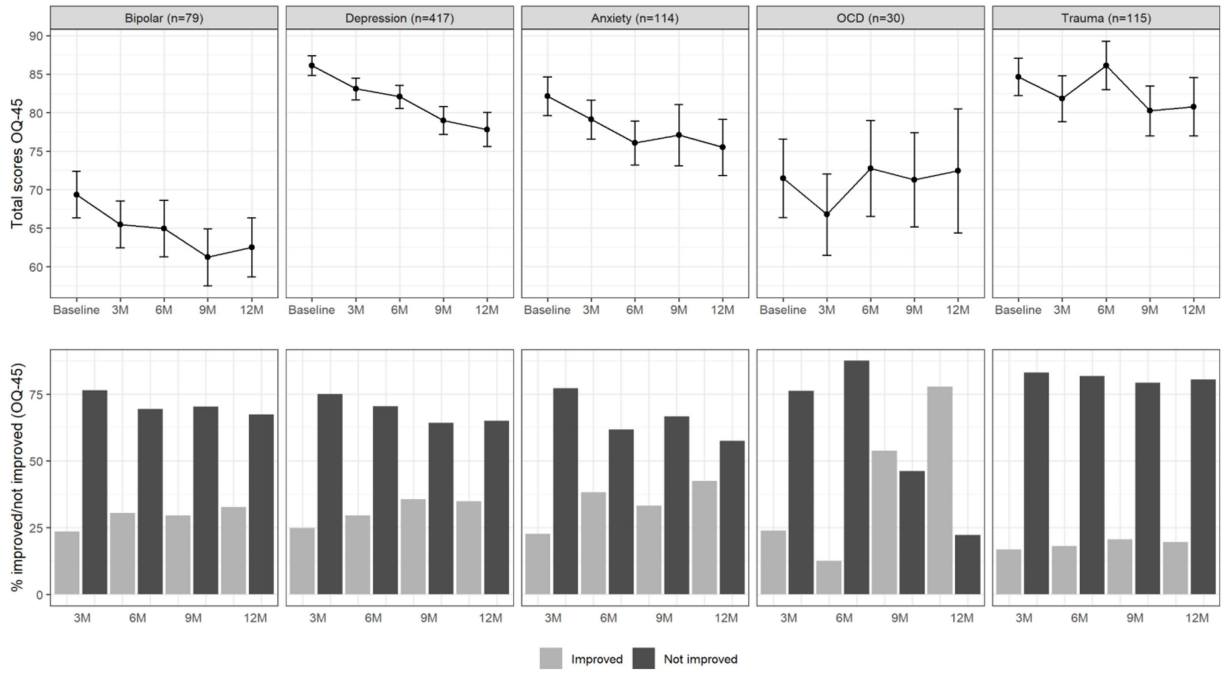



FIGURE 1
Total OQ-45 scores (upper panel) and percentage of improved and not improved patients (lower panel) per diagnosis group and over time. The error bars in the upper panel represent 95% confidence intervals.

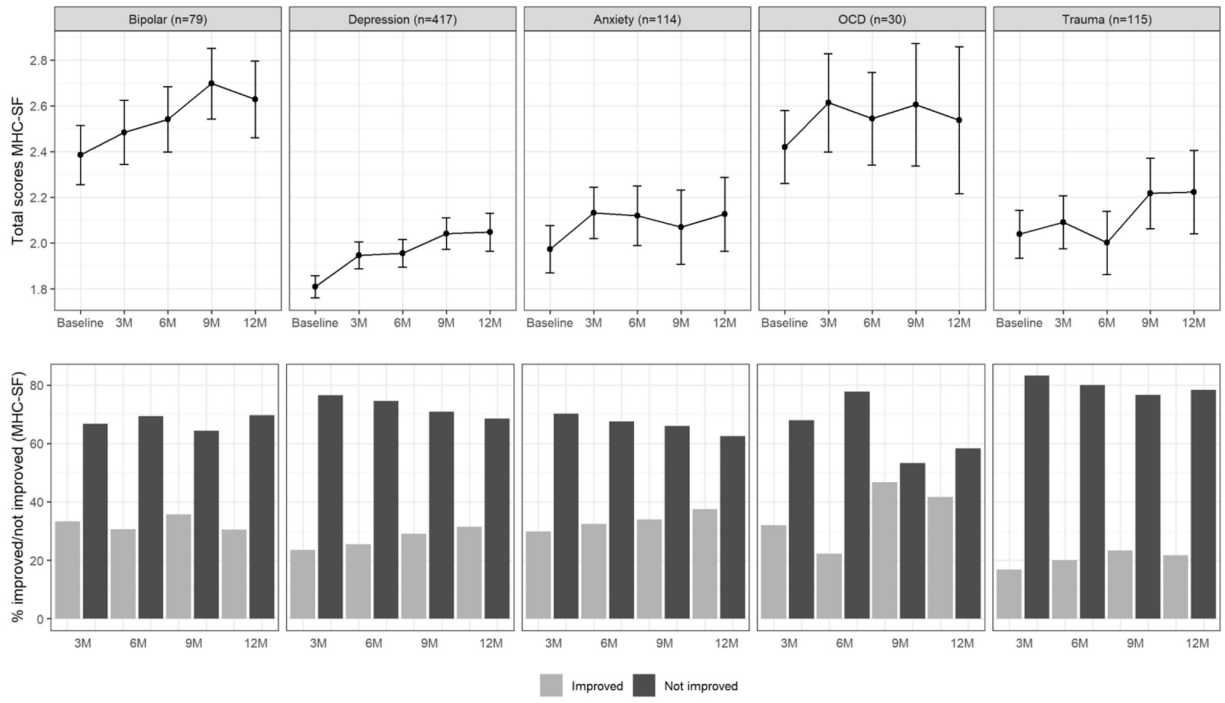



FIGURE 2
Total MHC-SF scores (upper panel) and percentage of improved and not improved patients (lower panel) per diagnosis group and over time. The error bars in the upper panel represent 95% confidence intervals.

(> 0.5 SD) in well-being. Overall, a similar picture emerged. Improvements in well-being appeared to happen mainly within the first 3 months, while the increase in improvements after this point remained rather small.

## 3.3. Feature reduction

Table 3 gives an overview of features that were included in the models without change scores and with change scores after LASSO regression was applied as additional preparatory step. In models without change scores, the only psychological feature of the 16 remaining features after feature reduction was the baseline total score of the OQ-45, while all others were demographic, diagnostic and clinical features. In models with change scores of the remained 13, psychological features of both the OQ-45 and the MHC-SF turned out to be of interest.

## 3.4. Predicting improvement at 6 months

In the training set, 70% of cases did not improve and in the hold-out sample 71% of cases did not improve. An overview of the performance of all models under the four different settings can be found in Table 4. Overall, the models performed best when change scores were included. In settings in which early change scores were included (from 0 to 3 months), the highest overall performance on the training set was obtained (AUC range: 0.79–0.84). The models in this setting also performed best on the hold-out sample (AUC range: 0.69–0.73). The best performing overall model in the hold-out sample in settings with change scores included was gradient boosting (AUC = 0.73). The models performed relatively poor in settings without change scores. In the training set, modest AUC values were found in these settings, ranging from 0.67 to 0.73. The best performance in the hold-out sample when no change scores were included was found for logistic regression (AUC = 0.63). Overall, these findings suggest that including change score substantially improves model performance in this dataset. An overview of all final hyperparameters after model training can be found in Table 5.

Another important comparison included settings in which reduced sets of features were used versus settings in which no reduced sets were used. Overall, the findings suggest that using a reduced set of features seemed to somewhat improve the performance in the training set. Yet, when validating the models on the hold-out sample it seems that using a reduced set of features does not substantially contribute the performance of the models. This indicates that using a reduced set of features does not decrease performance of the models to a relevant degree, suggesting that a reduced set of features might have a similar predictive ability compared with the full set of baseline features. The confusion matrices of the best performing models in the hold-out sample within each setting can be found in Table 6.

TABLE 3 Overview of features that were included after feature reduction was applied using LASSO regression.

| Model without change scores ($k$ = 16) | Model with change scores ($k$ = 13) |
|---|---|
| OQ-45 symptomatic distress | OQ-45 symptomatic distress |
| Gender | Change score OQ-45 interpersonal relations |
| Working problems | Change score OQ-45 somatic complaints |
| Living problems | Change score OQ-45 symptomatic distress |
| Log-transformed treatment intensity[a] | Change score MHC-SF total score |
| Education: moderate | Change score MHC-SF emotional well-being |
| Living situation: no partner | Main diagnosis: trauma |
| Living situation: other | Main diagnosis: anxiety |
| Comorbidity | Second comorbidity |
| Second comorbidity | Living situation: other |
| Main diagnosis: trauma | Working problems |
| Main diagnosis: anxiety | Living problems |
| Sum of previous enrollments in years: 0–2 | Social problems |
| Sum of previous enrollments in years: 5+ | – |
| Number of treatments in the past: 1–4 | – |
| Number of treatments in the past: 5–10 | – |

MHC-SF, Mental Health Continuum-Short Form; OQ-45, Outcome Questionnaire. All change scores refer to change from baseline to 3-month follow-up.

## 3.5. Feature importance

To allow for some interpretation of the models, one last step was to identify the most important features from the models that showed the best performance on the hold-out sample in each setting. An overview of these five most important features can be found in Table 7. It is noteworthy that change scores seem to play a crucial role in the models that include change scores. This, again, suggests that including information about change within the beginning of treatment seems to be valuable when aiming to improve model accuracy. Furthermore, in all settings, except the second setting, both psychopathology and well-being are among the most important features. This indicates that not only psychopathology seems to be of importance when predicting improvement in symptoms, but also well-being.

# 4. Discussion

## 4.1. Main findings

The goal of the current study was to evaluate and compare the performance of different machine learning (ML) models in predicting non-improvement in an observational sample of patients treated in routine specialized mental healthcare. Below, the results are critically discussed in the light of previous research and opportunities for future research.

First, the ML models applied in the current study showed only modest performance in predicting treatment outcomes. Although some previous prediction studies show relatively good predictive results [e.g., (98–100)], most previous studies also indicate modest performance [e.g., (30, 53, 57, 70, 73, 101, 102)]. Some explanations for the modest performance in the current study should be considered. Firstly, 'confounding by indication' could have introduced a bias into the observed association of observed features and non-improvement (103). The decision to assign (intensity of) treatment or adjustments along the way can be influenced by various factors, such as disease severity, previous

10.3389/fpsyt.2023.1236551

TABLE 4 Model performance metrics of the six algorithms under different conditions in the training and hold-out sample.

| | | Training sample ($n$ = 344) | | | | Hold-out sample ($n$ = 146) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Setting | Algorithm | $ACC_{Bal}$ | Sens | Spec | AUC | $ACC_{Bal}$ | Sens | Spec | AUC |
| No change scores, not reduced | Logistic regression | 0.61 | 0.66 | 0.56 | 0.68 | 0.59 | 0.64 | 0.54 | 0.63 |
| | Random forest | 0.62 | 0.67 | 0.57 | 0.67 | 0.52 | 0.59 | 0.44 | 0.58 |
| | SVM (linear) | 0.63 | 0.66 | 0.60 | 0.68 | 0.58 | 0.65 | 0.51 | 0.60 |
| | SVM (radial) | 0.62 | 0.70 | 0.54 | 0.69 | 0.56 | 0.65 | 0.47 | 0.62 |
| | SVM (polynomial) | 0.62 | 0.69 | 0.56 | 0.69 | 0.54 | 0.59 | 0.49 | 0.58 |
| | Gradient boosting | 0.61 | 0.68 | 0.54 | 0.67 | 0.54 | 0.71 | 0.37 | 0.58 |
| No change scores, reduced | Logistic regression | 0.66 | 0.69 | 0.64 | 0.73 | 0.59 | 0.63 | 0.56 | 0.62 |
| | Random forest | 0.65 | 0.68 | 0.62 | 0.71 | 0.56 | 0.61 | 0.51 | 0.58 |
| | SVM (linear) | 0.66 | 0.67 | 0.65 | 0.73 | 0.58 | 0.58 | 0.58 | 0.61 |
| | SVM (radial) | 0.66 | 0.70 | 0.63 | 0.73 | 0.56 | 0.61 | 0.51 | 0.59 |
| | SVM (polynomial) | 0.67 | 0.67 | 0.67 | 0.73 | 0.61 | 0.60 | 0.63 | 0.60 |
| | Gradient boosting | 0.65 | 0.67 | 0.63 | 0.72 | 0.59 | 0.67 | 0.51 | 0.62 |
| Change scores, not reduced | Logistic regression | 0.70 | 0.76 | 0.64 | 0.79 | 0.65 | 0.77 | 0.53 | 0.69 |
| | Random forest | 0.68 | 0.88 | 0.47 | 0.80 | 0.65 | 0.93 | 0.37 | 0.71 |
| | SVM (linear) | 0.72 | 0.76 | 0.67 | 0.80 | 0.67 | 0.76 | 0.58 | 0.69 |
| | SVM (radial) | 0.72 | 0.77 | 0.67 | 0.80 | 0.63 | 0.70 | 0.56 | 0.71 |
| | SVM (polynomial) | 0.72 | 0.75 | 0.68 | 0.81 | 0.63 | 0.73 | 0.53 | 0.68 |
| | Gradient boosting | 0.73 | 0.77 | 0.69 | 0.81 | 0.66 | 0.77 | 0.56 | 0.71 |
| Change scores, reduced | Logistic regression | 0.74 | 0.78 | 0.70 | 0.83 | 0.65 | 0.74 | 0.56 | 0.69 |
| | Random forest | 0.74 | 0.81 | 0.66 | 0.83 | 0.64 | 0.74 | 0.54 | 0.69 |
| | SVM (linear) | 0.74 | 0.77 | 0.71 | 0.84 | 0.66 | 0.74 | 0.58 | 0.69 |
| | SVM (radial) | 0.73 | 0.76 | 0.69 | 0.81 | 0.63 | 0.72 | 0.54 | 0.70 |
| | SVM (polynomial) | 0.74 | 0.76 | 0.71 | 0.84 | 0.67 | 0.76 | 0.58 | 0.70 |
| | Gradient boosting | 0.74 | 0.78 | 0.71 | 0.83 | 0.64 | 0.77 | 0.52 | 0.73 |

treatments, or patient preferences. It is possible that the predictors that drive treatment assignment, in this case confounding features, could have effected the treatment outcome and have made it difficult to assess the true predictive nature of the features considered in this study (103). Secondly, in real-world scenarios, external factors or sources of noise could have affected the outcome and introduced unpredictability. These factors may not be captured by the available features. Accounting for such factors or acquiring additional relevant data might help improve performance. Feature selection, domain expertise, or acquiring additional relevant features can potentially enhance the model's performance. The challenge remains to add the right features predicting treatment success (104). Thirdly, in the current study treatment success is assessed based on subjective self-reported measures. The patient's responses to outcome measures might be influenced by their desire to align their responses with the clinician's expectations. This can result in inflated self-reported outcomes, leading to reduced accuracy in predicting treatment success. People respond inconsistently over time, but algorithms assume no response bias (105). These potential errors undermine prediction. ML techniques *per se* aren't a panacea for higher accuracy without a quality dataset of informative and relative features and domain-specific considerations (106, 107).

Second, more complicated and flexible ML models did not perform substantially better than logistic regression. This is in line with a review of 71 clinical prediction modeling studies (108) and with a recent prediction study of eating disorder treatment response by Espel-Huynh et al. (98). One explanation for this finding might be that the feature set in the current study was not large enough for the more complex models to have an advantage over logistic regression. ML algorithms lead to better performance including in the prevention of the risk of overfit with a greater number of predictors than traditional statistical methods (109). More studies have to be conducted to investigate which model works best in which circumstances (60, 108, 110). Further research into the possibilities of ML methods is still warranted since traditional regression-related approaches have various potential limitations, such as the assumption of straightforward linearity, which may render them less suitable for investigating the complex relational patterns between varied predictors for treatment success in mental healthcare (58, 111).

Third, although still modest, models that included change scores showed the highest overall performance in the hold-out sample, with the gradient boosting model achieving the best overall performance. Models without change scores performed poorly overall. These findings suggest that including change scores substantially improves

TABLE 5 Final hyperparameters used for prediction in the hold-out sample after model training.

| Setting | Algorithm | Hyperparameter |
|---|---|---|
| No change scores, not reduced | Logistic regression | NA |
| | Random forest | mtry = 1 |
| | SVM (linear) | C = 0.01 |
| | SVM (radial) | C = 0.5, sigma = 0.02 |
| | SVM (polynomial) | C = 0.25, degree = 3, scale = 0.01 |
| | Gradient boosting | nTrees = 150, ID = 1, shrinkage = 0.1, NT = 10 |
| No change scores, reduced | Logistic regression | NA |
| | Random forest | mtry = 1 |
| | SVM (linear) | C = 0.01 |
| | SVM (radial) | C = 0.25, sigma = 0.04 |
| | SVM (polynomial) | C = 0.25, degree = 2, scale = 0.01 |
| | Gradient boosting | nTrees = 150, ID = 1, shrinkage = 0.1, NT = 10 |
| Change scores, not reduced | Logistic regression | NA |
| | Random forest | mtry = 2 |
| | SVM (linear) | C = 0.01 |
| | SVM (radial) | C = 0.25, sigma = 0.01 |
| | SVM (polynomial) | C = 0.25, degree = 1, scale = 0.01 |
| | Gradient boosting | nTrees = 50, ID = 1, shrinkage = 0.1, NT = 10 |
| Change scores, reduced | Logistic regression | NA |
| | Random forest | mtry = 1 |
| | SVM (linear) | C = 0.01 |
| | SVM (radial) | C = 0.5, sigma = 0.06 |
| | SVM (polynomial) | C = 0.5, degree = 1, scale = 0.01 |
| | Gradient boosting | nTrees = 100, ID = 1, shrinkage = 0.1, NT = 10 |

C, C-parameter; ID, Interaction depth; mtry, number of features used at each split; NA, Not applicable; nTrees, Number of trees; NT, number of training set samples in a node to commence splitting. For all gradient boosting models, shrinkage and NT were held constant at 0.1 and 10, respectively.

TABLE 6 Confusion matrices of the best performing models in the hold-out sample within each setting.

| | | Reference | |
|---|---|---|---|
| Setting 1: Logistic regression | | Non-improvement | Improvement |
| Predicted | Non-improvement | 66 | 20 |
| | Improvement | 37 | 23 |
| Setting 2: Gradient boosting | | | |
| Predicted | Non-improvement | 69 | 2 |
| | Improvement | 34 | 22 |
| Setting 3: Gradient boosting | | | |
| Predicted | Non-improvement | 79 | 19 |
| | Improvement | 24 | 24 |
| Setting 4: Gradient boosting | | | |
| Predicted | Non-improvement | 79 | 21 |
| | Improvement | 24 | 22 |

prediction performance in this setting. Improvement in the first months has often been found to be related to later treatment success in other studies as well (93, 94) and early change predicts outcome even better than patient characteristics (92, 112, 113). This underscores the relevance of continuous treatment effect monitoring and treatment adjustments in clinical practice.

TABLE 7 Five most important features of the best performing models in each setting.

| | Setting 1: Logistic regression | Setting 2: Gradient boosting | Setting 3: Gradient boosting | Setting 4: Gradient boosting |
|---|---|---|---|---|
| Feature 1 | OQ-45 symptomatic distress | OQ-45 symptomatic distress | Change score OQ-45 symptomatic distress | Change score OQ-45 symptomatic distress |
| Feature 2 | Treatment intensity | Treatment intensity | Change score OQ-45 somatic complaints | Change score OQ-45 somatic complaints |
| Feature 3 | OQ-45 social role performance | Number of previous treatments: 5–10 | OQ-45 symptomatic distress | Change score MHC-SF total score |
| Feature 4 | GAF score | Working problems | Change OQ-45 interpersonal relations | OQ-45 symptomatic distress |
| Feature 5 | MHC-SF total score | Main diagnosis: anxiety | Change score MHC-SF total score | Living problems |

Fourth, the feature-reduced models demonstrated no relevant decrease in performance for predicting treatment outcomes at 6 months in the hold-out sample. Feature-reduced models potentially prevent overfitting and increase generalizability. A trade-off exists between interpretability and accuracy when choosing algorithms. Reducing features also improves the explainability of ML based prediction models. Additional, if a reduced set of features performs equally well (or even better) in predicting non-improvement, it would also increase the practical value and implementability of such a model in daily clinical practice.

Finally, analysis of the feature importance across the different model settings suggested that the most relevant features were the 0–3 month change scores in symptomatic distress, somatic complaints, and well-being, as well as baseline symptomatic distress. The importance of monitoring both the level of psychopathology and well-being in patients with mental health problems has been demonstrated more often (81, 114–118). Crucial predictors found in prior research, including chronicity, comorbidity, interpersonal functioning and familial problems (119), seemed less relevant for predicting non-response in the current study.

For practice, past and present findings underline the importance of searching for additional features to better predict treatment effect in real-world treatment context. Hilbert et al. (73) argued previously that prediction models developed within a diagnostically homogeneous sample are not necessarily superior to a more diverse sample that includes different diagnostic groups. The current study shows that the specific main diagnosis has less predictive value than, for example, early change in treatment effect. After all, where psychiatric patients differ enormously in severity, duration or symptoms of psychopathology and in risk of recurrence, treatments in daily care differ in used methods, assumed mechanisms and appointment frequency. Even within a specific diagnostic group, tailoring psychotherapeutic interventions specifically to the circumstances and characteristics of the patient can improve treatment outcomes (16, 120, 121). Depending on the context and goal of a ML model, one might want to adjust the probability cut-off for predicting non-improvement. We decided to use a probability cut-off of 50% for predicting non-improvement, because we did assume the cost of misspecification to be equal for the positive and negative class. For example, if one wants to aim for a model that has higher sensitivity, lowering the threshold could be desirable.

## 4.2. Strengths and limitations

The current study is one of the first to explore the potential of different machine learning models to predict treatment outcomes in a real-world mental healthcare context using a wide range of routinely available sociodemographic, clinical and patient-reported outcome data. There are however some limitations to the current study that need to be considered.

First, although the current study used a cross-validation approach by randomly splitting the dataset into a training and a test sample, which is the common approach in ML, it should be noted that the study is still exploratory in nature, Although common practice in ML, the test set consisted of a random subset from the same overall patient sample and therefore the study was still limited in its ability to test the generalizability of the final models. Confirmatory studies in independent datasets from different contexts are still necessary to further examine the robustness of the prediction models (122).

Second, in the context of the routine collection of patient-reported outcome data, data is often missing during the course of the treatment process because patients have already improved sufficiently or, on the contrary, have not improved. This missing data is not at random, resulting in the ML algorithms to ultimately relate to a select and biased subpopulation that continues to receive treatment for at least a certain period of time.

Third, the features available in this study consisted largely of self-report data. For the future it would be interesting to incorporate more objective features such as psychological measurements into ML models (123, 124). Future ML studies could improve mental health predictions by adding a unique source of high-frequent and continuous data collecting using multi-modal assessment tools during the period of treatment. mHealth (mobile health) provides individuals real-time biofeedback via sensor apps in everyday devices such as smartphones or wearables on physiological or self-reported behavioral and state parameters, such as heart rate, sleep patterns, physical activity or stress levels (124–126). The combination of ML and mHealth, despite challenges in dimensionality, ethics, privacy and security, shows promise as a clinical tool for monitoring populations at risk and forms the basis for the next generation of mHealth interventions (124, 125).

Finally, though the chosen criterion of 0.5 SD for non-improvement is often used [e.g., (127–131)], a disadvantage is

that this cutoff is sample-dependent. Also, an improvement of 0.5 SD does not necessarily mean that a patient has recovered in such a way that (s)he no longer has clinically relevant complaints. Future research could consider to use the Jacobson-Truax concept of the Reliable Change Index (RCI), which considers the reliability of the improvement in the context of the overall distribution that the patient is likely to belong to post-treatment (132). Patients moving reliably into the functional distribution are *recovered*. Patients are considered to have *improved* if they have made a reliable change but remain in the dysfunctional population, *unchanged* if they have not made a reliable change, and *deteriorated* if they have reliably worsened (132).

### 4.3. Clinical implications and recommendations for future research

Some recommendations can be made for future research. On the one hand, the use of sophisticated psychological data with relevant features according to the latest theoretical models may increase predictions and thereby improve decision-making on therapy indication. This could include the therapeutic relationship as a known predictor of interest (133) diagnosis specific questionnaires in addition to generics, which could mean that the case for a transdiagnostic approach may not yet have been settled, or program-specific questionnaires, appropriate to the therapy offered. On the other hand, the development of more advanced tools is necessary to detect predictors for treatment response based on high-dimensional patient data (134). Based on current research, practitioners might decide to stop or adjust a treatment. In the future, it is desirable that patients can be indicated in a more targeted manner. After all, at present ML approaches cannot yet contribute to specific individualized clinical judgments (135). We would encourage future studies to develop predictors over rather broad diagnostic patient groups and not exclude features in advance, but use the full potential of information available in patient EHRs (136). Interestingly, ML techniques offer the opportunity to study patients who are underrepresented in RCTs.

Additionally, ML has the potential to benefit mental healthcare as it can account for the interaction between many features (137). The ML techniques are suitable to detect features with the strongest predictive influence in different contexts and mutual interactions, thereby providing a combined measure of both individual and multivariate impact of each feature (138). Subsequently, based on findings, the number of features to be implemented in daily care can be substantially reduced.

To reduce response bias, improve the predictive performance of the model, and provide a more comprehensive picture of treatment success, it may be helpful to consider multiple perspectives and assessment sources. In addition, it is important to recognize and address the potential discrepancies between the assessments of different stakeholders (e.g., clinician and patient) when defining the criterion for treatment success in predictive studies.

As change scores in both psychopathology and well-being proved relevant, implementing change measurements in ML applications could be more standardized. Therefore, for future studies, we recommend that in addition to predicting changes in psychopathology, algorithms to predict non-improvements in well-being and other domain/construction should be included. Also, adding multiple change scores, such as living conditions in daily activities and social relationships, or compliance with homework-related adherence could be relevant (139). Adding other data modalities, such as the relationship with the patient's life story, or test data could also improve prediction performance (140, 141). In any case, it is advisable to closely monitor changes in psychopathology and well-being in clinical practice and decision making from the very beginning, so that timely adjustments can be made in the therapy of non-responders. Tiemens et al. (142) recommend doing this at least 4 weeks after starting treatment. The measurement of change scores is also important because the use of feedback based on these evaluations in itself has a positive effect on complaint reduction and it can shorten the duration of treatment (143, 144).

Finally, applying both ML and traditional statistical approaches in the same study allows for comparisons (109, 145). By learning from unique strengths and limitations of different ML algorithms, future ML research can contribute to increasingly accurate predictions (146).

## 5. Conclusion

In the current study we applied ML techniques in a real-world mental healthcare patient population to predict non-improvement using sociodemographic, psychological, diagnostic and clinical data. The overall conclusion is that working with a reduced set of data, and implementing early change scores and relatively simple models gives the best results, both in terms of accuracy and broader in interpretability and applicability. Our results show that ML can be used as a step to indicate treatment change in an early stage of treatment, where it seems to be important to use psychopathology and well-being as important features. The results are encouraging and provide an important step to use patient specific and routine collected patient-level outcome data in clinical practice to help individual patients and clinicians select the right treatments. ML may help to bridge the gap between science and practice. None of the ML applications were developed to replace the clinician, but instead were designed to advance the clinicians' skills and treatment outcome (147). ML might become part of evidence-based practice, as a source of valuable information in addition to clinical knowledge and existing research evidence.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Data, syntax and output files can be found on the Open Science Framework website (https://osf.io/xwme4/).

## Ethics statement

Ethical approval was not required for the studies involving humans because as data were collected in the context of regular care and only anonymized data were analyzed, the study did not require medical ethical approval according to Dutch law. 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

KF collected the data, organized the database, and wrote the first draft of the manuscript. KF, JK, and PK contributed substantial to the conception, design and manuscript draft, and ensuring that the work was appropriately investigated and resolved. JK implemented machine learning algorithms and statistical analysis and wrote the first method section of the manuscript. EB and GW supervised and critically reviewed the manuscript. 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. World Health Organization. *World mental health report: Transforming mental health for all*. Geneva: World Health Organization (2022).

2. Whiteford HA, Ferrari AJ, Degenhardt L, Feigin V, Vos T. The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. *PLoS One*. (2015) 10:e0116820. doi: 10.1371/journal.pone.0116820

3. Chisholm D, Sweeny K, Sheehan P, Rasmussen B, Smit F, Cuijpers P, et al. Scaling-up of treatment of depression and anxiety–Authors' reply. *Lancet Psychiatry*. (2016) 3:603–4. doi: 10.1016/S2215-0366(16)30131-6

4. Jennings C, Singh B, Oni H, Mazzacano A, Maher C. A needs assessment for self-management services for adults awaiting community-based mental health services. *BMC Public Health*. (2023) 23:1–10. doi: 10.1186/s12889-023-15382-8

5. Deisenhofer AK, Delgadillo J, Rubel JA, Boehnke JR, Zimmermann D, Schwartz B, et al. Individual treatment selection for patients with posttraumatic stress disorder. *Depress Anxiety*. (2018) 35:541–50. doi: 10.1002/da.22755

6. Delgadillo J, Huey D, Bennett H, McMillan D. Case complexity as a guide for psychological treatment selection. *J Consult Clin Psychol*. (2017) 85:835–53. doi: 10.1037/ccp0000231

7. Lambert MJ, Hansen NB, Finch AE. Patient-focused research: using patient outcome data to enhance treatment effects. *J Consult Clin Psychol*. (2001) 69:159–72. doi: 10.1037/0022-006X.69.2.159

8. Ægisdóttir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, et al. The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction. *Couns Psychol*. (2006) 34:341–82. doi: 10.1177/0011000005285875

9. Carvalho A, McIntyre R. *Treatment-resistant mood disorders*. Oxford: Oxford Psychiatry Library (2015).

10. Cuijpers P, Karyotaki E, Weitz E, Andersson G, Hollon SD, van Straten A. The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *J Affect Disord*. (2014) 159:118–26. doi: 10.1016/j.jad.2014.02.026

11. De Vos JA, LaMarre A, Radstaak M, Bijkerk CA, Bohlmeijer ET, Westerhof GJ. Identifying fundamental criteria for eating disorder recovery: a systematic review and qualitative meta-analysis. *J Eat Disord*. (2017) 5:1–14. doi: 10.1186/s40337-017-0164-0

12. Driessen E, Van HL, Don FJ, Peen J, Kool S, Westra D, et al. The efficacy of cognitive-behavioral therapy and psychodynamic therapy in the outpatient treatment of major depression: a randomized clinical trial. *Am J Psychiatr*. (2013) 170:1041–50. doi: 10.1176/appi.ajp.2013.12070899

13. Barth J, Munder T, Gerger H, Nuesch E, Trelle S, Znoj H, et al. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLoSMedicine*. (2013) 10:e1001454. doi: 10.1371/journal.pmed.1001454

14. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJ. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry*. (2016) 15:245–58. doi: 10.1002/wps.20346

15. Mangolini VI, Andrade LH, Lotufo-Neto F, Wang Y-P. Treatment of anxiety disorders in clinical practice: a critical overview of recent systematic evidence. *Clinics*. (2019) 74:e1316. doi: 10.6061/clinics/2019/e1316

16. Norcross JC, Lambert MJ. Psychotherapy relationships that work III. *Psychotherapy*. (2018) 55:303–15. doi: 10.1037/pst0000193

17. Bolton D. Overdiagnosis problems in the DSM-IV and the new DSM-5: can they be resolved by the distress–impairment criterion? *Can J Psychiatry*. (2013) 58:612–7. doi: 10.1177/070674371305801106

18. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry*. (2018) 3:223–30. doi: 10.1016/j.bpsc.2017.11.007

19. 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. *Psychother Res*. (2021) 31:52–62. doi: 10.1080/10503307.2020.1839140

20. Haug T, Nordgreen T, Öst L-G, Kvale G, Tangen T, Andersson G, et al. Stepped care versus face-to–face cognitive behavior therapy for panic disorder and social anxiety disorder: predictors and moderators of outcome. *Behav Res Ther*. (2015) 71:76–89. doi: 10.1016/j.brat.2015.06.002

21. Cloitre M, Petkova E, Su Z, Weiss BJ. Patient characteristics as a moderator of posttraumatic stress disorder treatment outcome: combining symptom burden and strengths. *BJPsych open*. (2016) 2:101–6. doi: 10.1192/bjpo.bp.115.000745

22. Cohen M, Beard C, Björgvinsson T. Examining patient characteristics as predictors of patient beliefs about treatment credibility and expectancies for treatment outcome. *J Psychother Integr*. (2015) 25:90–9. doi: 10.1037/a0038878

23. Flückiger C, Del Re A, Wlodasch D, Horvath AO, Solomonov N, Wampold BE. Assessing the alliance–outcome association adjusted for patient characteristics and treatment processes: a meta-analytic summary of direct comparisons. *J Couns Psychol*. (2020) 67:706–11. doi: 10.1037/cou0000424

24. Goddard E, Wingrove J, Moran P. The impact of comorbid personality difficulties on response to IAPT treatment for depression and anxiety. *Behav Res Ther*. (2015) 73:1–7. doi: 10.1016/j.brat.2015.07.006

25. Gregertsen EC, Mandy W, Kanakam N, Armstrong S, Serpell L. Pre-treatment patient characteristics as predictors of drop-out and treatment outcome in individual and family therapy for adolescents and adults with anorexia nervosa: a systematic review and meta-analysis. *Psychiatry Res*. (2019) 271:484–501. doi: 10.1016/j.psychres.2018.11.068

26. Hamilton KE, Dobson KS. Cognitive therapy of depression: pretreatment patient predictors of outcome. *Clin Psychol Rev*. (2002) 22:875–93. doi: 10.1016/S0272-7358(02)00106-X

27. Hoyer J, Wiltink J, Hiller W, Miller R, Salzer S, Sarnowsky S, et al. Baseline patient characteristics predicting outcome and attrition in cognitive therapy for social phobia: results from a large multicentre trial. *Clin Psychol Psychother*. (2016) 23:35–46. doi: 10.1002/cpp.1936

28. Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner L, Ebert D, et al. Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiol Psychiatr Sci*. (2017) 26:22–36. doi: 10.1017/S2045796016000020

29. Knopp J, Knowles S, Bee P, Lovell K, Bower P. A systematic review of predictors and moderators of response to psychological therapies in OCD: do we have enough empirical evidence to target treatment? *Clin Psychol Rev*. (2013) 33:1067–81. doi: 10.1016/j.cpr.2013.08.008

30. Lutz W, Lambert MJ, Harmon SC, Tschitsaz A, Schürch E, Stulz N. The probability of treatment success, failure and duration—what can be learned from empirical data to support decision making in clinical practice? *Clin Psychol Psychother*. (2006) 13:223–32. doi: 10.1002/cpp.496

31. Mululo SCC, de Menezes GB, Vigne P, Fontenelle LF. A review on predictors of treatment outcome in social anxiety disorder. *Braz J Psychiatry*. (2012) 34:92–100. doi: 10.1590/S1516-44462012000100016

32. Salomonsson S, Santoft F, Lindsäter E, Ejeby K, Ingvar M, Öst L-G, et al. Predictors of outcome in guided self-help cognitive behavioural therapy for common mental disorders in primary care. *Cogn Behav Ther*. (2020) 49:455–74. doi: 10.1080/16506073.2019.1669701

33. Sarter L, Heider J, Witthöft M, Rief W, Kleinstäuber M. Using clinical patient characteristics to predict treatment outcome of cognitive behavior therapies for individuals with medically unexplained symptoms: a systematic review and meta-analysis. *Gen Hosp Psychiatry*. (2022) 77:11–20. doi: 10.1016/j.genhosppsych.2022.03.001

34. Eskildsen A, Hougaard E, Rosenberg NK. Pre-treatment patient variables as predictors of drop-out and treatment outcome in cognitive behavioural therapy for social phobia: a systematic review. *Nord J Psychiatry*. (2010) 64:94–105. doi: 10.3109/08039480903426929

35. Haby MM, Donnelly M, Corry J, Vos T. Cognitive behavioural therapy for depression, panic disorder and generalized anxiety disorder: a meta-regression of factors that may predict outcome. *Aust N Z J Psychiatry*. (2006) 40:9–19. doi: 10.1080/j.1440-1614.2006.01736.x

36. Aguinis H, Beaty JC, Boik RJ, Pierce CA. Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review. *J Appl Psychol*. (2005) 90:94–107. doi: 10.1037/0021-9010.90.1.94

37. Luedtke A, Sadikova E, Kessler RC. Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clin Psychol Sci*. (2019) 7:445–61. doi: 10.1177/2167702618815466

38. Ribeiro DC, Milosavljevic S, Abbott JH. Sample size estimation for cluster randomized controlled trials. *Musculoskeletal Sci Pract*. (2018) 34:108–11. doi: 10.1016/j.msksp.2017.10.002

39. Rothwell JC, Julious SA, Cooper CL. A study of target effect sizes in randomised controlled trials published in the health technology assessment journal. *Trials*. (2018) 19:1–13. doi: 10.1186/s13063-018-2886-y

40. Tam W, Lo K, Woo B. Reporting sample size calculations for randomized controlled trials published in nursing journals: a cross-sectional study. *Int J Nurs Stud*. (2020) 102:103450. doi: 10.1016/j.ijnurstu.2019.103450

41. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Stat Methods Med Res*. (2016) 25:1057–73. doi: 10.1177/0962280215588241

42. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard–lessons from the history of RCTs. *Mass Medical Soc*. (2016) 374:2175–81. doi: 10.1056/NEJMms1604593

43. Frieden TR. Evidence for health decision making—beyond randomized, controlled trials. *N Engl J Med*. (2017) 377:465–75. doi: 10.1056/NEJMra1614394

44. Rothwell PM. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *Lancet*. (2005) 365:82–93. doi: 10.1016/S0140-6736(04)17670-8

45. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. *Prev Sci*. (2015) 16:475–85. doi: 10.1007/s11121-014-0513-z

46. Saunders R, Cape J, Fearon P, Pilling S. Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. *J Affect Disord*. (2016) 197:107–15. doi: 10.1016/j.jad.2016.03.011

47. Stochl J, Soneson E, Stuart F, Fritz J, Walsh AE, Croudace T, et al. Determinants of patient-reported outcome trajectories and symptomatic recovery in improving access to psychological therapies (IAPT) services. *Psychol Med*. (2022) 52:3231–40. doi: 10.1017/S0033291720005395

48. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. (2019) 380:1347–58. doi: 10.1056/NEJMra1814259

49. Theobald O. *Machine learning for absolute beginners: A plain English introduction*, vol. *157*. London, UK: Scatterplot Press (2017).

50. Coppersmith G, Ngo K, Leary R, Wood A. *Exploratory analysis of social media prior to a suicide attempt*. Paper presented at the Proceedings of the third workshop on computational linguistics and clinical psychology. (2016).

51. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. *Discovering shifts to suicidal ideation from mental health content in social media*. Paper presented at the Proceedings of the 2016 CHI conference on human factors in computing systems. (2016).

52. Galatzer-Levy IR, Karstoft K-I, Statnikov A, Shalev AY. Quantitative forecasting of PTSD from early trauma responses: a machine learning application. *J Psychiatr Res*. (2014) 59:68–76. doi: 10.1016/j.jpsychires.2014.08.017

53. Haynos AF, Wang SB, Lipson S, Peterson CB, Mitchell JE, Halmi KA, et al. Machine learning enhances prediction of illness course: a longitudinal study in eating disorders. *Psychol Med*. (2021) 51:1392–402. doi: 10.1017/S0033291720000227

54. Jaques N, Taylor S, Nosakhare E, Sano A, Picard R. *Multi-task learning for predicting health, stress, and happiness*. Paper presented at the NIPS Workshop on Machine Learning for Healthcare. (2016).

55. Shatte AB, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. *Psychol Med*. (2019) 49:1426–48. doi: 10.1017/S0033291719000151

56. Mitchell TM. Does machine learning really work? *AI Mag*. (1997) 18:11–1.

57. Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *J Psychiatr Res*. (2016) 78:94–102. doi: 10.1016/j.jpsychires.2016.03.016

58. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. (2015) 349:255–60. doi: 10.1126/science.aaa8415

59. Chen EE, Wojcik SP. A practical guide to big data research in psychology. *Psychol Methods*. (2016) 21:458–74. doi: 10.1037/met0000111

60. 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

61. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability machines. *Methods Inf Med*. (2012) 51:74–81. doi: 10.3414/ME00-01-0052

62. 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 Psychiatr*. (2017) 174:154–62. doi: 10.1176/appi.ajp.2016.16010077

63. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci*. (2017) 5:457–69. doi: 10.1177/2167702617691560

64. 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

65. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect Psychol Sci*. (2017) 12:1100–22. doi: 10.1177/1745691617693393

66. Lim B, Van der Schaar M. *Disease-atlas: Navigating disease trajectories using deep learning*. Paper presented at the Machine Learning for Healthcare Conference. (2018).

67. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. (2019) 572:116–9. doi: 10.1038/s41586-019-1390-1

68. Alaa AM, van der Schaar M. Prognostication and risk factors for cystic fibrosis via automated machine learning. *Sci Rep*. (2018) 8:11242. doi: 10.1038/s41598-018-29523-2

69. Athreya AP, Neavin D, Carrillo-Roa T, Skime M, Biernacka J, Frye MA, et al. Pharmacogenomics-driven prediction of antidepressant treatment outcomes: a machine-learning approach with multi-trial replication. *Clin Pharmacol Ther*. (2019) 106:855–65. doi: 10.1002/cpt.1482

70. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. (2016) 3:243–50. doi: 10.1016/S2215-0366(15)00471-X

71. Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I. *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India*. (2018).

72. Gong X, Hu M, Basu M, Zhao L. Heterogeneous treatment effect analysis based on machine-learning methodology. *CPT Pharmacometrics Syst Pharmacol*. (2021) 10:1433–43. doi: 10.1002/psp4.12715

73. Hilbert K, Kunas SL, Lueken U, Kathmann N, Fydrich T, Fehm L. Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: a machine learning approach. *Behav Res Ther*. (2020) 124:103530. doi: 10.1016/j.brat.2019.103530

74. Van Mens K, De Schepper C, Wijnen B, Koldijk SJ, Schnack H, De Looff P, et al. Predicting future suicidal behaviour in young adults, with different machine learning techniques: a population-based longitudinal study. *J Affect Disord*. (2020) 271:169–77. doi: 10.1016/j.jad.2020.03.081

75. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. *Circ Cardiovasc Qual Outcomes*. (2020) 13:e006556. doi: 10.1161/CIRCOUTCOMES.120.006556

76. De Jong K, Nugter MA, Polak MG, Wagenborg JE, Spinhoven P, Heiser WJ. The outcome questionnaire (OQ-45) in a Dutch population: a cross-cultural validation. *Clin Psychol Psychother*. (2007) 14:288–301. doi: 10.1002/cpp.529

77. Keyes CL, Wissing M, Potgieter JP, Temane M, Kruger A, Van Rooy S. Evaluation of the mental health continuum–short form (MHC–SF) in setswana-speaking south Africans. *Clin Psychol Psychother*. (2008) 15:181–92. doi: 10.1002/cpp.572

78. Lamers SMA, Westerhof GJ, Bohlmeijer ET, ten Klooster PM, Keyes CL. Evaluating the psychometric properties of the mental health continuum-short form (MHC-SF). *J Clin Psychol*. (2011) 67:99–110. doi: 10.1002/jclp.20741

79. De Jong K, Nugter A. De Outcome Questionnaire: psychometrische kenmerken van de Nederlandse vertaling. *Ned Tijdschr Psychol Grensgebieden*. (2004) 59:77–80. doi: 10.1007/BF03062326

80. De Jong K, Spinhoven P. De Nederlandse versie van de Outcome Questionnaire (OQ-45): een crossculturele validatie. *Psychol Gezond*. (2008) 36:35–45. doi: 10.1007/BF03077465

81. Franken K, Lamers SM, ten Klooster PM, Bohlmeijer ET, Westerhof GJ. Validation of the mental health continuum-short form and the dual continua model of well-being and psychopathology in an adult mental health setting. *J Clin Psychol*. (2018) 74:2187–202. doi: 10.1002/jclp.22659

82. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. (2003) 41:582–92. doi: 10.1097/01.MLR.0000062554.74615.4C

83. Cohen J. Statistical power analysis for the behavioral sciences: Jacob Cohen. *J Am Stat Assoc*. (1988) 84:19–74.

84. IBM Corp. *Released 2013. IBM SPSS statistics for windows, Version 27.0*. Armonk, NY: IBM Corp. (2020).

85. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (2020).

86. Kuhn M. *Caret: Classification and regression training Version 6.0–86*. (2020). Available at: https://CRAN.R-project.org/package=caret.

87. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. (2012) 28:112–8. doi: 10.1093/bioinformatics/btr597

88. Bishop CM, Nasrabadi NM. *Pattern recognition and machine learning*, vol. *4*. Berlin: Springer (2006).

89. 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

90. Kuhn M, Johnson K. *Applied predictive modeling*, vol. *26*. Berlin: Springer (2013).

91. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*, vol. *103*. Berlin: Springer (2013).

92. Rubel J, Lutz W, Schulte D. Patterns of change in different phases of outpatient psychotherapy: a stage-sequential pattern analysis of change in session reports. *Clin Psychol Psychother*. (2015) 22:1–14. doi: 10.1002/cpp.1868

93. Schibbye P, Ghaderi A, Ljótsson B, Hedman E, Lindefors N, Rück C, et al. Using early change to predict outcome in cognitive behaviour therapy: exploring timeframe, calculation method, and differences of disorder-specific versus general measures. *PLoS One*. (2014) 9:e100614. doi: 10.1371/journal.pone.0100614

94. Wilson GT. Rapid response to cognitive behavior therapy. *Clin Gastroenterol Hepatol*. (1999) 6:289–92. doi: 10.1093/clipsy.6.3.289

95. Fonti V, Belitser E. *Feature selection using lasso*. VU Amsterdam research paper in business analytics, pp. 1–25. (2017).

96. Muthukrishnan R, Rohini R. *LASSO: a feature selection technique in predictive modeling for machine learning*. Paper presented at the 2016 IEEE international conference on advances in computer applications (ICACA). (2016).

97. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*, vol. *398*. New York: John Wiley and Sons (2013).

98. Espel-Huynh H, Zhang F, Thomas JG, Boswell JF, Thompson-Brenner H, Juarascio AS, et al. Prediction of eating disorder treatment response trajectories via machine learning does not improve performance versus a simpler regression approach. *Int J Eat Disord*. (2021) 54:1250–9. doi: 10.1002/eat.23510

99. Lenhard F, Sauer S, Andersson E, Månsson KN, Mataix-Cols D, Rück C, et al. Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: a machine learning approach. *Int J Methods Psychiatr Res*. (2018) 27:e1576. doi: 10.1002/mpr.1576

100. Tymofiyeva O, Yuan JP, Huang C-Y, Connolly CG, Blom EH, Xu D, et al. Application of machine learning to structural connectome to predict symptom reduction in depressed adolescents with cognitive behavioral therapy (CBT). *NeuroImage*. (2019) 23:101914. doi: 10.1016/j.nicl.2019.101914

101. Ball TM, Stein MB, Ramsawh HJ, Campbell-Sills L, Paulus MP. Single-subject anxiety treatment outcome prediction using functional neuroimaging. *Neuropsychopharmacology*. (2014) 39:1254–61. doi: 10.1038/npp.2013.328

102. Lutz W, Saunders SM, Leon SC, Martinovich Z, Kosfelder J, Schulte D, et al. Empirically and clinically useful decision making in psychotherapy: differential predictions with treatment response models. *Psychol Assess*. (2006) 18:133–41. doi: 10.1037/1040-3590.18.2.133

103. Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. *JAMA*. (2016) 316:1818–9. doi: 10.1001/jama.2016.16435

104. Flach P. *Machine learning: The art and science of algorithms that make sense of data*. Cambridge, UK: Cambridge University Press (2012).

105. Eikelenboom M, Smit JH, Beekman AT, Kerkhof AJ, Penninx BW. Reporting suicide attempts: consistency and its determinants in a large mental health study. *Int J Methods Psychiatr Res*. (2014) 23:257–66. doi: 10.1002/mpr.1423

106. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and machine learning forecasting methods: concerns and ways forward. *PLoS One*. (2018) 13:e0194889. doi: 10.1371/journal.pone.0194889

107. Zhang GP. Avoiding pitfalls in neural network research. *IEEE Trans Syst Man Cybern Part C Appl Rev*. (2006) 37:3–16. doi: 10.1109/TSMCC.2006.876059

108. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. (2019) 110:12–22. doi: 10.1016/j.jclinepi.2019.02.004

109. Schwartz B, Cohen ZD, Rubel JA, Zimmermann D, Wittmann WW, Lutz W. Personalized treatment selection in routine care: integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychother Res*. (2021) 31:33–51. doi: 10.1080/10503307.2020.1769219

110. Jacobucci R, Littlefield AK, Millner AJ, Kleiman E, Steinley D. *Pairing machine learning and clinical psychology: How you evaluate predictive performance matters*. United States: OSF Storage (2020).

111. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: applications and ethics. *Behav Sci Law*. (2019) 37:214–22. doi: 10.1002/bsl.2392

112. Cooper M. *Essential research findings in counselling and psychotherapy*. Thousand Oaks, CA: Sage Publication, pp. 1–256. (2008).

113. Haas E, Hill RD, Lambert MJ, Morrell B. Do early responders to psychotherapy maintain treatment gains? *J Clin Psychol*. (2002) 58:1157–72. doi: 10.1002/jclp.10044

114. Bolier L, Haverman M, Westerhof G, Riper H, Smit F. Positive psychology interventions: a meta-analysis of randomized controlled studies. *BMC Public Health*. (2013) 13:20. doi: 10.1186/1471-2458-13-119

115. Iasiello M, Van Agteren J, Schotanus-Dijkstra M, Lo L, Fassnacht DB, Westerhof GJ. Assessing mental wellbeing using the mental health continuum—short form: a systematic review and meta-analytic structural equation modelling. *Clin Psychol Sci Pract*. (2022) 29:442–56. doi: 10.1037/cps0000074

116. Kraiss JT, Peter M, Moskowitz JT, Bohlmeijer ET. The relationship between emotion regulation and well-being in patients with mental disorders: a meta-analysis. *Compr Psychiatry*. (2020) 102:152189. doi: 10.1016/j.comppsych.2020.152189

117. Slade M. Mental illness and well-being: the central importance of positive psychology and recovery approaches. *BMC Health Serv Res*. (2010) 10:1–14. doi: 10.1186/1472-6963-10-26

118. Trompetter H, Lamers S, Westerhof GJ, Fledderus M, Bohlmeijer ET. Both positive mental health and psychopathology should be monitored in psychotherapy: confirmation for the dual-factor model in acceptance and commitment therapy. *Behav Res Ther*. (2017) 91:58–63. doi: 10.1016/j.brat.2017.01.008

119. Vall E, Wade TD. Predictors of treatment outcome in individuals with eating disorders: a systematic review and meta-analysis. *Int J Eat Disord*. (2015) 48:946–71. doi: 10.1002/eat.22411

120. Norcross JC, Wampold BE. What works for whom: tailoring psychotherapy to the person. *J Clin Psychol*. (2011) 67:127–32. doi: 10.1002/jclp.20764

121. Vermote R, Lowyck B, Luyten P, Verhaest Y, Vertommen H, Vandeneede B, et al. Patterns of inner change and their relation with patient characteristics and outcome in a psychoanalytic hospitalization-based treatment for personality disordered patients. *Clin Psychol Psychother*. (2011) 18:303–13. doi: 10.1002/cpp.713

122. Sammut C, Webb GI. *Encyclopedia of machine learning*. Berlin: Springer Science and Business Media (2011).

123. Derks YP. *Alexithymia in borderline personality pathology: From theory to a biosensor application*. Netherlands: University of Twente (2022).

124. Terhorst Y, Knauer J, Baumeister H. Smart sensing enhanced diagnostic expert systems In: C Montag and H Baumeister, editors. *Digital phenotyping and Mobile sensing: New developments in Psychoinformatics*. Berlin: Springer (2022). 413–25.

125. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu Rev Clin Psychol*. (2017) 13:23–47. doi: 10.1146/annurev-clinpsy-032816-044949

126. Moshe I, Terhorst Y, Opoku Asare K, Sander LB, Ferreira D, Baumeister H, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. *Front Psych*. (2021) 12:625247. doi: 10.3389/fpsyt.2021.625247

127. Azizoddin DR, Jolly M, Arora S, Yelin E, Katz P. Longitudinal study of fatigue, stress, and depression: role of reduction in stress toward improvement in fatigue. *Arthritis Care Res*. (2020) 72:1440–8. doi: 10.1002/acr.24052

128. Chan KS, Friedman LA, Bienvenu OJ, Dinglas VD, Cuthbertson BH, Porter R, et al. Distribution-based estimates of minimal important difference for hospital anxiety and depression scale and impact of event scale-revised in survivors of acute respiratory failure. *Gen Hosp Psychiatry*. (2016) 42:32–5. doi: 10.1016/j.genhosppsych.2016.07.004

129. Mao F, Sun Y, Wang J, Huang Y, Lu Y, Cao F. Sensitivity to change and minimal clinically important difference of Edinburgh postnatal depression scale. *Asian J Psychiatr*. (2021) 66:102873. doi: 10.1016/j.ajp.2021.102873

130. Mauskopf JA, Simon GE, Kalsekar A, Nimsch C, Dunayevich E, Cameron A. Nonresponse, partial response, and failure to achieve remission: humanistic and cost burden in major depressive disorder. *Depress Anxiety*. (2009) 26:83–97. doi: 10.1002/da.20505

131. Strandberg RB, Graue M, Wentzel-Larsen T, Peyrot M, Rokne B. Relationships of diabetes-specific emotional distress, depression, anxiety, and overall well-being with HbA1c in adult persons with type 1 diabetes. *J Psychosom Res*. (2014) 77:174–9. doi: 10.1016/j.jpsychores.2014.06.015

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

133. Flückiger C, Del Re A, Wampold BE, Horvath AO. Alliance in adult psychotherapy In: JC Norcross and MJ Lambert, editors. *Psychotherapy relationships that work: Evidence-based therapist contributions*. Oxford: Oxford University Press (2019)

134. Bica I, Alaa AM, Lambert C, Van Der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin Pharmacol Ther*. (2021) 109:87–100. doi: 10.1002/cpt.1907

135. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. (2019) 28:231–7. doi: 10.1136/bmjqs-2018-008370

136. Selby JV, Fireman BH. Building predictive models for clinical care—where to build and what to predict? *JAMA Netw Open*. (2021) 4:e2032539–9. doi: 10.1001/jamanetworkopen.2020.32539

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

138. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods*. (2009) 14:323–48. doi: 10.1037/a0016973

139. Simpson HB, Marcus SM, Zuckoff A, Franklin M, Foa EB. Patient adherence to cognitive-behavioral therapy predicts long-term outcome in obsessive-compulsive disorder. *J Clin Psychiatry*. (2012) 73:1265–6. doi: 10.4088/JCP.12l07879

140. Smink WA. *What works when for whom?: A methodological reflection on therapeutic change process research*. Netherlands: University of Twente (2021).

141. Smink WA, Sools AM, Postel MG, Tjong Kim Sang E, Elfrink A, Libbertz-Mohr LB, et al. Analysis of the emails from the Dutch web-based intervention "alcohol de baas": assessment of early indications of drop-out in an online alcohol abuse intervention. *Front Psych*. (2021) 12:575931. doi: 10.3389/fpsyt.2021.575931

142. Tiemens B, Kloos M, Spijker J, Ingenhoven T, Kampman M, Hendriks G-J. Lower versus higher frequency of sessions in starting outpatient mental health care and the risk of a chronic course; a naturalistic cohort study. *BMC Psychiatry*. (2019) 19:1–12. doi: 10.1186/s12888-019-2214-4

143. De Jong K, Conijn JM, Gallagher RA, 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

144. Rognstad K, Wentzel-Larsen T, Neumer S-P, Kjøbli J. A systematic review and meta-analysis of measurement feedback systems in treatment for common mental health disorders. *Adm Policy Ment Health Ment Health Serv Res*. (2023) 50:269–82. doi: 10.1007/s10488-022-01236-9

145. Cohen ZD, Kim TT, Van HL, Dekker JJ, Driessen E. A demonstration of a multi-method variable selection approach for treatment selection: recommending cognitive–behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychother Res*. (2020) 30:137–50. doi: 10.1080/10503307.2018.1563312

146. Cho G, Yim J, Choi Y, Ko J, Lee S-H. Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investig*. (2019) 16:262–9. doi: 10.30773/pi.2018.12.21.2

147. Beaulieu-Jones BK, Finlayson SG, Yuan W, Altman RB, Kohane IS, Prasad V, et al. Examining the use of real-world evidence in the regulatory process. *Clin Pharmacol Ther*. (2020) 107:843–52. doi: 10.1002/cpt.1658

frontiers | Frontiers in Public Health

TYPE Brief Research Report
PUBLISHED 07 November 2023
DOI 10.3389/fpubh.2023.1302794



OPEN ACCESS

EDITED BY
Uli Niemann,
Otto von Guericke University Magdeburg, Germany

REVIEWED BY
Delia Virga,
West University of Timişoara, Romania
Seyed-Ali Sadegh-Zadeh,
Staffordshire University, United Kingdom

*CORRESPONDENCE
Hyunsuk Kim
✉ hyskim@etri.re.kr
Cheong Hee Park
✉ cheonghee@cnu.ac.kr

RECEIVED 27 September 2023
ACCEPTED 25 October 2023
PUBLISHED 07 November 2023

CITATION
Kim H, Kim M, Park K, Kim J, Yoon D, Kim W and Park CH (2023) Machine learning-based classification analysis of knowledge worker mental stress. *Front. Public Health* 11:1302794. doi: 10.3389/fpubh.2023.1302794



# Machine learning-based classification analysis of knowledge worker mental stress

Hyunsuk Kim[1]*, Minjung Kim[1], Kyounghyun Park[1], Jungsook Kim[1], Daesub Yoon[1], Woojin Kim[1] and Cheong Hee Park[2]*

[1]Mobility UX Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea, [2]Division of Computer Convergence, Chungnam National University, Daejeon, Republic of Korea

The aim of this study is to analyze the performance of classifying stress and non-stress by measuring biosignal data using a wearable watch without interfering with work activities at work. An experiment is designed where participants wear a Galaxy Watch3 to measure HR and photoplethysmography data while performing stress-inducing and relaxation tasks. The classification model was constructed using k-NN, SVM, DT, LR, RF, and MLP classifiers. The performance of each classifier was evaluated using LOSO-CV as a verification method. When the top 9 features, including the average and minimum value of HR, average of NNI, SDNN, vLF, HF, LF, LF/HF ratio, and total power, were used in the classification model, it showed the best performance with an accuracy of 0.817 and an F1 score of 0.801. This study also finds that it is necessary to measure physiological data for more than 2 or 3 min to accurately distinguish stress states.

KEYWORDS

**heart rate, machine learning, mental stress, knowledge worker, photoplethysmography, pulse rate variability**

## 1. Introduction

Low-moderate levels of perceived stress have been shown to be associated with increased Working Memory (WM)-related neural activation, resulting in more optimal WM behavioral performance (1). However, higher stress scores are associated significantly with lower productivity scores (2). Stress can affect health directly through autonomic and neuroendocrine responses, but it can also affect health indirectly through changes in health behaviors (3). Mental stress in workers can reduce the quality of labor and increase a nation's economic and industrial losses due to high medical costs and related insurance payments.

Recent studies have aimed to objectively quantify mental stress by analyzing physiological responses to stress using wearable sensors (4–6). Lee et al. (4) measured Electrocardiogram (ECG) and Electroencephalogram (EEG) data while the participants played money games, and they analyzed the effects of stress on human physiological response. The ECG sensors were attached based on the bipolar limb leads and 14 EEG channels were attached to the scalps of the participants. In a study by Acerbi et al. (5), ECG information was collected using a wearable Bluetooth chest belt, and Galvanic Skin Responses (GSR) were collected using a finger-type GSR sensor. Their analyzes of the ECG and GSR data highlighted significant differences between stressed and non-stressed individuals. In a study conducted by Chalmers et al., Heart Rate (HR) was measured using a wearable Fitbit Versa 2 device on the nondominant wrist, and HR Variability (HRV) was measured using a three-lead ECG on the

chest. In the stress state, the HR and the Low-Frequency (LF) and High-Frequency (HF) increase significantly (6). However, it is disruptive for workers to wear these devices and measure their biosignal information at work.

In addition, researches are being conducted to collect data using wearable watches and then apply machine learning techniques to measure mental stress (7–9). Arsalan and Majid (7) used electroencephalography, GSR, and Photoplethysmography (PPG) signal data acquired during the resting state and public speaking activities to classify stressed and non-stressed groups. The classification was performed using five different classifiers. Dalmeida and Masala (8) collected HR from four Apple Watch users during a break while listening to relaxing music and after an 8-h workday. After extracting and normalizing HRV features from HR, they split the training and testing datasets 80:20 and used the Multilayer Perceptron (MLP) classifier. Can et al. (9) collected heart activity, skin conductance and accelerometer signals using Empatica E4 and Samsung Gear from algorithm programming competition participants. They discriminated contest stress, relatively higher cognitive load (lecture) and relaxed time activities by using different machine learning methods.

However, to develop a system that can monitor and identify the current mental stress of knowledge workers at work, it is necessary to measure and analyze physiological data by simulating their work and rest behaviors. Additionally, noninvasive methods that can quickly measure biosignals to classify and predict mental stress without disrupting work are required. Thus, we set the following research questions and designed an experiment to measure the mental stress state of knowledge workers by performing stressful tasks and relax tasks.

- Is it possible to classify stressed and non-stressed states using biosignals data measured by a wearable watch?
- For the prediction of stressed and non-stressed states, how long is it appropriate to measure biosignal data with a wearable watch?

# 2. Experimental environments for data collection

## 2.1. Experiment environment

The experiment in this study were approved by the Korean Public Institutional Bioethics Committee (http://public.irb.or.kr/; approval number: P01-202109-13-002). The 80 participants were involved in the experiment and data from 13 subjects were excluded from the analysis for reasons including device malfunction, missing some data, and abnormal data collection due to Bluetooth communication errors. The 67 participants used in the analysis were 39 men (58%) and 28 women (42%), with an average age of 36.5 years (standard deviation 8.6 years).

The top left of Figure 1 represents the data collection environment. We developed the WellMind Application (App) and installed on the Samsung Watch3 to collect the HR and peak to peak interval (PPI) data from the Watch3 and to transmit the data to the Galaxy Tablet. We developed an application called WellMind Space (WSpace), installed it on a tablet, connected the Watch3 and tablet via Bluetooth, and collected data using the app. The WSpace possesses a labeling function that permits the annotation of stressful and relaxing task data as stress and non-stress labels, respectively. All data were stored on a computer installed with PostgreSQL (10).

## 2.2. Experimental procedure

The experimental procedure was as follows.

*Preparation*: The participants completed the consent form and profile questionnaire and then placed the Watch3 on their wrists. The operator established a Bluetooth connection between the Watch3 and the WSpace.

*Stress task*: The participants followed the operator's instructions and performed a stress task for 5 min, that is, the operator sent the participants three emails which asked to search for information on the specific topics at 1 min intervals and each replied separately to three emails. This stress-inducing task was chosen following (11), where email writing was used as a stress-inducing task. In addition, to keep the participants' stress level during the physiological data measurement, they were asked to memorize contents of the email for later presentation.

*Measurement of PPG data after completing the stress task*: The operator measured the participants' HR and PPI data using the Watch3.

*Announce email contents*: Participants had to announce the contents of the email; this was done to keep participants stress state after the stress task before the PPG data measurements.

*Survey about stress task*: After announcing the email content, the participants completed a survey on their experiences with stress.

*Relaxation task*: Participants then performed one of three relaxation tasks: closing their eyes, stretching, or using a massager. The participants were divided into three groups to account for counterbalancing, and each group performed the relaxation tasks in a different order.

*Measurement of PPG data after completing the relaxation task*: The operator measured the participants' HR and PPI data.

*Survey about relaxation task*: After performing the relaxation task, the participants completed a questionnaire about their relaxation task.

Participants repeated the above procedure three times. The bottom of Figure 1 represents the experiment procedure diagram.

# 3. Data manipulation

## 3.1. Photoplethysmography

PPG sensor uses a photodetector to measure the intensity of light reflected from the tissue, and changes in blood volume can be measured depending on the amount of light detected. Similar to ECG, PPG exhibits stable cardiac and respiratory activity. PPI defined as the time interval between successive peaks of the PPG waveform, can be utilized to derive the Pulse Rate Variability (PRV), which shares similarities with the ECG-derived HRV (12).

Because mental stress affects the Autonomic Nervous System (ANS), PRV is a means to observe ANS responses indirectly. Therefore, studies are being conducted to classify and predict the presence or absence of stress state using PPG signals (7, 13). HRV data can be used for stress detection by analyzing the time- and frequency-domain features (7, 13, 14). In particular, the Standard Deviation of

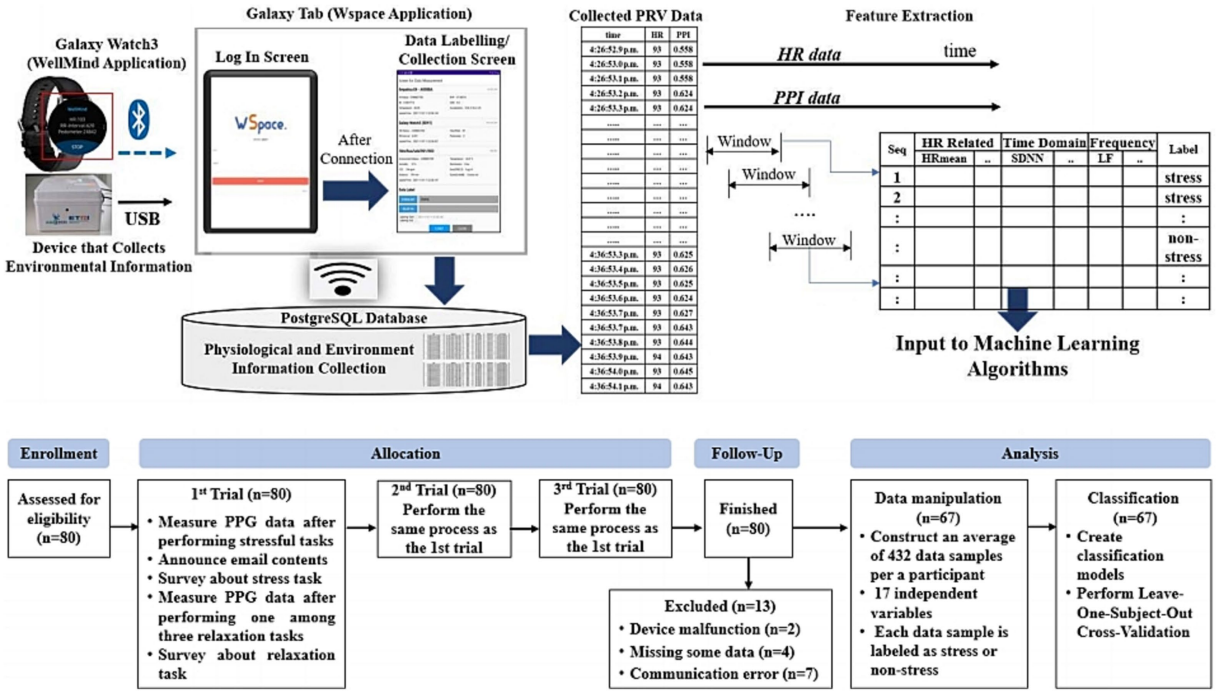



FIGURE 1
Experiment design and data feature extraction.

Normal-to-Normal intervals (SDNN) and Root Mean Square of Successive Differences between normal heartbeats (RMSSD) which are related to the interval between consecutive heartbeats (the interbeat interval) and LF/HF in the frequency domain appear to be the primary factors that differentiate stress states. These features can also be used when analyzing PRV (12).

## 3.2. Feature variables

To analyze the mental stress of knowledge workers during working hours, it is necessary to acquire biosignals from these workers without disturbing them. Therefore, wrist-worn devices are more user-friendly in daily life than chest-worn devices. The Watch3, which integrates a PPG sensor to measure light intensity changes in the microvascular tissue and derive HR and PPI information, is worn on the wrist and offers a convenient and noninvasive approach for HR and PPI measurements (15). Hence, this study employed a Watch3 to collect these data.

The top right of Figure 1 illustrates the process of extracting features from HR and PPI data sequences. After completing each task, participants had 5 to 7 minutes of physiological data measurements. Moving a 3-min window forward with a shift size of 10 s in the HR and PPI data sequences collected from each participant, a total of 17 independent features were extracted from the data within each window to form a data sample. The minimum, mean, median, and maximum values were calculated from a window in the HR data sequence. Time-domain and frequency-domain features were calculated from a window in the PPI data sequence. Time-domain features include the average NN Intervals (NNI), RMSSD, SDNN, Standard Deviation of Differences between adjacent NN intervals (SDSD), Percentage of successive NN intervals that differ by more than 50 ms (PNN50), and PNN20 values. Frequency domain features include LF, HF, LF/HF ratio, LF power in normalized units (LFnu), HF power in normalized units (HFnu), total power, and very Low Frequency (vLF).

The label "stress" was assigned to data samples which were constructed from physiological data measured when performing the stress task, and the label "non-stress" was assigned to data samples obtained from the relaxation task. Since one participant performs 6 tasks and the measurements were made over 5 min for each task, an average of 432 data samples per a participant can be obtained. Physiological data varies depending on each subject's personal health status. Subsequently, min–max normalization was applied to each feature of each participant to generate the final data features for analysis. The collection of all data samples from all participants was used as an input to machine learning algorithms for binary classification of stress and non-stress.

# 4. Classification results

## 4.1. Classification analysis

In this study, k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), MLP, Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR) classifiers of the scikit-learn library was used (16). To achieve the highest performing classification model, hyperparameter tuning was performed using GridsearchCV function for each algorithm used. To evaluate the classifiers, Leave-One-Subject-Out Cross-Validation (LOSO-CV) were performed. In LOSO-CV, from the 67 participants, the data for 66 people were used

as the training set, and the data from one participant was used as the test set. This process was repeated 67 times to measure the performance and to calculate the average to determine the overall performance.

The classification models were evaluated using the accuracy, precision, recall, and F1 scores as evaluation measures, as shown in Eqs. 1–4 (17), based on the confusion matrix. True Positive (TP) is the number of data samples predicted to be positive when belonging to the positive class. False Positive (FP) is the number of data samples predicted to be positive when belonging to the negative class. True Negative (TN) and False Negative (FN) are defined similarly. Matthews Correlation Coefficient (MCC) as expressed in Eq. 5 can also be used to evaluate the performance of the classification model (18).

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \quad (5)$$

In this study, the positive corresponds to the stress state and the negative corresponds to the normal state. To accurately predict the stress state, it is important to optimize the performance measures of TP and TN and minimize the occurrence of FP and FN. In particular, a high TP rate (correct identification of stressed cases) and a low FN rate (correct identification of unstressed cases) are crucial. It is necessary to find a model with a high-recall value to effectively predict stressed knowledge workers and guide them to take breaks. High F1 scores indicate that the corresponding classification model effectively predicts stressed workers.

This study constructs six machine-learning models and conducted a classification analysis to determine whether mental health state was categorized as either stress or non-stress.

Table 1 lists the results of the classification analysis using the LOSO-CV for the data generated using a 3-min window. The results showed that the LR classifiers achieved the best performance with accuracy of 0.814, precision of 0.843, recall of 0.805, F1 score of 0.796, and MCC of 0.643. The k-NN classifier achieved the lowest performance with an accuracy of 0.719 and an F1 score of 0.692.

## 4.2. Window size

Further analysis was conducted to determine the optimal window size required for measuring physiological data to predict the stress experienced by knowledge workers during working hours. In the analysis by LOSO-CV, the LR classifier was used as the prediction model owing to its best performance, as shown in Table 2, and various window sizes ranging from 30 s to 300 s (with 30-s intervals) were tested. This analysis aimed to identify the most appropriate time for measuring physiological data during working hours to accurately predict the stress status of workers.

The classification accuracy significantly improved when the window size was greater than 2 min. The highest performance was achieved when the window size was set to 150, with an accuracy of 0.816, precision of 0.843, recall of 0.807, F1 score of 0.8, and an MCC value of 0.646. It can be suggested that measuring physiological data for at least 2–3 min is necessary to accurately distinguish between stressed and non-stressed states in knowledge workers.

## 4.3. Feature selection

Performance improvements in classification models typically depend on the selection of a suitable set of features. Gioia et al. (19) used a feature selection strategy based on Recursive Feature Elimination (RFE).

TABLE 1 Classification analysis using leave-one-subject out CV for data generated using a 3-min window.

| Classifier | Accuracy | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|
| k-NN | 0.719 | 0.729 | 0.700 | 0.692 | 0.426 |
| SVM | 0.743 | 0.766 | 0.730 | 0.723 | 0.491 |
| MLP | 0.741 | 0.760 | 0.725 | 0.718 | 0.482 |
| DT | 0.756 | 0.787 | 0.739 | 0.730 | 0.519 |
| RF | 0.788 | 0.821 | 0.770 | 0.766 | 0.585 |
| LR[a] | 0.814 | 0.843 | 0.805 | 0.796[a] | 0.643 |

[a]Highest F1 score.

TABLE 2 Results of logistic regression classification analysis after changing the window size.

| Size | Accuracy | Precision | Recall | F1 | MCC |
|---|---|---|---|---|---|
| 30 s | 0.762 | 0.792 | 0.758 | 0.747 | 0.545 |
| 60 s | 0.775 | 0.806 | 0.770 | 0.761 | 0.572 |
| 90 s | 0.795 | 0.825 | 0.791 | 0.782 | 0.612 |
| 120 s | 0.801 | 0.832 | 0.796 | 0.787 | 0.624 |
| 150 s[a] | 0.816 | 0.843 | 0.807 | 0.800[a] | 0.646 |
| 180 s | 0.814 | 0.843 | 0.805 | 0.796 | 0.643 |
| 210 s | 0.805 | 0.842 | 0.792 | 0.784 | 0.628 |
| 240 s | 0.808 | 0.846 | 0.792 | 0.784 | 0.630 |
| 270 s | 0.815 | 0.848 | 0.791 | 0.787 | 0.632 |
| 300 s | 0.826 | 0.862 | 0.793 | 0.792 | 0.646 |

[a]Highest F1 score.

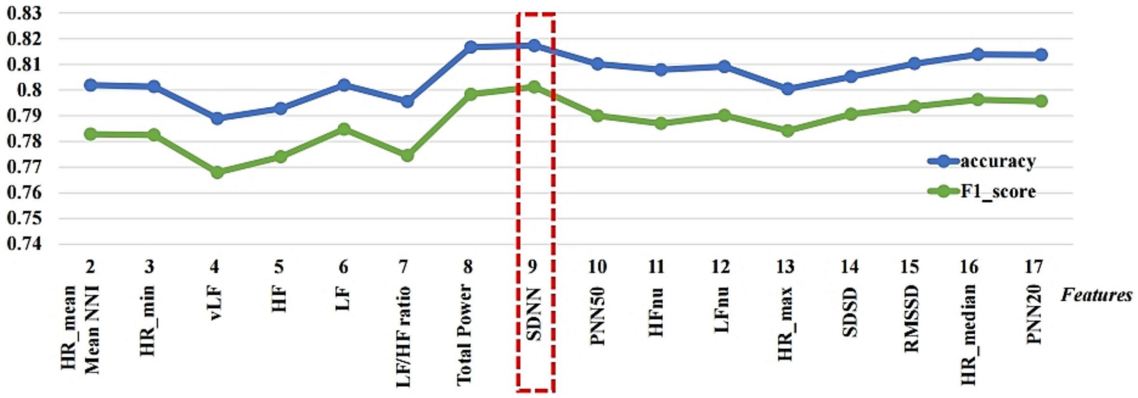



FIGURE 2
Comparison of accuracy and F1-scores when feature subsets are used based on feature ranking.

This study employs the LR-RFE model to determine how performance varies depending on the features utilized. Figure 2 compares the performance of LOSO-CV using the selected features after the feature rank is determined by applying RFE with LR to the entire dataset. The features are displayed on the x-axis based on rank, and the accuracy and F1 score are measured by adding the features in the top rank individually. The best performance was obtained with an accuracy value of 0.817 and an F1 score of 0.801 when nine top-ranked features were used, including 2 HR-related features, HR_mean and HR_min; two time-domain features, Mean_NNI and SDNN; and five frequency-domain features, vLF, HF, LF, LF/HF ratio, and Total Power.

## 5. Discussion and conclusions

Since mental stress can reduce the quality of work and worsen health conditions, the technology needed for a mental health management system that monitors the mental stress of knowledge workers in the workplace must continue to be researched. In this study, HR and PPI data were measured using the Galaxy Watch3 rather than a chest-worn ECG device. The classification model was constructed using k-NN, SVM, DT, LR, RF, and MLP classifiers. The performance of each classifier was evaluated using LOSO-CV as a verification method.

To determine the optimal duration for measuring biosignals to classify and predict the mental stress, the HR and PRV data features were calculated using varying window sizes. The window size was varied from 30 to 300 s. The best performance was achieved with an accuracy of 0.816, precision of 0.843, recall of 0.807, F1 score of 0.8, and MCC of 0.646, using an LR classifier with 17 features extracted by setting the window size to 150. The results of this study show that it is possible to analyze mental stress using PPG data obtained over a sufficiently short period of time of 2 to 3 min that does not interfere with work activities at work.

Additionally, the LR-RFE model was utilized to investigate how performance changes depending on the type of features used. The best performance exhibited an accuracy value of 0.817 and an F1 score of 0.801 when the nine top-ranked features were used. The feature selection results can be used to achieve a classification model suitable for highest performance.

To develop a system that can measure mental stress during working hours and guide rest in the event of stress, it is necessary to obtain biosignals without disturbing workers. In our study, HR and PPI data were collected using a Watch3 to noninvasively measure biosignals. However, it should be noted that PPG signals have a limitation in that they are sensitive to motion artifacts caused by hand movements (20). Recently, studies on HR measurement methods based on remote PPG detection using deep learning-based facial videos have also been implemented (21, 22). In order to minimize worker inconvenience and simultaneously improve prediction performance, research should be conducted on stress analysis through the combination of facial images and biosignal information through wearable watches. Additionally, we plan to conduct research on the development of stress monitoring and intervention apps that incorporate these technologies.

## Data availability statement

The datasets presented in this article are not readily available because the data analyzed in this study is subject to the following licenses/restrictions: the data is the property of the Mobility UX Research Section, Electronics and Telecommunications Research Institute, Republic of Korea. Requests to access the datasets should be directed to HK, hyskim@etri.re.kr.

## Author contributions

HK: Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. MK: Conceptualization, Resources, Writing – original draft. KP: Conceptualization, Resources, Writing – original draft. JK: Conceptualization, Resources, Writing – original draft. DY: Funding acquisition, Project administration, Writing – original draft. WK: Funding acquisition, Project administration, Writing – original draft. CP: Methodology, Visualization, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was financially supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program (Project No. P0011894) and has been achieved in the European ITEA project "Mental Wellbeing Management and Productivity Boosting in the Workplace (18033 Mad@Work)." This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) under grant funded by the Korea government (MSIT) (2022-0-00501).

## 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. Oshri A, Cui Z, Owens MM, Carvalho CA, Sweet L. Low-to-moderate level of perceived stress strengthens working memory: testing the hormesis hypothesis through neural activation. *Neuropsychologia*. (2022) 176:108354. doi: 10.1016/j.neuropsychologia.2022.108354

2. Bui T, Zackula R, Dugan K, Ablah E. Workplace stress and productivity: a cross-sectional study. *Kans J Med*. (2021) 14:42–5. doi: 10.17161/kjm.vol1413424

3. O'Connor DB, Thayer JF, Vedhara K. Stress and health: a review of psychobiological processes. *Annu Rev Psychol*. (2021) 72:663–88. doi: 10.1146/annurev-psych-062520-122331

4. Lee J, Kim C, Lee KC. An empirical approach to analyzing the effects of stress on individual creativity in business problem-solving: emphasis on the electrocardiogram, electroencephalogram methodology. *Front Psychol*. (2022) 13:705442. doi: 10.3389/fpsyg.2022.705442

5. Acerbi G, Rovini E, Betti S, Tirri A, Rónai JF, Sirianni A, et al. A wearable system for stress detection through physiological data analysis, ambient assisted living. *Lect Notes Electr Eng*. (2017) 426:31–50. doi: 10.1007/978-3-319-54283-6_3

6. Chalmers T, Hickey BA, Newton P, Lin C-T, Sibbritt D, McLachlan CS, et al. Stress watch: the use of heart rate and heart rate variability to detect stress: a pilot study using smart watch wearables. *Sensors*. (2021) 22:151. doi: 10.3390/s22010151

7. Arsalan A, Majid M. Human stress classification during public speaking using physiological signals. *Comput Biol Med*. (2021) 133:104377. doi: 10.1016/j.compbiomed.2021.104377

8. Dalmeida KM, Masala GL. HRV features as viable physiological markers for stress detection using wearable devices, sensors 21. *Sensors (Basel)*. (2021) 8:2873. doi: 10.3390/s21082873

9. Can YS, Chalabianloo N, Ekiz D, Ersoy C. Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study. *Sensors*. (2019) 19:1849. doi: 10.3390/s19081849

10. The PostgreSQL Global Development Group. PostgreSQL. Available at: https://www.postgresql.org/ (accessed April 12, 2023).

11. Koldijk S. (2014). The SWELL knowledge work dataset for stress and user modeling research. Proceedings of the 16th International Conference on Multimodal Interaction.

12. Pinheiro N., Couceiro R., Henriques J., Muehlsteff J., Quintal I., Goncalves L., et al. (2016). Can PPG be used for HRV analysis? 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

13. Ahmadi N, Sasangohar F, Nisar T, Danesh V, Larsen E, Sultana I, et al. Quantifying occupational stress in intensive care unit nurses: an applied naturalistic study of correlations among stress, heart rate, electrodermal activity, and skin temperature. *Hum Factors*. (2022) 64:159–72. doi: 10.1177/00187208211040889

14. Hao T, Zheng X, Wang H, Xu K, Chen S. Linear and nonlinear analyses of heart rate variability signals under mental load. *Biomed. Signal Process. Control*. (2022) 77:103758. doi: 10.1016/j.bspc.2022.103758

15. TIZEN Project. Human Activity Monitor. Available at: https://docs.tizen.org/application/web/guides/sensors/ham/ (accessed April 2023).

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

17. Tan PN, Steinbach M, Kumar V. *Introduction to data mining*. Ney York, NY: Pearson Addison Wesley (2006).

18. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. (2020) 21:6. doi: 10.1186/s12864-019-6413-7

19. Gioia F, Greco A, Callara AL, Scilingo EP. Towards a contactless stress classification using thermal imaging. *Sensors*. (2022) 22:976. doi: 10.3390/s22030976

20. Castaneda D, Esparza A, Ghamari M, Soltanpur C, Nazeran H. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int J Biosens Bioelectron*. (2018) 4:195–202. doi: 10.15406/ijbsbe.2018.04.00125

21. Cheng CH, Wong KL, Chin JW, Chan TT, So RHY. Deep learning methods for remote heart rate measurement: a review and future research agenda. *Sensors*. (2021) 21:6296. doi: 10.3390/s21186296

22. Przybyło J. A deep learning approach for remote heart rate estimation. *Biomed Signal Process Control*. (2022) 74:103457. doi: 10.1016/j.bspc.2021.103457

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 08 January 2024
DOI 10.3389/fpubh.2023.1348803



OPEN ACCESS

EDITED BY
Peter ten Klooster,
University of Twente, Netherlands

REVIEWED BY
Seyed-Ali Sadegh-Zadeh,
Staffordshire University, United Kingdom
Zhiang Niu,
West China Hospital, Sichuan University, China

*CORRESPONDENCE
Chunguang Liang
✉ liangchunguang@jzmu.edu.cn

RECEIVED 03 December 2023
ACCEPTED 22 December 2023
PUBLISHED 08 January 2024

CITATION
Li E, Ai F and Liang C (2024) A machine learning model to predict the risk of depression in US adults with obstructive sleep apnea hypopnea syndrome: a cross-sectional study.
*Front. Public Health* 11:1348803.
doi: 10.3389/fpubh.2023.1348803



# A machine learning model to predict the risk of depression in US adults with obstructive sleep apnea hypopnea syndrome: a cross-sectional study

Enguang Li, Fangzhu Ai and Chunguang Liang*

Department of Nursing, Jinzhou Medical University, Jinzhou, China

**Objective:** Depression is very common and harmful in patients with obstructive sleep apnea hypopnea syndrome (OSAHS). It is necessary to screen OSAHS patients for depression early. However, there are no validated tools to assess the likelihood of depression in patients with OSAHS. This study used data from the National Health and Nutrition Examination Survey (NHANES) database and machine learning (ML) methods to construct a risk prediction model for depression, aiming to predict the probability of depression in the OSAHS population. Relevant features were analyzed and a nomogram was drawn to visually predict and easily estimate the risk of depression according to the best performing model.

**Study design:** This is a cross-sectional study.

**Methods:** Data from three cycles (2005–2006, 2007–2008, and 2015–2016) were selected from the NHANES database, and 16 influencing factors were screened and included. Three prediction models were established by the logistic regression algorithm, least absolute shrinkage and selection operator (LASSO) algorithm, and random forest algorithm, respectively. The receiver operating characteristic (ROC) area under the curve (AUC), specificity, sensitivity, and decision curve analysis (DCA) were used to assess evaluate and compare the different ML models.

**Results:** The logistic regression model had lower sensitivity than the lasso model, while the specificity and AUC area were higher than the random forest and lasso models. Moreover, when the threshold probability range was 0.19–0.25 and 0.45–0.82, the net benefit of the logistic regression model was the largest. The logistic regression model clarified the factors contributing to depression, including gender, general health condition, body mass index (BMI), smoking, OSAHS severity, age, education level, ratio of family income to poverty (PIR), and asthma.

**Conclusion:** This study developed three machine learning (ML) models (logistic regression model, lasso model, and random forest model) using the NHANES database to predict depression and identify influencing factors among OSAHS patients. Among them, the logistic regression model was superior to the lasso and random forest models in overall prediction performance. By drawing the nomogram and applying it to the sleep testing center or sleep clinic, sleep technicians and medical staff can quickly and easily identify whether OSAHS patients have depression to carry out the necessary referral and psychological treatment.

KEYWORDS
machine learning, depression, OSAHS, prediction models, NHANES

# Introduction

Depression is a widespread mental health disorder that seriously limits the patient's psychological and social function, which reduces their quality of life. At the same time, depression also brings severe financial and emotional stress to the families of patients. Its main features include persistent fatigue, depression, low mood, reduced interest, and poor concentration (1). Depression is related to mental health and is now the main reason for the global burden of disease. In addition to the severe influence on personal emotion and psychological state, it may seriously impact work and personal relationships (2). In 2015, the WHO announced that, globally, depression affects more than 300 million people, or 4.4% of the world's population, and is the leading cause of disability globally (3), with about 1 million people dying of depression each year (4). At the same time, depression also imposes significant socioeconomic costs. The annual cost of treating depression in the US is reported to be as high as $210 billion (5). However, in high-income countries, nearly half of people with depression are not diagnosed or treated. In low and middle-income countries, the proportion is as high as 80%–90%. Early detection and prevention of depression is, therefore, essential to reduce the global burden. Society needs to take action early in life and in adversity and the impact of inequality (6).

Obstructive Sleep Apnea Hypopnea Syndrome (OSAHS) is a chronic disease characterized by recurrent upper airway collapse and obstruction during sleep (7), resulting in periodic reduction or cessation of ventilation, which causes hypoxia, hypercapnia and sleep arousal (8). Cross-sectional studies have found that OSAHS may increase the risk of depression (9). In addition, a dose–response relationship has been found between the severity of OSAHS and the risk of depression (10). That is, the more severe the OSAHS, the higher the risk of depression. This means that the presence of OSAHS may cause more significant difficulties in the treatment of depression. A cross-sectional study of community and clinical populations found a relatively high prevalence of depression in patients with OSAHS of 17% (11). In contrast, the prevalence of depression among patients with a definite diagnosis of OSAHS in the sleep clinic showed a wide variability, ranging from 5% to 63% (2). To verify the causal relationship between OSAHS and depression, according to a prospective longitudinal study of depression a year later with OSAHS between cause and effect (12). Therefore, we suggest that screening for psychiatric in patients with OSAHS timely find depression to effectively prevent and treat depression and reduce the impact on the quality of life and social economy.

Currently, one of the early screening methods for depression is to determine the presence or absence of depression by using the Depression Self-Rating Scale (DSRS). However, there is no self-assessment scale for depression for patients with OSAHS. Although several self-rating depression scales have been shown to have reliable reliability and validity in OSAHS patients (13–16), these scales still do not accurately predict the risk of depression in OSAHS patients. Clinical predictive modeling was introduced to solve this problem. It is a mathematical formula to estimate the probability that a particular individual is currently suffering from a disease or experiencing a specific outcome. In this study, a prediction model was used to estimate the likelihood of depression in OSAHS patients to more accurately assess the risk of depression in patients and take appropriate interventions.

However, traditional statistical methods are only suitable for solving simple linear problems rather than for dealing with complex nonlinear relationships. Secondly, traditional models lack adaptive learning capabilities and require manual selection and extraction of variables. This process requires specialized knowledge and experience and relies on prior knowledge or specific rules to build and adapt. Moreover, the traditional model can only deal with small-scale data sets, and the processing effect on high-dimensional data is poor (17).

Therefore, introducing machine learning (ML), a powerful and intelligent tool, can solve all these problems. ML models can adaptively learn and adjust models from data without manually specifying model parameters or rules. In addition, ML models also have good generalization ability. It can effectively generalize the patterns learned from the training data to the new data. In addition, ML models can handle large-scale data, automatically extract features, and build models from the data. Finally, the ML model also has high interpretability, can through the way of visualization and explanatory, help researchers to understand the behavior of the model and the decision-making process (18, 19).

At present, the ML has been widely applied in the depression risk prediction model was constructed. For example, Dai Su et al. used ML algorithms to construct a risk prediction model for depression in older Chinese adults (20). Fang Xia et al. developed a prediction model for depression caused by heavy metals in older people using the ML method based on the National Health and Nutrition Examination Survey (NHANES) database (21). The research result shows that ML, which improves the prediction accuracy of depression, reduces error and mass data processing, and so on, shows great potential.

After a comprehensive literature search, we found that most of the previous studies focused on exploring the correlation between OSAHS and depression. At the same time, there are a large number of predictive studies of patients with OSAHS (9, 22–26). However, no studies have been found to predict whether patients with OSAHS will develop depression. This includes studies using traditional statistical methods (such as logistic regression) and ML methods (such as random forest, SVM, etc.) to construct depression risk prediction models for OSAHS patients. Therefore, this study selected a large data sample from the NHANES database and screened for influencing factors associated with depression. To construct a risk prediction model that can predict whether OSAHS patients will have depression using ML methods.

# Materials and methods

## Data and sample

### Description of National Health and Nutrition Examination Survey data

The data used in this study come from the NHANES database published by the Centers for Disease Control and Prevention (CDC). NHANES, a population-based cross-sectional survey, aims to collect

information about relevant American adults' and children's diet, nutrition, health, and health behavior (5). A representative sample of households across the United States was selected using multistage stratified random sampling. Since 1999, the NHANES program has conducted a nationally representative sample every two years. Each year, NHANES investigators conduct home visits and in-person interviews with a nationally representative sample of about 5,000 people of all ages. They collect data on basic information, family structure, health status, and eating habits of the respondents. After the face-to-face survey, participants were invited to a temporary examination center for various physical measurements, physical function tests, and laboratory tests. Finally, the collected data will be collated, coded, and anonymized before being stored in the NHANES database. It was also approved by the Research Ethics Review Board of the National Center for Health Statistics (NCHS). Each participant was asked to sign a consent form, including all the questionnaires, and check. For participants younger than 18 years of age, they were required to complete data collection with informed consent from their parents or guardians (27, 28). In this study, the OSAHS population was selected based on participants' self-report of the question "How often do you snort/stop breathing?" on the sleep questionnaire. Therefore, we excluded data periods that did not include this question in the sleep questionnaire and finally selected data from the three 2-year periods that had this question (2005–2006, 2007–2008, and 2015–2016). These data will be used to construct a prediction model for depression in the OSAHS population. All data were downloaded from the official NHANES website.[1]

## Outcome variable

The 9-item Patient Health Questionnaire-9 (PHQ-9) was used in this study to assess depression in patients with OSAHS. The questionnaire used a four-point Likert scale, with options for each item including 0 (not at all), 1 (a few days), 2 (more than half a day), and 3 (almost every day). Each item is scored from 0 to 3, and the total score ranges from 0 to 27 (29). Patients with a PHQ-9 total score $\geq 5$ were considered to have depression according to study criteria (30). It is worth noting that the ultimate purpose and significance of this study is to estimate the probability of depression in OSAHS patients by selecting the best prediction model and constructing a nomogram based on the relevant influencing factors. The application of this nomogram in sleep testing centers or sleep clinics can help sleep technicians and medical staff quickly and easily identify whether OSAHS patients have depression and make necessary referrals. Therefore, patients were divided into two groups with and without depression only according to whether they would develop depression. However, Kroenke noted that significant clinical significance was often seen in moderate to severe cases and suggested concomitant antidepressants to improve sleep (31). Therefore, if considering the practical application value of psychiatric clinical practice, it is recommended that future research be able to divide the severity of depression in detail and construct multivariate dependent variable prediction models. Such studies are expected to improve the accuracy and predictive power of the model and thus better provide clinical assistance to psychiatrists.

## Predictor variables

In this study, we categorized the OSAHS population based on participants' responses to the question "How often do you stop breathing?" on a sleep questionnaire. In answer to this question, we recorded responses of 0 (never) as indicating the non-OSAHS population, while responses of 1 (rarely, 1–2 nights per week), 2 (occasionally, 3–4 nights per week), and 3 (often, five or more nights per week) were defined as indicating the OSAHS population.

Data on demographic characteristics of NHANES from 2005–2006, 2007–2008, and 2015–2016 were selected for this study. Data on age, gender, race, education level, marital status, ratio of family income to poverty (PIR), body mass index (BMI), and sleep hours were included. Of these, we selected adults aged 18 years and older for the study. For race, we categorized them into five categories: Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, and Other Race. Education level was categorized into five groups: Less than 9th Grade, 9th–11th Grade, High School Grad/GED or Equivalent, Some College or AA degree, and College Graduate or above. Marital status included married, widowed, divorced, separated, unmarried, and living with a partner. Income status was divided into two categories by using PIR: low-income (PIR $\leq 1.3$) and non-low-income (PIR $> 1.3$). BMI was categorized into four types: underweight (BMI $< 18.5$), normal weight (18.5–24.99), overweight (25.0–29.99), and obese (BMI $\geq 30.0$). Sleep hours were also categorized into three categories: short sleep hours (<7 h), normal sleep hours (7–9 h), and long sleep hours (>9 h).

Lifestyle variables include smoking and alcohol drinking. Smoking status was determined based on respondents' self-reports to two questions: "Have you ever smoked more than 100 cigarettes in your lifetime?" and "Do you currently smoke?." Smoking status was categorized into three categories: never smoker (lifetime never smoked more than 100 cigarettes, current never smoked), former smoker (lifetime smoked more than 100 cigarettes, current never smoked), and now smoker (lifetime smoked more than 100 cigarettes, current daily smoker or current smoker for a few days).

Alcohol drinking was determined by self-report of respondents on the following questions: "In the past 12 months, how often did you drink any type of alcoholic beverage (measured in days)?" Based on their responses, we defined drinking as three types: never drinking (0), low drinking (1–36 days), and heavy drinking ($\geq$37 days).

Health information variables included general health condition, hypertension, diabetes, asthma, coronary heart disease, and OSAHS severity. General health condition was determined by self-report of respondents on the following questions: "I have some general questions about your health." and "Would you say your health in general is?" Of these, "excellent," "very good," and "good" responses were redefined as "good general health condition." "Fair" is defined as "General health." "Poor" is defined as "bad general health condition."

The presence of the four conditions, hypertension, diabetes, asthma, and coronary heart disease, was determined using a "yes" or "no" response. For diabetes problems, it is essential to note that if respondents answer "Border," it has also been defined as no diabetes.

1 https://www.cdc.gov/nchs/nhanes/index.htm

OSAHS severity was determined by the subject's response to the question, "How often do you snort/stop breathing?." Specifically, "1–2 nights per week" was defined as "mild," and "3–4 nights per week" was described as "moderate." "5 or more nights per week" was defined as "severe." Table 1 provides details of the assignment of each influence factor.

TABLE 1 Predictor variable assignment.

| Predictive factors | Variable type | Assignment |
|---|---|---|
| Gender | Categorical variables | "Male" = 1, "Female" = 2 |
| Age | Continuous variables | Original value entry |
| Race | Categorical variables | "Mexican American" = 1, "Other Hispanic" = 2, "Non-Hispanic White" = 3 , "Non-Hispanic Black" = 4 , "Other Race" = 5 |
| Education level | Categorical variables | "Less than 9th grade" = 1, "9-11th grade" = 2, "High school graduate" = 3, "Some college or AA degree" = 4 , "College graduate or above" = 5 |
| Marital status | Categorical variables | "Married" = 1, "Widowed" = 2 , "Divorced" = 3, "Separated" = 4, "Never married" =5, "Living with a partner" = 6 |
| PIR | Categorical variables | "Low-income" = 1, "Non-low-income" = 2 |
| General health condition | Categorical variables | "Good" = 1, "General" = 2, "Bad" = 3 |
| Sleep hours | Categorical variables | "Short" = 1, "Normal" = 2, "Long" = 3 |
| BMI | Categorical variables | "Underweight" = 0, "Normal weight" = 1, "Overweight" = 2, "Obese" = 3 |
| Alcohol drinking | Categorical variables | "Never drinking" = 0, "Small amount" = 1, "Large amount" = 2 |
| Smoking | Categorical variables | "Never smoker" = 0, "Former smoker" = 1, "Now smoker" = 2 |
| Hypertension | Categorical variables | "Yes" = 1, "No" = 2 |
| Diabetes | Categorical variables | "Yes" = 1, "No" = 2 |
| Asthma | Categorical variables | "Yes" = 1, "No" = 2 |
| Coronary heart disease | Categorical variables | "Yes" = 1, "No" = 2 |
| OSAHS severity | Categorical variables | "Mild" = 1, "Moderate" = 2 , "Severe" = 3 |

PIR, ratio of family income to poverty; BMI, body mass index.

## Statistical analysis

### Data description

Stata 17.0 software was applied to extract and clean the NHANES data, and SPSS 25.0 and R Studio software were used for statistical analysis and description. Measurements that conformed to normal distribution were expressed as M ± SD (Mean ± standard error), and comparisons between groups were made using the independent samples *t*-test. If it did not meet, it was expressed as M (P25, P75), and comparisons between groups were made using the Mann–Whitney U test. Count data were expressed as *n* (%), and comparisons between groups were made using the $\chi^2$ test, with $p < 0.05$ being considered statistically significant.

### ML models

Before ML model training and evaluation, we used the set. seed(123) function in R studio software to split the dataset into training and validation sets at a 7:3 ratio. The training set is used for training multiple models, while the validation set is used to verify the performance and generalization ability of the model.

### Logistic regression model

In R Studio software, the functions and commands of the mlr package were used for univariate logistic regression, followed by multiple collinearity diagnoses in SPSS software, and the variables with statistically significant differences were included in multivariate logistic regression for analysis. The final selected variables were used in R Studio software to draw the nomogram and establish the logistic regression prediction model by the plotLearnerPrediction() function.

### LASSO model

In this study, we used the mlr package and glmnet package in R Studio software for training and fitting the lasso model. We performed 10-fold cross-validation with the cv.glmnet() function to select the best lambda value. Then, we retrained the lasso model based on the best lambda value and used the coef() function to obtain the coefficients of the model to complete the training of the lasso model. Given that there may be some covariance and correlation between the independent variables, in order to avoid overfitting the model, we performed dimensionality reduction on the independent variables to screen out the influencing factors related to OSAHS depression. Based on the above dimension reduction analysis, the lasso method was used to analyze all the independent variables included in the model, as shown in Figure 1. In this process, the model can be started from the initial to join the independent variable coefficient of compression gradually until the part of the independent variable coefficient is compressed to 0 to avoid the model's overfitting problem.

### Random forest model

Using the randomForest package in the R Studio for training the random forest model. Based on MeanDecreaseGini, We ranked the 16 independent variables and used the random forest feature importance assessment algorithm to derive the importance of each influencing factor (32), selection of important variables with high impact on depression in OSAHS patients. The optimal number of features of the random forest model was chosen according to the out-of-bag error rate. To better understand the relationship between variables and improve the model's prediction performance. Among the model

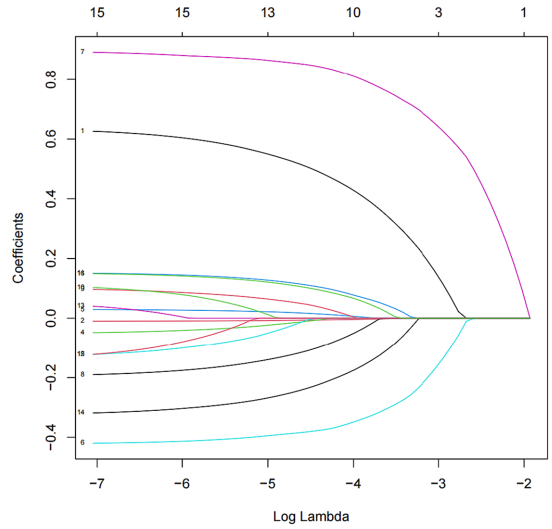



FIGURE 1
Lasso screening variable dynamic plot.

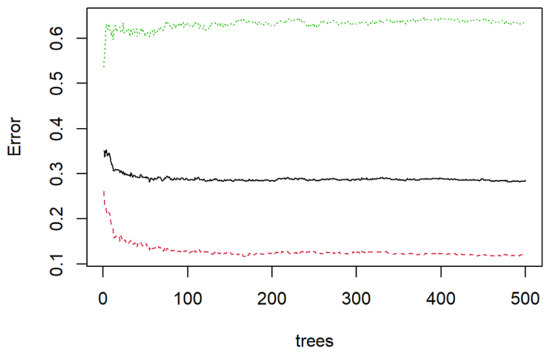



FIGURE 2
Dynamic relationship between prediction error of random forest and the number of random trees.

parameters, there are two key parameters to consider: the number of predicted evaluation indicators (mtry) and the number of random trees (ntree). Among them, mtry is the number of randomly selected evaluation indicators used to construct a random tree, usually the square root of the number of all evaluation indicators in the sample. The tree represents the number of random trees built in the model. When mtry = 5, the minimum error rate outside the package. When ntree = 500, the error is basically stable, and the dynamic relationship between the prediction error of random forest and the number of random trees is shown in Figure 2. Therefore, the parameters of the optimal model are mtry = 5 and ntree = 500. The final selected variables were included in the multivariate logistic regression analysis, and the random forest model was finally constructed.

### Model comparison

Adopt the receiver operating characteristic curve (ROC curve) and the area under the curve (AUC), specificity, sensitivity, Youden index, and DCA comparison of the model to evaluate and compare the performance of the forecasting model. Specifically, this study used the pROC package of R studio software to draw the ROC curve of the prediction model. Then, calculate the AUC, specificity, sensitivity, and Youden index. Subsequently, the ROC curves of the three models were compared using the DeLong test to judge whether the ROC curves of the three models were significantly different. Finally, the "rmda" package and the "decision_curve" function algorithm were used to draw and compare the differences between the DCA curves of different models.

## Results

### Patient screening and statistical analysis process

After strict data cleaning, we selected the three NHANES data cycles: 2005–2006, 2007–2008, and 2015–2016. In the process, we finally chose to include 2,453 patients in the standard. All eligible patients were randomly divided into a training set and a validation set at a ratio of 7:3, with 1718 patients in the training set and 735 in the validation set. Such a dataset partitioning is consistent with the approach Yalong Zhang et al. adopted in their study of ML prediction models (33). Meanwhile, Jianping Lv et al., in their research, used a ML model designed to predict the risk of bullying victimization among adolescents in the same way that our dataset was partitioned (34). The detailed screening process is shown in Figure 3.

### Comparison of baseline information

Through the statistical analysis, this study found no significant difference in baseline characteristics between the training and validation sets ($p > 0.05$). This shows no deviation between the two groups caused by the uneven distribution of the dependent variable, as shown in Table 2. In addition, we divided the training set into non-depressed and depressed groups and compared baseline information between the two groups. Specific comparative results can be found in Table 3.

### Models predict performance in depressed patients with OSAHS

#### Logistic regression model

In the training set, we divided 1718 OSAHS patients into depressed and non-depressed groups. Through the single variable analysis, we found a statistically significant ($p < 0.05$) result of 12 factors involved, including gender, age, education level, marital status, PIR, general health condition, BMI, smoking, hypertension, diabetes, asthma, and OSAHS severity. Subsequently, statistically significant variables in the univariate analysis were included in the multicollinearity diagnosis. According to the analysis results, all variance inflation factors (VIF) involved in the binary logistic regression analysis were less than 5, and the tolerance index was greater than 0.1. This indicates that there is no case of multicollinearity between covariates. All variables are included in the logistic model as a predictor. The results of multicollinearity diagnosis are shown in Table 4. Statistically, there is a difference in the univariate analysis, and there is no multicollinearity of a variable in binary logistic regression

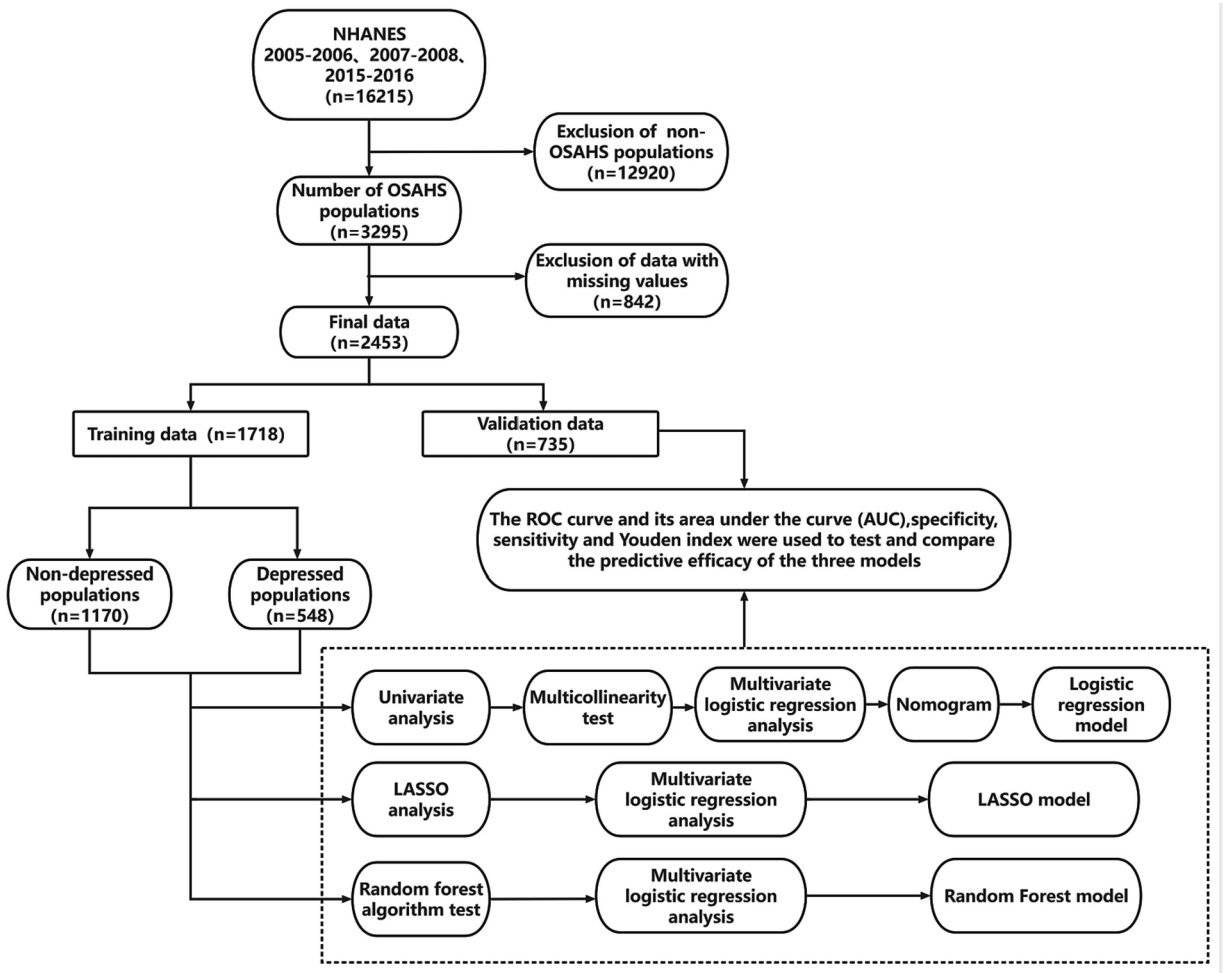



FIGURE 3
Researchers and statistical flowchart.

analysis. Adopting the positive method step by step and likelihood ratio test, the method of removing confounding factors, finally got into the model's variables. The results showed that gender, general health condition, BMI, smoking, OSAHS severity, age, education level, PIR, and asthma were significant influencing factors for depression in OSAHS patients. Among these influencing factors, Gender, General health condition, BMI, Smoking, and OSAHS severity were identified as independent risk factors for depression in OSAHS patients. The factors associated with depression in univariate and multivariate analyses are shown in Table 5.

Based on the factors included in the above regression analysis and the corresponding regression coefficients of each element, a risk prediction model for depression in OSAHS patients was constructed, and a nomogram was drawn. According to the influencing factors in the nomogram and the corresponding scores of each variable, the prediction probability corresponding to the total score was the probability of depression in OSAHS patients when the scores were summed. Points are the individual scores, total points are the full scores, and risk of depression is the incidence of depression corresponding to the total scores, as shown in Figure 4. The nomogram assignment method of relevant factors is shown in Table 6.

## LASSO model

Depression was used as the dependent variable, and a total of gender, age, race, education level, marital status, PIR, general health condition, sleep hours, BMI, alcohol drinking, smoking, hypertension, diabetes, asthma, coronary heart disease, and OSAHS severity, a total of 16 independent variables. From Figure 5, the optimal model was obtained by selecting the λ value with the minor error (0.005586744) through ten-fold cross-validation. On this basis, we choose the associated with OSAHS depression 14 of the most promising of the independent variables, including gender, age, education level, marital status, PIR, general health condition, sleep hours, BMI, alcohol drinking, smoking, hypertension, asthma, coronary heart disease, and OSAHS severity. We conducted the binary logistic regression analysis based on the selection of variables, and the results were obtained. After analysis, it was found that marital status, PIR, general health condition, sleep hours and smoking are the independent influencing factors of depression in OSAHS patients ($p < 0.05$).

## Random forest model

According to the results, age, general health condition, race, education level, marital status, OSAHS severity, BMI, smoking, and

10.3389/fpubh.2023.1348803

TABLE 2 Comparison of baseline data between the two groups.

| Predictive factors | Training data ($n$ = 1,718) | Validation data ($n$ = 735) | $\chi^2/t$ | $p$ value |
|---|---|---|---|---|
| Age | 50.01 ± 15.984 | 50.35 ± 16.274 | −0.471 | 0.341 |
| Gender | | | 3.835 | 0.052 |
| Male | 1,091(63.5) | 436(59.3) | | |
| Female | 627 (36.5) | 299(40.7) | | |
| Race | | | 5.913 | 0.206 |
| Mexican American | 286(16.6) | 129(17.6) | | |
| Other Hispanic | 178(10.4) | 61(8.3) | | |
| Non-Hispanic White | 811(47.2) | 328(44.6) | | |
| Non-Hispanic Black | 326(19.0) | 159(21.6) | | |
| Other race | 117(6.8) | 58(7.9) | | |
| Education level | | | 3.250 | 0.517 |
| Less than 9th grade | 167(9.7) | 63(8.6) | | |
| 9–11th grade | 263(15.3) | 124(16.9) | | |
| High school graduate | 426(24.8) | 171(23.3) | | |
| Some college or AA degree | 528(30.7) | 219(29.8) | | |
| College graduate or above | 334(19.4) | 158(21.5) | | |
| Marital status | | | 5.048 | 0.410 |
| Married | 997(58.0) | 435(59.2) | | |
| Widowed | 79(4.6) | 40(5.4) | | |
| Divorced | 183(10.7) | 74(10.1) | | |
| Separated | 54(3.1) | 27(3.7) | | |
| Unmarried | 218(12.7) | 98(13.3) | | |
| Living with a partner | 187(10.9) | 61(8.3) | | |
| PIR | | | 0.645 | 0.422 |
| Low-income | 483(28.1) | 195(26.5) | | |
| Non-low-income | 1,235(71.9) | 540(73.5) | | |
| General health condition | | | 3.340 | 0.188 |
| Good | 1,233(71.8) | 514(69.9) | | |
| General | 360(21.0) | 176(23.9) | | |
| Bad | 125(7.3) | 45(6.1) | | |
| Sleep hours | | | 2.070 | 0.355 |
| Short sleep hours | 652(38.0) | 262(35.6) | | |
| Normal sleep hours | 962(56.0) | 434(59.0) | | |
| Long sleep hours | 104(6.1) | 39(5.3) | | |
| BMI | | | 1.597 | 0.660 |
| Underweight | 14(0.8) | 10(1.4) | | |
| Normal weight | 319(18.6) | 135(18.4) | | |
| Overweight | 548(31.9) | 235(32.0) | | |
| Obese | 837(48.7) | 355(48.3) | | |
| Alcohol drinking | | | 0.334 | 0.846 |
| Never drinking | 398(23.2) | 165(22.4) | | |
| Low drinking | 1,303(75.8) | 564(76.7) | | |
| Heavy drinking | 17(1.0) | 6(0.8) | | |
| Smoking | | | 1.392 | 0.499 |
| Never smoker | 741(43.1) | 335(45.6) | | |
| Former smoker | 519(30.2) | 208(28.3) | | |
| Now smoker | 458(26.7) | 192(26.1) | | |
| Hypertension | | | 0.078 | 0.781 |
| Yes | 728(42.4) | 307(41.8) | | |
| No | 990(57.6) | 428(58.2) | | |
| Diabetes | | | 0.015 | 0.901 |
| Yes | 284(16.5) | 123(16.7) | | |
| No | 1,434(83.5) | 612(83.3) | | |
| Asthma | | | 0.130 | 0.719 |
| Yes | 312(18.2) | 129(17.6) | | |
| No | 1,406(81.8) | 606(82.4) | | |
| Coronary heart disease | | | 1.277 | 0.258 |
| Yes | 104(6.1) | 36(4.9) | | |
| No | 1,614(93.9) | 699(95.1) | | |
| OSAHS severity | | | 1.547 | 0.461 |
| Mild | 777(45.2) | 339(46.1) | | |
| Moderate | 492(28.6) | 221(30.1) | | |
| Severe | 449(26.1) | 175(23.8) | | |

PIR, ratio of family income to poverty; BMI, body mass index.

sleep hours were the top nine critical factors for predicting depression in OSAHS patients. Figure 6 demonstrates the ranking of importance of these indicators. Subsequently, the above nine variables in binary logistic regression analysis, the final results showed that marital status, general health condition, and smoking were independent influencing factors for depression in OSAHS patients ($p < 0.05$).

## Comparison of model prediction performance

### Comparison of ROC curve prediction performance

To compare the performance of the three models in predicting depression in OSAHS patients, we used the test set data for evaluation. Results show that compared to the lasso model, the sensitivity of the logistic regression model is low, but its specificity and AUC area are higher. This means that the logistic regression model performs better in accurately identifying non-depressed patients, while the lasso model is more sensitive in capturing depressed patients. AUC was used as the preferred index to judge the model's prediction performance. Therefore, the prediction performance of the logistic

10.3389/fpubh.2023.1348803

TABLE 3 Comparison of baseline data between depression group and non-depression group in training data.

| Predictive factors | Non-depression group (*n* = 1,170) | Depression group (*n* = 548) |
|---|---|---|
| Age | 50.65 ± 16.332 | 48.65 ± 15.139 |
| Gender | | |
| Male | 801(68.5) | 290(52.9) |
| Female | 369(31.5) | 258(47.1) |
| Race | | |
| Mexican American | 200(17.1) | 86(15.7) |
| Other Hispanic | 116(9.9) | 62(11.3) |
| Non-Hispanic White | 540(46.2) | 271(49.5) |
| Non-Hispanic Black | 222(19.0) | 104(19.0) |
| Other race | 92(7.9) | 25(4.6) |
| Education level | | |
| Less than 9th grade | 96(8.2) | 71(13.0) |
| 9–11th grade | 168(14.4) | 95(17.3) |
| High school graduate | 283(24.2) | 143(26.1) |
| Some college or AA degree | 354(30.3) | 174(31.8) |
| College graduate or above | 269(23.0) | 65(11.9) |
| Marital status | | |
| Married | 722(61.7) | 275(50.2) |
| Widowed | 52(4.4) | 27(4.9) |
| Divorced | 114(9.7) | 69(12.6) |
| Separated | 29(2.5) | 25(4.6) |
| Never married | 126(10.8) | 92(16.8) |
| Living with a partner | 127(10.9) | 60(10.9) |
| PIR | | |
| Low-income | 259(22.1) | 224(40.9) |
| Non-low-income | 911(77.9) | 324(59.1) |
| General health condition | | |
| Good | 944(80.7) | 289(52.7) |
| General | 189(16.2) | 171(31.2) |
| Bad | 37(3.2) | 88(16.1) |
| Sleep hours | | |
| Short sleep hours | 407(34.8) | 245(44.7) |
| Normal sleep hours | 709(60.6) | 253(46.2) |
| Long sleep hours | 54(4.6) | 50(9.1) |
| BMI | | |
| Underweight | 6(0.5) | 8(1.5) |
| Normal weight | 239(20.4) | 80(14.6) |
| Overweight | 385(32.9) | 163(29.7) |
| Obese | 540(46.2) | 297(54.2) |
| Alcohol drinking | | |
| Never drinking | 260(22.2) | 138(25.2) |
| Low drinking | 899(76.8) | 404(73.7) |
| Heavy drinking | 11(0.9) | 6(1.1) |
| Smoking | | |
| Never smoker | 541(46.2) | 200(36.5) |
| Former smoker | 356(30.4) | 163(29.7) |
| Now smoker | 273(23.3) | 185(33.8) |
| Hypertension | | |
| Yes | 466(39.8) | 262(47.8) |
| No | 704(60.2) | 286(52.2) |
| Diabetes | | |
| Yes | 174(14.9) | 110(20.1) |
| No | 996(85.1) | 438(79.9) |
| Asthma | | |
| Yes | 176(15.0) | 136(24.8) |
| No | 994(85.0) | 412(75.2) |
| Coronary heart disease | | |
| Yes | 63(5.4) | 41(7.5) |
| No | 1,107(94.6) | 507(92.5) |
| OSAHS severity | | |
| Mild | 560(47.9) | 217(39.6) |
| Moderate | 330(28.2) | 162(29.6) |
| Severe | 280(23.9) | 169(30.8) |

PIR, ratio of family income to poverty; BMI, body mass index.

regression model was better than that of the lasso and random forest models. The comparative results are shown in Table 7. The ROC curve is shown in Figure 7.

## Comparison of DCA prediction performance

Clinical decision curve analysis of the prediction model found that when the probability threshold was in the range of 0.19 to 0.82, the prediction model had an excellent net benefit in predicting depression in OSAHS patients. The decision curve analysis results show that the net benefits of the three models were similar for thresholds probability ranging from 0.25 to 0.45. When the threshold probability range was 0.19–0.25 and 0.45–0.82, respectively, the net benefit of the logistic regression model was the most significant. Therefore, the logistic regression model showed better clinical utility than the random forest and lasso models as shown in Figure 8.

Considering the above indicators, the logistic regression model has better predictive performance than the lasso and random forest models in predicting depression in OSAHS patients. And analysis of the influence of related factors, including gender, general health condition, BMI, smoking, OSAHS severity, age, education level, PIR, and asthma.

## Clinical utility

Figure 9 shows an example of a patient's nomogram. The patients who are 35 years of age, female, have a bachelor's degree, low income, have general health, obesity, and smoking in the past, now give up

10.3389/fpubh.2023.1348803

TABLE 4 Multi-collinearity analysis results of predictive variables of depression of OSAHS patients.

| Coefficients[a] | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Non-standardized coefficient | | Standardized coefficient | | | Colinearity statistics | |
| Model | B | Standard error | Beta | *t* | Significance | Allowance | VIF |
| (Constant) | 0.174 | 0.134 | | 1.295 | 0.195 | | |
| Gender | 0.121 | 0.022 | 0.125 | 5.462 | 0.000 | 0.960 | 1.042 |
| Age | −0.002 | 0.001 | −0.070 | −2.679 | 0.007 | 0.745 | 1.342 |
| Education level | −0.008 | 0.009 | −0.022 | −0.868 | 0.385 | 0.817 | 1.224 |
| Marital status | 0.005 | 0.006 | 0.022 | 0.883 | 0.377 | 0.828 | 1.208 |
| PIR | −0.086 | 0.026 | −0.083 | −3.339 | 0.001 | 0.820 | 1.219 |
| General health condition | 0.194 | 0.019 | 0.254 | 10.079 | 0.000 | 0.789 | 1.267 |
| BMI | 0.019 | 0.014 | 0.032 | 1.373 | 0.170 | 0.896 | 1.117 |
| Smoking | 0.031 | 0.013 | 0.055 | 2.319 | 0.021 | 0.897 | 1.115 |
| Hypertension | −0.028 | 0.023 | −0.029 | −1.176 | 0.240 | 0.809 | 1.236 |
| Diabetes | 0.012 | 0.031 | 0.010 | 0.397 | 0.691 | 0.832 | 1.202 |
| Asthma | −0.069 | 0.028 | −0.057 | −2.482 | 0.013 | 0.962 | 1.039 |
| OSAHS severity | 0.031 | 0.013 | 0.055 | 2.394 | 0.017 | 0.945 | 1.059 |

PIR, ratio of family income to poverty; BMI, body mass index.[a]Dependent variable: depression.

TABLE 5 Factors associated with depression in univariable and multivariable analyses in the training set.

| Predictive factors | Univariable model | | Multivariable model | |
|---|---|---|---|---|
| | OR(95%CI) | *p* value | OR(95%CI) | *p* value |
| Gender | 1.93 (1.57, 2.38) | <0.001 | 1.86 (1.48, 2.33) | <0.001 |
| Age | 0.99 (0.99, 1.00) | 0.016 | 0.99 (0.98, 1.00) | 0.011 |
| Race | 0.96 (0.87, 1.05) | 0.363 | | |
| Education level | 0.80 (0.74, 0.87) | <0.001 | 0.95 (0.86, 1.05) | 0.005 |
| Marital status | 1.11 (1.05, 1.17) | <0.001 | 1.03 (0.97, 1.09) | 0.397 |
| PIR | 0.41 (0.33, 0.51) | <0.001 | 0.66 (0.51, 0.86) | <0.001 |
| General health condition | 2.85 (2.41, 3.38) | <0.001 | 2.43 (2.00, 2.94) | <0.001 |
| Sleep hours | 0.85 (0.71, 1.01) | 0.072 | | |
| BMI | 1.22 (1.07, 1.39) | 0.004 | 1.11 (0.96, 1.28) | 0.025 |
| Alcohol drinking | 0.87 (0.69, 1.09) | 0.217 | | |
| Smoking | 1.35 (1.19, 1.53) | <0.001 | 1.17 (1.02, 1.35) | 0.047 |
| Hypertension | 0.72 (0.59, 0.89) | 0.002 | 0.86 (0.67, 1.10) | 0.062 |
| Diabetes | 0.70 (0.53, 0.91) | 0.007 | 1.06 (0.77, 1.46) | 0.758 |
| Asthma | 0.54 (0.42, 0.69) | <0.001 | 0.71 (0.54, 0.94) | 0.011 |
| Coronary heart disease | 0.70 (0.47, 1.06) | 0.091 | | |
| OSAHS severity | 1.25 (1.10, 1.41) | <0.001 | 1.18 (1.03, 1.35) | 0.024 |

PIR, ratio of family income to poverty; BMI, body mass index.

smoking, do not admit to a history of asthma, suffer from severe OSAHS. According to the diagram model, the patients with a total score of 163.5 points have a probability of about 57% of depression.

# Discussion

The main strength of this study is the use of the NHANES extensive sample database and the use of ML models to predict and identify potential influencing factors of depression in OSAHS patients. In this study, we constructed and trained a model based on the depression of OSAHS adults in the US. We considered variables such as demographic characteristics, lifestyle, and health factors and weighted them during the construction of the model. Through our selection model and can be output in probability of OSAHS adult depression in the United States. These results improve the intelligence of the mental health care system and have a positive impact. Sleep-related healthcare providers can use these ML algorithms to identify OSAHS patients who are potentially at risk for depression, which in turn can detect early if they are suffering from depression and help them with early intervention. The research also further increases the chances of OSAHS patient's access to mental health services. This study will explore the incidence of depression in American adults with OSAHS, the influencing factors, and the predictive power of risk prediction models. The practical application of these findings to real life will also be discussed.

## Depression among OSAHS adults in the US

The present study was based on three cycles of NHANES data (2005–2006, 2007–2008, and 2015–2016) and included variables that included demographic characteristics, lifestyle, and health information. A total of 2,453 OSAHS patients were screened, including 1,671 non-depressed patients and 782 depressed patients. The incidence of depression was 31.9%, slightly lower than the findings of Houda Gharsalli and Siddharth Bajpai et al. (4, 13). The difference in the incidence of depression may be due to the different inclusion criteria used in the OSAHS population. The study by Houda

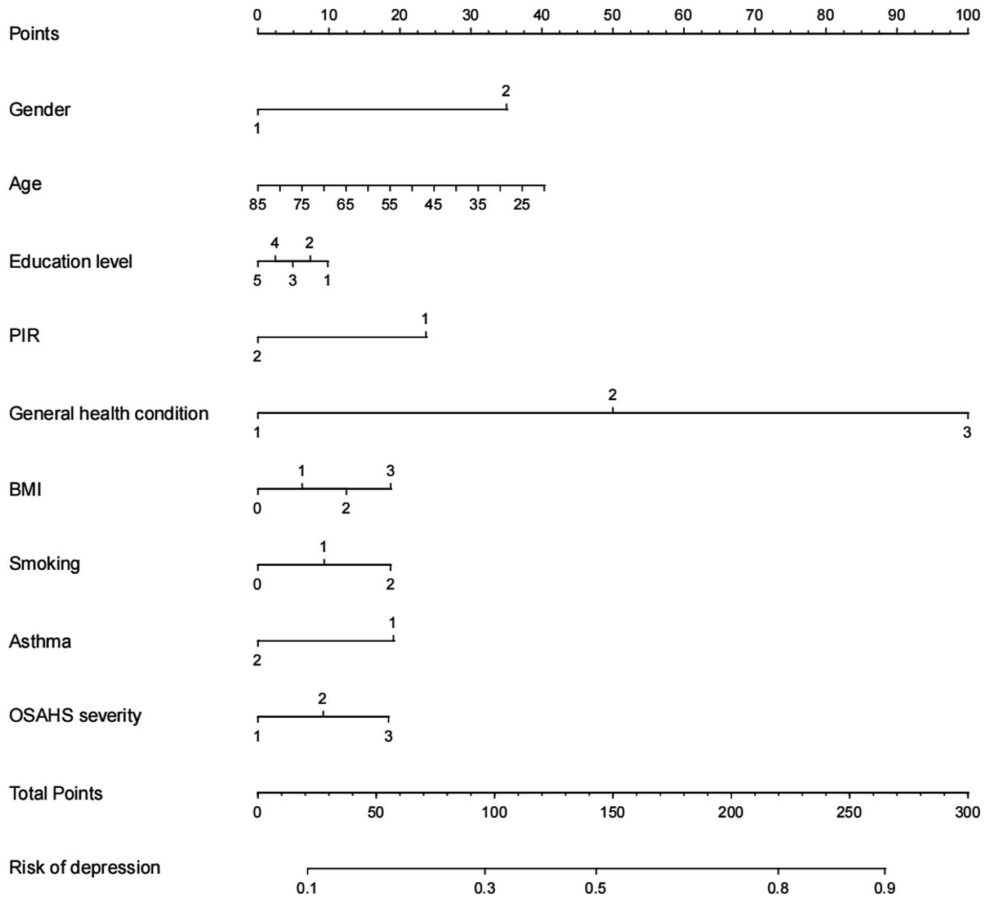



FIGURE 4
Nomogram prediction model for logistic risk of depression in OSAHS patients. PIR, ratio of family income to poverty; BMI, body mass index.

TABLE 6 Nomogram of relevant factors in the assignment method.

| Risk factors | Assignment |
|---|---|
| Gender | "Male" = 1, "Female" = 2 |
| Age | Original value entry |
| Education level | "Less than 9th grade" = 1, "9-11th grade" = 2, "High school graduate" = 3 , "Some college or AA degree" = 4, "College graduate or above" = 5 |
| PIR | "Low-income" = 1, "Non-low-income" = 2 |
| General health condition | "Good" = 1, "General" = 2, "Bad" = 3 |
| BMI | "Underweight" = 0, "Normal weight" = 1 , "Overweight" = 2, "Obese" = 3 |
| Smoking | "Never smoker" = 0, "Former smoker" = 1, "Now smoker" = 2 |
| Asthma | "Yes" = 1, "No" = 2 |
| OSAHS severity | "Mild" = 1, "Moderate" = 2, "Severe" = 3 |

PIR, ratio of family income to poverty; BMI, body mass index.

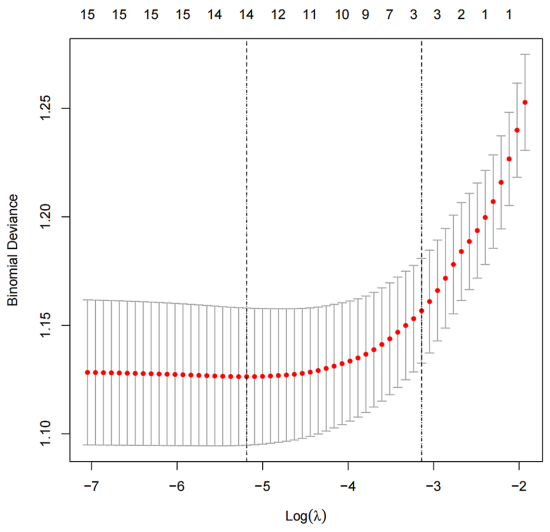



FIGURE 5
Lasso model ten-fold cross-validation method to screen predictors process diagram.

Gharsalli and Siddharth Bajpai et al. employed specialized diagnostic equipment, such as polysomnography (PSG), with an apnea-hypopnea index (AHI) ≥5 as the diagnostic criteria for OSAHS. In contrast, this study was judged and included only based on patients' self-reported results on "How often do you snore/stop breathing?." At the same time, temporal and geographic differences may also contribute to lower rates of depression in the OSAHS population than the results of other studies.

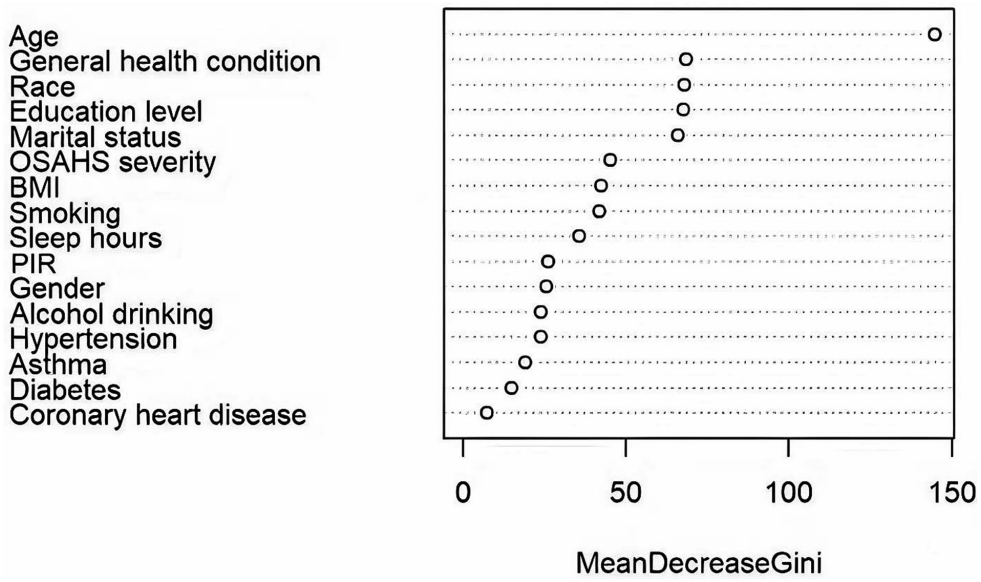



FIGURE 6
Variable importance plot. PIR, ratio of family income to poverty; BMI, body mass index.

TABLE 7 Comparison of prediction performance of three kinds of models.

| Model | AUC | 95%CI | Sensitivity | Specificity | Youden index |
|---|---|---|---|---|---|
| Random forest | 0.710 | (0.669, 0.752) | 0.624 | 0.727 | 0.350 |
| Lasso | 0.727 | (0.687, 0.767) | 0.756 | 0.599 | 0.355 |
| Logistic regression | 0.746 | (0.707, 0.785) | 0.603 | 0.764 | 0.367 |

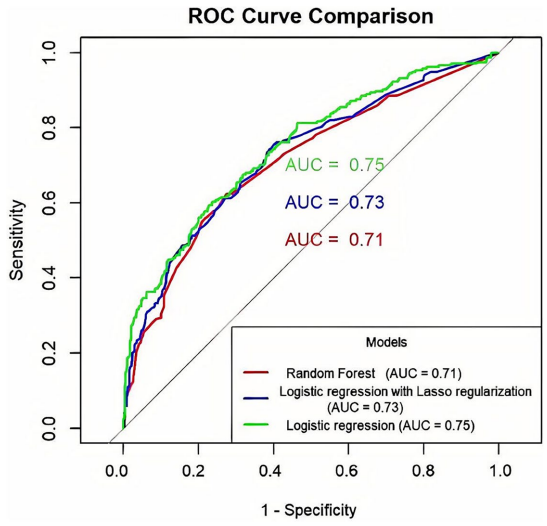



FIGURE 7
Comparison of ROC curve prediction performance of three prediction models for OSAHS patients with depression (The x-axis indicates the false positive rate, and the y-axis represents sensitivity.).

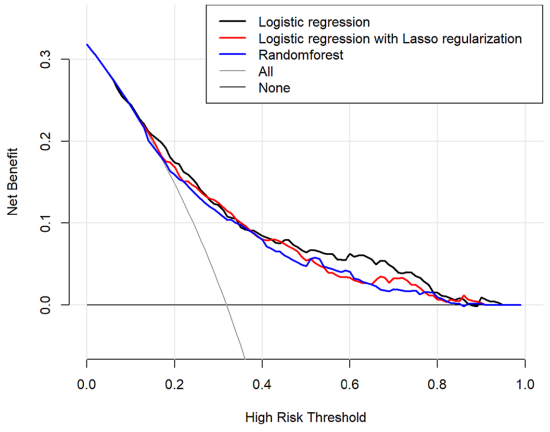



FIGURE 8
Comparison of the predictive performance of three predictive models decision-making curve analysis (DCA) predictions (The x-axis indicates the high risk threshold, and the y-axis represents net benefit.).

## Factors influencing the development of depression in OSAHS adults in the US

This study used three ML algorithms, logistic regression, lasso, and random forest, to construct a predictive model of depression in the US OSAHS population. Results show that the logistic regression model is better than the random forest and logistic regression models in terms of specificity, Youden index, and AUC. However, its sensitivity is lower. Therefore, according to the results of this study, among the three prediction models, the logistic regression model performs better than the lasso model and the random forest model. In predicting the occurrence of depression in the OSAHS population, the model is

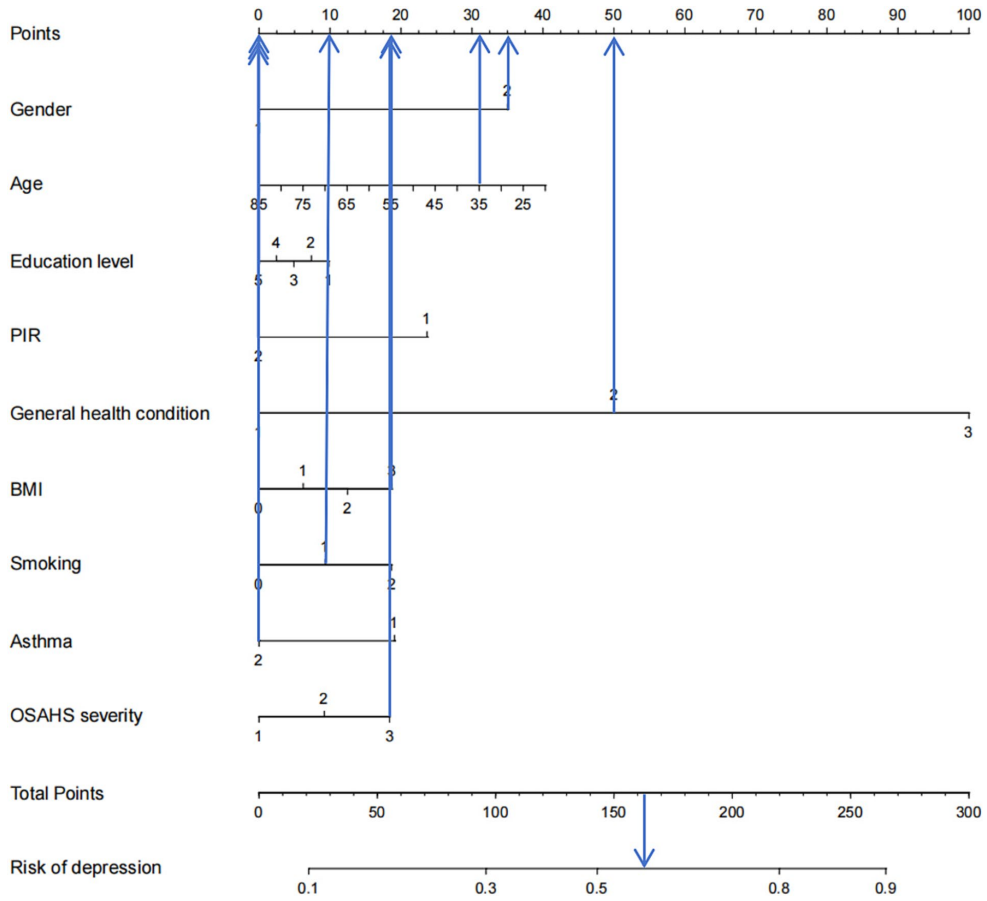



FIGURE 9
Example of nomogram. PIR, ratio of family income to poverty; BMI, body mass index.

affected by factors such as gender, age, education level, PIR, general health condition, BMI, smoking, asthma, and OSAHS severity. In addition, the model results are visually presented in Figure 4 in this study.

In terms of sociodemographic characteristics， according to the findings of Alimohamad Asghari, Magali Saint Martin, and Yaozhang Dai, as well as Rachel H. Salk, women are more prone to depression relative to men (35–37). This difference exists not only in China but also in most countries and cultures worldwide. In general, women are twice as likely to suffer from depression (OR = 1.95) than men, which may be related to frequent hormonal disturbances due to genetic and physiological factors. Some studies have shown that women are more likely to experience depression during menopause (38, 39). According to the results of this study, obese individuals are more likely to suffer from depression compared to normal weight or underweight individuals, which is consistent with the findings of Ashley Wendell Kranjac and Tuula H. Heiskanen (40, 41). According to Ashley Wendell Kranjac et al., after taking into account the combined effects of gender and BMI, obese women were 43% more likely to experience depression than normal-weight women. Tuula H. Heiskanen's study, a 6-year prospective study of outpatients, found that subjects with significant weight gain were more likely to develop major depression (41). This may be because obese people often suffer from associated chronic inflammation. Immune cells in adipose tissue produce signaling proteins related to inflammation, and some of these proteins, such as cytokines, are strongly related to mental health problems and have even been used as biomarkers for depression (42).

Compared to older people, teens are at higher risk of depression. This is consistent with the findings of Stephanie Wagner (43). In the study of Stefanie Wagner, hospitalized patients with depression were divided into four different age groups. The results showed that the aged 18 to 29 years old young patients are more likely to show extreme behavior, such as suicide and drug abuse. In contrast, middle-aged and older patients aged between 50 and 65 years were more likely to show mild depression, such as decreased sexual interest. This may be related to adolescent adolescence body hormone disorder and lack of mental toughness, leading to mental instability (44).

The higher the level of education, the lower the risk of depression. Early studies have pointed out that depression is associated with low levels of education (45), and education has a significant effect on the development of depression. That is, illiterates are more likely to have more severe depression (46). This may be due to the low level of education, which leads to narrower social contacts and fewer avenues for problem solving when experiencing negative life events and is more likely to cause heavier negative emotions, which can lead to depression.

The findings suggest that low-income people are more likely to suffer from depression than non-low-income people, which was confirmed in the study of Glaesmer (47). It is estimated that while around half of people with depression in high-income countries are

not diagnosed or treated, in low-income and middle-income countries, the proportion may be as high as 80%–90%. According to a December 2020 commentary in the journal *Science*, there is a causal interaction between poverty and mental illness. This means that people with low incomes are more vulnerable to the threat of mental illness, and at the same time, mental illness is also one of the vital causes of people with low incomes (13). This phenomenon may be because low-income people usually face financial difficulties in their daily lives and lack sufficient funds to meet basic needs such as food, housing, and healthcare. This financial pressure may make them feel helpless, anxious, depressed, and more prone to depression.

The poorer the general health condition, the greater the probability of the risk of depression. This finding is consistent with results from the World Health Survey published by Saba Moussavi in The *Lancet*. This study showed that depression was associated with the lowest health scores, both in isolation and in co-occurrence with other chronic diseases (48). A research study by Nicolas Zdanowicz also indicated that physical health and its improvement are related to the level of depression (49), and Érica Dorigatti de Ávila clarified that patients without depression have a higher level of mental health (50). The possible cause is physical factors. Bad health condition is accompanied by chronic illness, pain, or discomfort that may affect an individual's emotional and psychological state, thereby increasing the risk of depression. In addition, due to limited physical function and lower quality of life, people with depression may harm their self-worth and self-esteem. This psychological stress may further aggravate depression. Chronic physical illnesses and health problems may cause individuals to develop negative feelings and increase the risk of depression. And overall, poor health increases the personal burden of life. People may need more time and money to treat their diseases, which may lead to economic stress and anxiety, thus increasing the risk of depression.

According to the study, smoking is considered to be one of the factors that predict the high risk of depression in OSAHS patients. Specific studies have shown higher rates of depression among current and former smokers compared with never smokers. According to a survey of the U.S. population, LUIS G. ESCOBEDO found that former smokers are more likely to develop depression, especially those who have a history of major depression (51). The cross-sectional study conducted by Tana M. Luger also noted that current smokers were more likely to develop depression than never-smokers. In contrast, current smokers were more likely to develop depression than former smokers (52). This phenomenon may be because nicotine intake from smoking can bring short-term pleasure and relaxation. Still, long-term smoking may lead to neurotransmitter disorders, thereby affecting emotional stability and increasing the risk of depression.

People with asthma are more likely to suffer from depression than people without asthma. This idea is supported by a biological linkage study by Mingdi Jiang et al. It implies that the inflammatory response may be a critical factor in regulating the common pathways of depression and asthma (53). The results of Mahima Akula's study also confirmed the correlation between the two (54). Furthermore, a bidirectional association between asthma and depression was observed in Hyo Geun Choi's study (55). In a clinical practice study of adolescents with asthma, 11.5% had depression (42). This may be because people with asthma may feel negative emotions such as low self-esteem, anxiety, and fear. Due to asthma having wave properties, some patients may need to avoid social situations and activities, which may lead to individual patients being isolated and isolated, which will affect their psychological health. In addition, patients with asthma often face physical discomfort such as dyspnea and chest tightness, which may affect the individual's emotional and psychological state, which in turn exacerbates negative emotions and increases the risk of depression.

As the severity of OSAHS increases, so does the risk of depression. Cass Edwards et al. Research confirmed the results and found that with the rise in the severity of OSAHS, PHQ score and the incidence of depression also gradually increased (56). This result may be due to the OSAHS patients during apnea oxygen supply is insufficient and hypoxemia. With the deterioration of OSAHS, hypoxemia has a more serious negative impact on brain function and emotion regulation, which leads to the occurrence of depression. In addition, patients with OSAHS may experience symptoms such as fatigue, lethargy, and difficulty concentrating during the day due to decreased sleep quality. These limitations in daily functioning may negatively affect an individual's psychological state and increase the risk of depression.

Severe OSAHS can also cause sleep disturbances, which in turn can lead to a decline in social activities and work ability. This situation is further exacerbated by negative emotions such as anxiety, low self-esteem, and depression. Therefore, it can be concluded that there is a strong correlation between OSAHS and depression and that its severity is positively related to the risk of depression.

## Evaluation and application of risk prediction models for depression

Research results show that the logistic regression model is better than the random forest and lasso models regarding specificity, Youden index, and AUC area. To validate the model in the clinical application value of this study, the DCA was used, and the net income of the model was used in the comparison. The results showed that the logistic regression model had the most significant net benefit within the vast majority of threshold probability ranges (0.19 to 0.25 and 0.45 to 0.82) and had a good effect on clinical application. Therefore, the comprehensive prediction ability of the comparison results shows that the logistic regression model is superior to the lasso and random forest models. It should be noted that the lasso model may ignore some relevant features due to the high correlation between features. In addition, selecting the appropriate regularization parameter needs experience, cross validation, and other methods. This increases the complexity of model tuning (57). The random forest model is composed of multiple decision trees. Although feature importance can be used to understand the contribution of each feature to the model, the overall model is less explanatory than the logistic regression model. In addition, due to the random forest model to build a decision tree and perform multiple feature selection and integration of the operation, its training time is relatively long (58). In contrast, the logistic regression model, as a generalized linear model, uses the least squares method to fit the model and thus has high accuracy (59). Logistic regression models can make predictions and explore the direction and degree of influence between independent and dependent variables, so they have better explanatory power (49). Logistic regression models can be quantified and visualized by nomograms, which have outstanding advantages in auxiliary diagnosis in the medical field (60).

Before applying a risk prediction model for depression in the OSAHS population, it is necessary to select the most appropriate model and adjust the parameters to maximize the prediction effect of the model to improve the accuracy and sensitivity of identifying OSAHS patients at high risk for depression. Subsequently, predictive models need to be translated into forms applicable to the community and clinic, such as nomograms or mobile applications that allow physicians and sleep technologists to calculate the probability of depression easily and quickly. This will help to protect OSAHS patients in advance and effectively prevent the occurrence of depression.

In this study, ML models, especially logistic regression models, demonstrated excellent depression prediction and recognition capabilities in a large dataset. In contrast to traditional statistical methods, ML methods no longer require the researcher to specify the relevant variables subjectively but can automatically identify the variables associated with the outcome variables in the data set. This is precisely one of the advantages of ML in building clinical prediction models. Future research could apply ML methods to model and compare in longitudinal studies to obtain basic information such as the prevalence of depression. For example, as far as studies in predicting depression are concerned, studies like those done by Dai Su et al. in a longitudinal study of the older adult population in China are a good example (20). In addition, it is possible to cross-combine multiple models in ML to form a hybrid model and verify whether the hybrid model outperforms the traditional single ML model in terms of predictive performance (61). To better psychological doctors and health care at all levels, provide appropriate information and services.

In a clinical sense, physicians can assess whether a patient is at risk for depression based on gender, general health condition, BMI, smoking, OSAHS severity, age, education level, PIR, and asthma. Once a patient is identified as being at risk for depression, interventions can be implemented, including medication, psychotherapy, and behavioral changes. Early intervention and treatment can help patients reduce the symptoms of depression, improve their sleep quality and quality of life, and improve the effect of OSAHS treatment.

When using the logistic regression model to predict depression in OSAHS patients, this study suggests that specificity, sensitivity, and Youden index should be considered comprehensively, and the choice of specificity and sensitivity should be weighed according to the specific situation. In addition, in clinical practice, demonstrating the effectiveness of these indicators is also necessary. To ensure the applicability and reliability of the model, it is recommended to continue to collect larger scale, diverse data and use these data to validate and replicate the findings. By expanding the dataset's scope, the model's predictive performance in different populations and contexts can be more fully evaluated. Such efforts can help improve the accuracy and generalization ability of the model and provide a more reliable basis for future clinical practice. When the model is applied in clinical practice, it needs to be comprehensively evaluated by combining clinical experience and individual patient differences. This process involves interpretation and interpretation of the model predictions and a comprehensive consideration of the patient's situation. At the same time, it is necessary to continuously update and optimize the model to improve its predictive performance and clinical utility. Finally, for the research on the influencing factors, the specific mechanism of each factor's influence on the occurrence of depression can be further explored in depth. This includes detailed studies of biological, psychosocial, and other factors to reveal their associations with depression. At the same time, how to prevent and treat depression by intervening in these factors can also be studied.

The results of this study can be incorporated into the development and implementation of relevant public health policies. Government departments can develop prevention and intervention strategies for depression in OSAHS patients according to the logistic regression model and influencing factors identified in this study. By formulating strategies based on these models and influencing factors, patients' mental health can be effectively improved. Department of Public Health and medical institutions can reasonably allocate resources to strengthen the prevention and treatment of depression in patients with OSAHS. This could include making more counselors or psychotherapists available and improving screening and diagnostic facilities for depression, among others. In related health education activities, the findings of this study should be disseminated to improve the public's awareness and attention to depression in OSAHS patients. This helps reduce the social discrimination against depression to promote social support and understanding. It can also guide medical practice. By applying a logistic regression model, the medical personnel can more accurately identify the existence of the risk of depression in patients with OSAHS, which can carry on the intervention and treatment in a timely manner. This method helps to improve the early diagnostic rate of depression and to provide more effective personalized treatment options for patients to enhance their mental health. On prevention strategies, the results of this study are to develop in OSAHS patients with depression prevention strategy provides an essential basis. Based on the logistic regression model, the influence factors of medical personnel can be targeted to carry out the intervention measures, including strengthening the OSAHS patients' psychological health education and psychological support services and establishing health management programs.

It should be noted that our results may have been affected by various potential biases in the NHANES database. One of the main biases is the low participation rate of specific populations, which may introduce sampling bias. Therefore, the data representation may be poor and cannot fully reflect the OSAHS group. This may affect the generalizability of the findings. Therefore, it is suggested that a multicenter study be carried out to expand the sample representativeness and increase the external data validity and generalization ability. Second, it should be noted that some of the data in the NHANES database rely on participant self-reports, such as diagnostic information for patients with OSAHS. This dependence may be limited by the deviation of subjective evaluation and memory bias, the influence of such factors to assess the severity of OSAHS symptoms, or inaccuracy. This may affect the reliability and accuracy of the findings. Therefore, to improve the objectivity and accuracy of diagnosis, this study suggested introducing objective measurement tools, such as more sleep figures (PSG). Using these objective measurement tools can reduce the dependence on self-reported respondents and help assess the symptoms and severity of OSAHS more accurately. Suggestions in the study consider integrating other data sources, such as clinical records and medical insurance databases, to get more comprehensive and multi-angle OSAHS data. By combining these diverse data sources, the reliability and generalizability of the findings can be increased.

In addition, future research should further optimize the ML model's performance to improve the prediction accuracy of depression in patients with OSAHS. The introduction of other machine learning algorithms or deep learning methods can also be considered to explore better predictive models. These methods can model and analyze the data from different perspectives and improve the stability and accuracy of the model. By exploring effective intervention strategies and conducting intervention trials, the effects of other methods can be evaluated, and their long-term effects can be followed up to reduce the incidence of depression and improve the quality of life of patients with OSAHS.

## Conclusion

This research uses the NHANES database to establish three ML models, the logistic regression model, the lasso model, and the random forest model, to predict depression in the OSAHS group and identify the related factors. Among them, the logistic regression model was superior to the lasso and random forest models' overall prediction performance. By drawing the nomogram and applying it to the sleep testing center or sleep clinic, sleep technicians and medical staff can quickly and easily identify whether OSAHS patients have depression to carry out the necessary referral and psychological treatment.

## Limitations

The data used in this study were obtained from the NHANES database, which includes a variety of relevant variables, such as smoking history and sleep disorders. Most of these variables are based on patient self-reporting and may have subjective bias, affecting the data's objective accuracy.

Due to the lack of relevant variables in the NHANES database, such as the AHI index, minimum oxygen saturation, etc., including these predictors may allow for better model prediction performance.

In this study, due to the limited sample size of the screening, we did not perform external validation to verify the effect of this predictive model. However, we hope that future studies will externally validate this model by conducting a multicenter study with an increased sample size and applying it to a community population for screening further to test the generalization ability and robustness of the model.

## 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 NCHS Research Ethics Review Board (ERB) reviewed and approved the studies involving human participants. Written informed consent for participation was not required for this study by the national legislation and the institutional requirements.

## Author contributions

EL: Writing – original draft, Conceptualization, Data curation, Formal analysis, Resources, Visualization. FA: Writing – review & editing, Investigation, Methodology, Software. CL: Writing – review & editing, Funding acquisition, Project administration, Supervision, Validation.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the Social Science Planning Foundation of Liaoning Province (L21CSH005).

## Acknowledgments

Thanks to all the authors for their hard work on this study.

## 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



## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1348803/full#supplementary-material

## References

1. Aalbers S, Fusar-Poli L, Freeman RE, Spreen M, Ket JCF, Vink AC, et al. Music therapy for depression. *Cochrane Database Syst Rev*. (2017) 2017:CD004517. doi: 10.1002/14651858.CD004517.pub3

2. Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips MR, et al. No health without mental health. *Lancet*. (2007) 370:859–77. doi: 10.1016/s0140-6736(07)61238-0

3. World Health Organization. *Depression and other common mental disorders: global health estimates*. Geneva: World Health Organization (2017).

4. Douglas N, Young A, Roebuck T, Ho S, Miller BR, Kee K, et al. Prevalence of depression in patients referred with snoring and obstructive sleep apnoea. *Intern Med J*. (2013) 43:630–4. doi: 10.1111/imj.12108

5. 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 Psychiatr*. (2017) 174:154–62. doi: 10.1176/appi.ajp.2016.16010077

6. Herrman H, Patel V, Kieling C, Berk M, Buchweitz C, Cuijpers P, et al. Time for united action on depression: a lancet–World psychiatric association commission. *Lancet*. (2022) 399:957–1022. doi: 10.1016/s0140-6736(21)02141-3

7. Rundo JV. Obstructive sleep apnea basics. *Cleve Clin J Med*. (2019) 86:2–9. doi: 10.3949/ccjm.86.s1.02

8. Veasey SC, Solomon CG, Rosen IM. Obstructive Sleep Apnea in Adults. *N Engl J Med*. (2019) 380:1442–9. doi: 10.1056/NEJMcp1816152

9. Edwards C, Almeida OP, Ford AH. Obstructive sleep apnea and depression: a systematic review and meta-analysis. *Maturitas*. (2020) 142:45–54. doi: 10.1016/j.maturitas.2020.06.002

10. Peppard PE, Szklo-Coxe M, Hla KM, Young T. Longitudinal association of sleep-related breathing disorder and depression. *Arch Intern Med*. (2006) 166:1709–15. doi: 10.1001/archinte.166.16.1709

11. Harris M, Glozier N, Ratnavadivel R, Grunstein RR. Obstructive sleep apnea and depression. *Sleep Med Rev*. (2009) 13:437–44. doi: 10.1016/j.smrv.2009.04.001

12. Chen Y-H, Keller JK, Kang J-H, Hsieh H-J, Lin H-C. Obstructive sleep apnea and the subsequent risk of depressive disorder: a population-based follow-up study. *J Clin Sleep Med*. (2013) 09:417–23. doi: 10.5664/jcsm.2652

13. Gharsalli H, Harizi C, Zaouche R, Sahnoun I, Saffar F, Maalej S, et al. Prevalence of depression and anxiety in obstructive sleep apnea. *Tunis Med*. (2022) 100:525–33.

14. Mihalj M. Depression and fatigue are due to obstructive sleep apnea in multiple sclerosis. *Acta Clin Croat*. (2022) 61:599–604. doi: 10.20471/acc.2022.61.04.05

15. Lang CJ, Appleton SL, Vakulin A, McEvoy RD, Wittert GA, Martin SA, et al. Co-morbid OSA and insomnia increases depression prevalence and severity in men. *Respirology*. (2017) 22:1407–15. doi: 10.1111/resp.13064

16. Haddock N, Wells ME. The association between treated and untreated obstructive sleep apnea and depression. *Neurodiagn J*. (2018) 58:30–9. doi: 10.1080/21646821.2018.1428462

17. 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/j.jad.2018.08.073

18. Hatton CM, Paton LW, McMillan D, Cussens J, Gilbody S, Tiffin PA. Predicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare. *J Affect Disord*. (2019) 246:857–60. doi: 10.1016/j.jad.2018.12.095

19. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. *Psychol Med*. (2019) 49:1426–48. doi: 10.1017/s0033291719000151

20. Su D, Zhang X, He K, Chen Y. Use of machine learning approach to predict depression in the elderly in China: a longitudinal study. *J Affect Disord*. (2021) 282:289–98. doi: 10.1016/j.jad.2020.12.160

21. Xia F, Li Q, Luo X, Wu J. Machine learning model for depression based on heavy metals among aging people: a study with National Health and Nutrition Examination Survey 2017–2018. *Front Public Health*. (2022) 10:10. doi: 10.3389/fpubh.2022.939758

22. Li M, Zou X, Lu H, Li F, Xin Y, Zhang W, et al. Association of sleep apnea and depressive symptoms among US adults: a cross-sectional study. *BMC Public Health*. (2023) 23:427. doi: 10.1186/s12889-023-15358-8

23. Yan X, Wang L, Liang C, Zhang H, Zhao Y, Zhang H, et al. Development and assessment of a risk prediction model for moderate-to-severe obstructive sleep apnea. *Front Neurosci*. (2022) 16:936946. doi: 10.3389/fnins.2022.936946

24. Mazzotti DR, Keenan BT, Lim DC, Gottlieb DJ, Kim J, Pack AI. Symptom subtypes of obstructive sleep apnea predict incidence of cardiovascular. *Am J Respir Crit Care Med*. (2019) 200:493–506. doi: 10.1164/rccm.201808-1509OC

25. Schmickl CN, Orr JE, Kim P. Point-of-care prediction model of loop gain in patients with obstructive sleep. *BMC Pulm Med*. (2022) 22:158. doi: 10.1186/s12890-022-01950-y

26. Keshavarz Z, Rezaee R. Obstructive sleep apnea: a prediction model using supervised machine learning. *Stud Health Technol Inform*. (2020) 272:387–90. doi: 10.3233/SHTI200576

27. Curtin LR, Mohadjer LK, Dohrmann SM, Kruszon-Moran D, Mirel LB, Carroll MD, et al. National Health and nutrition examination survey: sample design, 2007–2010. *Vital Health Stat 2*. (2013) 160:1–23.

28. Johnson CL, Dohrmann SM, Burt VL, Mohadjer LK. National health and nutrition examination survey: sample design, 2011-2014. *Vital Health Stat 2*. (2014) 162:1–33.

29. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. (2001) 19:708–9. doi: 10.1046/j.1525-1497.2001.016009606.x

30. McIntyre RS, Lee Y, Rong C, Rosenblat JD, Brietzke E, Pan Z, et al. Ecological momentary assessment of depressive symptoms using the mind.me application: Convergence with the Patient Health Questionnaire-9 (PHQ-9). *J Psychiatr Res*. (2021) 135:311–317. doi: 10.1016/j.jpsychires.2021.01.012

31. Kroenke K. PHQ-9: global uptake of a depression scale. *World Psychiatry*. (2021) 20:135–6. doi: 10.1002/wps.20821

32. Luo L, Sun W, Han Y, Zhang W, Liu C, Yin S. Importance evaluation based on random Forest algorithms: insights into the relationship between negative air ions variability and environmental factors in urban green spaces. *Atmos*. (2020) 11:706. doi: 10.3390/atmos11070706

33. Zhang Y, Zhang Z. Construction and validation of nomograms combined with novel machine learning algorithms to predict early death of patients with metastatic colorectal cancer. *Front Public Health*. (2022) 10:1008137. doi: 10.3389/fpubh.2022.1008137

34. Lv J, Ren H, Guo X, Meng C, Fei J, Mei H, et al. Nomogram predicting bullying victimization in adolescents. *J Affect Disord*. (2022) 303:264–72. doi: 10.1016/j.jad.2022.02.037

35. Asghari A, Mohammadi F, Kamrava SK, Tavakoli S, Farhadi M. Severity of depression and anxiety in obstructive sleep apnea syndrome. *Eur Arch Otorhinolaryngol*. (2012) 269:2549–53. doi: 10.1007/s00405-012-1942-6

36. Sforza E, Saint Martin M, Barthélémy JC, Roche F. Mood disorders in healthy elderly with obstructive sleep apnea: a gender effect. *Sleep Med*. (2016) 19:57–62. doi: 10.1016/j.sleep.2015.11.007

37. Arias-Carrion O, Dai Y, Li X, Zhang X, Wang S, Sang J, et al. Prevalence and predisposing factors for depressive status in Chinese patients with obstructive sleep apnoea: a large-sample survey. *PLoS One*. (2016) 11:e0149939. doi: 10.1371/journal.pone.0149939

38. Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: meta-analyses of diagnoses and symptoms. *Psychol Bull*. (2017) 143:783–822. doi: 10.1037/bul0000102

39. Saunamäki T, Jehkonen M. Depression and anxiety in obstructive sleep apnea syndrome: a review. *Acta Neurol Scand*. (2007) 116:277–88. doi: 10.1111/j.1600-0404.2007.00901.x

40. Kranjac AW, Nie J, Trevisan M, Freudenheim JL. Depression and body mass index, differences by education: evidence from a population-based study of adult women in the U.S. Buffalo-Niagara region. *Obes Res Clin Pract*. (2017) 11:63–71. doi: 10.1016/j.orcp.2016.03.002

41. Heiskanen TH, Koivumaa-Honkanen HT, Niskanen LK, Lehto SM, Honkalampi KM, Hintikka JJ, et al. Depression and major weight gain: a 6-year prospective follow-up of outpatients. *Compr Psychiatry*. (2013) 54:599–604. doi: 10.1016/j.comppsych.2013.02.001

42. Licari A, Castagnoli R, Ciprandi R, Brambilla I, Guasti E, Marseglia GL, et al. Anxiety and depression in adolescents with asthma: a study in clinical practice. *Acta Biomed*. (2022) 93:e2022021. doi: 10.23750/abm.v93i1.10731

43. Wagner S, Wollschläger D, Dreimüller N, Engelmann J, Herzog DP, Roll SC, et al. Effects of age on depressive symptomatology and response to antidepressant treatment in patients with major depressive disorder aged 18 to 65 years. *Compr Psychiatry*. (2020) 99:99. doi: 10.1016/j.comppsych.2020.152170

44. Luo Y, Wang A, Zeng Y, Zhang J. A latent class analysis of resilience and its relationship with depressive symptoms in the parents of children with cancer. *Support Care Cancer*. (2022) 30:4379–87. doi: 10.1007/s00520-022-06860-7

45. Wickersham A, Sugg HVR, Epstein S, Stewart R, Ford T, Downs J. Systematic review and meta-analysis: the association between child and adolescent depression and later educational attainment. *J Am Acad Child Adolesc Psychiatry*. (2021) 60:105–18. doi: 10.1016/j.jaac.2020.10.008

46. da Costa Dias FL, Teixeira AL, Guimarães HC, Santos APB, Resende EPF, Machado JCB, et al. The influence of age, sex and education on the phenomenology of depressive symptoms in a population-based sample aged 75+ years with major depression: the Pietà study. *Aging Ment Health*. (2019) 25:462–7. doi: 10.1080/13607863.2019.1698517

47. Glaesmer H, Riedel-Heller S, Braehler E, Spangenberg L, Luppa M. Age- and gender-specific prevalence and risk factors for depressive symptoms in the elderly: a population-based study. *Int Psychogeriatr*. (2011) 23:1294–300. doi: 10.1017/s1041610211000780

48. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet*. (2007) 370:851–8. doi: 10.1016/s0140-6736(07)61415-9

49. AJ VD, Pasupathy KS, Huschka TR, Heaton HA, Hellmich TR, Sir MY. Extended patient alone time in emergency department leads to increased risk of 30-day hospitalization. *J Patient Saf*. (2021) 17:e1458–64. doi: 10.1097/PTS.0000000000000545

50. de Ávila ÉD, de Molon RS, Loffredo LCM, Massucato EMS, Hochuli-Vieira E. Health-related quality of life and depression in patients with dentofacial deformity. *Oral Maxillofac Surg*. (2012) 17:187–91. doi: 10.1007/s10006-012-0338-5

51. Escobedo LG, Kirch DG. Depression and smoking initiation among US Latinos. *Addiction*. (1996) 91:113–9.

52. Weinberger AH, Mazure CM, Morlett A, McKee SA. Two decades of smoking cessation treatment research on smokers with depression: 1990-2010. *Nicotine Tob Res*. (2012) 15:1014–31. doi: 10.1093/ntr/nts213

53. Jiang M, Qin P, Yang X. Comorbidity between depression and asthma via immune-inflammatory pathways: a meta-analysis. *J Affect Disord*. (2014) 166:22–9. doi: 10.1016/j.jad.2014.04.027

54. Akula M, Kulikova A, Khan DA, Brown ES. The relationship between asthma and depression in a community-based sample. *J Asthma*. (2018) 55:1271–7. doi: 10.1080/02770903.2017.1418885

55. Choi HG, Kim J-H, Park J-Y, Hwang YI, Jang SH, Jung K-S. Association between asthma and depression: a National Cohort Study. *J Allergy Clin Immunol Pract*. (2019) 7:1239–1245.e1. doi: 10.1016/j.jaip.2018.10.046

56. Edwards C, Mukherjee S, Simpson L, Palmer LJ, Almeida OP, Hillman DR. Depressive symptoms before and after treatment of obstructive sleep apnea in men and women. *J Clin Sleep Med*. (2015) 11:1029–38. doi: 10.5664/jcsm.5020

57. Dai P, Chang W, Xin Z, Cheng H, Ouyang W, Luo A. Retrospective study on the influencing factors and prediction of hospitalization expenses for chronic renal failure in China based on random forest and LASSO regression. *Front Public Health*. (2021) 9:9. doi: 10.3389/fpubh.2021.678276

58. Zhang C, Ma Y. *Ensemble machine learning*. Berlin: Springer Science & Business Media (2012).

59. Nguyen PTT, Hoang DV, Pham KM, Nguyen HT. A multiple logistic regression model based on gamma-Glutamyl transferase as a biomarker for early prediction of drug-induced liver injury in Vietnamese patients. *J Clin Pharmacol*. (2021) 62:110–7. doi: 10.1002/jcph.1955

60. Hu H, Lai X, Tan C, Yao N, Yan L. Factors associated with in-patient mortality in the rapid assessment of adult earthquake trauma patients. *Prehosp Disaster Med*. (2022) 37:299–305. doi: 10.1017/s1049023x22000693

61. Liu Y. Prediction of depression in the elderly based on machine learning [bachelor]: Shandong University (2023).

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 12 March 2024
DOI 10.3389/fpubh.2024.1305746



OPEN ACCESS

EDITED BY
Jorge Piano Simoes,
University of Twente, Netherlands

REVIEWED BY
Paola Longo,
University of Turin, Italy
Dan Qiu,
Central South University, China

*CORRESPONDENCE
Lihua Jiang
✉ lhjiang@scu.edu.cn
Li Lu
✉ luli@scu.edu.cn

RECEIVED 02 October 2023
ACCEPTED 23 February 2024
PUBLISHED 12 March 2024

CITATION
Guo X, Wang L, Li Z, Feng Z, Lu L, Jiang L and Zhao L (2024) Factors and pathways of non-suicidal self-injury in children: insights from computational causal analysis. *Front. Public Health* 12:1305746. doi: 10.3389/fpubh.2024.1305746



# Factors and pathways of non-suicidal self-injury in children: insights from computational causal analysis

Xinyu Guo[1], Linna Wang[2], Zhenchao Li[1], Ziliang Feng[2], Li Lu[2]*, Lihua Jiang[1,3]* and Li Zhao[1]

[1]Department of Health Policy and Management, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China, [2]College of Computer Science, Sichuan University, Chengdu, Sichuan, China, [3]Teaching and Research Section of General Practice, The General Practice Medical Center, West China Hospital of Sichuan University, Chengdu, China

**Background:** Non-suicidal self-injury (NSSI) has become a significant public health issue, especially prevalent among adolescents. The complexity and multifactorial nature of NSSI necessitate a comprehensive understanding of its underlying causal factors. This research leverages the causal discovery methodology to explore these causal associations in children.

**Methods:** An observational dataset was scrutinized using the causal discovery method, particularly employing the constraint-based approach. By integrating machine learning and causal inference techniques, the study aimed to determine direct causal relationships associated with NSSI. The robustness of the causal relationships was evaluated using three methods to construct and validate it: the PC (Peter and Clark) method, Fast Causal Inference (FCI) method, and the GAE (Graphical Autoencoder) method.

**Results:** Analysis identified nine nodes with direct causal relationships to NSSI, including life satisfaction, depression, family dysfunction, sugary beverage consumption, PYD (positive youth development), internet addiction, COVID-19 related PTSD, academic anxiety, and sleep duration. Four principal causal pathways were identified, highlighting the roles of lockdown-induced lifestyle changes, screen time, positive adolescent development, and family dynamics in influencing NSSI risk.

**Conclusions:** An in-depth analysis of the factors leading to Non-Suicidal Self-Injury (NSSI), highlighting the intricate connections among individual, family, and pandemic-related influences. The results, derived from computational causal analysis, underscore the critical need for targeted interventions that tackle these diverse causative factors.

KEYWORDS

**NSSI, causal discovery, mental health, artificial intelligence, risk factors, COVID-19**

## 1 Introduction

Non-suicidal self-injury (NSSI) has become a pressing public health issue, with rising prevalence both in developed and developing countries (1). Rates of NSSI fluctuate between 11.5 and 33.8%, contingent on sample type and study design, but there is an undeniable uptrend worldwide, even in developing nations. Adolescence is the peak risk period for NSSI (2), with about 23% of adolescents, 13.4% of young adults, and 5.5% of adults being affected (3).

Alarmingly, up to 24.7% of Chinese adolescents report experiencing NSSI (4), which necessitates further attention (5).

The repercussions of NSSI in children and adolescents are severe and long-lasting (6, 7). It is closely linked to heightened suicidal ideation and attempts (8). Even when accounting for suicidal thoughts, NSSI remains a potent predictor of suicidal actions (9–11). Specifically, 39.6% of those who have self-harmed report suicidal behaviors, and of that cohort, 66.3% have a history of NSSI (12). Besides, NSSI correlates with several psychological challenges, like depression, anxiety, and post-traumatic stress disorder(PTSD), and negatively impacts familial and interpersonal bonds (13, 14).

The risk factors for non-suicidal self-injury in children and adolescents are diverse and multifaceted. While much emphasis has been placed on the psychological, family, and school levels—including psychological disorders and symptoms, adverse family experiences, and victimization (4, 15, 16)—the underlying mechanisms driving NSSI behaviors are complex. Many studies to date have been constrained by methodological limitations, failing to grasp the problem from a comprehensive and systematic viewpoint. Moreover, the emergence of the COVID-19 pandemic and its associated lockdown measures has further complicated the scenario (17, 18). Factors such as changes in sleep time (19), physical activity, screen time (20), and increased stress due to isolation and academic pressures have heightened the mental and emotional issues among children and adolescents, possibly leading to a surge in NSSI behaviors (21–23).

Protective factors against Non-Suicidal Self-Injury (NSSI), including life satisfaction and Positive Youth Development (PYD), have been highlighted as promising avenues for intervention. Research highlights that higher levels of life satisfaction can serve as a significant buffer against self-injurious behaviors, while the presence of PYD qualities has been shown to not only reduce the risk of NSSI but also lessen the impact of depressive symptoms on such behaviors (24–26). This underscores the PYD perspective's shift from focusing on youth deficits or psychopathology to emphasizing their strengths, skills, and assets, which can be nurtured and improved (27, 28). These factors are theorized to buffer individuals from the deleterious effects of risk factors like depression and family dysfunction. Enhancing these protective mechanisms may mitigate the risk of NSSI, providing a critical strategy for supporting at-risk adolescents.

Based on the literature review above and the four-function model of self-injury (29, 30), we developed a comprehensive hypothetical framework (Figure 1) for the emergence of NSSI in adolescents. we have formulated a series of causal pathway hypotheses to be explored within our comprehensive hypothetical framework.

- Changes in sleep times, physical activity, consumption of sugar-sweetened beverages, and screen time during COVID-19 lockdowns are hypothesized to augment the probability of developing familial and psychological dysfunctions, subsequently enhancing the risk of NSSI.
- COVID-19-related PTSD is presumed to escalate the risk of psychological issues, further contributing to NSSI.
- Family dysfunction may indirectly lead to NSSI by exacerbating psychological distress.
- PYD qualities could potentially modulate the relationship between psychological distress and NSSI.

Building on the comprehensive review and the hypothetical framework outlined above, this research adopts causal discovery methods within a robust theoretical framework to elucidate the complex web of factors influencing NSSI among adolescents. Traditional research methodologies, which are predominantly observational and correlational, have contributed valuable insights yet frequently encounter limitations in establishing causality (31). Moreover, there is a tendency within existing research to adopt a narrow focus on isolated variables, neglecting the broader constellation of contributing factors. Recognizing the limitations of conventional research methodologies in capturing the multifaceted nature of NSSI, our study utilizes a causal discovery approach applied to a unique observational dataset, aiming to identify causal factors associated with NSSI in children.

Causal discovery, a methodological paradigm dedicated to unearthing cause-and-effect relationships among variables, emerges as a formidable tool in scientific inquiry and data analysis. It enables researchers to transcend the boundaries of correlation, probing the underlying "why" behind observed phenomena (32–34). In the domain of NSSI, discerning the causal dynamics behind this complex behavior is crucial for crafting effective prevention and intervention measures. Through this approach, our study aims to contribute a nuanced understanding of NSSI, fostering the development of more precise and efficacious strategies to combat this pressing issue among youth.

# 2 Materials and methods

## 2.1 Study design and participants

Data for this research was sourced from the Chengdu Positive Child Development (CPCD) survey (35). Launched in December 2019 in Chengdu, this school-based longitudinal study targeted students aged 6–16 years, drawn from five primary and secondary schools. Information was collected using a mix of paper and electronic questionnaires completed by children and caregivers, as well as through direct interviews. The study aimed to explore connections between students' sociodemographic factors, health, lifestyle behaviors, and academic performance, with supplementary data obtained from caregiver health evaluations and school records. The first follow-up of the survey, conducted between June 16, 2020, and July 8, 2020, aligned with the resumption of classroom activities following the COVID-19 lockdowns.

A meticulous vetting process was utilized to refine the dataset, ultimately including 5,807 students to ensure high-quality data for robust causal inference. This process involved identifying and eliminating statistical outliers, removing variables with a missing rate exceeding 90%, and excluding samples lacking caregiver questionnaires or COVID-19 related responses. The "Sample Selection and Refinement Process for NSSI Study" is illustrated in Figure 2.

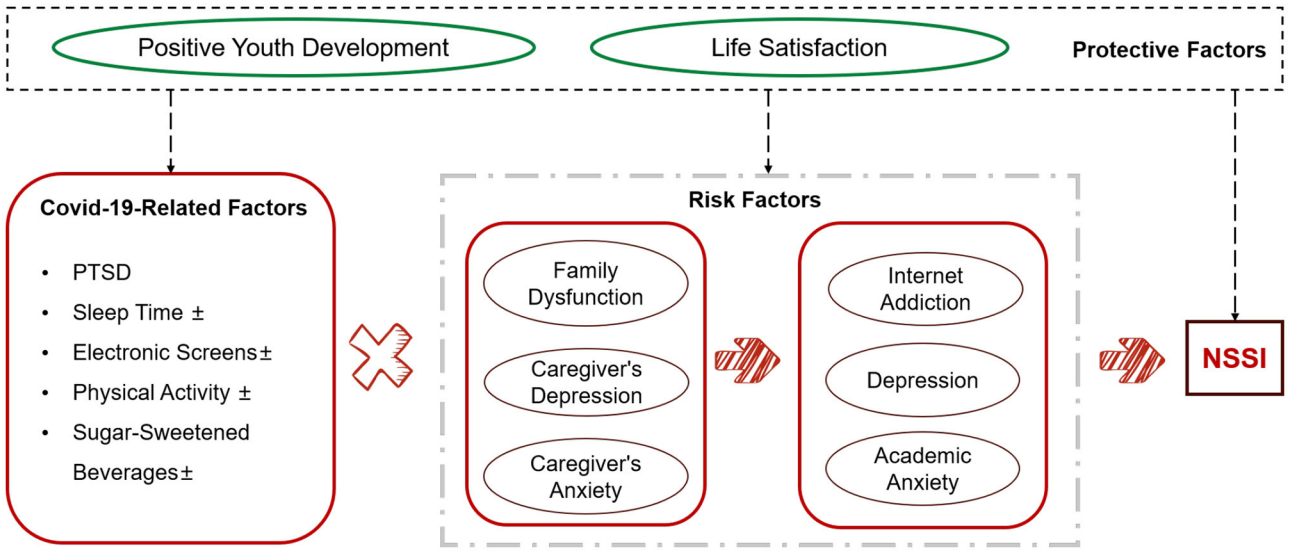



FIGURE 1
NSSI comprehensive hypothetical framework.

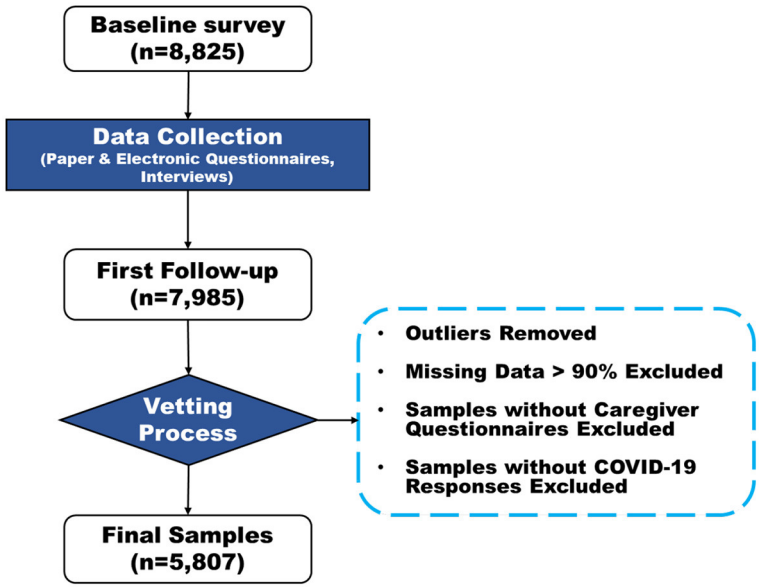



FIGURE 2
Sample selection and refinement process for NSSI study.

## 2.2 Ethics statement

The confidentiality of all data was meticulously maintained by the research team, with no disclosure of any personal information within the study's findings. This research adhered to the principles set forth in the Helsinki Declaration, secured approval from the Ethics Committee of Sichuan University (Approval No. K2020025), and obtained consent from the relevant school authorities, parents, and students.

## 2.3 Measures

### 2.3.1 Non-suicidal self-injury

The assessment of Non-suicidal self-injury (NSSI) in our study was conducted using the 9-item Intentional Self-Injury Scale, developed and validated by Gratz (36). This scale evaluates eight specific NSSI behaviors, including cutting, burning, biting, stabbing, hitting, pinching, and ingesting non-food items, along with a self-assessment for NSSI-related hospitalizations. Each behavior is rated on a 4-point scale, ranging from 1 (never) to

4 (three or more times). Participants reporting a score of 1 or higher on any item were identified as exhibiting NSSI behaviors. This measurement approach has been previously validated in studies involving Chinese adolescents (37).

### 2.3.2 Explanatory variables

In our research, we carefully selected 21 variables collected during the first follow-up period of the CPCD Survey. This selection was strategically guided by the existing literature (4, 8–10, 15, 38), and the aim of our study was to explore causal relationships with non-suicidal self-injury (NSSI). We focus on various sociodemographic, behavioral, and psychological factors that have been previously associated with NSSI behaviors. This approach aimed to create a manageable and comprehensive analytical framework that aligned our study with established findings and theory in NSSI research.

Sociodemographic data, including gender, age, grade, and BMI, were collected, prioritizing "Grade" for its relevance to educational context and peer interactions. Mental health status and psychological characteristics were evaluated through scales measuring depression, academic anxiety, positive child development, life satisfaction, family dysfunction, and internet addiction, aiming to create a comprehensive analytical framework that aligns with established NSSI research findings and theories.

Moreover, the study evaluated COVID-19-related behavioral factors like PTSD, changes in sleep patterns, physical activity, screen time, and sugary beverage consumption to understand the pandemic's impact on students' routines and its potential link to NSSI behaviors. These evaluations used precise binary coding for behavioral changes, offering detailed insights into lifestyle adjustments during the pandemic. Additional considerations included internet addiction and caregivers' mental health status, with a comprehensive overview of all variables, definitions, and measurement methods provided in the Supplementary material.

## 2.4 Data pre-processing

We conducted essential data preprocessing steps to ensure that we had a well-prepared dataset for subsequent causal discovery analysis. To harmonize the numerical and categorical features within the dataset, we employed standardized procedures. For the categorical variables, which included ordinal data with a discernible degree of ordering, we applied the LabelEncoder technique to transform them into numerical representations while preserving the ordinal relationships. Simultaneously, we standardized the numerical features using the StandardScaler method, which centered the variables around a mean of zero and scaled them to unit variance. This preprocessing not only alleviated potential issues arising from varying scales but also facilitated the compatibility of our data with the causal discovery algorithm, ensuring that it operated effectively in uncovering causal relationships within our dataset.

## 2.5 Causal discovery model application

It is crucial to recognize that causality and correlation are distinct concepts with profound implications for scientific research and data analysis. While correlation measures the strength and direction of statistical association between variables, it does not elucidate the direction of influence or establish a cause-and-effect relationship. Causality, on the other hand, delves into the intricate web of cause-and-effect connections, emphasizing that changes in one variable lead to changes in another. In this study, we aimed to identify the complex causality hidden in NSSI based on social-demographic characteristics, psychological characteristics, COVID-related behavioral factors, behavioral factors, mental health status and caregiver's mental health status.

The primary methods for learning the causal graph structure among variables include constraint-based, score-based, and gradient-based approaches. Constraint-based methods, such as PC (Peter and Clark) and FCI (39), identify causal relationships by recognizing patterns of conditional independence. Score-based methods, including GES (40), assign numerical scores to different causal structures and search for the structure with the highest score. Gradient-based methods, including GAE (41), have emerged as notable alternatives, exhibiting superior accuracy and computational efficiency. In comparison to constraint-based and score-based methods, gradient-based methods directly learn the causal graph through end-to-end optimization, eliminating the need for explicit detection of conditional independence or the use of scoring functions in intermediate steps.

Consequently, having recognized the strengths of each approach, we integrate a combination of these methods to enhance the accuracy and robustness of our causal Directed Acyclic Graph (DAG) learning. We employed three approaches for constructing and validating the DAG: the PC method (34), FCI (42), and the GAE. This is aimed at fortifying the reliability and credibility of our DAG, ensuring its resilience and accuracy across diverse analytical frameworks. The framework of our prediction work is presented in Figure 3.

**PC algorithm** represents a fundamental approach to causality analysis that is grounded in conditional independence testing. Its fundamental logic hinges upon the principle that genuine causal relationships manifest as conditional independence relationships in observational data. **FCI algorithm** is an extension of PC, designed to handle latent variables more effectively. As both PC and FCI are Constraint-based methods, they first use conditional independence tests to learn the skeleton of the underlying casual graph, and then orient the edges based on a series of orientation rules. It's important to note that the order in which variables are considered can influence the outcomes of these two algorithms.

To mitigate the impact of order-dependency, we conducted multiple runs of both PC and FCI with different random orders of variables. This ensures that the algorithm are exposed to various variable orders, mitigating the influence of a specific sequence on the final results. In our study, for FCI and the three variations of PC (original, stable, and parallel), we performed 1,000 runs for each, employing random variable orders. We identified the edges that appear most frequently across the multiple runs. The final structure, represented by edges that consistently appear, can be considered

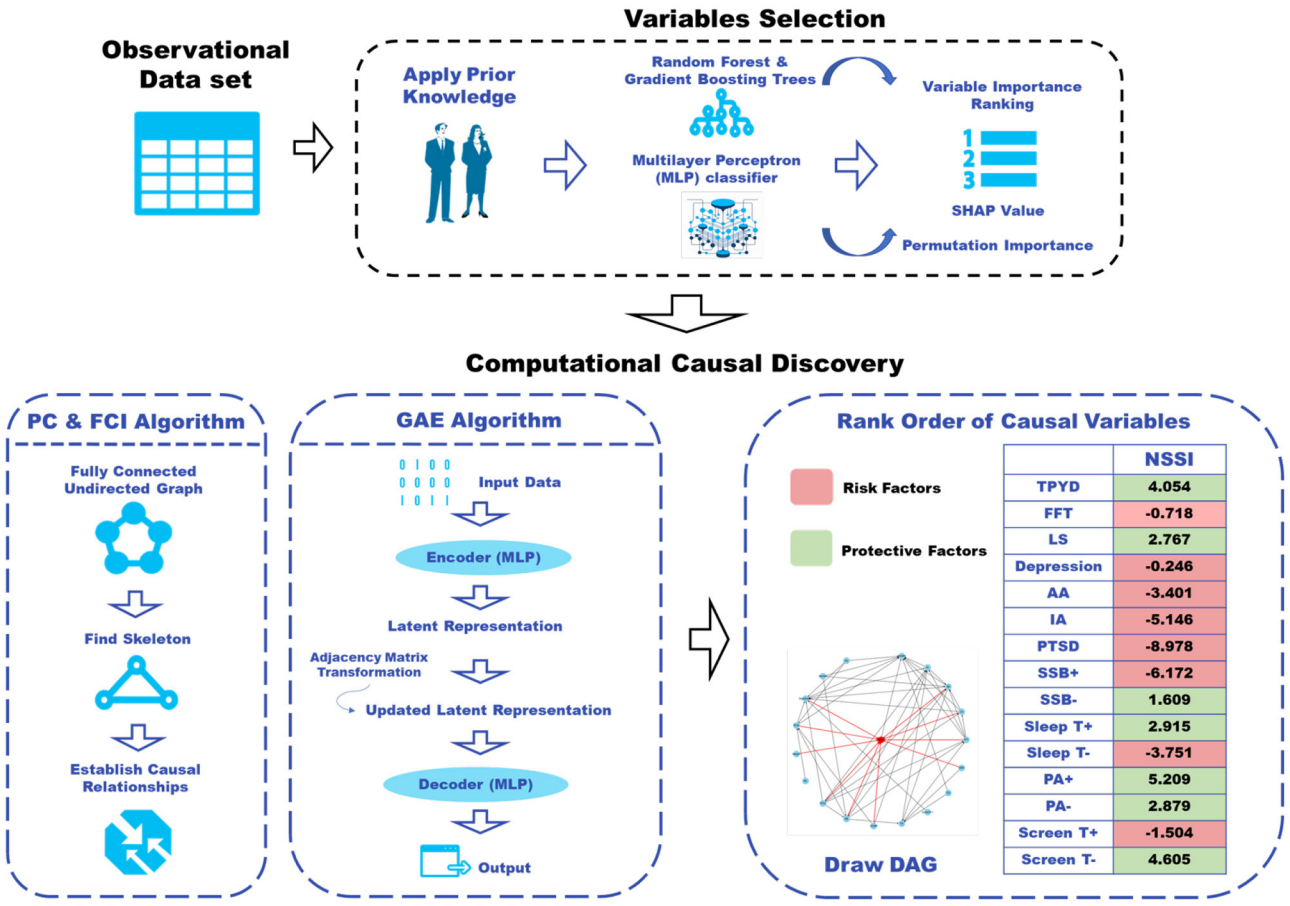



FIGURE 3
The workflow for computational casual discovery.

as the more robust DAG. The general steps for PC and FCI are described as follows:

1. Prior knowledge on variable selection. It is crucial to note that the PC and FCI algorithm assume causal sufficiency, and therefore, all potential confounding factors should be present in the data for a comprehensive causal analysis. We conscientiously leveraged prior knowledge in the face of a large number of total features to identify the variables relevant to NSSI.

2. Prior knowledge on variable relationship. We added prior knowledge to guide our causal modeling by informing the anticipated forbidden causal pathways. The rationale behind the prohibition of each path is detailed below.

- Prohibition of paths leading to "Grade" and "Gender" variables. Within our causal modeling framework, we have enforced a stringent constraint by prohibiting any causal pathways that lead to the "Grade" and "Gender" variables. This restriction is grounded in the fundamental premise that "Grade" and "Gender" are considered intrinsic attributes of individuals, impervious to external causal influences.
- Prohibition of paths leading to "COVID-Related Behavioral Factors". We also prohibited any causal pathways leading to the "COVID-Related Behavioral Factors". This constraint arose from the distinctive temporal characteristics and underlying nature of these factors. Unlike other variables in our analysis, which were sourced from a single survey conducted during the COVID-19 pandemic, the "COVID-Related Behavioral Factors" represented the dynamic changes observed between this COVID-19 period survey and another before-COVID-19 period survey. The timeline of these behavioral factors is inherently distinct, encompassing the evolving responses of individuals to the unique circumstances brought about by the pandemic. Thus, we enforced the prohibition of all paths leading to "COVID-Related Behavioral Factors".

3. Causal discovery process. Our application of the PC and FCI algorithms followed these essential steps:

- Skeleton discovery: Begin with a fully connected undirected graph that includes all variables. Utilized conditional independence tests to uncover the initial skeleton of the causal graph, representing potential pairwise relationships between variables. We performed independence test using Fisher-$z$'s test.

- Edge Orientation: PC orients edges based on conditional independence tests. FCI orients edges within Markov equivalence classes, considering latent variables and capturing more complex causal relationships.
- Refinement: PC refines the graph iteratively by applying conditional independence tests and edge removals. FCI refines the graph by considering additional conditional independence tests and making decisions about edge orientations within Markov equivalence classes.
- Multiple runs and final structure: For FCI and three variations of PC, we performed 1,000 runs with random variable orders. Identified edges that consistently appeared across multiple runs to establish a final structure.

**GAE algorithm** represents Graph Autoencoder for causal structure learning. It is an alternative generalization of NOTEARS (43) to handle nonlinear causal relations. After decoder, GAE can generate an adjacency matrix that captures the relationships between nodes. The adjacency matrix typically includes positive values for positive edges and negative values for negative edges. Positive value might represent a positive relationship or interaction, while negative value might represent a negative or inhibitory relationship. We used the learned latent representations from the GAE to identify factors associated with protective effects and those contributing to an increased risk, shown in Figure 3.

# 3 Results

## 3.1 Sample characteristics

Among the 5,307 participants included in the final analysis, 1,394 individuals (26.27%) reported engaging in non-suicidal self-injury (NSSI) in the past year. Among the 1,394 participants who reported engaging in NSSI, the NSSI detection rate was 24.82% for males ($n$ = 667) and 27.75% for females ($n$ = 727). The NSSI detection rate among primary school students was 23.51% ($n$ = 886), while among middle school students, it was 33% ($n$ = 508). The characteristics of the sample are presented in Table 1.

## 3.2 Direct causes and effects of NSSI (local causal network)

We utilize a directed acyclic graph (DAG) to exhibit the complexity of variables influencing adolescent and child non-suicidal self-injury (NSSI) behaviors during the COVID-19 pandemic.Within this illustrative graph, the structure of the local causal network model, fundamental to our study, is evident. It includes an array of variables, each having its significance in the broader scope. The DAG aptly captures the "upstream" and "downstream" variables (31). Here, "upstream" variables act as forerunners or primary triggers that might affect succeeding variables. Conversely, the "downstream" variables are those likely influenced or modified due to the preceding ones.

The DAG offers a comprehensive visualization, revealing both direct neighbors and secondary neighbors within two "steps" of NSSI behaviors. This graphical representation paints an in-depth portrait of the relationships between various factors during the COVID-19 pandemic's acute phase for adolescents. Our application of the GAE algorithm not only corroborates these findings but also enriches them by providing a nuanced understanding of the factors contributing to NSSI. This advanced method, represented in Figure 3, facilitated the quantification of each variable's impact by calculating their polarity and magnitude. The GAE algorithm's unique ability to model nonlinear relationships allowed us to further delineate factors with protective effects (denoted in green) from those associated with increased risk (indicated in red). The results, as depicted in Figure 4, confirm and strengthen our causal network model, providing a nuanced understanding of the variables that directly or indirectly contribute to NSSI behaviors.

The nodes with direct causal relationship with NSSI are life satisfaction, depression, family dysfunction, sugary beverage consumption, PYD, Internet addiction, COVID-19 related PTSD, academic anxiety and sleep duration. The analysis identified four key causal pathways influencing NSSI: (1) Lockdown-induced reductions in physical activity escalate academic anxiety, potentially leading to PTSD, heightening internet addiction and NSSI risks. (2) Enhanced screen time and sugary beverage consumption are linked to elevated depression risks and increased NSSI likelihood. (3) Positive adolescent development acts as a buffer, mitigating the adverse effects of family dysfunction and internet addiction on NSSI. (4) Family dysfunction negatively impacts life satisfaction and fosters depression, directly contributing to NSSI.The detailed DAG illustrating these relationships is provided in Supplementary Figure S1.

The graph further enlightens us about individual attributes that could influence this dynamic, underscoring the phenomena's multifaceted nature. The focus on NSSI behavior in the DAG, and its potential links with these various factors, is particularly noteworthy. Moreover, the graph showcases the intricate ties between these variables, elucidating the ripple effects and possible feedback mechanisms within the system. Ultimately, the DAG serves as a comprehensive visual tool, elucidating the intricate connections and potential causal pathways pertaining to NSSI behavior in adolescents amid the COVID-19 pandemic.

# 4 Discussion

Our research harnessed machine learning and causal discovery techniques to understand the intricate web of causal links surrounding NSSI in children and adolescents. Aiming to shed light on influential factors related to individual characteristics, family dynamics, and the repercussions of the COVID-19 pandemic, we strived to provide actionable insights that could spearhead innovative prevention and intervention approaches. Given the complex nature of NSSI, dictated by numerous variables, and the limitations of current research methodologies, we sought a fresh approach to discern causation, ultimately guiding more effective strategies. The PC and FCI algorithms exhibit sensitivity to the order of variables, implying that the obtained results may vary based on the sequence in which variables are considered during execution. To address this, we conducted multiple runs of these algorithms using different orders of features. Additionally, to

10.3389/fpubh.2024.1305746

TABLE 1 Sample characteristics.

| | Total ($n$ = 5,307) | NSSI ($n$ = 1,394) | Incidence of NSSI (%) |
|---|---|---|---|
| **Gender** | | | |
| Male | 2,687 (50.63) | 667 (47.85) | 24.82 |
| Female | 2,620 (49.37) | 727 (52.15) | 27.75 |
| **Grade** | | | |
| 1 | 239 (4.50) | 33 (2.37) | 13.81 |
| 2 | 312 (5.88) | 49 (3.52) | 15.71 |
| 3 | 852 (16.05) | 200 (14.35) | 23.47 |
| 4 | 1,019 (19.20) | 238 (17.07) | 23.36 |
| 5 | 664 (12.51) | 188 (13.49) | 28.31 |
| 6 | 682 (12.85) | 178 (12.77) | 26.10 |
| 7 | 807 (15.21) | 247 (17.72) | 30.61 |
| 8 | 732 (13.79) | 261 (18.72) | 35.66 |
| Primary school | 3,768 (71.00) | 886 (63.56) | 23.51 |
| Junior high school | 1,539 (29.00) | 508 (36.44) | 33.01 |
| **BMI** | | | |
| Normal | 4,169 (78.56) | 1,103 (79.12) | 26.46 |
| Overweight | 454 (8.55) | 126 (9.04) | 27.75 |
| Obesity | 684 (12.89) | 165 (11.84) | 24.12 |
| **COVID-19-related behavioral factors** | | | |
| Sleep time+ | 1,401 (26.40) | 363 (26.04) | 25.91 |
| Sleep time− | 2,470 (46.54) | 714 (51.22) | 28.91 |
| Physical activity+ | 1,796 (33.84) | 470 (33.72) | 26.17 |
| Physical activity− | 2,233 (42.08) | 618 (44.33) | 27.68 |
| Screen time+ | 1,387 (26.14) | 398 (28.55) | 28.70 |
| Screen time− | 2,521 (47.50) | 692 (49.64) | 27.45 |
| SSB+ | 727 (13.70) | 247 (17.72) | 33.98 |
| SSB− | 1,963 (36.99) | 570 (40.89) | 29.04 |
| **Mental health status** | | | |
| No depression | 3,449 (64.99) | 507 (36.37) | 14.70 |
| Depression | 1,858 (35.01) | 887 (63.63) | 47.74 |
| No anxiety | 4,472 (84.27) | 896 (64.28) | 20.04 |
| Anxiety | 835 (15.73) | 498 (35.72) | 59.64 |
| Academic anxiety (<4) | 3,366 (63.43) | 787 (56.46) | 23.38 |
| Academic anxiety (≥4) | 1,941 (36.57) | 607 (43.54) | 31.27 |
| No PTSD | 3,468 (65.35) | 710 (50.93) | 20.47 |
| PTSD | 1,839 (34.65) | 684 (49.07) | 37.19 |
| **Internet addiction** | | | |
| Average | 4,514 (85.06) | 987 (70.8) | 21.87 |
| Occasional to frequent | 722 (13.60) | 363 (26.04) | 50.28 |
| Significant | 71 (1.34) | 44 (3.16) | 61.97 |

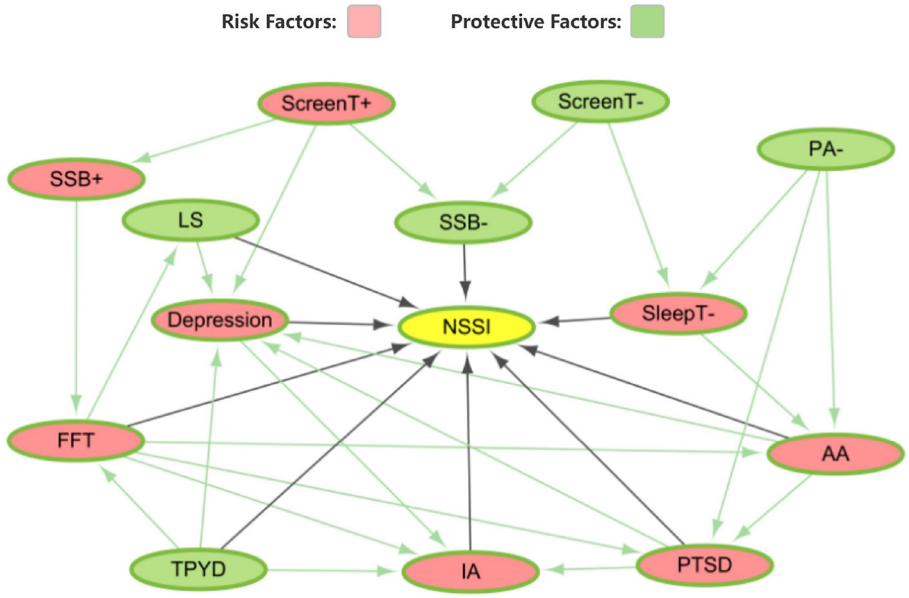



FIGURE 4
The DAG for local variable in NSSI.

enhance the robustness of the causal structure, we incorporated the GAE method into our approach. Subsequently, the most frequently identified edges were retained for further in-depth analysis.

Our study has delineated a network of nine pivotal variables that exhibit direct causal associations with NSSI behavior in children, as visualized in Figure 4 (for the full names and assignments corresponding to the abbreviations of the variables shown in Supplementary Table S1). Central among these findings is the pervasive influence of the COVID-19 pandemic, which manifest as a marked decrease in sugary beverage intake, diminished sleep quality, and the emergence of COVID-19-triggered PTSD. Individual psychological drivers encompassed themes of academic stress and manifestations of depression. In tandem, key psychological attributes were identified: the holistic life satisfaction measure, the nuanced interplays of family dynamics, the embodiment of Positive Youth Development (PYD) (26), and the grip of internet addiction. The coherence of these findings with existing literature underscores the robustness of our methodological approach (4, 44, 45).

Significantly, the COVID-19 pandemic has altered the sleep patterns of children and adolescents, a crucial aspect of mental health. Prolonged lockdowns have disrupted regular sleep schedules, heightening stress and anxiety levels. These sleep disruptions, combined with the stress of new learning modes and social isolation, have intensified academic anxieties, affecting family dynamics and increasing household tensions. Moreover, the pandemic has triggered PTSD in young people, a concerning development given its long-term mental health implications. This emergence of PTSD, fueled by constant pandemic-related news and personal experiences, adds a critical dimension to our understanding of the pandemic's impact on youth mental health.

Digging deeper into these variables, the interconnectedness within this network is profound. Academic stress finds its roots in the diminished sleep and curtailed physical activity patterns during the pandemic, as well as the challenges posed by family dynamics. The haunting presence of COVID-19-related PTSD is sculpted by the restrictive physical activity regimes, academic stressors, and family dysfunction. Depression's emergence is further amplified by excessive screen time, both for academic and recreational purposes, interwoven with academic anxieties, the nurturing aspects of Positive Youth Development, and overarching life satisfaction. Completing this intricate web, internet addiction bears the imprints of PTSD linked to the pandemic, the shadows of depression, family struggles, and the protective or exacerbating elements of Positive Youth Development.

Emerging from our findings is a comprehensive understanding that provides invaluable insights for safeguarding the mental wellbeing of children and adolescents in the wake of unforeseen public health crises, such as the COVID-19 pandemic. This understanding emphasizes the potential of our data in guiding practical interventions and preventive measures. The notable reduction in sugary beverage consumption and sleep duration, as well as the emergence of pandemic-induced PTSD, highlight the profound physiological and psychological shifts induced by prolonged lockdowns and associated societal changes (46, 47). Our findings advocate for prevention strategies that are not only trauma-informed but also adaptive to the evolving public health landscape.

The established causal relationship between sugary beverage consumption and NSSI echoes previous studies associating unhealthy diets with increased depressive symptoms in adolescents (48, 49). Inflammatory diets can intensify mental health problems, possibly through obesity and inflammation (50, 51). Thus, the importance of balanced diets is underscored. Emphasizing a diet rich in fruits, vegetables, and anti-inflammatory foods, combined

with strategies like restoring regular sleep patterns and trauma-informed interventions, may offer a comprehensive approach to improving the mental wellbeing of children and adolescents.

Furthermore, understanding the intricate network of causal factors—including academic anxieties (52), family dynamics (53), and individual psychological attributes—means that interventions can be more targeted and precise. This precision is vital in the clinical setting, where tailored interventions can lead to more effective outcomes. Rather than applying generic measures, strategies can be devised to specifically counteract or augment identified causal agents. This approach enhances the practical utility of our findings, offering a roadmap for clinicians and policymakers in developing targeted interventions. The identified drivers, such as academic stress and depression, could be targeted through school-based programs emphasizing coping mechanisms, emotional regulation, and peer support. Family-centered interventions might focus on strengthening familial bonds and improving communication, reducing the chance of family dysfunction exacerbating mental health issues (54). The correlation between internet addiction and NSSI emphasizes the importance of digital literacy programs that equip adolescents with skills to navigate the online world safely.

A key distinction of our study lies in the use of computational causal discovery. Traditional methodologies often restrict themselves to observational correlations, which, although informative, don't offer a genuine window into the underlying causative structures. Causal discovery goes beyond merely identifying these associations, allowing us to pinpoint the drivers of adverse outcomes such as NSSI. The superiority of this approach lies in its potential to tailor prevention and intervention strategies based on the actual causes, rather than mere symptoms or correlated factors. This means that initiatives informed by our findings can be significantly more effective, as they strike at the heart of the issue, directly addressing and mitigating the root causes. As we move forward in our collective endeavor to nurture the mental health of our younger generation, leveraging advanced methodologies like computational causal discovery will be paramount in ensuring our strategies are not only well-informed but also impactful.

## 5 Strengths and limitations

A significant strength of our study is the use of multiple causal discovery algorithms, enhancing the robustness and interpretability of the results. This approach marks a departure from traditional approaches that predominantly rely on correlations. This novel approach facilitates a nuanced understanding of the intricate "cause-and-effect" dynamics underlying NSSI behaviors in children and adolescents, particularly in the unique context of the COVID-19 pandemic. By pinpointing fundamental causative elements, our study lays the foundation for more targeted and efficacious interventions that address root causes, providing both immediate and sustained psychological health benefits.

Nevertheless, the study has limitations, notably the influence of variable ordering on model outcomes. We've addressed this by running numerous iterations, lending consistency to our findings despite potential variable ordering effects.

While the adoption of causal algorithms marks an advancement in our analysis, they cannot fully negate the impact of unobserved confounders or bidirectional relationships. The cross-sectional data limits temporal causality claims, necessitating further validation with longitudinal studies.

Additionally, while AI methods enhance efficiency and aid in the application of causal inference to health data, they are not infallible. The necessity for human judgment remains a key component in the interpretative process, not least because different AI algorithms might yield varying interpretations or conclusions (55). This involves using different AI algorithms for mutual validation and conducting repeated experiments to test their robustness. This dual approach of leveraging both AI and human judgment facilitates a more nuanced and robust analysis than could be achieved by either one alone.

In terms of academic recommendations and future directions, there is a compelling need for further exploration of the identified mechanisms underpinning NSSI behaviors. Using longitudinal datasets with time series information, in conjunction with computational causal discovery, can offer more robust and definitive insights into causality. forging interdisciplinary collaborations that meld psychological, societal, and technological insights could provide a more holistic understanding and usher in innovative intervention strategies tailored to the multifaceted challenges of the modern era.

## 6 Conclusions

Drawing on unique computational causal discovery and machine learning methods, this study illuminated the intricate causal network of factors influencing NSSI in children during the COVID-19 pandemic. Our findings underscore nine critical variables intricately interwoven, reflecting the profound effects of the pandemic, academic stress, family dynamics, and individual psychological attributes. The study's insights offer a fresh perspective for devising impactful interventions, emphasizing the significance of addressing root causes, particularly in the wake of unprecedented global challenges.

## Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the dataset can only be used for non-commercial, academic research purposes. Requests to access these datasets should be directed to Rui Hu, hu2857911896@163.com.

## Ethics statement

The studies involving humans were approved by Ethics Committee of Sichuan University (Approval No. K2020025). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

## Author contributions

XG: Conceptualization, Data curation, Investigation, Writing – original draft. LW: Formal analysis, Methodology, Writing – original draft. ZF: Writing – review & editing. ZL: Data curation, Formal analysis, Visualization, Writing – original draft. LL: Supervision, Writing – review & editing. LJ: Supervision, Writing – review & editing. LZ: 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 Natural Science Foundation of China (NSFC) (No. 82273748).

## Acknowledgments

We would like to express our gratitude to ChatGPT, a language model based on the GPT-4 architecture developed by OpenAI, for assisting with the translation of portions of this manuscript.

## 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



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2024.1305746/full#supplementary-material

## References

1. Mannekote Thippaiah S, Shankarapura Nanjappa M, Gude JG, Voyiaziakis E, Patwa S, Birur B, et al. Non-suicidal self-injury in developing countries: a review. *Int J Soc Psychiatry*. (2021) 67:472–82. doi: 10.1177/0020764020943627

2. Lim KS, Wong CH, McIntyre RS, Wang J, Zhang Z, Tran BX, et al. Global lifetime and 12-month prevalence of suicidal behavior, deliberate self-harm and non-suicidal self-injury in children and adolescents between 1989 and 2018: a meta-analysis. *Int J Environ Res Public Health*. (2019) 16:4581. doi: 10.3390/ijerph16224581

3. Rodriguez-Blanco L, Carballo JJ, Baca-Garcia E. Use of ecological momentary assessment (EMA) in non-suicidal self-injury (NSSI): a systematic review. *Psychiatry Res*. (2018) 263:212–9. doi: 10.1016/j.psychres.2018.02.051

4. Qu D, Wen X, Liu B, Zhang X, He Y, Chen D, et al. Non-suicidal self-injury in Chinese population: a scoping review of prevalence, method, risk factors and preventive interventions. *Lancet Reg Health West Pac*. (2023) 37:100794. doi: 10.1016/j.lanwpc.2023.100794

5. Wang C, Zhang P, Zhang N. Adolescent mental health in China requires more attention. *Lancet Public Health*. (2020) 5:e637. doi: 10.1016/S2468-2667(20)30094-3

6. Borschmann R, Becker D, Coffey C, Spry E, Moreno-Betancur M, Moran P, et al. 20-year outcomes in adolescents who self-harm: a population-based cohort study. *Lancet Child Adolesc Health*. (2017) 1:195–202. doi: 10.1016/S2352-4642(17)30007-X

7. Mars B, Heron J, Crane C, Hawton K, Lewis G, Macleod J, et al. Clinical and social outcomes of adolescent self harm: population based birth cohort study. *BMJ*. (2014) 349:g5954. doi: 10.1136/bmj.g5954

8. Poudel A, Lamichhane A, Magar KR, Khanal GP. Non suicidal self injury and suicidal behavior among adolescents: co-occurrence and associated risk factors. *BMC Psychiatry*. (2022) 22:96. doi: 10.1186/s12888-022-03763-z

9. Mars B, Heron J, Klonsky ED, Moran P, O'Connor RC, Tilling K, et al. Predictors of future suicide attempt among adolescents with suicidal thoughts or non-suicidal self-harm: a population-based birth cohort study. *Lancet Psychiatry*. (2019) 6:327–37. doi: 10.1016/S2215-0366(19)30030-6

10. Hawton K, Bale L, Brand F, Townsend E, Ness J, Waters K, et al. Mortality in children and adolescents following presentation to hospital after non-fatal self-harm in the multicentre study of self-harm: a prospective observational cohort study. *Lancet Child Adolesc Health*. (2020) 4:111–20. doi: 10.1016/S2352-4642(19)30373-6

11. Groschwitz RC, Kaess M, Fischer G, Ameis N, Schulze UM, Brunner R, et al. The association of non-suicidal self-injury and suicidal behavior according to DSM-5 in adolescent psychiatric inpatients. *Psychiatry Res*. (2015) 228:454–61. doi: 10.1016/j.psychres.2015.06.019

12. Voss C, Hoyer J, Venz J, Pieper L, Beesdo-Baum K. Non-suicidal self-injury and its co-occurrence with suicidal behavior: an epidemiological-study among adolescents and young adults. *Acta Psychiatr Scand*. (2020) 142:496–508. doi: 10.1111/acps.13237

13. Bentley KH, Cassiello-Robbins CF, Vittorio L, Sauer-Zavala S, Barlow DH. The association between nonsuicidal self-injury and the emotional disorders: a meta-analytic review. *Clin Psychol Rev*. (2015) 37:72–88. doi: 10.1016/j.cpr.2015.02.006

14. Taylor PJ, Jomar K, Dhingra K, Forrester R, Shahmalak U, Dickson JM, et al. meta-analysis of the prevalence of different functions of non-suicidal self-injury. *J Affect Disord*. (2018) 227:759–69. doi: 10.1016/j.jad.2017.11.073

15. Liang K, Zhao L, Lei Y, Zou K, Ji S, Wang R, et al. Nonsuicidal self-injury behaviour in a city of China and its association with family environment, media use and psychopathology. *Compr Psychiatry*. (2022) 115:152311. doi: 10.1016/j.comppsych.2022.152311

16. Wang Y, Luo B, Hong B, Yang M, Zhao L, Jia P. The relationship between family functioning and non-suicidal self-injury in adolescents: a structural equation modeling analysis. *J Affect Disord*. (2022) 309:193–200. doi: 10.1016/j.jad.2022.04.124

17. Murata S, Rezeppa T, Thoma B, Marengo L, Krancevich K, Chiyka E, et al. The psychiatric sequelae of the COVID-19 pandemic in adolescents, adults, and health care workers. *Depress Anxiety*. (2021) 38:233–46. doi: 10.1002/da.23120

18. De Luca L, Giletta M, Nocentini A, Menesini E. Non-suicidal self-injury in adolescence: the role of pre-existing vulnerabilities and COVID-19-related stress. *J Youth Adolesc*. (2022) 51:2383–95. doi: 10.1007/s10964-022-01669-3

19. Wright KP, Linton SK, Withrow D, Casiraghi L, Lanza SM, de la Iglesia H, et al. Sleep in university students prior to and during COVID-19 stay-at-home orders. *Curr Biol*. (2020) 30:R797–8. doi: 10.1016/j.cub.2020.06.022

20. Brito LMS, da Silva Boguszewski MC, de Souza MTR, Martins F, Mota J, Leite N. Indoor physical activities, eating and sleeping habits among school adolescents during COVID-19 pandemic. *Rev Bras Atividade Física Saúde*. (2020) 25:1–6. doi: 10.12820/rbafs.25e0117

21. Schwartz-Mette RA, Duell N, Lawrence HR, Balkind EG. COVID-19 distress impacts adolescents' depressive symptoms, NSSI, and suicide risk in the rural, northeast US. *J Clin Child Adolesc Psychol*. (2022) 52:2042697. doi: 10.1080/15374416.2022.2042697

22. John A, Eyles E, Webb RT, Okolie C, Schmidt L, Arensman E, et al. The impact of the COVID-19 pandemic on self-harm and suicidal behaviour: update of living systematic review. *F1000Research*. (2020) 9:1097. doi: 10.12688/f1000research.25522.2

23. Xiao J, Wang R, Hu Y, He T, Ruan Z, Chen Q, et al. Impacts of the psychological stress response on nonsuicidal self-injury behavior in students during the COVID-19 epidemic in China: the mediating role of sleep disorders. *BMC Psychol*. (2022) 10:87. doi: 10.1186/s40359-022-00789-6

24. Kress VE, Newgent RA, Whitlock J, Mease L. Spirituality/religiosity, life satisfaction, and life meaning as protective factors for nonsuicidal self-injury in college students. *J Coll Counsel*. (2015) 18:160–74. doi: 10.1002/jocc.12012

25. Chen Y, Luo J, Jiang L, Shi W, Jia P, Zhang J, et al. Association between positive youth development and non-suicidal self-injury: a longitudinal survey of children and adolescents in southwest China, 2019-21. *J Affect Disord*. (2024) 350:755–60. doi: 10.1016/j.jad.2024.01.072

26. Zhu X, Shek D. The predictive effect of depression on self-injury: positive youth development as a moderator. *Appl Res Qual Life*. (2023) 18:2877–94. doi: 10.1007/s11482-023-10211-x

27. Shek DT, Dou D, Zhu X, Chai W. Positive youth development: current perspectives. *Adolesc Health Med Ther*. (2019) 10:131–41. doi: 10.2147/AHMT.S179946

28. Tolan P, Ross K, Arkin N, Godine N, Clark E. Toward an integrated approach to positive development: Implications for intervention. *Appl Dev Sci*. (2016) 20:214–36. doi: 10.1080/10888691.2016.1146080

29. Nock MK. Why do people hurt themselves? New insights into the nature and functions of self-injury. *Curr Direct Psychol Sci*. (2009) 18:78–83. doi: 10.1111/j.1467-8721.2009.01613.x

30. Hepp J, Carpenter RW, Störkel LM, Schmitz SE, Schmahl C, Niedtfeld I. A systematic review of daily life studies on non-suicidal self-injury based on the four-function model. *Clin Psychol Rev*. (2020) 82:101888. doi: 10.1016/j.cpr.2020.101888

31. Saxe GN, Ma S, Morales LJ, Galatzer-Levy IR, Aliferis C, Marmar CR. Computational causal discovery for post-traumatic stress in police officers. *Transl Psychiatry*. (2020) 10:233. doi: 10.1038/s41398-020-00910-6

32. Zanga A, Ozkirimli E, Stella F. A survey on causal discovery: theory and practice. *Int J Approx Reason*. (2022) 151:101–29. doi: 10.1016/j.ijar.2022.09.004

33. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. *Front Genet*. (2019) 10:524. doi: 10.3389/fgene.2019.00524

34. Zhang K, Zhu S, Kalander M, Ng I, Ye J, Chen Z, et al. gcastle: a python toolbox for causal discovery. *arXiv*. (2021). doi: 10.48550/arXiv.2111.15155

35. Zhao L, Shek DT, Zou K, Lei Y, Jia P. Cohort profile: Chengdu positive child development (CPCD) survey. *Int J Epidemiol*. (2022) 51:e95–e107. doi: 10.1093/ije/dyab237

36. Gratz KL. Measurement of deliberate self-harm: preliminary data on the deliberate self-harm inventory. *J Psychopathol Behav Assess*. (2001) 23:253–63. doi: 10.1037/t04163-000

37. Wu J, Liu S, Luo J, Li X, You J. The effects of childhood abuse, depression, and self-compassion on adolescent nonsuicidal self-injury: a moderated mediation model. *Child Abuse Neglect*. (2023) 136:105993. doi: 10.1016/j.chiabu.2022.105993

38. Hu R, Peng Ll, Du Y, Feng Yw, Xie Ls, Shi W, et al. Reciprocal effect between depressive symptoms and adolescent non-suicidal self-injury before and after COVID-19: a longitudinal study. *Preprint*. (2023). doi: 10.21203/rs.3.rs-2529545/v1

39. Spirtes P, Glymour CN, Scheines R. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press (2000).

40. Chickering DM. Optimal structure identification with greedy search. *J Mach Learn Res*. (2002) 3:507–54. doi: 10.1162/153244303321897717

41. Ng I, Zhu S, Chen Z, Fang Z. A graph autoencoder approach to causal structure learning. *arXiv* [preprint]. (2019). doi: 10.48550/arXiv.1911.07420

42. Spirtes PL, Meek C, Richardson TS. Causal inference in the presence of latent variables and selection bias. *arXiv* [preprint]. (2013). doi: 10.48550/arXiv.1302.4983

43. Zheng X, Aragam B, Ravikumar PK, Xing EP. Dags with no tears: Continuous optimization for structure learning. *Adv Neur Inf Process Syst*. (2018) 31:9492–503. doi: 10.48550/arXiv.1803.01422

44. Xiong A, Liao S, Luo B, Luo S, Tong Y, Li Z. Associations between problematic internet use, life satisfaction, and deliberate self-harm among Chinese adolescents: a multi-centered longitudinal study. *Addict Behav*. (2023) 147:107808. doi: 10.1016/j.addbeh.2023.107808

45. Zhang Y, Jin Z, Li S, Xu H, Wan Y, Tao F. Relationship between chronotype and mental behavioural health among adolescents: a cross-sectional study based on the social ecological system. *BMC Psychiatry*. (2023) 23:1–12. doi: 10.1186/s12888-023-04879-6

46. Du N, Ouyang Y, Xiao Y, Li Y. Psychosocial factors associated with increased adolescent non-suicidal self-injury during the COVID-19 pandemic. *Front Psychiatry*. (2021) 12:743526. doi: 10.3389/fpsyt.2021.743526

47. Fegert JM, Vitiello B, Plener PL, Clemens V. Challenges and burden of the Coronavirus 2019 (COVID-19) pandemic for child and adolescent mental health: a narrative review to highlight clinical and research needs in the acute phase and the long return to normality. *Child Adolesc Psychiatry Ment Health*. (2020) 14:1–11. doi: 10.1186/s13034-020-00329-3

48. Orlando L, Savel KA, Madigan S, Colasanto M, Korczak DJ. Dietary patterns and internalizing symptoms in children and adolescents: a meta-analysis. *Aust New Zeal J Psychiatry*. (2022) 56:617–41. doi: 10.1177/00048674211031486

49. Chopra C, Mandalika S, Kinger N. Does diet play a role in the prevention and management of depression among adolescents? A narrative review. *Nutr Health*. (2021) 27:243–63. doi: 10.1177/0260106020980532

50. Sureda A. Bibiloni MdM, Julibert A, Bouzas C, Argelich E, Llompart I, et al. Adherence to the mediterranean diet and inflammatory markers. *Nutrients*. (2018) 10:62. doi: 10.3390/nu10010062

51. Oddy WH, Allen KL, Trapp GS, Ambrosini GL, Black LJ, Huang RC, et al. Dietary patterns, body mass index and inflammation: pathways to depression and mental health problems in adolescents. *Brain Behav Immun*. (2018) 69:428–39. doi: 10.1016/j.bbi.2018.01.002

52. Chen H, Guo H, Chen H, Cao X, Liu J, Chen X, et al. Influence of academic stress and school bullying on self-harm behaviors among Chinese middle school students: the mediation effect of depression and anxiety. *Front Public Health*. (2023) 10:1049051. doi: 10.3389/fpubh.2022.1049051

53. Cassels M, van Harmelen AL, Neufeld S, Goodyer I, Jones PB, Wilkinson P. Poor family functioning mediates the link between childhood adversity and adolescent nonsuicidal self-injury. *J Child Psychol Psychiatry*. (2018) 59:881–7. doi: 10.1111/jcpp.12866

54. Kelada L, Hasking P, Melvin G. Adolescent NSSI and recovery: the role of family functioning and emotion regulation. *Youth Soc*. (2018) 50:1056–77. doi: 10.1177/0044118X16653153

55. Hunter DJ, Holmes C. Where medical statistics meets artificial intelligence. *N Engl J Med*. (2023) 389:1211–9. doi: 10.1056/NEJMra2212850

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 13 March 2024
DOI 10.3389/fpsyt.2024.1275236



OPEN ACCESS

EDITED BY
Mosad Zineldin,
Linnaeus University, Sweden

REVIEWED BY
Trine Theresa Holmberg Sainte-Marie,
Mental Health Services in the Region of Southern Denmark, Denmark
Brian Schwartz,
University of Trier, Germany

*CORRESPONDENCE
Jannis T. Kraiss
j.t.kraiss@utwente.nl

RECEIVED 09 August 2023
ACCEPTED 08 February 2024
PUBLISHED 13 March 2024

CITATION
Huisman SM, Kraiss JT and de Vos JA (2024) Examining a sentiment algorithm on session patient records in an eating disorder treatment setting: a preliminary study. *Front. Psychiatry* 15:1275236. doi: 10.3389/fpsyt.2024.1275236



# Examining a sentiment algorithm on session patient records in an eating disorder treatment setting: a preliminary study

Sophie M. Huisman[1], Jannis T. Kraiss[1]* and Jan Alexander de Vos[2,3]

[1]Department of Psychology, Health and Technology, Centre for eHealth and Wellbeing Research, University of Twente, Enschede, Netherlands, [2]Department of Research, GGZ Friesland Mental Healthcare Institution, Leeuwarden, Netherlands, [3]Human Concern, Centrum Voor Eetstoornissen, Amsterdam, Netherlands

**Background:** Clinicians collect session therapy notes within patient session records. Session records contain valuable information about patients' treatment progress. Sentiment analysis is a tool to extract emotional tones and states from text input and could be used to evaluate patients' sentiment during treatment over time. This preliminary study aims to investigate the validity of automated sentiment analysis on session patient records within an eating disorder (ED) treatment context against the performance of human raters.

**Methods:** A total of 460 patient session records from eight participants diagnosed with an ED were evaluated on their overall sentiment by an automated sentiment analysis and two human raters separately. The inter-rater agreement (IRR) between the automated analysis and human raters and IRR among the human raters was analyzed by calculating the intra-class correlation (ICC) under a continuous interpretation and weighted Cohen's kappa under a categorical interpretation. Furthermore, differences regarding positive and negative matches between the human raters and the automated analysis were examined in closer detail.

**Results:** The ICC showed a moderate automated-human agreement (ICC = 0.55), and the weighted Cohen's kappa showed a fair automated-human (k = 0.29) and substantial human-human agreement (k = 0.68) for the evaluation of overall sentiment. Furthermore, the automated analysis lacked words specific to an ED context.

**Discussion/conclusion:** The automated sentiment analysis performed worse in discerning sentiment from session patient records compared to human raters and cannot be used within practice in its current state if the benchmark is considered adequate enough. Nevertheless, the automated sentiment analysis does show potential in extracting sentiment from session records. The automated analysis should be further developed by including context-specific ED words, and a more solid benchmark, such as patients' own mood, should be established to compare the performance of the automated analysis to.

KEYWORDS

**eating disorders, automated sentiment analysis, session patient records, validation, sentiment extraction**

# 1 Introduction

Eating disorders (EDs) are serious psychological disorders characterized by disturbed eating patterns that can lead to (severe) somatic and psychological complications (1, 2). The three most common EDs are anorexia nervosa (AN), bulimia nervosa (BN), and binge-eating disorder (BED) (1). EDs that do not meet the criteria of one of the aforementioned disorders but do create significant distress or functional impairment are classified under the category of "other specified feeding and eating disorders" (OSFED) (2). The lifetime prevalence of all EDs is 8.4% for women and 2.2% for men, which has increased during the last decades (3–5).

Despite the different types of therapy available for EDs, they remain challenging to treat and are followed by high levels of relapse, reflecting the often chronic nature of these disorders (6–9). Hence, it is essential to better understand and monitor the recovery process to protect individuals against relapse. One way to facilitate recovery is by monitoring the responsiveness of patients to treatment with routine outcome monitoring (ROM) (10). The ROM is an instrument to periodically evaluate patients' progress through using diagnostic indicators and severity scales (11, 12). ROM can alert therapists when treatment is ineffective, indicate a worsening of symptoms, or reassure patients by providing insight into slight improvements in their situation (13).

However, ROM requires patients to fill out self-report questionnaires, which may lead to subjective bias resulting in an over- or underestimation of patient's states (14). Furthermore, ROM is supposed to be administered at fixed time intervals during treatment, which is burdensome for patients and time-consuming for therapists, making it costly and not always feasible within clinical settings (11, 15–17). As a result, ROM is often only completed at the beginning and end of therapy, leading to a limited representation of patients' treatment progress (16, 18). The limitations of the ROM demonstrate that therapists could benefit from a less burdensome procedure and data utilization to continuously monitor patients' treatment progress.

Therapists already collect information about patients' treatment progress within session-by-session patient records (session records) (19). Clinicians write the session records after therapy sessions that may contain valuable information, such as patients' reactivity to and states during treatment, details of therapeutic conversations, and clinicians' impressions of the patient (20, 21). Session records are essential to treatment as they improve patient care by ensuring effective communication between clinicians and can support the substantiation of treatment choices (22, 23). Exhaustive evaluation of the session records could yield insightful information into patients' treatment process and progress.

However, the utilization of session records in research is limited due to the records being lengthy and complex, requiring more advanced and customized approaches to manage the difficulties in extracting information from such texts (21, 24–26). The session records are classified as unstructured data, meaning that the qualitative texts are not stored in an organized predefined format, making them challenging to analyze with conventional analysis techniques (27, 28). One conventional method to analyze such texts is by using human raters. However, this task is demanding and time-consuming and is often not feasible when large amounts of text data are involved (29). Throughout the last few years, new techniques have emerged that allow for more cost-effective and efficient analysis of unstructured text data (30). One such method is natural language processing (NLP) in which computer programs attain the ability to understand natural language in text or spoken words (31). A subfield within NLP is automated sentiment analysis, aiming to analyze natural language by using an algorithm operating through a set of rules to identify sentiment encompassing attitudes, emotions, appraisals, and the emotional tone within a text (32). Hence, automated sentiment analysis could be particularly suited to analyze session records because these often contain sentiment.

Sentiment analysis has become increasingly popular and was mainly used for the mining of sentiment from online customer reviews. However, prior research has started to examine the sentiment of patients' medical records, which showed potential regarding the mining of sentiment from such texts (33–36). Despite this, sentiment analysis applications within clinical practice remain limited; especially, the sentiment within session records has hardly been examined.

A few sentiment analysis studies have been executed within a clinical setting. A study by Provoost et al. (37) investigated the performance of an automated sentiment analysis on texts from online behavioral therapy interventions regarding different psychological disorders against a set of human raters. They found that the sentiment analysis performed similarly to the human raters in discerning sentiment from such mental health texts. Furthermore, a study investigating the performance of four different sentiment analyses on healthcare-related texts against a human baseline found three sentiment analyses to have fair agreement and one to have moderate agreement with the human raters (38). Moreover, a study evaluating the sentiment on videos and comments about AN found a fair agreement between the automated sentiment analysis and human raters (39). However, to date, only one study has investigated the performance of an automated sentiment analysis on written statements from patients diagnosed with anorexia nervosa regarding their body perception (40). This study showed that a relationship existed between patients' vocabulary in written texts and their mental states, Furthermore, the texts could be categorized in one of the six predefined subcategories related to AN (40).

Despite these studies showing promising results, a challenge within this type of research is that there is no solid benchmark to compare the performance of automated sentiment analyses with, because research regarding the analysis of sentiment from session records within the mental healthcare domain is very limited. For example, Provoost and colleagues (37) used the agreement among the human raters as benchmark to compare the performance of the automated sentiment analysis too. Their research suggested that the automated sentiment analysis performed similar to the human raters. However, the aforementioned study showed a moderate human-human agreement, meaning that the human raters differed in many cases regarding the sentiment of the texts. Hence, because of a lack of consensus between raters, it cannot be determined with certainty whether the performance of the automated sentiment analysis is either "good" or "bad". Another

point is that this research is conducted within the field of clinical psychology; therefore, thorough research is required on new technologies before they can actually be applied in practice (37, 41). Furthermore, automated sentiment analyses can be highly context-specific, as texts within different contexts may require different vocabulary and language, such as analyzing social media texts in contrast to clinical documents (42–45). Thus, the vocabulary within an ED context may differ from the vocabulary used within other domains of mental healthcare.

In all, limited evidence exists on the performance of automated sentiment analyses on session patient records within an ED treatment context. The automated sentiment analysis is not tailored to an ED context; however, because of the context specificity of such analyses, it is not clear whether an automated sentiment analysis (without tailoring) can extract sentiment reliably and validly from session records within such a context. Furthermore, because of little understanding about the application of an automated sentiment analyses within clinical practice, it must be thoroughly researched and validated before such analyses can be applied within the clinical field. The session records are readily available to examine patients' treatment progress; therefore, efficient analysis of these records by an automated sentiment analysis may provide a less burdensome method for both patients and clinicians to monitor treatment progression over time and be used on different texts related to EDs. Therefore, this study will examine how an existing Dutch automated sentiment analysis evaluates unstructured text data from session patient records compared to human raters.

# 2 Materials and methods

## 2.1 Participants

Participants were Dutch patients with the criteria of having a minimum age of 17 at the time of providing an informed consent and an ED diagnosis during data collection. A total of 149 patients were asked to sign the consent form, of which 12.1% rejected. A total of 131 patients provided consent. A random selection was made for this preliminary study, including patients with different ED diagnoses and a minimum of forty session records.

The sample consisted of eight patients: two patients diagnosed with AN, three patients with BN, one with BED, and two with OSFED. Five patients were between the ages of 21 and 25, two between the ages of 26 and 30, and one between the ages of 31 and 35. The average duration of patients' treatment up to the start of the study was approximately 10 months (SD = 4.8).

## 2.2 Procedure

Patients' session records were evaluated on their sentiment by an automated sentiment analysis and separately by two human raters. The two human raters examined each session patient record and allocated a sentiment score to each record individually.

Data collection occurred between February 2019 and April 2022, during which participants received outpatient treatment at a specialized ED treatment institution in The Netherlands (46). Patients were diagnosed with an ED by a psychiatrist or clinical psychologist in collaboration with an intake team. Participants visited their therapist once or twice a week for individual face-to-face treatment sessions, which were partly online due to the restrictions regarding the COVID-19 pandemic in The Netherlands (47). Therapy sessions concerned topics regarding recovery, autonomy, and decreasing problematic eating behavior using cognitive behavioral therapy and insight-giving therapy. Patients also received homework after the sessions to apply what they had learned (46). Furthermore, at the start of treatment, each patient received an account for an eHealth environment in which questionnaires and exercises were offered, where patients were provided with a brochure explaining the aim of the research as well. Patients were able to contact the researchers for further information and signed an informed consent form which they could withdraw from when they no longer wished to participate (see Appendix A and B).

The client advisory board of Human Concern advised on the execution of the study regarding adherence to ethical principles concerning patient privacy, possible risk, and harm and clarity of the study brochure. The study protocol was approved by the board of directors at Human Concern and the Ethical Committee of the University of Twente (EC-220422).

## 2.3 Materials

### 2.3.1 Session patient record data

The data utilized for this study were session patient record data. The session records were written electronically within the used system by the clinicians during treatment; they were free to use their own format in writing the records and could include any information they deemed important. The records included information from therapy sessions, treatment progression, ROM results, and patients' background information. The records varied in length, language, and format. However, not all session records were suited for the analysis. Some records only contained brief information about arranged appointments with other clinicians or institutions or descriptions of actions taken by the clinician(s) regarding administrative activities. Therefore, records that included one (or several) of the aforementioned actions or contained less than five words were excluded from the analysis by the human raters. In contrast, the automated analysis only excluded records with less than five words or records that did not include sentiment words.

### 2.3.2 Anonymization

The model "deduce" tailored to the Dutch language was executed on the pseudonymized session patient records to anonymize the data (48). First, patient and postal codes, addresses, email addresses, telephone numbers, URLs, and other contact information, including those of relatives, clinicians, and other care providers and institutions, were excluded. Second, the session records were tokenized; names and initials were changed to (NAME-1) and dates to (DATE-1); and dates indicating the start or

end of treatment were transformed to a month and year, ages to (AGE), and locations or cities to (LOCATION-1).

### 2.3.3 Automated sentiment analysis

To analyze the sentiment within the session records, an automated sentiment analysis from 6Gorillas tailored to the Dutch language and mental healthcare domain was used (49). Before analyzing the data, the sentiment analysis automatically pre-processed the data by transforming capital letters to lowercase letters and removing stop words, numbers, words with only one character, and underscores to improve the data mining functionality and prevent misleading results (50). The automated sentiment analysis employed a top-down lexicon-based approach, using three lexicons to extract sentiment. The primary lexicon used was from NRC Word-Emotion Association containing English sentiment words translated into Dutch; furthermore, a healthcare-specific lexicon created by 6Gorillas and an adjustment dictionary from Ynformed (a data science company) changed or removed words with multiple meanings within a text (51).

The lexicon indicated whether a positive or negative sentiment score was awarded to a sentiment-bearing word within a session record. Furthermore, the automated sentiment analysis searched for words prior to a sentiment-bearing word to examine the semantic context by using N-grams, including bigrams (a two-word sequence) and trigrams (a three-word sequence). Consequently, the automated analysis could account for negations that reverse the polarity of a sentence (e.g., "not good") and strengthening words ("extremely good") (52, 53). The sentiment score of a bigram was calculated by scoring the sentiment-bearing word with either "0," "+1," or "−1," which was multiplied by two when the preceding word was a reinforcer, and the sentiment score was inverted when the preceding word was a negation. The final score was calculated by adding all the bigram scores of a session record divided by the total number of bigrams (49). For trigrams, the same approach was used; the sentiment-bearing word determined the sentiment, and the two preceding words indicated whether the score was inverted or reinforced. The final score was calculated by adding all the trigrams scores of a session record divided by the total number of trigrams.

A final overall sentiment score was awarded to each session record, which was an average of all the sentiment scores within a record ranging between an interval of −1 and 1. Higher (positive) scores indicated greater positive sentiment, scores close to zero indicated a neutral sentiment, and lower (negative) scores indicated a greater negative sentiment of the record.

### 2.3.4 Human sentiment analysis

The procedure of Provoost and colleagues (37) was followed for the human sentiment analysis as a guideline because this was the only study examining the extraction of sentiment from texts within a Dutch mental health context.

Two human raters were involved in the human sentiment analysis; the first author was considered the first human rater, and the last author the second human rater. First, the human raters rated the first 20 session records together to explore variations in their ratings. After individually rating a session record, they discussed their reasoning and justifications for their scores. This collaborative approach served as the foundation for the preliminary protocol. Subsequently, they independently rated the next eighty session records. After, a feedback session was arranged to discuss issues and difficulties concerning the sentiment rating, upon which the protocol was refined and finalized. Hereafter, the new protocol was used to evaluate the overall sentiment of the remaining session (see Appendix C). Every record was rated on a scale from 1 to 7, with "1" indicating very negative, "2" indicating negative, "3" indicating somewhat negative, "4" indicating neutral, "5" indicating somewhat positive, "6" indicating positive, and "7" indicating very positive.

The category "neutral" was assigned when a record was considered objective (including no sentiment) or contained about the same number of positive and negative sentiments. Furthermore, a separate category "mixed" was created to indicate that a session record contained both an equal number of positive and negative sentiment. Because the automated sentiment analysis frequently scored such records as "neutral," the category "mixed" was created to explore the frequency of this occurrence.

## 2.4 Data analysis

Analyses were performed within the statistical program R (54) and Statistical Package of the Social Sciences (SPSS) 28 (55). The alpha level was set at 0.05.

### 2.4.1 Data preparation

The raw sentiment scores from the automated sentiment analysis and scores from the human raters were standardized in order to compare the automated and human sentiment analysis.

#### 2.4.1.1 Automated sentiment analysis

Categories were created for the standardized sentiment scores on the session records from the automated analysis. For the standardized sentiment scores, no score of zero existed indicating the category "neutral," given the wide range of scores generated by the automated analysis. Therefore, the category "neutral" was defined as a range bounded by the first positive and first negative standardized sentiment score. The category "negative" was defined by the scores below the first negative standardized sentiment score, and the category "positive" was defined by the scores above first positive standardized sentiment scores. Consequently, the categories for the standardized overall sentiment scores from the automated analysis were defined as follows: negative for values smaller than −0.03 and positive for values larger than 0.11.

#### 2.4.1.2 Human sentiment analysis

Categories were created for the raw sentiment scores of each human rater as these are similar to the standardized sentiment scores. Values smaller than 4 were categorized as negative, values larger than 4 as positive, and scores equal to 4 as neutral.

Furthermore, the sentiment scores of each human rater were standardized. An overall human sentiment score was calculated by

taking the average of both raters' sentiment score on each record, which was standardized and is referred to as the average human rating. A contingency table was created, including both human raters' raw sentiment scores and a frequency distribution of negative, neutral, and positive scores between the human raters.

### 2.4.2 Human-automated agreement

#### 2.4.2.1 Categorical interpretation

A weighted Cohen's kappa was calculated to assess the inter-rater agreement (IRR), which measured the extent that two (or more) examiners agreed on their assessment decisions (56). The weighted Cohen's kappa accounted for ordinal categorical data and was used to measure a text's polarity in terms of its direction (category). The weighted Cohen's kappa was calculated to examine the IRR between the standardized categorical sentiment scores of the automated analysis and categorical scores of rater 1 and rater 2 (57, 58). Values for the weighted Cohen's kappa range between −1 and 1; the degree of agreement was interpreted as none (<0), slight (0 to 0.20), fair (0.21 to 0.4), moderate (0.41 to 0.60), substantial (0.61 to 0.80), or almost perfect reliability (> 0.80) (59).

#### 2.4.2.2 Continuous interpretation

The intra-class correlation (ICC) can be used to assess the IRR on continuous data and data with missing values (58). The ICC correlated the standardized sentiment scores of the automated analysis against the standardized sentiment scores of rater 1 and rater 2 to measure the intensity of the agreement between the two analyses, accounting for a two-way mixed effect model based on an absolute agreement (60). Values for the ICC ranged between 0 and 1; the degree of agreement was interpreted as poor (<0.50), moderate (0.50 to 0.75), good (0.75 to 0.90), and excellent reliability (>0.90) (60).

### 2.4.3 Human-human agreement

#### 2.4.3.1 Categorical interpretation

A weighted Cohen's kappa was calculated to assess the IRR between the categorical scores of the human raters. The Cohen's kappa was interpreted as aforementioned.

#### 2.4.3.2 Continuous interpretations

The ICC was calculated to assess the IRR between the raw sentiment scores of the human raters. The ICC was interpreted as aforementioned.

### 2.4.4 Human-automatic agreement per individual patient

#### 2.4.4.1 Continuous interpretation

The ICC was calculated to assess the IRR between the standardized scores of the automated sentiment analysis and each human rater for each patient individually. The ICC was interpreted as aforementioned.

#### 2.4.4.2 Differences between the automated sentiment analysis and human sentiment analysis

A line graph was created for each patient to visualize the differences between the automated and human sentiment analysis, illustrating a patient's sentiment score over time. The graphs included the standardized automated sentiment analysis's and average human sentiment scores on each session record (y-axis) and the number of records (x-axis). The average human sentiment rating was used due to the good (ICC = 0.89) and substantial (k = 0.68) human-human agreement. Furthermore, deviations in sentiment scores between the automated and human raters were examined and reflected upon. The sentiment-bearing words and its assigned positive or negative match by the automated sentiment analysis and human raters were explored in closer detail. Accordingly, a word list was created for words specific to an ED context, which were not considered during the automated analysis. Furthermore, a word list was created for words considered of positive or negative sentiment by the automated analysis, which were not considered or considered of the opposite sentiment by the human raters.

## 3 Results

### 3.1 Patient session records

Out of the total 460 session patient records with an average of 57.50 (SD = 48.02) records per patient, 268 (58.3%) records were deemed relevant for the analysis by the first human rater and 263 (57.1%) by the second rater, whereas the automated analysis scored 315 (68.5%) records as relevant for the analysis.

### 3.2 Categorical comparison between the human raters and automated sentiment analysis

The automated sentiment analysis rated more session records as positive compared to the human raters, whereas the scores for the categories neutral and negative from the automated analysis and human raters are closer to each other (see Table 1). The human raters showed similar ratings for each category, with the largest difference for the category "positive" (see Table 1).

Furthermore, the human raters showed the most consensus on the scoring of the session records in the "positive" category, followed by the "negative" category (see Table 2). The lowest consensus was observed for the category "neutral" where, when one human rater categorized a record as "neutral," the other human rater more often categorized the record in one of the other two categories.

### 3.3 Automated-human agreement

#### 3.3.1 Categorical interpretation

The weighted Cohen's kappa indicated a fair agreement, k = 0.29 (95% CI, 0.199 to 0.387, p < 0.001), between the automated sentiment analysis and rater 1 regarding overall sentiment of the session records.

The weighted Cohen's kappa indicated a fair agreement, k = 0.29 (95% CI, 0.191 to 0.378, p < 0.001), between the automated

TABLE 1 Comparison of categorical sentiment evaluations on the session patient records from the human raters and automated sentiment analysis.

| | Rater 1 *N* (%) | Rater 2 *N* (%) | Automated Analysis *N* (%) |
|---|---|---|---|
| Negative (%) | 126 (47.0%) | 127 (48.3%) | 135 (36.8%) |
| Neutral (%) | 64 (23.9%) | 70 (26.6%) | 64 (20.3%) |
| Positive (%) | 78 (29.1%) | 66 (25.1%) | 116 (42.9%) |
| Total | 268 | 263 | 315 |

sentiment analysis and rater 2 regarding overall sentiment of the session records.

### 3.3.2 Continuous interpretation

The ICC analysis revealed a moderate IRR [ICC = 0.51, CI = 0.37–0.61, F(267, 267) = 2.02, $p < 0.001$] between the automated analysis and rater 1 regarding overall sentiment on the session records.

The ICC analysis revealed a moderate IRR [ICC = 0.57, CI = 0.43–0.65, F(262, 262) = 2.245 $p < 0.001$] between the automated analysis and rater 2 regarding overall sentiment on the session records.

## 3.4 Human-human agreement

### 3.4.1 Categorical interpretation

The weighted Cohen's kappa indicated a substantial agreement [k = 0.68 (95% CI, 0.62 to 0.75), p = 0.000] between rater 1 and rater 2 regarding overall sentiment on the session records.

### 3.4.2 Continuous interpretation

The ICC analysis revealed a good IRR [ICC = 0.89, CI = 0.86–0.91, F(262, 262) = 9.02, $p < 0.001$] between rater 1 and rater 2 regarding overall sentiment on the session records.

## 3.5 Automated-human agreement per individual patient

### 3.5.1 Continuous interpretation

The ICC revealed a poor IRR for participants 1 (OFSED), 4 (AN), and 6 (BN) for rater 1 (see Table 3). The ICC revealed a poor IRR for participants 1, 4, and 5 (BED) for rater 2 (see Table 4). Moderate ICC values were found for the remaining participants for both raters. The values were significant for four cases for rater 1 and five cases for rater 2 (see Tables 3, 4).

### 3.5.2 Differences between the automated and human sentiment analysis

The visualizations of the sentiment over time per patient regarding sentiment scores from the automated analysis and human raters can be seen in Figures 1–8. Figure 1 shows a large difference between the average human rating and the automated analysis on session record 106 of participant 1, where the automated analysis showed a sentiment score of 4.0; however, the human raters identified this record as irrelevant. Likewise, in Figure 2, the automated sentiment analysis peaked at record 34 of participant 2, whereas the human raters considered this record irrelevant. Participants 4, 5, and 6 illustrate this occurrence as well, showing a larger peak of the automated analysis without the human raters having assigned a sentiment score to the record in question, such as on record, 10, 13, and 17, respectively, in Figures 4-6. The automated sentiment analysis presenting a considerably larger sentiment score compared to the human rater is often paired with the human raters evaluating the session record as irrelevant.

#### 3.5.2.1 Sentiment words specific to ED context.

The automated sentiment analysis did not consider words specific to an ED context. An example can be seen from participant 4 diagnosed with AN in session record 37, where the average human rating showed a sentiment score of 1.97 and the automated sentiment analysis a score of 0.40 (see Figure 4). When examining the positive and negative matches from the automated analysis on the record, it was observed that the automated analysis did not rate certain context-specific positive ED words or expressions. For instance, the automated analysis did not rate the expression "beautiful recovery line," "feeling more," or "taking space," which are of positive sentiment within the context of EDs. The aforementioned examples are not the only ones encountered when examining the differences between positive and negative matches of the human raters and the automated sentiment analysis. Therefore, a list with context-specific ED words and the different diagnoses can be found in Appendix D.

Lastly, the automated sentiment analysis categorized certain words to have a positive or negative polarity, which were not considered or considered of the opposite sentiment within the human analysis. For example, the automated analysis indicated

TABLE 2 Comparison between the human raters' categorical sentiment evaluations on the patient session records.

| | | Rater 2 | | | |
|---|---|---|---|---|---|
| Rater 1 | | Negative *N* (%) | Neutral N (%) | Positive N (%) | Total N (%) |
| | Negative | 106 (83.5%) | 14 (20.0%) | 5 (7.6%) | 125 (47.5%) |
| | Neutral | 14 (11.0%) | 43 (61.4%) | 4 (6.1%) | 61 (23.2%) |
| | Positive | 7 (5.5%) | 13 (18.6%) | 57 (86.4%) | 77 (29.3%) |
| | Total | 127 | 70 | 66 | 263 |

TABLE 3 Intra-class correlation value for the agreement between the first human rater and the automated sentiment analysis per participant.

| | ICC | 95% CI | | F-statistics | | |
|---|---|---|---|---|---|---|
| | | Lower | Upper | Value | df1 | df2 |
| Participant 1 (OFSED) | 0.13 | −0.47 | 0.48 | 1.14 | 56 | 56 |
| Participant 2 (AN) | 0.63 | 0.34 | 0.79 | 2.65** | 49 | 49 |
| Participant 3 (BN) | 0.50 | −0.29 | 0.82 | 2.10 | 15 | 15 |
| Participant 4 (AN) | 0.37 | −0.18 | 0.67 | 1.58 | 41 | 41 |
| Participant 5 (BED) | 0.60 | 0.25 | 0.78 | 2.42** | 43 | 43 |
| Participant 6 (BN) | 0.38 | −0.60 | 0.75 | 1.57 | 20 | 20 |
| Participant 7 (BN) | 0.69 | 0.19 | 0.87 | 3.05* | 19 | 19 |
| Participant 8 (OFSED) | 0.60 | −0.11 | 0.85 | 2.41* | 17 | 17 |

ICC, intra-class correlation; CI, confidence intervals, * < 0.05, ** < 0.01.

"exercising" or "compensating" as a positive match on a record with a patient diagnosed with AN when, in fact, these expressions are mostly not of a positive polarity within such a context. Moreover, the words "emotion regulation" and "body experience" were categorized as a negative match. However, these were not considered sentiment-bearing words in the human analysis. Further differences regarding the positive and negative matches between the automated analysis and human raters can be found in Appendix E.

# 4 Discussion

The aim of this study was to examine the performance of an automated sentiment analysis at extracting sentiment from session patient records within an ED treatment context compared to human raters. In addition, the purpose of this study was to provide feedback to the designers of the automated sentiment analysis to optimize the analysis' future utilization potential. The results showed a fair automated-human agreement with rater 1 and rater 2 (k = 0.29) under categorical interpretation and a moderate automated-human agreement with rater 1 (ICC = 0.51) and rater 2 (ICC = 0.55) under continuous interpretation regarding the extraction of overall sentiment from the session records. The human-human agreement regarding overall sentiment was substantial under the categorical interpretation (k = 0.68) and good (ICC = 0.89) under the continuous interpretation. Furthermore, the automated analysis scored the sentiment of the session records more positive than the human raters. The automated analysis did not demonstrate increased difficulties in assessing sentiment related to specific types of EDs, despite its challenges with disorder-specific vocabulary.

## 4.1 Automated-human agreement

The findings of the automated-human agreement are partly in line with other studies. While this study found a moderate continuous automated-human agreement and a fair categorical agreement for both human raters, the exemplary study by Provoost et al. (37) found a moderate automated-human agreement under both continuous and categorical interpretations. Furthermore, a study investigating the performance of four different sentiment analyses compared to a baseline of multiple human raters

TABLE 4 Intra-class correlation value for the agreement between the second human rater and the automated sentiment analysis per participant.

| | ICC | 95% CI | | F-statistics | | |
|---|---|---|---|---|---|---|
| | | Lower | Upper | Value | df1 | df2 |
| Participant 1 (OFSED) | 0.30 | −0.18 | 0.59 | 1.44 | 55 | 55 |
| Participant 2 (AN) | 0.72 | 0.51 | 0.84 | 3.52** | 49 | 49 |
| Participant 3 (BN) | 0.66 | −0.05 | 0.88 | 2.93* | 15 | 15 |
| Participant 4 (AN) | 0.41 | −0.09 | 0.68 | 1.69* | 41 | 41 |
| Participant 5 (BED) | 0.40 | 0.10 | 0.67 | 1.66* | 43 | 43 |
| Participant 6 (BN) | 0.69 | 0.21 | 0.88 | 3.24* | 18 | 18 |
| Participant 7 (BN) | 0.68 | 0.14 | 0.88 | 3.07* | 17 | 17 |
| Participant 8 (OFSED) | 0.54 | −0.26 | 0.83 | 2.12 | 17 | 17 |

ICC, intra-class correlation; CI, confidence intervals, * < 0.05, ** < 0.01.

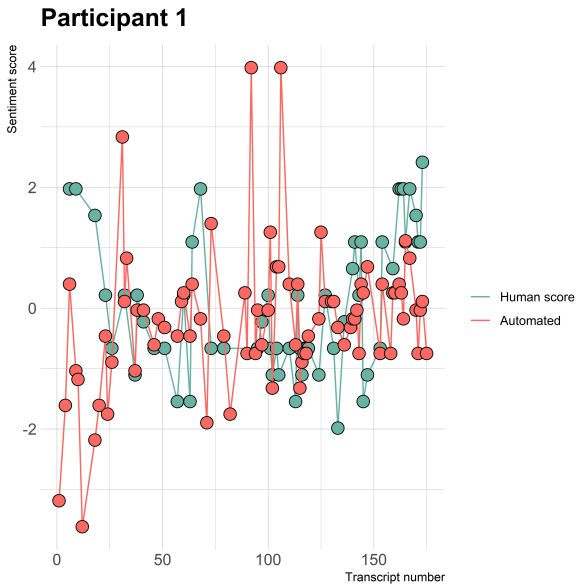



FIGURE 1
Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant 1 (OFSED) (N = 175).

found a fair automated-human agreement for three sentiment analyses and one moderate agreement, all under a categorical interpretation (38). Similarly, a study by Oksanen et al. (39) found a fair automated-human categorical agreement between an automated sentiment analysis and each of its three human raters, rating the sentiment of videos and comments related to AN.

Some shortcomings of the automated analysis could explain the findings of the automated-human agreement. The automated analysis' lexicon did not include vocabulary of sentiment specific to an ED context and labeled negative words as positive and vice versa. Besides, the automated analysis assigned a sentiment score to words that were not sentiment-bearing and not considered by the human raters. Literature has shown that sentiment analyses are often domain and context-specific, accordingly, a word's polarity may have been altered due to the context and domain within which it occurred and labeled as of the opposite sentiment (42–45). Furthermore, the automated analysis used "n-grams," which only considered words before a sentiment-bearing word and not after; as a result, it may have overlooked the context of certain words and labeled them incorrectly. A study investigating the performance of different machine and deep learning methods showed that the accuracy of n-grams was best for unigrams (one-word sequences) and decreased with bigrams and trigrams, as these may contain more complex human language (61). These shortcomings could have led to a discrepancy in sentiment scores between the two analyses, leading to a lower automated-human agreement and potential more positive rating of the records' sentiment opposed to the human raters.

Another explanation that may cause a variance in the sentiment scores between the two analyses is the difference in approach regarding the rating of the session records. The automated analysis' word-by-word analysis with use of two and three-word combinations in comparison the human raters' holistic interpretation of the records' sentiment may result in diverging sentiment scores on the session records. This effect was amplified when only one or two words were rated by the automated analysis within a record compared to the human raters considering the entire record and, hence, caused a difference in the observed sentiment scores.

Furthermore, other possible explanations may be due to the characteristics of the session records. The records included occasional misspellings or incorrect sentences, implicit statements of sentiment, or varied in their length, content, and written language due to differences in writing of clinicians. This will make the extraction of sentiment from the records more complex and misinterpretation more likely by the automated analysis, whereas human raters possess the ability and intelligence to comprehend difficult and ambiguous sentences and to extract sentiment from these more precisely (40, 62). The automated sentiment analysis rated more records than the human raters due to its inability to consistently identify and exclude "irrelevant" records. This occasionally resulted in the algorithm rating records with minimal sentiment content, leading to outliers often paired with the human raters rating the records as "irrelevant." Furthermore, the session records often contained a summary of patients' difficulties and successes from the past days or weeks in between therapy sessions. Seventy percent of the session records classified as "neutral" within the human analysis were also categorized as "mixed," meaning that the records contained both an equal positive and negative polarity. Furthermore, the automated analysis' sentiment scores were mostly centered around zero, whereas the majority of the human raters' sentiment scores were mostly centered around slightly positive or slightly negative, meaning that sentiment may be difficult to extract from session records, often containing sentiment from both polarities.

Furthermore, the sentiment within the session records does not directly stem from the patients; rather, it is a clinician's

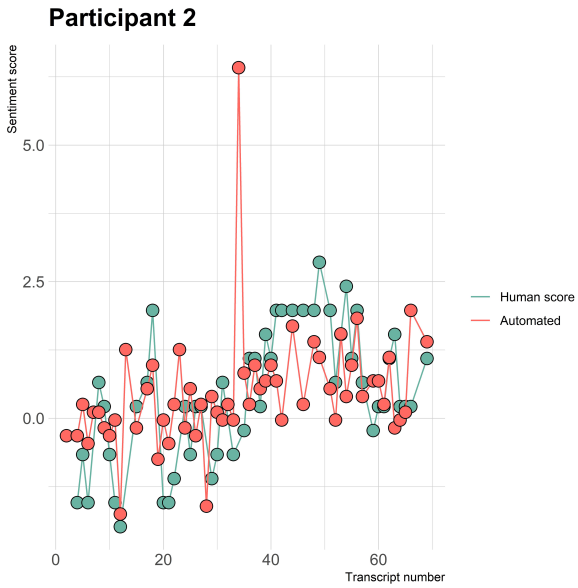



FIGURE 2
Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant 2 (AN) (N = 68).

interpretation of patients' sentiment and may, therefore, contain a subjective view of clinicians. The human raters agreed to only score sentiment stemming from the patients. Whereas human raters are able to distinguish between sentiment stemming from the patient or the clinician, the automated analysis could not. The human raters were able to take this into account when scoring the records that could have resulted in the observed difference in sentiment ratings.

In summary, the automated analysis performed worse in discerning sentiment from session patient records as opposed to the human raters, meaning that the automated sentiment analysis cannot be used within practice in its current state, assuming that the gold standard of the human-human agreement is considered good enough.

## 4.2 Agreement between human raters

The finding of the substantial categorical human-human agreement is in line with previous research, which investigated the performance of an automated sentiment analysis against two or

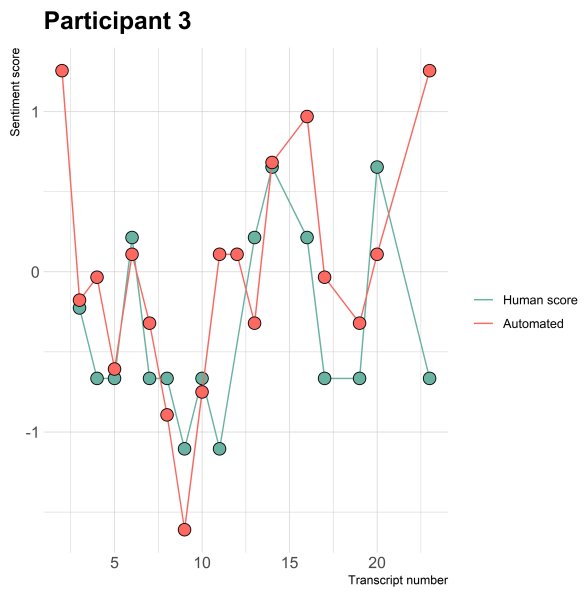



FIGURE 3
Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant 3 (BN) (N = 22).

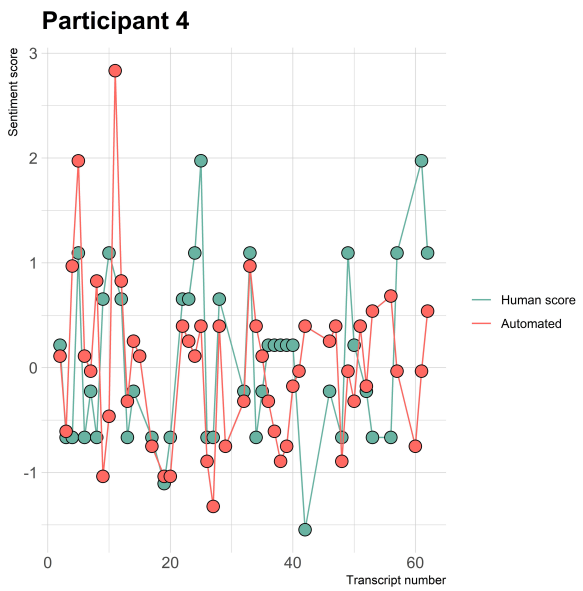



FIGURE 4
Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant 4 (AN) (N = 61).

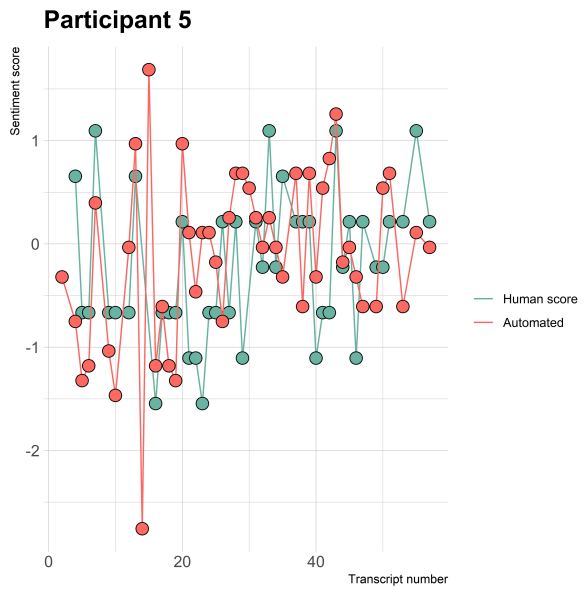



FIGURE 5
Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant 5 (BED) (N = 56).

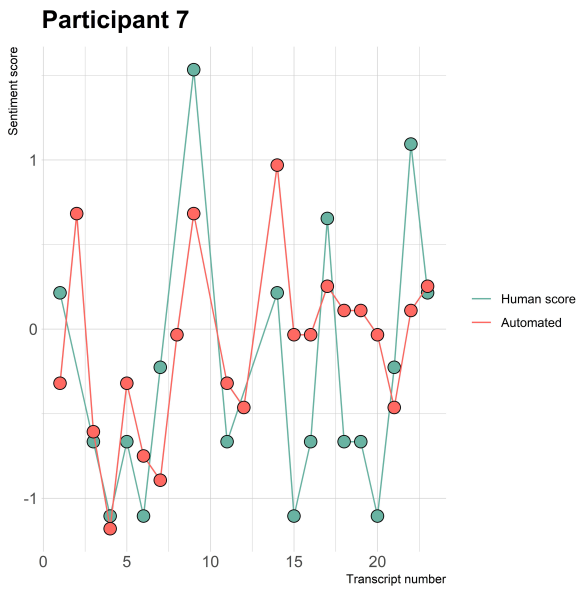



FIGURE 7
Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant 7 (BN) (N = 22).

three human raters and found a substantial agreement as well, under the categorical interpretation (42, 63, 64). In contrast, a lower (moderate) categorical and continuous human-human agreement was found in the study of Provoost and colleagues (37), who used an average of eight human raters per text.

A possible explanation for the findings could be due to the human raters' utilization of a feedback session and clear protocol. Likewise, a study by Moreno-Ortiz et al. (64) incorporated a feedback session to optimize the followed protocol. They found a significant increase in the human-human agreement between the first and second trial, ensuring that the session records were rated similarly. Furthermore, both raters of this study possessed knowledge of EDs as they were both educated within the field of psychology. Hence, they may have similarly interpreted words or expressions specific to an ED context and whether these were of positive or negative sentiment.

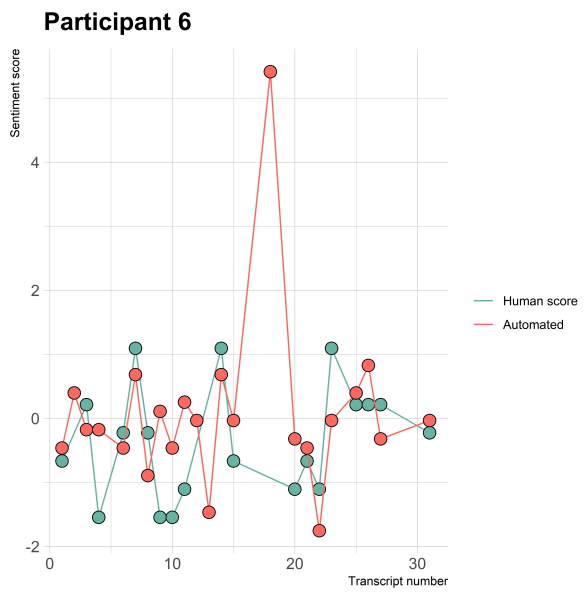



FIGURE 6
Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant 6 (BN) (N = 30).

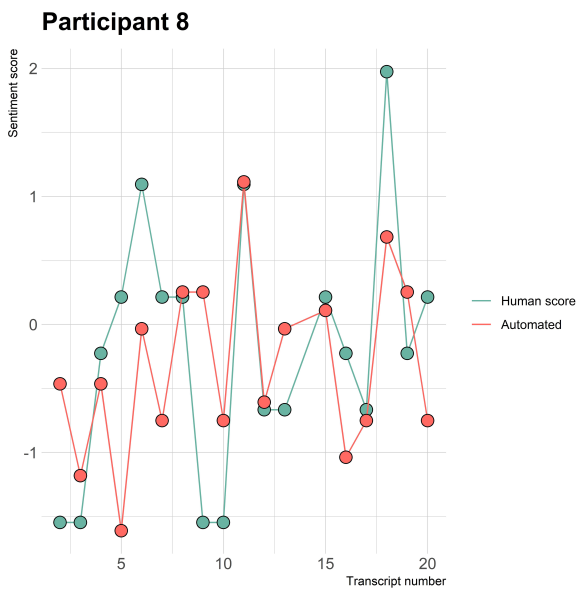



FIGURE 8
Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant 8 (OFSED) (N = 19).

The human-human agreement within this study was chosen as the "gold standard" to compare the performance of the automated analysis with. Nevertheless, no perfect agreement has been found within literature regarding human-human agreement for the validation of an automated sentiment analysis within a mental healthcare context, meaning that human raters still lack consensus regarding the rating texts' sentiment (42). For this reason, it cannot be determined with certainty whether the automated analysis performed either "good" or "bad" as there is no solid benchmark.

## 4.3 Qualitative differences between the automated sentiment analysis and human raters

The automated sentiment analysis was tailored to the Dutch language and mental healthcare context but not to the context of EDs. Hence, because of the context of the records and limited domain-specificity of the used lexicons, differences in positive and negative matches between the automated analysis and human raters were identified. Furthermore the automated analysis does not seem to encounter more difficulties with rating the sentiment within the context of a specific diagnosis, as each diagnosis once showed to have a lower automated-human agreement in comparison to the overall automated-human agreement, except for BN which showed a lower ICC value two times.

## 4.4 Strengths and limitations

A strength of this study is that the session records were written by trained clinicians providing real contextual data from patients from an actual ED treatment center. The findings of this study will also be provided as feedback to the developers of the automated sentiment analysis to improve its performance for future usage. Furthermore, the utilization of a feedback session may have supported that the records were rated similarly by the human raters (64). A limitation of this study was that there is no solid benchmark to compare the performance of the automated analysis with. The human raters were chosen as gold standard; however, the human raters still lack solid consensus when rating the session records. Hence, the results should be interpreted with caution. Furthermore, this study used fewer texts for the analysis than other research investigating the performance of automated sentiment analyses, as more than 40% of the records within this study were not suitable for the analysis, decreasing the reliability of the results and possibly leading to a selective sample of records (37, 39, 42, 65). The human raters only evaluated sentiment related to the patient, whereas the automated analysis rated an entire session record, which may have led to a discrepancy in the content evaluated by the human raters and the automated analysis. Therefore, the interpretation of the IRR between the human raters and the automated analysis requires caution. Another limitation is that the human raters may have been subjected to emotional bias, which is a distortion in one's cognitions due to emotional factors such as personal feelings at the time of decision-making (66). Consequently, the affective state of the human raters at the time of rating the session records could have influenced the sentiment score that was given to a certain text. Furthermore, this study only included two human raters, which makes for a less representative interpretation of the overall sentiment within the session records compared to using multiple raters (67). Lastly, the method for the standardized sentiment scores regarding the category "neutral" differed between the automated analysis and human raters, as establishing a clear median or "neutral" point was challenging. The decision to use a range for the algorithm was made to accommodate the nuances and variability inherent in an automated sentiment analysis to represent the category "neutral." However, this may have resulted in differences within the category "neutral" between the automated analysis and human raters.

## 4.5 Future research and implications

The findings suggest that the automated analysis performs worse than human raters in discerning sentiment from session records. However, it is questionable whether the human-human agreement can be considered the gold standard to determine the performance of the automated analysis. Nevertheless, no clinically relevant IRR values that would allow methods to be applied within practice could be identified within the literature sufficient enough to apply such methods within practice, and, therefore, although excellent reliability should be strived for, it is of interest to investigate what IRR values are sufficient enough to apply such methods within clinical practice.

This research is among the first to assess the performance of automated sentiment analysis on contextual patient data. Its potential application in clinical practice could serve as a feedback system, allowing for quick analysis of patients' sentiment over time. This could be especially long-term treatments, where subtle changes in sentiment might be challenging to discern through manual review alone. Consequently, this approach could reduce the burden on both clinicians and patients and, importantly, aid in identifying when treatment adjustments are necessary or detect deterioration in patients' conditions. Such an application could be a significant step forward in optimizing mental healthcare delivery.

For future research, it is recommended to increase the number of human raters and examine the differences between the raters' sentiment scores in closer detail to improve the gold standard. Moreover, because of limited evidence regarding the utilization of human raters as the gold standard, patients' ratings of their own moods after or before therapy sessions or utilization of patients' diaries and accompanying mood ratings could make an additional benchmark to validate the automated sentiment analysis to. Furthermore, the sentiment scores of the automated analysis could be compared to therapists' sentiment ratings of the session records, which may not only yield insightful information about the efficacy of the tool but also identify sentiments that might not be

immediately apparent to the therapist and could give an additional layer of insight into patient progress.

Another key recommendation is to update the automated analysis lexicon with context-specific ED words and investigate its performance again on texts or session records within an ED treatment setting to improve its accuracy (68). Furthermore, potential confounding variables should be investigated by operating the automated sentiment analysis on more homogenized samples of texts with controlled participant demographics such as specific age groups and types of EDS to investigate the impact of different variables on the sentiment analysis.

Furthermore, the usability of session records for the extraction of patients' sentiment can be questioned because of its characteristics and it is primarily an account by the clinician of the patients' sentiment. Therefore, the sentiment of the session records and whether these could give an accurate representation of the patients' sentiment should be further investigated. In addition, future research could focus on exploring novel procedures to document patients' sentiment more directly, such as, by requesting the patient to summarize their feelings about the past week(s) in a few sentences at the beginning or end of a session, which could be used for the monitoring of patients' sentiment over time.

However, despite the session records including complex and ambiguous information, which makes them difficult to analyze, the records do contain valuable information about processes and underlying patterns contributing to EDs. Hence, it may be particularly interesting to use an open coding, through which the session records are examined on recurring ED themes, which may be beneficial for the understanding of the mechanisms exhibited by individuals with an ED disorder. Furthermore, it would be particularly interesting to explore session records capturing both sentiment from patients and clinicians to investigate the therapeutic alliance and dynamic, as this is a contributing factor within treatment and may yield insightful information about such processes.

## 5 Conclusion

To conclude, this study suggests that the current automated sentiment analysis tool does not perform as well as human raters in discerning sentiment from session patient records within a Dutch ED treatment context when compared against the human-human agreement standard. However, it is crucial to acknowledge the limitations of this benchmark. The lack of a solid consensus among human raters on sentiment evaluation indicates a need for alternative benchmarks in future research to more accurately assess the efficacy of automated sentiment analysis tools in clinical practice, such as patients' own mood ratings. Furthermore, this study showed that the sentiment of patients extracted from session records can be portrayed over time. Moreover, the automated sentiment analysis must be optimized by including context-specific ED terms and expressions within its lexicon to increase the analysis' accuracy, requiring further investigation. Lastly, it remains uncertain whether the patient session records are suitable for the extraction of patients' sentiments due to their complex and ambiguous nature containing both an equal number of positive and negative sentiment.

## Data availability statement

The raw data supporting the conclusion will be made available without undue reservation. However, the session patient records cannot be given without undue reservation due to the privacy of the participants. The patient session records of the participants can be made available upon reasonable request.

## Ethics statement

The studies involving humans were approved by Commission Ethics Psychology University of Twente. 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

SH: Data curation, Formal analysis, Investigation, Software, Writing – original draft. JK: Investigation, Software, Supervision, Validation, Visualization, Writing – review & editing. JdV: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Acknowledgments

The authors would like to express their sincere appreciation to the center of eating disorders Human Concern for acquiring the data to be evaluated. Further, the authors would like to express their sincere appreciation to Anton Kuijer and Wessel Sandtke working at 6Gorillas for providing the automated sentiment analysis to evaluate the texts with and their support regarding upcoming issues of the automated sentiment analysis. This manuscript has previously appeared online as a master's thesis (69).

## 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.2024.1275236/full#supplementary-material

## References


1. Keel PK, Brown TA, Holland LA, Bodell LP. Empirical classification of eating disorders. *Annu Rev Clin Psychol.* (2012) 8:381–404. doi: 10.1146/annurev-clinpsy-032511-143111

2. American Psychiatric Association. Feeding and eating disorders. In: *Diagnostic and Statistical Manual of Mental Disorders,* 5th (2013) (Washington, DC: American psychiatric publishing). doi: 10.1176/appi.books.9780890425596

3. Hoek HW. Review of the worldwide epidemiology of eating disorders. *Curr Opin Psychiatry.* (2016) 29:336–9. doi: 10.1097/yco.0000000000000282

4. 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. doi: 10.1093/ajcn/nqy342

5. Bagaric M, Touyz S, Heriseanu A, Conti J, Hay P. Are bulimia nervosa and binge eating disorder increasing? Results of a population-based study of lifetime prevalence and lifetime prevalence by age in South Australia. *Eur Eating Disord Rev.* (2020) 28:260–8. doi: 10.1002/erv.2726

6. Muzio LL, Russo LL, Massaccesi C, Rapelli G, Panzarella V, di Fede O, et al. Eating disorders: A threat for women's health. Oral manifestations in a comprehensive overview. *Minerva Stomatol.* (2007) 56:281–92.

7. Anderson LK, Reilly EE, Berner L, Wierenga CE, Jones MD, Brown TA, et al. Treating eating disorders at higher levels of care: Overview and challenges. *Currr Psychiatry Rep.* (2017) 19:1–9. doi: 10.1007/s11920-017-0796-4

8. von Holle A, Pinheiro AP, Thornton LM, Klump KL, Berrettini WH, Brandt H, et al. Temporal patterns of recovery across eating disorder subtypes. *Aust New Z J Psychiatry.* (2008) 42(2):108–17. doi: 10.1080/00048670903118465

9. Berends T, Boonstra N, van Elburg A. Relapse in anorexia nervosa: A systematic review and meta-analysis. *Curr Opin Psychiatry.* (2018) 31:445–55. doi: 10.1097/YCO.0000000000000453

10. Boswell JF, Kraus DR, Miller SD, Lambert MJ. Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychother Res.* (2015) 25:6–19. doi: 10.1080/10503307.2013.817696

11. de Beurs E, den Hollander-Gijsman ME, van Rood YR, van der Wee NJ, Giltay EJ, van Noorden MS, et al. Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clin Psychol Psychother.* (2011) 18:1–12. doi: 10.1002/cpp.696

12. Schulte-van Maaren YW, Carlier IV, Zitman FG, van Hemert AM, de Waal MW, van der Does AW, et al. Reference values for major depression questionnaires: the Leiden Routine Outcome Monitoring Study. *J Affect Disord.* (2013) 149:342–9. doi: 10.1016/j.jad.2013.02.009

13. Youn SJ, Kraus DR, Castonguay LG. The Treatment Outcome Package: Facilitating practice and clinically relevant research. *Psychotherapy.* (2012) 49:115–22. doi: 10.1037/a0027932

14. Karpen SC. The social psychology of biased self-assessment. *Am J Pharm Educ.* (2018) 82:441–8. doi: 10.5688/ajpe6299

15. Gilbody SM, House AO, Sheldon TA. Outcome measures and needs assessment tools for schizophrenia and related disorders. *Cochrane Database System Rev.* (2003) No. 1:1–14. doi: 10.1002/14651858.CD003081

16. Norman S, Dean S, Hansford L, Ford T. Clinical practitioner's attitudes towards the use of Routine Outcome Monitoring within Child and Adolescent Mental Health Services: A qualitative study of two Child and Adolescent Mental Health Services. *Clin Child Psychol Psychiatry.* (2014) 19:576–95. doi: 10.1177/1359104513492348

17. Kuo PB, Tanana MJ, Goldberg SB, Caperton DD, Narayanan S, Atkins DC, et al. Machine-learning-based prediction of client distress from session recordings. *Clin psychol Sci.* (2023), 1–12. doi: 10.1177/21677026231172694

18. Wampold BE. Routine outcome monitoring: Coming of age—With the usual developmental challenges. *Psychotherapy.* (2015) 52:458–62. doi: 10.1037/pst0000037

19. Swinkels ICS, Van den Ende CHM, De Bakker D, van der Wees P, Hart DL, Deutscher D, et al. Clinical databases in physical therapy. *Physiother Theory Pract.* (2007) 23:153–67. doi: 10.1080/09593980701209097

20. Maio JE. HIPAA and the special status of psychotherapy notes. *Prof Case Manag.* (2003) 8:24–9. doi: 10.1097/00129234-200301000-00005

21. Percha B. Modern clinical text mining: a guide and review. *Annu Rev Biomed Data Sci.* (2021) 4:165–87. doi: 10.1146/annurev-biodatasci-030421-030931

22. Patel VL, Kushniruk AW, Yang S, Yale JF. Impact of a computer-based patient record system on data collection, knowledge organization, and reasoning. *J Am Med Inf Assoc.* (2000) 7:569–85. doi: 10.1136/jamia.2000.0070569

23. Ledbetter CS, Morgan MW. Toward best practice: leveraging the electronic patient record as a clinical data warehouse. *J Healthc Inf Manag.* (2001) 15:119–31.

24. Raja U, Mitchell T, Day T, Hardin JM. Text mining in healthcare. Applications and opportunities. *J Healthc Inf Manage.* (2008) 22:52–6.

25. Berndt DJ, McCart JA, Finch DK, Luther SL. A case study of data quality in text mining clinical progress notes. *ACM Trans Manage Inf System.* (2015) 6:1–21. doi: 10.1145/2669368

26. Lee S, Xu Y, D'Souza AG, Martin EA, D Doktorchik C, Zhang Z, et al. Unlocking the potential of electronic health records for health research. *Int J Popul Data Sci.* (2020) 5:1–9. doi: 10.23889/ijpds.v5i1.1123

27. Boulton D, Hammersly M. Analysis of unstructured data. In: Sapsford R, Jupp V, editors. *Data Collection and Analysis.* Sage Publications, London, UK (2006). p. 243–66. doi: 10.4135/9781849208802

28. Nikhil R, Tikoo N, Kurle S, Pisupati HS, Prasad GR. A survey on text mining and sentiment analysis for unstructured web data. *J Emerg Technol Innovative Res.* (2015) 2:1292–6.

29. Basit T. Manual or electronic? The role of coding in qualitative data analysis. *Educ Res.* (2003) 45:143–54. doi: 10.1080/0013188032000133548

30. Smink WAC, Sools AM, van der Zwaan JM, Wiegersma S, Veldkamp BP, Westerhof GJ. Towards text mining therapeutic change: A systematic review of text-based methods for Therapeutic Change Process Research. *PloS One.* (2019) 14: e0225703. doi: 10.1371/journal.pone.022570

31. Chowdhary KR. Natural language processing. In: *Fundamentals of Artificial Intelligence.* Springer, Cham (2020). p. 603–49. doi: 10.1007/978-81-322-3972-7_19

32. Iliev R, Dehghani M, Sagi E. Automated text analysis in psychology: Methods, Applications, and Future Developments. *Lang Cognit.* (2015) 7:265–90. doi: 10.1017/langcog.2014.30

33. Hoerbst A, Ammenwerth E. Electronic health records. *Methods Inf Med.* (2010) 49:320–36. doi: 10.3414/ME10-01-0038

34. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study. *PloS One.* (2015) 10:1–10. doi: 10.1371/journal.pone.0136341

35. Carrillo-de-Albornoz J, Rodriguez Vidal J, Plaza L. Feature engineering for sentiment analysis in e-health forums. *PloS One.* (2018) 13:e0207996. doi: 10.1371/journal.pone.0207996

36. Mäntylä MV, Graziotin D, Kuutila M. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Comput Sci Rev.* (2018) 27:16–32. doi: 10.1016/j.cosrev.2017.10.002

37. Provoost S, Ruwaard J, van Breda W, Riper H, Bosse T. Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: An exploratory study. *Front Psychol.* (2019) 10:1065. doi: 10.3389/fpsyg.2019.01065

38. Georgiou D, MacFarlane A, Russell-Rose T. Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools. In: *Science and*

*Information Conference*. IEEE, London, UK (2015). p. 352–61. doi: 10.1109/SAI.2015.7237168

39. Oksanen A, Garcia D, Sirola A, Näsi M, Kaakinen M, Keipi T, et al. Pro-anorexia and anti-pro-anorexia videos on YouTube: Sentiment analysis of user responses. *J Med Internet Res*. (2015) 17:e256. doi: 10.2196/jmir.5007

40. Spinczyk D, Nabrdalik K, Rojewska K. Computer aided sentiment analysis of anorexia nervosa patients' vocabulary. *BioMed Eng Online*. (2018) 17:1–11. doi: 10.1186/s12938-018-0451-2

41. Ben-Zeev D. Technology in mental health: creating new knowledge and inventing the future of services. *Psychiatr Services*. (2017) 68:107–8. doi: 10.1176/appi.ps.201600520

42. Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. In: *Proceedings of human language technology conference and conference on empirical methods in natural language processing*. Association for Computational Linguistics, Vancouver, Canada (2005). p. 375–54. doi: 10.3115/1220575.1220619

43. Goeuriot L, Na JC, Min Kyaing WY, Khoo C, Chang YK, Theng YL, et al. Sentiment lexicons for health-related opinion mining. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. Association for Computer Machinery, New York, NY (2012). p. 219–26. doi: 10.1145/2110363.2110390

44. Denecke K, Deng Y. Sentiment analysis in medical settings: New opportunities and challenges. *Artif Intell Med* (2015) 64(1):17–27. doi: 10.1016/j.artmed.2015.03.006

45. Mohammad SM. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement* (2016), 201–37. doi: 10.1016/B978-0-08-100508-8.00009-6

46. Human Concern. Ambulante behandeling. Available online at: https://humanconcern.nl/ambulante-behandeling/ (Accessed July 26, 2023).

47. Daemen DM. De groep in tijden van corona. *Tijdschrift voor Groepsdynamica groepspsychotherapie*. (2020) 15:4–12.

48. Menger V, Scheepers F, van Wijk LM, Spruit M. Deduce: A pattern matching for automatic de-identification of Dutch medical text. *Telemat Inform*. (2017) 35:727–36. doi: 10.1016/j.tele.2017.08.002

49. 6Gorillas. Het innovatieve dataplatform voor de zorg(2021). Available online at: https://6gorillas.nl/ (Accessed July 26, 2023).

50. Hemalatha I, Varma GS, Govardhan A. Preprocessing the informal text for efficient sentiment analysis. *Int J Emerging Trends Technol Comput Sci (IJETTCS)*. (2012) 1(2):58–61.

51. Royal HaskoningDHV. (2018). Available online at: https://www.royalhaskoningdhv.com/ (Accessed July 27, 2023).

52. Dadvar M, Hauff C, de Jong F. Scope of negation detection in sentiment analysis. In: *DIR 2011: Dutch-Belgian Information Retrieval Workshop*. University of Amsterdam, Amsterdam (2011). p. 16–9.

53. Farooq U, Mansoor H, Nongaillard A, Ouzrout Y, Qadir AM. Negation handling in sentiment analysis at sentence level. *J Comput*. (2017) 12:470–8. doi: 10.17706/jcp.12.5.470-478

54. R Core Team. *R: A Language and Environment for Statistical Computing* (2016). Available online at: https://www.r-project.org (Accessed July 27, 2023).

55. IBM Corp. IBM SPSS Statistics for Macintosh, Version 28.0(2021). Available online at: https://hadoop.apache.org (Accessed July 26, 2023).

56. Lange RT. Interrater reliability. In: Kreutzer JS, DeLuca J, Caplan B, editors. *Encyclopedia of Clinical Neuropsychology*. Springer, New York, NY (2011). p. 1348. doi: 10.1007/978-0-387-79948-3

57. Cohen J. A coefficient of agreement for nominal scales. *Educ psychol Meas*. (1960) 20:37–46. doi: 10.1177/001316446002000104

58. Devitt A, Ahmad K. Sentiment polarity identification in financial news: a cohesion-based approach. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague: Association for Computational Linguistics (2007). p. 984–91.

59. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. (1977) 33:159–74. doi: 10.2307/2529310

60. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. (2016) 15:155–63. doi: 10.1016/j.jcm.2016.02.012

61. Ojo OE, Gelbukh A, Calvo H, Adebanji OO. Performance study of N-grams in the analysis of sentiments. *J Nigerian Soc Phys Sci*. (2021) 3:477–83. doi: 10.46481/jnsps.2021.201

62. Pang B, Lee L. Opinion mining and sentiment analysis. *Found Trends Inf Retr*. (2008) 2:1–135. doi: 10.1561/1500000001

63. Mukhtar N, Khan MA, Chiragh N. Effective use of evaluation measures for the validation of best classifier in Urdu sentiment analysis. *Cogn Computation*. (2017) 9:446–56. doi: 10.1007/s12559-017-9481-5

64. Moreno-Ortiz A, Salles-Bernal S, Orrequia-Barea A. Design and validation of annotation schemas for aspect-based sentiment analysis in the tourism sector. *Inf Technol Tourism*. (2019) 21:535–57. doi: 10.1007/s40558-019-00155-0

65. Charter RA. Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *J Clin Exp Neuropsychol*. (1999) 21:559–66. doi: 10.1076/jcen.21.4.559.889

66. Yuan J, Tian Y, Huang X, Fan H, Wei X. Emotional bias varies with stimulus type, arousal and task setting: Meta-analytic evidences. *Neurosci Biobehav Rev*. (2019) 107:461–72. doi: 10.1016/j.neubiorev.2019.09.035

67. Stappen L, Schumann L, Sertolli B, Baird A, Weigell B, Cambria E, et al. Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox. In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. Association for Computing Machinery, Changu, China (2021). p. 75–82. doi: 10.1145/3475957.3484451

68. Islam MR, Zibran MF. (2017). Leveraging automated sentiment analysis in software engineering, in: *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, Piscataway, NJ: Institute of Electrical and Electronics Engineers. pp. 203–14. doi: 10.1109/MSR.2017.9

69. Huisman SM. A Preliminary Study Examining an Automated Sentiment Analysis on Extracting Sentiment from Session Patient Records in an Eating Disorder Treatment Setting [Master Thesis]. University of Twente, Enschede (Netherlands (2022).

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 10 July 2024
DOI 10.3389/fpsyg.2024.1345406



OPEN ACCESS

EDITED BY
Patrick K. A. Neff,
University of Zurich, Switzerland

REVIEWED BY
Jung Yeon Park,
George Mason University, United States
Larry R. Price,
Texas State University, United States

*CORRESPONDENCE
Miljan Jović
✉ m.jovic@utwente.nl

RECEIVED 13 December 2023
ACCEPTED 17 June 2024
PUBLISHED 10 July 2024

CITATION
Jović M, Haeri MA, Whitehouse A and van den Berg SM (2024) Harmonizing the CBCL and SDQ ADHD scores by using linear equating, kernel equating, item response theory and machine learning methods. *Front. Psychol.* 15:1345406. doi: 10.3389/fpsyg.2024.1345406



# Harmonizing the CBCL and SDQ ADHD scores by using linear equating, kernel equating, item response theory and machine learning methods

Miljan Jović[1]*, Maryam Amir Haeri[1], Andrew Whitehouse[2] and Stéphanie M. van den Berg[1]

[1]Department of Learning, Data Analytics and Technology, University of Twente, Enschede, Netherlands, [2]Telethon Kids Institute, University of Western Australia, Perth, WA, Australia

**Introduction:** A problem that applied researchers and practitioners often face is the fact that different institutions within research consortia use different scales to evaluate the same construct which makes comparison of the results and pooling challenging. In order to meaningfully pool and compare the scores, the scales should be harmonized. The aim of this paper is to use different test equating methods to harmonize the ADHD scores from Child Behavior Checklist (CBCL) and Strengths and Difficulties Questionnaire (SDQ) and to see which method leads to the result.

**Methods:** Sample consists of 1551 parent reports of children aged 10-11.5 years from Raine study on both CBCL and SDQ (common persons design). We used linear equating, kernel equating, Item Response Theory (IRT), and the following machine learning methods: regression (linear and ordinal), random forest (regression and classification) and Support Vector Machine (regression and classification). Efficacy of the methods is operationalized in terms of the root-mean-square error (RMSE) of differences between predicted and observed scores in cross-validation.

**Results and discussion:** Results showed that with single group design, it is the best to use the methods that use item level information and that treat the outcome as interval measurement level (regression approach).

KEYWORDS

**data harmonization, test equating, machine learning, IRT, linear equating, kernel equating, ADHD**

## 1 Introduction

When researchers work with data from different institutions, they often encounter situations where different scales are used for the evaluation of the same construct. This makes pooling of data and comparison of the results challenging. Nevertheless, combining data from different groups of participants who filled in different questionnaires is often necessary to obtain (a) sufficiently large sample sizes, (b) to be able to make comparisons across subpopulations, or (c) to increase generalizability and validity of research results (Smith-Warner et al., 2006; Thompson, 2009; Fortier et al., 2010, 2011; Hamilton et al., 2011; van den Berg et al., 2014).

Two mental health instruments that are widely used by different institutions for assessing the same constructs are the Child Behavior Checklist (CBCL) and the Strengths and Difficulties

Questionnaire (SDQ). Both assess mental health problems among children and adolescents, but they differ both quantitatively (different number of items) and content wise (e.g., different phrasing of items). The CBCL consists of 113 items and operationalizes childhood problem behavior on eight subscales/dimensions (social withdrawal, somatic complaints, anxiety/depression, social problems, thought problems, attention problems, delinquent behavior, and aggressive behavior; Achenbach et al., 1991; Achenbach and Ruffle, 2000). The SDQ consists of 25 items equally divided across five scales, also called dimensions (Emotional, Conduct, Hyperactivity, Peer, and Prosocial problems; Goodman, 1997, 2001).

Both instruments are already well-established and widely used for assessing psychopathology in general, but also for assessing specific mental health problems (Achenbach, 1991; Allen and Prior, 1995; Caspi et al., 1995; Muris et al., 2003; Ortuno-Sierra et al., 2015).

One relatively common mental health problem in children is Attention-Deficit Hyperactivity Disorder (ADHD). Both CBCL and SDQ contain items that address ADHD-related symptoms and various research studies proved that both of those instruments perform well in the context of screening for ADHD problems (Chen et al., 1994; Algorta et al., 2016; Hall et al., 2019).

Even though they are both valid, there are differences in the content and number of items. SDQ has a hyperactivity scale that also includes items that measure concentration problems (SDQ; 5 ADHD items in total), while CBCL is more focused on attention problems but also contains hyperactivity items (CBCL; 11 ADHD items in total). Both measure ADHD in broader sense, but they do not completely overlap. Those differences make it difficult to compare scores on SDQ and CBCL scales directly, as they have different distributions (e.g., different means and different variance). In order to make the scores obtained by those instruments comparable, it is necessary to harmonize them, that is, to put them on the same scale. Such a scale could be the SDQ scale, where CBCL scores are transformed in some way into an SDQ scale, or vice versa. Alternatively, both SDQ and CBCL scale scores could be translated into a third, normalised scale score, with for instance mean 0 and variance 1.

There are different methodologies that can be used for data harmonization, and the most common one is test equating, also known as test linking or scaling (Mislevy, 1992; Holland et al., 2006; Kolen and Brennan, 2014). It is applied mostly in the context of educational measurement, where the test scores from one exam need to be harmonized somehow with test scores from a similar but different exam.

The type of method used for test equating depends on what information is available or used. For example, if we only have test scores on exam A in pupils from school I and test scores on exam B in pupils from school II, we can either use only the mean test scores, we use both means and standard deviations, or we use the entire distributions in terms of quantiles. The respective methods that are based on these statistics are mean equating, linear equating and equipercentile equating. The strong assumption in these methods is that the scores provide all the necessary information (sufficiency) and that the two tests measure exactly the same trait, conceptually. Although reasonable for exam versions in education, in the context of psychopathology this may be too strong an assumption.

In the case if scales do not measure exactly the same thing, we need data to link those two scales. In that case, we need either at least some common items in both scales (Common Items study design with so-called anchor items), or the same group of persons that filled in both scale versions (Common Persons study design, a.k.a. single group design).

In Common Items design, there are two different samples of participants. One sample filled in scale A while another filled in scale B. Majority of items in those scales are different, but there is a certain number of items that is common for both scales. Accordingly, those overlapping (common) items that are present in both scales can be used to obtain harmonized scores.

Another design is Common Persons (single group) design in which we have only one sample of participants, but they filled in both scales at the same time. Since we have responses of all participants on both scales, we can use them to harmonize the scores.

This type of information (raw data at the item level), when it is available, could be used in a powerful way by implementing Item Response Theory (IRT) models. These models take into account not only differences between test takers, but also between the items, for instance their relative difficulty (van den Berg et al., 2014; Jabrayilov et al., 2016; Sansivieri et al., 2017; Jović et al., 2022; Mansolf et al., 2022).

Still, the basic unidimensional IRT approach includes the assumption that exactly the same trait is measured. If we allow for the possibility that the scales only partly overlap, in that the constructs that are measured are only correlated, we could use either a more complex IRT model (more complex than Rasch models), or try a whole bunch of other methods. Rasch model is a simple IRT model in which participants` response to an item is determined by the latent trait level of the participant and difficulty/threshold parameter of the item. Threshold parameter is defined as the point on the latent trait continuum where the response probability for two adjacent response categories is equal (Wetzel and Carstensen, 2014). In more complex IRT models (e.g., Generalized Partial Credit Model), participants` responses are determined not only by their latent trait value and difficulty of the item, but also by discrimination parameter of the item (refers to the strength of the relationship between trait level and participants` responses on the item; see Embretson and Reise (2000) for more details). For example, van den Berg et al. (2014) used Generalized Partial Credit Model to harmonize neuroticism and extraversion scores, while Jović et al. (2022) used it to harmonize anxiety/depression and ADHD scores of CBCL and SDQ scales. Mansolf et al. (2022) used IRT to harmonize Internalizing, Externalizing, and Total Problems domains from CBCL and SDQ. Recently there has been attention for methods based on the machine learning (ML) literature. For example, Jiang et al. (2023) used ensemble learning and their results showed that ML based equating outperformed Mean, Linear and Equipercentile equating methods both in simulation and empirical studies (educational assessment). Tsutsumi et al. (2021) successfully combined deep learning and IRT for data harmonization of both simulated and actual datasets.

## 1.1 Existing research on CBCL and SDQ data harmonization

In the past few years, there were a few interesting research studies that aimed to harmonize CBCL and SDQ data and they used different methodologies for data harmonization. Stevens et al. (2021)

harmonized CBCL and SDQ total scores on a sample of 284 high-risk youth in a residential care facility. They used equipercentile equating. Mansolf et al. (2022) harmonized Internalizing, Externalizing, and Total Problems domains separately on a sample of 1,500 participants from general population between 2 and 17 years old. They used both equipercentile equating and IRT. They evaluated the quality of harmonization using the correlation between harmonized and observed scores: these were all higher than 0.82, except for Externalizing in the school-aged samples, which reached a minimum of about 0.75 for females ages 12–17 (Mansolf et al., 2022).

Jović et al. (2022) focused on harmonizing ADHD and anxiety/depression scores obtained by using CBCL and SDQ on a sample of 1,330 participants between 10 and 11.5 years old from Australia. Authors used IRT to harmonize CBCL and SDQ.

In all three research studies, the participants filled in both CBCL and SDQ scales, which is referred to as a common persons or single-group design, which is, according to Dorans (2007), ideal design for test linking.

They all had different samples. Stevens et al. (2021) used a high-risk population, while Mansolf et al. (2022) and Jović et al. (2022) harmonized data on a general population. They used different harmonization methods, equipercentile equating (Stevens et al., 2021) and IRT (Jović et al., 2022), while Mansolf et al. (2022) compared the results of both of those approaches. Also, they all harmonized CBCL and SDQ on a different level of granularity. Stevens et al. (2021) and Mansolf et al. (2022) harmonized externalising, internaling and total scores, while Jović et al. (2022) focused on more specific subscales (Anxiety/Depression and Hyperactivity/attention problems). In sum, it is unknown which harmonization method works the best in the case of CBCL and SDQ while harmonizing hyperactivity/attention problems.

## 1.2 The aim of the research

This study focuses on finding the most accurate approach for harmonizing hyperactivity/attention problem scores obtained by CBCL and SDQ scales. Our aim is to try out different data harmonization methods (that have different levels of complexity, different underlying assumptions and limitations) to see if there is one particular method that works best. We define best in the sense that a method helps us to put SDQ and CBCL on the same scale. For instance, the method should be able to translate a child's score based on SDQ items into a CBCL-like score, so that the child's level of ADHD related problems can be compared to those of its peers that only have CBCL item scores.

As mentioned above, the performance of the harmonization methods largely depends on the conceptual overlap between scales. If they completely overlap, the methods that use only mean and standard deviation or percentiles should be adequate enough. If, at the other hand, scales that we want to harmonize measure completely different constructs, we will need as much information as we can get. In the field of psychopathology it is hard to expect complete overlap between scales. Particularly a construct like ADHD where both attention problems and hyperactivity play a role, and different questionnaires put different emphases on these subdimensions. To see what works best, we will try out different methods of test equating, and compare their performance.

# 2 Methodology

## 2.1 Scales

The CBCL consists of 113 items and operationalizes childhood behavior on eight subscales/dimensions (social withdrawal, somatic complaints, anxiety/depression, social problems, thought problems, attention problems, delinquent behavior, and aggressive behavior; Achenbach et al., 1991; Achenbach and Ruffle, 2000). We used the attention problems subscale that includes both hyperactivity and attention problems.

The SDQ consists of 25 items equally divided across five scales, also called *dimensions* (Emotional, Conduct, Hyperactivity, Peer, and Prosocial problems; Goodman, 1997, 2001) and it is used for children aged 3–16 years. We used the hyperactivity-inattention scale that also includes items related to concentration problems.

## 2.2 Data collection design

Both CBCL and SDQ were administered to the same group of participants in the Raine study (McKnight et al., 2012). Accordingly, the Single-Group Design (Common Persons) was used to harmonize data in this study because we had responses of the same group of participants on both of scales. In the Single-Group Design, different scales that measure the same construct are administered to the same sample of participants. Different scales are filled in by the participants at the same time, so we assume that there were no changes in the measured construct that can affect the scores on different scales.

## 2.3 Sample

The Raine study is a prospective cohort of children that begun in 1989 and included 2,900 randomly assigned pregnant women who attended the public antenatal clinic at King Edward Memorial Hospital (KEMH; Perth, Western Australia) and nearby private clinics between May 1989 and November 1991 (Newnham et al., 1993; Chivers et al., 2010; Howard et al., 2011; McKnight et al., 2012). Those women completed questionnaires at 18 and 34 weeks of gestation, and follow-up investigations took place at birth, and at 1, 2, 3, 5, 8, 10, 14, 17, 18, and 20 years (Howard et al., 2011; McKnight et al., 2012). The study had two main aims: to investigate the hypothesis that complications of pregnancy might be prevented by frequent ultrasound scans and to develop a long-term cohort to study the role that early life events have on later health (McKnight et al., 2012). The subset of the dataset that we used for this study consists of both the CBCL and SDQ parent-filled questionnaires of 2,861 children ('Generation 2') aged between 10 and 11.5 years (1,417 girls, 1,444 boys). The 1991 Aseba version for the CBCL (age 4–18) by Achenbach (1991) and the 1997 SDQ version by Goodman (1997) were used. In the CBCL, the item scores consisted of either 0 – not true, 1 - omewhat/sometimes true, or 2 - very true/often true. In the SDQ, the responses are 0 – not true. 1 – somewhat true and 2 – certainly true. CBCL and SDQ data were collected at the same time. In this research, we used a subsample of 1,551 children whose mother provided responses on all attention problems/hyperactivity items (complete cases only). Which means only participants with complete answers on

all 11 CBCL and 5 SDQ attention problems/hyperactivity items were included in the analysis.

## 2.4 Data harmonization methodology

We harmonized data by using linear equating, kernel equating, IRT and various ML based methods and compared the quality of harmonization results. But first it was important to decide which scale to use as the target scale. In the case of harmonizing the SDQ and the CBCL scale scores, there are three options: either (1) we leave the SDQ scores as they are and transform the CBCL scores in such a way they can be interpreted as SDQ items, (2) we leave the CBCL scores as they are and transform the SDQ scores to CBCL scores, or (3) we define a new scale, and we translate both SDQ and CBCL scores to that new scale. Van den Berg et al. (2014) used the option to create a new scale, but for practitioners it seems more logical to choose either the SDQ scale or the CBCL scale as the target, as these scales are already familiar to them. But which scale should be chosen as the target scale? For harmonization in daily practice, it is important to keep as much of the original information as possible. When more cases with only SDQ scores are present than children with only CBCL scores, it makes sense to leave the SDQ data as they are and find a way to transform the CBCL scores into SDQ scores. However, any large differences in the reliability of the scores should also be considered (Mansolf et al., 2022). When the CBCL scale scores are substantially more reliable than the SDQ scores, there should be a preference to leave the CBCL scores intact and find a way to translate SDQ scores into CBCL scores. With the present RAINE data set, all children had both SDQ and CBCL scores, so relative number of cases was not a consideration. We found however that the CBCL scores we slightly more reliable than the SDQ items (based on Guttman's Lambda-2, see results), so we devised models to transform SDQ scores into CBCL scores.

For all methods we applied the same logic: we constructed a function or model that determines how to translate one scale score (SDQ) to an equivalent score on the other scale (CBCL). In all methods, the model was constructed based on a training set: one subset of the data based on a random selection of children. To check the effectiveness of each model, the model was applied to the remaining children using only the SDQ data as if the CBCL data were missing, predicting the CBCL score, and comparing it with the actual observed score.

### 2.4.1 Equating based on distributions only

The most common traditional (non-IRT) methods are mean, linear or equipercentile equating. Those methods are focused on the test level scores and they are described in detail by . In mean equating, the scores on test B are transformed such that the transformed scores have the same mean as the scores on test A. Linear equating takes into account not only the means but also the standard deviations. A linear function is estimated that translates the scores on test A such that they have a comparable mean and standard deviation as the scores on test B. We used the 'equate' package from R to conduct linear equating.

When not only the means and variances of two scales are different (the first two moments), but the whole shape of the distribution looks different, it is necessary to also make the higher moments equal (i.e., skewness, kurtosis). For that we can use equipercentile equating where, after a nonlinear transformation, the scores on tests A and B have equal percentile ranks.

### 2.4.2 Equating exploiting the single group design

Kernel equating is a more elaborate method to make the distribution of one score more like the one for another score. It also includes the possibility to use more information that is available when the scores are coming from the same individuals. In kernel equating, the scores are first converted from discrete to continuous using for example a Gaussian kernel distribution (von Davier et al., 2006; Liu and Low, 2008; Arikan and Gelbal, 2018). Kernel equating can be used in such a way that one exploits the single group design: the information of what CBCL scores go together with which SDQ scores in the same children. We used the 'kequate' package with the single group option (Andersson et al., 2013).

### 2.4.3 Item response theory (IRT) and other model based approaches

The IRT approach uses the responses to the individual items, rather than the total scores. A model is used that links a participant's response to an item to both the participant's trait level and the item parameters of that particular item (Embretson and Reise, 2000). One commonly used model is the Generalized Partial Credit Model (GPCM; e.g., van den Berg et al., 2014; Jović et al., 2022). This model contains one discrimination and several threshold parameters for each item (Embretson and Reise, 2000). The discrimination parameter represents the capability of an item to differentiate among respondents with similar trait levels (Embretson and Reise, 2000). It is conceptually similar to a factor loading in factor analysis (van den Berg et al., 2007). The threshold parameters are defined as the point on the latent trait continuum where the response probability for two adjacent response categories is equal (Wetzel and Carstensen, 2014). Accordingly, for a 3-point scale, we have two threshold parameters, between categories 1 and 2 and between categories 2 and 3 (Uto and Ueno, 2018). For the IRT approach we used the mirt package (Chalmers, 2012) to estimate the discrimination and threshold item parameters of the GPCM. The IRT harmonization approaches focuses mainly on the items: based on the SDQ item scores, an estimate is made of an individual's latent trait level, after which this latent trait estimate is used to make a prediction of the total score on the CBCL, conditional on the CBCL item parameters.

Apart from IRT, several other models were tried that are not traditionally using in test equating: linear and ordinal regression, support vector machines (SVM; Hearst et al., 1998; Noble, 2006; Awad and Khanna, 2015 and random forest).

Regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables and it plays a fundamental role in statistical modelling. It is widely used in a form of linear regression where dependent variable is continuous. There is also variant of regression for predicting responses on a categorical scale, ordinal regression. Ordinal regression objective is to classify patterns using a categorical scale which shows a natural order between the labels, and in the case when the scale is ordinal, the ordering consideration improves the performance in comparison to their nominal equivalents (Gutiérrez et al., 2015). You can find more details about ordinal regression and underlying formulas in (Tutz, 2022).

A support vector machine (SVM) is a computer algorithm that learns by example to assign labels to objects. In general, a SVM is an algorithm for maximizing a particular mathematical function with respect to a given collection of data (Noble, 2006).

For more details and underlying mathematical formulas behind SVM, check Noble (2006), Hearst et al. (1998), and Awad and Khanna (2015).

Random forest (RF) is a supervised learning algorithm that combines the output of multiple randomized decision/regression trees to reach a single result by averaging them (Biau and Scornet, 2016). Random Forests can be used for either a categorical response variable as "classification" or a continuous response, referred to as "regression" (Cutler et al., 2012; Qi, 2012).

Within these methods, there are several options on how to use them. The IRT approach tis fully focused on item level data, where the information on the SDQ items is used to make a prediction of the SDQ sum score through the latent trait level and the item parameters. In contrast, for the other methods we can choose whether to work with the raw item data or with the total scores, for both the target scale (CBCL) as the original scale (SDQ). We harmonized data in three different ways (these methods are illustrated in Figure 1):

a) Using the SDQ sum score to predict CBCL sum score (sum to sum).
b) Using SDQ item responses to predict the CBCL sum score (items to sum).
c) Using SDQ item responses to predict CBCL item responses and subsequently summing the predicted item responses (items to items).

Note that sum to sum prediction was only realistic in the case of linear and ordinal regression, but was not sensible in the case of Random Forest and Support Vector Machine since you then have only one predictor variable.

Apart from the choice whether to work with items or total scores, there is also the choice regarding the measurement level of the target variable. Regression approaches in ML regard the target as having interval measurement level, whereas classification approaches regard the target variable as a categorical variable (nominal measurement level). In between is the option to regard the target as ordinal. We therefore applied a linear regression (using the 'lm' function from the R stats package) and compared it to an ordinal regression, using the 'clm' function for ordinal regression model (R package 'ordinal'; Christensen, 2018). For the SVM we used both the regression and the classification version with the package 'e1071' (Dimitriadou et al., 2009). We used the 'svm' function both for classification (type = 'C-classification', kernel = 'linear') and regression (type = 'eps-regression', kernel = 'linear'). For random forest ordinal classification, we used the 'ordfor' function ('ordinalForest'; Hornung, 2020) and for random forest regression the 'randomForest' function ('randomForest'; Liaw and Wiener, 2002).

## 2.5 Evaluating the quality of harmonization

We evaluated the quality of harmonization by comparing the scores as predicted by the models (i.e., the harmonized scores) to the observed (true) ones by computing the root mean squared error (RMSE). First, we calculated the difference between the observed and predicted scores by certain method for every participant, squared them and summed them for all participants (or data points). After that, we divided the sum with the number of data points in order to get the mean value and calculated the square root of the mean value to get RMSE.

A small RMSE represents small differences between observed and predicted scores, and therefore high-quality harmonization. To avoid overfitting and to get a realistic idea of how well the methods would work in practice, we used 5-fold cross-validation. We randomly divided our sample (1,551 participant with complete responses on all 11 CBCL and all 5 SDQ items; no missing data) into 5 subsets (folds). We used 4 folds as a training set to estimate the model, and one-fold as a test set (80% training, 20% test), predicting the CBCL data on the basis of the SDQ data. Every fold was used once as the test set. We used the RMSEs across the five folds to construct boxplots. Next to these RMSE boxplots, we used scatterplots of observed and predicted scores to further illustrate differences between the methods.

# 3 Results

The CBCL scale had a slightly higher reliability in comparison with SDQ scale (0.82 vs. 0.80). Consequently, we decided to use the SDQ scale score as the predictor and CBCL scale as the criterion.

The RMSEs associated with the various methods are presented in Figure 2 and Table 1. The various methods showed a large variation in performance. The overall worst performance was seen in the ordinal regression with the sum score as the predictor. The other methods that used only the SDQ sum score as predictor were also relatively poor, compared to the methods that used individual SDQ items as predictors. Overall, the items to sum options (in green) performed better than the items to items options (in red), except for linear regression and random forest regression where they showed comparable success. Overall, it seems best to use individual SDQ items to predict the CBCL scale score directly. Another pattern is that the regression approaches perform better than the classification/ordinal approaches, that is, regarding the output as interval measurement level rather than ordinal or nominal.

Generally, we see that the IRT method, the linear regression and the random forest regression showed the best results, with very similar RMSEs.

Kernel equating performed better than linear equating, which is to be expected since it exploits the single group design whereas linear equating only uses the means and standard deviations of the SDQ and CBCL score distributions.

Figure 3 shows the relationship between predicted and observed scores for the linear and kernel equating, IRT and ML regression methods. For clarity, the line with intercept 0 and slope 1 is drawn where the dots should be in case of perfect prediction. In order to avoid presenting the mixture of 5 subsets (folds) in the same graph, we used the results from 1 fold as an example.

First thing that we can clearly see is that harmonization of SDQ scores close to 0 is more precise than harmonization of high scores (above 10). For scores close to 0, the predictions are relatively good, with only some overestimation. Scores above 10 are generally far

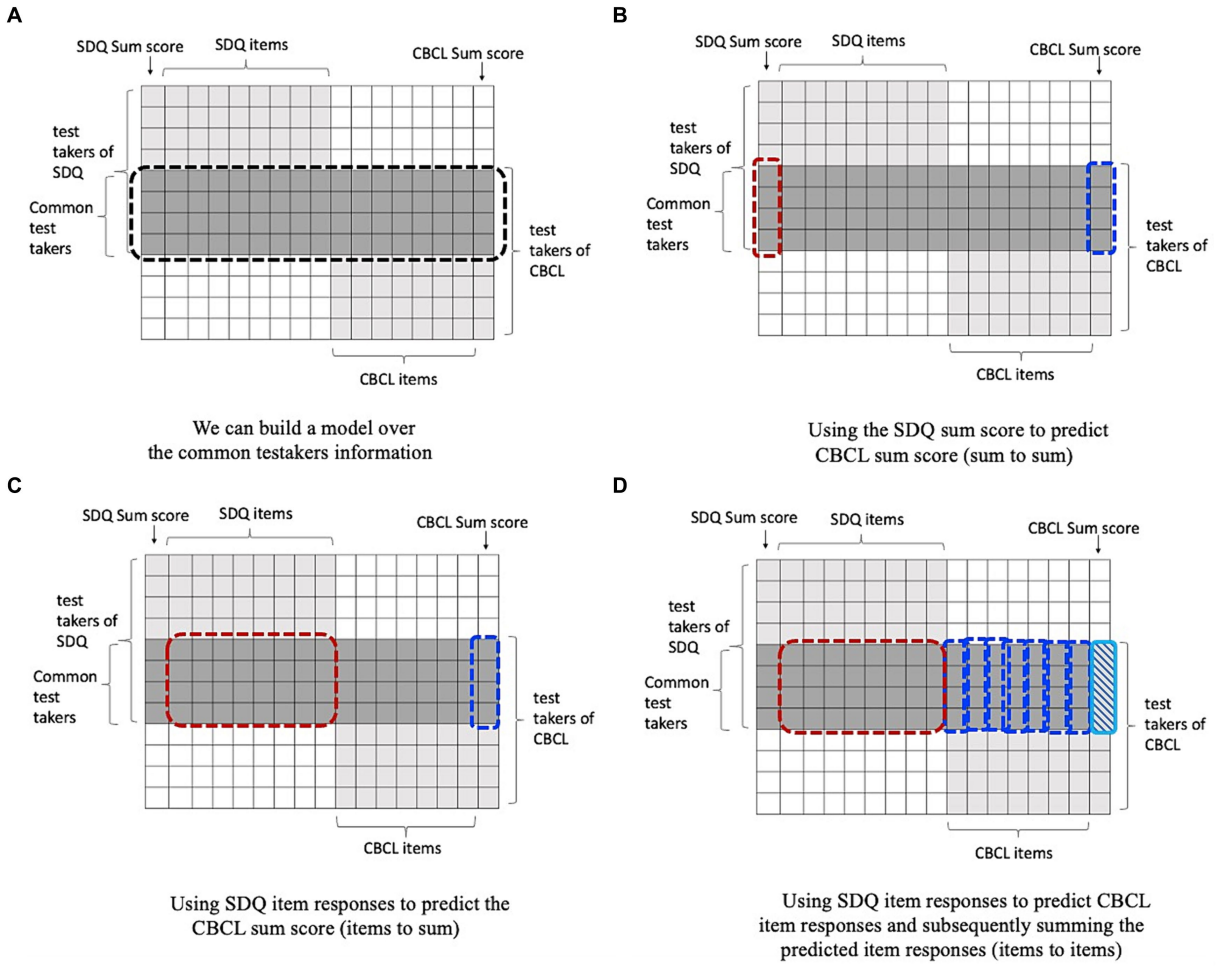



FIGURE 1
Illustration of using common persons design for harmonization. **(A)** shows the common person harmonization problem. We have two tests SDQ and CBCL; for the common person, we have the responses for both sets of items. Thus, we can train and validate a model for predicting the outcomes of one test from the other one over the information of common persons using three different ways. **(B)** shows the sum to sum method for harmonization using machine learning. **(C)** demonstrates the items to sum approach, and **(D)** depicts item to items approach.

from the perfect line and generally underestimated. It seems particularly hard to harmonize the relatively high scores. It is worth mentioning that in the case where majority of responses are skewed towards the lower side of the scale lead to underestimation of high scores which also affects RMSE values. That is, methods that underestimate the high scores more are expected to have higher RMSE.

## 4 Discussion

The aim of this paper was to harmonize ADHD scores measured by CBCL and SDQ scales. We used different data harmonization methods with different levels of complexity, different underlying theories and limitations. We compared the quality of harmonization obtained by linear and kernel equating, IRT and three different machine learning methods (regression, random Forest and SVM) by using both regression and classification approaches. The methods showed a large variety in performance. The best performing models were based on SDQ items rather than SDQ sum scores, and treated the outcome as interval measurement level (referred to as a regression approach in machine learning), rather than ordinal or nominal (classification). The IRT method, the random forest regression and the linear regression based on items showed the best overall performance in terms of RMSE.

Looking more closely at these three methods, the random forest regression and the linear regression showed very comparable patterns in the scatterplot of observed and predicted scores. The pattern was slightly different in the IRT approach with more overestimation for the low scores and less underestimation but more variability for the higher scores. It seems that the bias for the IRT model is less, but that the variance of the predictions is larger.

For all methods there was a bias in that low scores were overestimated and high scores were underestimated. This is due to three causes: impossibility of scores less than 0, regression towards the mean and lack of information on the high score end of the scale (sparsity). For instance, in the IRT approach and in the classification approaches, it is simply impossible to predict a sum score less than

10.3389/fpsyg.2024.1345406

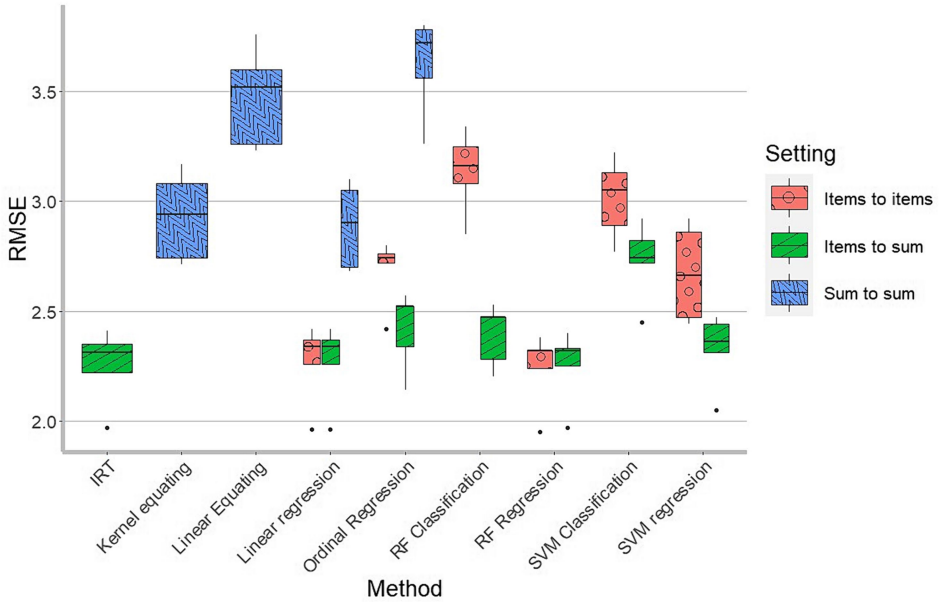



FIGURE 2
Boxplot of RMSEs in 5-fold cross-validation, as a function of harmonization method. Per method we see the spread of the RMSEs in the 5-fold cross-validation. There are in total 17 boxplots. 6 for items to items setting (Linear regression, Ordinal regression, RF classification, RF regression, SVM classification and SVM regression), 7 for items to sum setting (IRT, Linear regression, Ordinal regression, RF classification, RF regression, SVM classification and SVM regression) and 4 for sum to sum setting (Linear equating, Kernel equating, Linear regression and Ordinal regression).

0, so all misclassifications of CBCL scores of 0 are due to overestimation. In the linear regression approach, there is a natural regression to the mean since the correlation between a weighted sum of SDQ item scores and CBCL sum scores is never 100%. Because the relationship is forced to be linear, there will be overestimation on the low end of the scale and underestimation on the high end. An increase in sample size will never fix this problem. On the other hand, in the random forest approach, a nonlinear relation between the items and the outcome is possible. With increasing data on the higher end of the scale it is theoretically quite possible to get better results. In this study the number of children with high scores on attention problems/hyperactivity were relatively scarce since the sample was from the general population. Future research should look deeper into the relationship between sample size, sparsity, bias and variance for the IRT, linear regression, and random forest methods.

Machine learning methods performed better than linear and kernel equating in all the cases except for ordinal regression in sum to sum setting. The lower quality of harmonization in the case of ordinal regression in sum to sum setting could possibly be explained by the fact that we use limited amount of information (only one predictor, SDQ sum score) as in the linear and kernel equating, but on the top of it predicted scores are rounded to be on an ordinal scale in the case of ordinal regression. In that way, by rounding the scores we lose some information which is not the case with linear and kernel equating which scores are not necessarily round numbers. In the case when we use same small amount of information in different methods (linear and kernel equating and ordinal regression in sum to sum setting), losing information due to rounding the scores could make the difference in favour of method which does not round the scores and keeps more information (linear and kernel equating). That is something that would be good to pay attention to and investigate further in the future research. In all other cases, machine learning methods performed better than linear and kernel equating and were very close to IRT. That is in accordance to the findings of the previous studies. For example, Tsutsumi et al. (2021) showed that machine learning can be used for successful data harmonization (in their study they combined it with IRT), while in the study of Jiang et al. (2023) machine learning methods outperformed mean, linear and equipercentile equating. Machine learning methods are more advanced, more complex and take into account more information than the linear and kernel equating methods that are often used in the educational assessment field (i.e., item level data). Our results confirmed that machine learning has strong potential.

Because the best performing methods used item level data, it is not straightforward to construct crosswalk tables that provides researchers the information what SDQ score is equivalent with what CBCL score. Although quick and easy, we do recommend to instead use SDQ item level information to predict the equivalent CBCL, as that yields more reliable results.

We used a harmonization approach where we attempted to find a function that transforms an SDQ score into a CBCL-like score, in such a way that all scores can now be interpreted as CBCL scores. In practice that would mean that if you have several groups of children that were assessed using the CBCL and several groups of children that were assessed using the SDQ, you can keep the original CBCL scores, and only have to transform the data from the children with SDQ data. In this context, it is important to mention that we had the same sample sizes for CBCL and SDQ data, so we decided which scale to use as a predictor based on scale reliability (by using more reliable scale as a criterion to keep as much information as possible at the end). But in the case if samples are not equal and if we have larger sample size for,

TABLE 1 Summary of the results for different methods and settings.

| Method | Setting | Median RMSE (lower the better) | Mean RMSE | SD RMSE |
|---|---|---|---|---|
| IRT | Items to sum | 2.31 | 2.25 | 0.17 |
| Linear equating | Sum to sum | 3.52 | 3.47 | 0.23 |
| Kernel equating | Sum to sum | 2.94 | 2.93 | 0.20 |
| Linear regression | Items to items | 2.34 | 2.27 | 0.18 |
| | Items to sum | 2.34 | 2.27 | 0.18 |
| | Sum to sum | 2.9 | 2.89 | 0.19 |
| Ordinal regression | Items to items | 2.74 | 2.69 | 0.15 |
| | Items to sum | 2.52 | 2.42 | 0.18 |
| | Sum to sum | 3.72 | 3.62 | 0.22 |
| Random forest regression | Items to items | 2.32 | 2.24 | 0.17 |
| | Items to sum | 2.32 | 2.25 | 0.17 |
| Random forest classification | Items to items | 3.16 | 3.14 | 0.19 |
| | Items to sum | 2.47 | 2.39 | 0.14 |
| SVM regression | Items to items | 2.66 | 2.67 | 0.22 |
| | Items to sum | 2.36 | 2.33 | 0.17 |
| SVM classification | Items to items | 3.05 | 3.01 | 0.18 |
| | Items to sum | 2.74 | 2.73 | 0.18 |

Methods with the best results (the lowest RMSE) are highlighted.

for example, SDQ than for CBCL, then it would make more sense to use the scale with larger sample size as a criterion in order to keep more of the original information, especially in the case if scale reliability of two scales is very similar.

Our results showed that an approach based on the raw SDQ item scores works better than using the sum score only. Moreover, based on previous research we knew that attention problems/hyperactivity scores from CBCL and SDQ scales can be harmonized successfully by using IRT (Jović et al., 2022) in the sense that the IRT unidimensional model fitted reasonably well. In this research paper we showed that for attention problems/hyperactivity the IRT approach gave the best results: it is better at predicting what the CBCL score would look like than many of the other methods. This is surprising given the fact that the CBCL attention problems subscale and the SDQ hyperactivity subscale do not fully match content-wise. You would expect that the constructs assessed with these two scales overlap, but not 100%. Both measure ADHD in a broader sense, but with different emphasis to hyperactivity and attentional problems. SDQ is focused on hyperactivity, while CBCL is more focused on attention problems. They are correlated ($r = 0.43$), but they do not completely overlap conceptually. In that case, one would expect the IRT model that we applied here would not be ideal as the IRT model assumed there is only one dimension underlying all SDQ and CBCL items. But here we saw that the approach works is quite robust to model violations in that the other methods performed similarly or worse.

Stevens et al. (2021) successfully harmonized total SDQ and CBCL scores by using equipercentile equating. Mansolf et al. (2022) also harmonized scales on a more general level (Internalizing, Externalizing, Total problems). They also used a single group design and the results showed that both IRT and kernel equating led to similar quality of harmonization, which was quantified as a high correlation between predicted and observed scores. It should be noted Stevens et al. (2021) and Mansolf et al. (2022) did not use a cross-validation approach (at least not within age groups) so that it is unsure to what extent there was model overfit.

We zoomed in on the more specific subscales hyperactivity and attention problems where the conceptual overlap is less precise, the scale reliabilities are lower, and consequently, the correlation between the observed scores is lower. This inevitably results in lower quality of harmonization. The correlation between observed and predicted scores was between 0,43 and 0,73, depending on the method. The unidimensionality assumption is particularly pertinent in the IRT approach we took. It would therefore not be strange to find that one of the machine learning methods would perform better as these do not have this unidimensionality assumption. In our study, machine learning methods (especially with a regression approach) led to much higher quality of harmonization than linear and kernel equating but were comparable to IRT. Potential explanation is in the similarity of the underlying mechanisms of these methods. Namely, in the regression approach, every predictor (item) is weighted and contributes to the prediction of criterion in different degree, while in the IRT approach we have discrimination parameters that are doing the same by referring to the strength of the relationship between trait level (criterion) and participants` responses on the item (predictor). There is a strong case to believe that the IRT approach can often outperform linear regression. The IRT approach takes advantage of both linear and non-linear transformations. In the first part of IRT, there is nonlinear, more or less S-shaped, translation of item scores to latent trait level while in the next step the estimated latent trait values are translated to sum score using again a nonlinear S-shaped form (depends on the exact IRT model that was used). In contrast, with linear regression there is only a strictly linear relationship between the weighted SDQ sum score and the unweighted CBCL score. We expect IRT to outperform linear regression in cases where the threshold item parameters are very different across scales. Overall, we expect that with less sparsity in the top end of the scale, the random forest approach can outperform the IRT and the linear regression approaches.

At the end, it is worth mentioning that even though the quality of harmonization conducted by linear and kernel equating methods was not very good, there are situations in which those methods are only possible methods for data harmonization. That is the case if we do not have the full data obtained by the Single Group design (Common persons) but we only have summary statistics. In addition, in the context of quality of harmonization and machine learning it is important to mention that even though our results showed that regression approach gives better results in comparison to classification, treating categorical variable as continuous and using regression approach may not always be possible.

Concluding, when harmonizing data, different methods should be tested for a particular application, making use of cross-validation to avoid overfitting. Whenever data is available from the same individuals, one should make use of either IRT or regression based machine learning methods (in the case if regression based approach is suitable) that use the items as predictors.

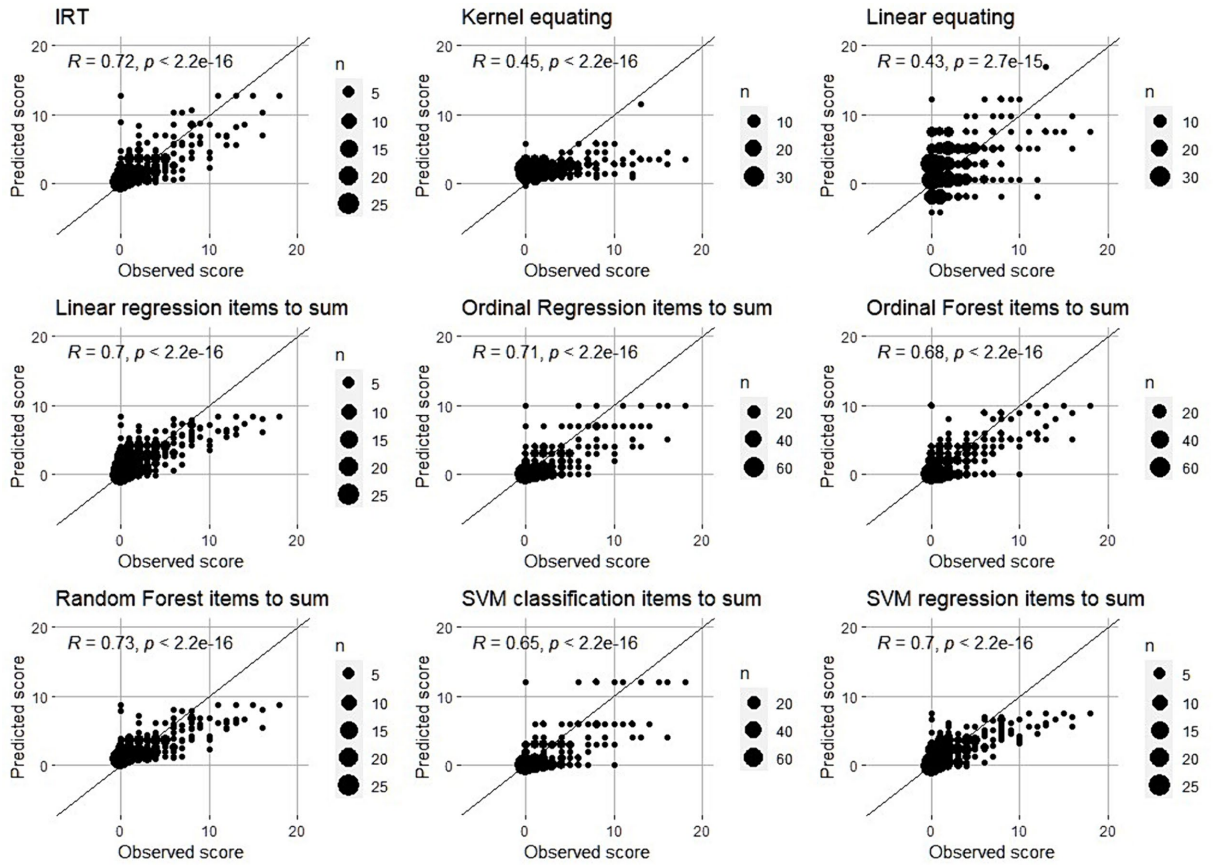



FIGURE 3
Observed vs. Predicted scores – Scatter plots. Black line is the line with intercept 0 and slope 1. In the case of perfect quality of harmonization, all points should be on that line which would mean that predicted and observed scores are exactly the same.

## Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Data is not publicly available because it is owned by Raine Study and permission for getting and analyzing it should be obtained from them. Requests to access these datasets should be directed to andrew.whitehouse@telethonkids.org.au.

## Ethics statement

The studies involving humans were approved by the broader Raine Study and have ethics approval from The University of Western Australia Human Research Ethics Committee. 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

MJ: Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing. MH: Conceptualization, Methodology, Supervision, Writing – review & editing. AW: Resources, Writing – review & editing. SB: Conceptualization, Methodology, Supervision, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## 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


Achenbach, T. M. (1991). Manual for the child behavior checklist/4–18 and 1991 profile. Burlington, VT: University of Vermont.

Achenbach, T. M., Howell, C. T., Quay, H. C., Conners, C. K., and Bates, J. E. (1991). National survey of problems and competencies among four- to sixteen-year-olds: parents' reports for normative and clinical samples. *Monogr. Soc. Res. Child Dev.* 56, i–130. doi: 10.2307/1166156

Achenbach, T. M., and Ruffle, T. M. (2000). The child behavior checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatr. Rev.* 21, 265–271. doi: 10.1542/pir.21.8.265

Algorta, G. P., Dodd, A. L., Stringaris, A., and Youngstrom, E. A. (2016). Diagnostic efficiency of the SDQ for parents to identify ADHD in the UK: a ROC analysis. *Eur. Child Adolesc. Psychiatry* 25, 949–957. doi: 10.1007/s00787-015-0815-0

Allen, K., and Prior, M. (1995). Assessment of the validity of easy and difficult temperament through observed mother-child behaviours. *Int. J. Behav. Dev.* 18, 609–630. doi: 10.1177/016502549501800403

Andersson, B., Branberg, K., and Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *J. Stat. Softw.* 55, 1–25. doi: 10.18637/jss.v055.i06

Arikan, Ç. A., and Gelbal, S. (2018). A comparison of traditional and kernel equating methods. *Int. J. Assess. Tools Educ.* 5, 417–427. doi: 10.21449/ijate.409826

Awad, M., and Khanna, R. (2015). "Support vector machines for classification" in *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Eds. M. Awad and R. Khanna (Berkeley, CA: Apress), 39–66.

Biau, G., and Scornet, E. (2016). A random forest guided tour. *TEST* 25, 197–227. doi: 10.1007/s11749-016-0481-7

Caspi, A., Henry, B., McGee, R. O., Moffitt, T. E., and Silva, P. A. (1995). Temperamental origins of child and adolescent behavior problems: from age three to age fifteen. *Child Dev.* 66, 55–68. doi: 10.2307/1131190

Chalmers, R. P. (2012). Mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48, 1–29. doi: 10.18637/jss.v048.i06

Chen, W. J., Faraone, S. V., Biederman, J., and Tsuang, M. T. (1994). Diagnostic accuracy of the child behavior checklist scales for attention-deficit hyperactivity disorder: a receiver-operating characteristic analysis. *J. Consult. Clin. Psychol.* 62, 1017–1025. doi: 10.1037/0022-006X.62.5.1017

Chivers, P., Hands, B., Parker, H., Bulsara, M., Beilin, L. J., Kendall, G. E., et al. (2010). Body mass index, adiposity rebound and early feeding in a longitudinal cohort (Raine study). *Int. J. Obes.* 34, 1169–1176. doi: 10.1038/ijo.2010.61

Christensen, R. H. B. (2018). Cumulative link models for ordinal regression with the R package ordinal. *Submitted J. Stat. Software* 35, 1–46.

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random forests. Ensemble machine learning: Methods and applications. Eds. C. Zhang and Y. Ma (New York, NY: Springer), 157–175.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., and Leisch, M. F. (2009). Package 'e1071'. R software package, Available at: http://cran.rproject.org/web/packages/e1071/index.html

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Qual. Life Res.* 16, 85–94. doi: 10.1007/s11136-006-9155-3

Embretson, S., and Reise, S. (2000). Item response theory for psychologists. Lawrence Eribaum Associates. New York, NY: Psychology Press.

Fortier, I., Burton, P. R., Robson, P. J., Ferretti, V., Little, J., L'Heureux, F., et al. (2010). Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int. J. Epidemiol.* 39, 1383–1393. doi: 10.1093/ije/dyq139

Fortier, I., Doiron, D., Little, J., Ferretti, V., L'Heureux, F., Stolk, R. P., et al. (2011). Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *Int. J. Epidemiol.* 40, 1314–1328. doi: 10.1093/ije/dyr106

Goodman, R. (1997). The strengths and difficulties questionnaire: a research note. *J. Child Psychol. Psychiatry* 38, 581–586. doi: 10.1111/j.1469-7610.1997.tb01545.x

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *J. Am. Acad. Child Adolesc. Psychiatry* 40, 1337–1345. doi: 10.1097/00004583-200111000-00015

Gutiérrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., and Hervas-Martinez, C. (2015). Ordinal regression methods: survey and experimental study. *IEEE Trans. Knowl. Data Eng.* 28, 127–146. doi: 10.1109/TKDE.2015.2457911

Hall, C. L., Guo, B., Valentine, A. Z., Groom, M. J., Daley, D., Sayal, K., et al. (2019). The validity of the strengths and difficulties questionnaire (SDQ) for children with ADHD symptoms. *PLoS One* 14:e0218518. doi: 10.1371/journal.pone.0218518

Hamilton, C. M., Strader, L. C., Pratt, J. G., Maiese, D., Hendershot, T., Kwok, R. K., et al. (2011). The PhenX toolkit: get the most from your measures. *Am. J. Epidemiol.* 174, 253–260. doi: 10.1093/aje/kwr193

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intell. Syst. Appl.* 13, 18–28. doi: 10.1109/5254.708428

Holland, P. W., Dorans, N. J., and Petersen, N. S. (2006). 6 equating test scores. *Handbook Stat.* 26, 169–203. doi: 10.1016/S0169-7161(06)26006-1

Hornung, R. (2020). Ordinal forests. *J. Classif.* 37, 4–17. doi: 10.1007/s00357-018-9302-x

Howard, A. L., Robinson, M., Smith, G. J., Ambrosini, G. L., Piek, J. P., and Oddy, W. H. (2011). ADHD is associated with a "Western" dietary pattern in adolescents. *J. Atten. Disord.* 15, 403–411. doi: 10.1177/1087054710365990

Jabrayilov, R., Emons, W. H., and Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Appl. Psychol. Meas.* 40, 559–572. doi: 10.1177/0146621616664046

Jiang, Z., Han, Y., Zhang, J., Xu, L., Shi, D., Liang, H., et al. (2023). Empirical ensemble equating under the NEAT design inspired by machine learning ideology. *Methodology* 19, 116–132. doi: 10.5964/meth.10371

Jović, M., Agarwal, K., Whitehouse, A., and van den Berg, S. M. (2022). Harmonized phenotypes for anxiety, depression, and attention-deficit hyperactivity disorder (ADHD). *J. Psychopathol. Behav. Assess.* 44, 663–678. doi: 10.1007/s10862-021-09925-9

Kolen, M. J., and Brennan, R. L. (2014). Test equating, scaling, and linking. *2nd* Edn. New York, NY: Springer.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R news* 2, 18–22.

Liu, J., and Low, A. C. (2008). A comparison of the kernel equating method with traditional equating methods using SAT® data. *Journal of Educational Measurement* 45, 309–323. doi: 10.1111/j.1745-3984.2008.00067.x

Mansolf, M., Blackwell, C. K., Cummings, P., Choi, S., and Cella, D. (2022). Linking the child behavior checklist to the strengths and difficulties questionnaire. *Psychol. Assess.* 34, 233–246. doi: 10.1037/pas0001083

McKnight, C. M., Newnham, J. P., Stanley, F. J., Mountain, J. A., Landau, L. I., Beilin, L. J., et al. (2012). Birth of a cohort—the first 20 years of the Raine study. *Med. J. Aust.* 197, 608–610. doi: 10.5694/mja12.10698

Mislevy, R. J. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: Educational Testing Service.

Muris, P., Meesters, C., and van den Berg, F. (2003). The strengths and difficulties questionnaire (SDQ). *Eur. Child Adolesc. Psychiatry* 12, 1–8. doi: 10.1007/s00787-003-0298-2

Newnham, J. P., Evans, S. F., Michael, C. A., Stanley, F. J., and Landau, L. I. (1993). Effects of frequent ultrasound during pregnancy: a randomised controlled trial. *Lancet* 342, 887–891. doi: 10.1016/0140-6736(93)91944-h

Noble, W. S. (2006). What is a support vector machine? *Nat. Biotechnol.* 24, 1565–1567. doi: 10.1038/nbt1206-1565

Ortuno-Sierra, J., Chocarro, E., Fonseca-Pedrero, E., Riba, S. S., and Muñiz, J. (2015). The assessment of emotional and behavioural problems: internal structure of the strengths and difficulties questionnaire. *Int. J. Clin. Health Psychol.* 15, 265–273. doi: 10.1016/j.ijchp.2015.05.005

Qi, Y. (2012). Random forest for bioinformatics. Ensemble machine learning: Methods and applications. Eds. C. Zhang and Y. Ma (New York, NY: Springer), 307–323.

Sansivieri, V., Wiberg, M., and Matteucci, M. (2017). A review of test equating methods with a special focus on IRT-based approaches. *Statistica* 77, 329–352.

Smith-Warner, S. A., Spiegelman, D., Ritz, J., Albanes, D., Beeson, W. L., Bernstein, L., et al. (2006). Methods for pooling results of epidemiologic studies: the pooling project of prospective studies of diet and Cancer. *Am. J. Epidemiol.* 163, 1053–1064. doi: 10.1093/aje/kwj127

Stevens, A. L., Ho, K. Y., Mason, W. A., and Chmelka, M. B. (2021). Using equipercentile equating to link scores of the CBCL and SDQ in residential youth. *Resid. Treat. Child. Youth* 38, 102–113. doi: 10.1080/0886571X.2019.1704670

Thompson, A. (2009). Thinking big: large-scale collaborative research in observational epidemiology. *Eur. J. Epidemiol.* 24, 727–731. doi: 10.1007/s10654-009-9412-1

Tsutsumi, E., Kinoshita, R., and Ueno, M. (2021). Deep item response theory as a novel test theory based on deep learning. *Electronics* 10:1020. doi: 10.3390/electronics10091020

Tutz, G. (2022). Ordinal regression: a review and a taxonomy of models. *Wiley Interdiscip. Rev. Comput. Stat.* 14:e1545. doi: 10.1002/wics.1545

Uto, M., and Ueno, M. (2018). Empirical comparison of item response theory models with rater's parameters. *Heliyon* 4:e00622. doi: 10.1016/j.heliyon.2018.e00622

van den Berg, S. M., De Moor, M. H., McGue, M., Pettersson, E., Terracciano, A., Verweij, K. J., et al. (2014). Harmonization of neuroticism and extraversion phenotypes across inventories and cohorts in the genetics of personality consortium: an application of item response theory. *Behav. Genet.* 44, 295–313. doi: 10.1007/s10519-014-9654-x

van den Berg, S. M., Glas, C. A., and Boomsma, D. I. (2007). Variance decomposition using an IRT measurement model. *Behav. Genet.* 37, 604–616. doi: 10.1007/s10519-007-9156-1

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., and Martin, K. (2006). An evaluation of the kernel equating method: a special study with pseudotests constructed from real test data. *ETS Res. Rep. Series* 2006, i–31. doi: 10.1002/j.2333-8504.2006.tb02008.x

Wetzel, E., and Carstensen, C. H. (2014). Reversed thresholds in partial credit models: a reason for collapsing categories? *Assessment* 21, 765–774. doi: 10.1177/1073191114530775

# Frontiers in Public Health

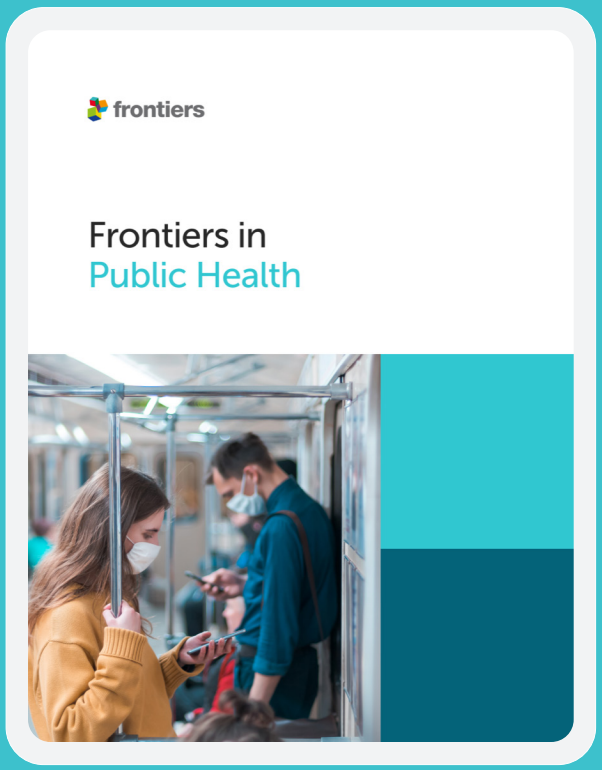
**Explores and addresses today's fast-moving healthcare challenges**

One of the most cited journals in its field, which promotes discussion around inter-sectoral public health challenges spanning health promotion to climate change, transportation, environmental change and even species diversity.

---

## Discover the latest Research Topics

See more →




**Frontiers**

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

**Contact us**

+41 (0)21 510 17 00
frontiersin.org/about/contact