Citation for published version (APA):
Ciharova, M., Amarti, K., Peng, X., van Breda, W., Lorente-Català, R., Funk, B., Hoogendoorn, M., Koutsouleris, N., Fusar-Poli, P., Karyotaki, E., & Cuijpers, P. (in press). Use of Machine-Learning Algorithms Based on Text, Audio and Video Data in the Prediction of Anxiety and Post-Traumatic Stress in General and Clinical Populations: A Systematic Review: Machine-Learning for Anxiety and Post-Traumatic Stress . *Biological psychiatry*.

Use of Machine-Learning Algorithms Based on Text, Audio and Video Data in the Prediction of Anxiety and Post-Traumatic Stress in General and Clinical Populations: A Systematic Review

Marketa Ciharova[1,2,*,†], Khadicha Amarti[1,†], Ward van Breda[3], Xianhua Peng[1,4], Rosa Lorente-Català[5], Burkhardt Funk[6], Mark Hoogendoorn[3], Nikolaos Koutsouleris[7,8], Paolo Fusar-Poli[7], Eirini Karyotaki[1,9], Pim Cuijpers[1,9,10] & Heleen Riper[1,11]

[1] Department of Clinical, Neuro- and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2] Black Dog Institute, University of New South Wales, Sydney, NSW, Australia

[3]Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[4] Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands

[5] Department of Basic and Clinical Psychology and Psychobiology, Universitat Jaume I, Castellon, Spain

[6] Institute of Information Systems, Leuphana University, Lüneburg, Germany

[7] Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neurosciences, King's College London, London, United Kingdom

[8] Max Planck Institute of Psychiatry, Munich, Germany

[9]WHO Collaborating Center for Research and Dissemination of Psychological Interventions, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[10] Babeș-Bolyai University, International Institute for Psychotherapy, Cluj-Napoca, Romania

[11]Amsterdam UMC, Vrije Universiteit Amsterdam, Psychiatry, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

*** Correspondence:** Marketa Ciharova, MSc, Department of Clinical, Neuro- and Developmental Psychology, Vrije Universiteit Amsterdam, Van der Boechorststraat 7-9, 1081 BT, Amsterdam, m.ciharova@vu.nl.

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

Short/running title: Machine-Learning for Anxiety and Post-Traumatic Stress

Keywords: anxiety, post-traumatic stress, machine learning, text, audio, video.

Manuscript length: 3999 words
Number of figures: 6
Number of tables: 4

Abstract

Research in machine-learning (ML) algorithms using natural behavior (i.e., text, audio, and video data) suggests that these techniques could contribute to personalization in psychology and psychiatry. However, a systematic review of the current state-of-the-art is missing. Moreover, individual studies often target ML experts, and may overlook potential clinical implications of their findings. In a narrative accessible to mental health professionals, we present a systematic review, conducted in 5 psychology and 2 computer-science databases. We included 128 studies assessing the predictive power of ML algorithms using text, audio, and/or video data in the prediction of anxiety and post-traumatic stress (PTSD). Most studies (*n* = 87) aimed at predicting anxiety, the remainder (*n* = 41) focused on PTSD. They were mostly published since 2019, in computer science journals, and tested algorithms using text (*n* = 72), as opposed to audio or video. They focused mainly on general populations (*n* = 92), less on laboratory experiments (*n* = 23) or clinical populations (*n* = 13). Methodological quality varied, as did reported metrics of the predictive power, hampering comparison across studies. Two thirds of studies, focusing on both disorders, reported acceptable to very good predictive power (including high-quality studies only). Results of 33 studies were uninterpretable, mainly due to missing information. Research into ML algorithms using natural behavior is in its infancy, but shows potential to contribute to diagnostics of mental disorders, such as anxiety and PTSD, in the future, if standardization of methods, reporting of results, and research in clinical populations are improved.

Introduction

Recently, medicine has moved towards personalization, meaning selecting an appropriate treatment for given individual based on their characteristics (1). Yet, in the psychotherapy field, there is room for improvement regarding diagnostic accuracy, and indication which treatment will work best for which patient in which situation (2). Simultaneously, research and clinical potential of machine-learning (ML) methods for psychology and psychiatry have grown thanks to theoretical breakthroughs and improved computational capacity (3, 4). These developments led to an increase of using ML models in the prediction of mental disorders, focused on diagnostics, prognosis or treatment outcome (5-7). These models could, if proven reliable, valid, and generalizable, act as a component of decision support systems in clinical practice, assist in the early identification of symptoms of mental disorders for epidemiological or prevention purposes, and provide a step towards personalization in psychotherapy and psychiatry (8).

A range of data sources has been successfully used to predict presence of mental disorders, such as depression (9), anxiety (10) or post-traumatic stress disorder (PTSD; 11). Such data sources may be subjective, e.g., self-reported symptoms or ecological momentary assessment ratings (12), or objective, e.g., socio-demographic characteristics (13), biological data (e.g., blood-based gene-expression biomarkers) (14), or neuroimaging data (e.g., magnetic resonance imaging) (15). Although these developments are promising, more research on the use and clinical applications of ML in the prediction of mental disorders is warranted. Single "markers" are not enough to improve diagnostics and treatment (16). Moreover, different data sources may provide insights into different stages of disorder development, from detection of early symptoms to prediction of full-blown onset (17).

Natural behavior, reflected in text, audio, or video data, is an innovative data source used in the ML prediction, which may be based on, for example, vocabulary of the text the individuals write, characteristics of their speech (e.g., pitch or articulation), facial expressions or bodily movements (18-21). During psychotherapy, such data may be session transcripts, or audio and video recordings of patients during sessions (22). Natural behavior data may be collected from individuals recruited in different settings, such as general populations and communities (23), social media platforms (24), clinics (19) or laboratory experiments (25). They may be gathered actively, meaning the individual is required to take action, e.g., narrate about their experience (26). Data may also be collected passively, where no user involvement is necessary, e.g., gait speed recording in university corridors to predict depressive or anxiety symptoms among students (27). Examples of prediction studies include use of messages sent by patients to their therapist in a digital therapy platform to predict severity of anxiety (28), presence of PTSD recognized based on audio recording of an interview with warzone-exposed veterans, or based on a combination of audio and video of patients admitted to a trauma unit (29, 30). The use of social media data could also assist in the early identification of symptoms of mental disorders, becoming thus a helpful prevention component (8).

Most of the studies on the ML prediction of mental states based on natural behavior data target specialists in artificial intelligence (AI) and are published in computer science journals, which are not often read by mental health professionals (31). Moreover, a systematic overview of the topic is so far missing. Therefore, we aimed to provide a systematic review of studies focusing on ML algorithms based on natural behavior in the prediction of anxiety and PTSD, including the level of achievable translation into clinical practice, and potential research gaps. We did so in a close collaboration between psychologists and ML experts, to ensure comprehension of the purpose and quality of included studies by mental health professionals.

Methods

This study was part of the IT4Anxiety project (INTERREG North-West Europe; 32) connecting research institutions to start-ups (i.e., small and medium-sized enterprises) that develop digital products to improve prevention and treatment of anxiety and PTSD. The overall aim of IT4Anxiety was to help the exchange of expertise between these two sectors. In the long term, it aimed to create a framework under which such innovations may be used in clinical settings. Our second aim within IT4Anxiety was the development and testing of an ML algorithm using audio and video data for stress detection, which will be published elsewhere.

The protocol of this study (https://osf.io/adeqk) refers to the whole project (i.e., recognition of depression, anxiety, and related outcomes). In the current publication, results only related to anxiety disorders and PTSD are presented. The remainder will be presented elsewhere.

Bibliographic databases (PubMed, Embase, APA PsycInfo, Web of Science, Scopus, ACM Digital Library Database and Dblp Computer Science Bibliography) were systematically searched (inception - 1st of January 2023). Studies were eligible if they (1) were written in English, (2) reported results of an original study, and (3) aimed at predicting symptoms of anxiety or PTSD using a ML algorithm based on text, audio or video data. The ML algorithm could be any model (e.g., regression, support vector machines, or neural networks) able to predict the content of one dataset based on knowledge learned from another dataset (with known or unknown presence of mental states).

Titles, abstracts, and full texts of identified studies were screened independently by two reviewers, and a senior researcher was consulted if disagreements arose. Similarly, data were extracted by two reviewers independently. We focused on study and sample characteristics, methodology, indicators of predictive power (for example, F1-measure, accuracy, or precision), and information about study quality. These data were narratively summarized.

Since no standardized quality assessment exists for ML studies, a tool was created specifically for the current study. Its creation was based on a review of existing validated quality assessment instruments, namely the PROBAST Tool for prediction studies (33), Cochrane Risk of Bias Tool 2.0 for randomized controlled trials (34), and ROBINS-I Tool for non-randomized studies (35), and previous reviews of ML prediction studies (36, 37). Study quality was evaluated using five criteria: (1) sample size (min. 100 participants); (2) sample balance (in the categorical prediction, no group of participants could be smaller than 1:10 in comparison to other categories, e.g., a group of anxious participants was not much smaller than a group of non-anxious participants); (3) algorithm validation, meaning that the parameters of the model were first tuned on one (i.e., training) dataset, and then evaluated using a dataset not used for algorithm training (i.e., testing), (4) outcome of the algorithm confirmed using a validated instrument, ensuring that the predicted outcome was indeed the disorder of interest (referred to as “ground truth”; 38), and (5) if an emotion-inducing task was used in an experiment, whether this task was validated in previous research. The included criteria are important to ensure generalizability by avoiding overfitting, meaning modelling which corresponds too closely to the training data, eventually resulting in failing to predict observations in different, previously unseen datasets (39). A study was considered high-quality if at least 4 out of these 5 criteria were met.

Please, see Supplementary materials (SM) 1-3, for the search string for PubMed, additional details about methods (e.g., more information about eligibility criteria), and the full explanation of the quality assessment items, including how these biases influence prediction.

Results

Selection of Studies

For a full overview, see Figure 1. Hundred twenty-eight studies (121 publications; 87 on the prediction of anxiety and 41 on PTSD) were included. Figures 2-6, and Tables 1-4 show summative results of the study characteristics, predictive power, and study quality. Most studies were published in North America and Asia (both *n* = 44), between 2019-2023 (*n* = 96), and targeted computer scientists (*n* = 93). Ninety-two studies (*n* = 62 for predicting anxiety, *n* = 30 for PTSD) recruited participants from general non-mental-health-care-seeking populations. Twenty-three studies were laboratory experiments, which either induced anxiety or stress in the participants, or asked them to mimic these states (anxiety: *n* = 20, PTSD: *n* = 3). Thirteen studies involved clinical, mental-health-care-seeking, populations (*n* = 5 for anxiety, *n* = 8 for PTSD). The SM provide a decision tree for the interpretation of the predictive power of the included studies (SM 4), a list of included studies (SM 5), terms used to describe predictive power of studies (Table S1), study-by-study characteristics (Tables S2 – S5), summative study quality reported by population (Table S6), sponsorship of included studies (Table S7), and a list of excluded studies with reasons (Table S8).

General populations

The aim of the general population studies was to predict anxiety or PTSD in an individual (*n* = 87), ranging from early studies into identification of variables to be fed into an ML model for prediction ("features"), to later-stage research focused on improvement of predictive power of existing models, or map symptoms in a target group (*n* = 9). For example, Solanki and Mittra (40) assessed anxiety in Twitter users with history of relocation (F1-measure: .31, poor predictive power). Seventy-seven studies applied a categorical (case-control) and only 11 studies a continuous outcome (cross-sectional). For example, in a case-control study, Sawalha and colleagues (41) predicted presence or absence of PTSD from transcripts of an interview with a virtual (i.e., artificial) clinician in a mixed sample of veterans and civilians (F1-measure: .75, acceptable predictive power). Wang and Zhao (42) predicted anxiety from posts on social media (Weibo) comparing the predicted severity with the continuous score on the Interaction Anxiousness Scale (*r* = .30, weak relationship; 43).

Most studies (*n* = 67) developed and tested text-based algorithms for which they used various data sources, such as social media posts (*n* = 54), answers to open questions in online surveys on mental health, or transcripts of interviews with participants. In audio and video studies, recordings were made during interviews with participants. Most of all studies collected data themselves (*n* = 64). Some (*n* = 23) applied analysis on existing data created for previous general population research into emotion detection, especially social media data (*n* = 16). For example, Buddhitha and Inkpen (44) applied a secondary analysis on Twitter data originally collected for the 2015 Computational Linguistics and Clinical Psychology Workshop. This dataset included data of Twitter users with self-declared anxiety and PTSD (among other disorders), and neurotypical controls, and the authors distinguished between these groups with good predictive power (F1-measure: .87).

Participants in all studies were adults, and could be social media users (e.g., Twitter or Facebook), veterans or other trauma survivors recruited through non-profit organizations, university students, employees or patients seeking help for a physical medical condition. For example, a text-based algorithm was applied in the study by Almeqren and colleagues (45), who found 955 tweets with hashtags related to COVID-19 pandemic on Twitter. These tweets were labelled as no, mild or moderate/strong anxiety by three independent annotators. These agreed subsequently on the categorization of each tweet. The ML algorithm then predicted this categorization with good predictive power (F1-measure: .87). An audio-based algorithm was assessed in the study by Marmar et al. (29), who predicted whether a warzone-exposed veteran had or did not have PTSD from speech features derived from interviews (good predictive power, accuracy: .89).

Similar predictive power for both anxiety and PTSD were found, as around 60% reported acceptable to very good predictions for both dichotomous and continuous outcomes. A similar pattern was seen in high-quality studies only (*n* = 58, 66%) and studies (*n* = 6, 66%) which validated the algorithm externally (i.e., with a previously unseen dataset). Results of almost a third of all studies (22% among high-quality) could not be interpreted due to missing information on the number of participants or imbalanced samples (quality assessment, item no. 2).

Almost two thirds (63%) were rated as high-quality (at least 4 out of 5 quality criteria met), meaning their methodological approach was robust. Regarding specific items of the quality assessment, around a half of studies also recruited a sufficient sample size and their sample was balanced. Nine out of 10 studies for both anxiety and PTSD applied a validation of the algorithm and compared the result with a comparative measure of anxiety/PTSD (i.e., "ground truth"), regardless of the algorithm investigated. Of these, 39% used a validated measure, such as a diagnostic interview or a self-report instrument. This means that more than half used no or a non-validated ground truth, such as self-declared diagnosis, annotation by reviewers, or hashtags used in posts on social media. Few studies (6%) validated their algorithm externally, meaning using a new dataset. This is an important quality as it increases generalizability (see Methods).

Laboratory experiments

Similarly, to most general population studies, the 23 studies conducting an experiment were interested in feature selection or model improvement for the prediction, mainly of anxiety (*n* = 20). All included studies used audio and/or video data, none used text.

With the exception of 2 studies in children (46, 47), all samples included adults, for example, university students, general populations recruited through community means (such as leaflets). In these studies, the authors induced anxiety symptoms in non-mental-health-seeking general population (*n* = 10), or in individuals who already reported anxiety symptoms, enhancing thus their manifestation (*n* = 5). Some studies used actors to mimic anxiety or PTSD (*n* = 8). For example, Zbancioc and Feraru (48) used an existing Emo-DB database (49), where 10 actors expressed seven emotions (neutral, anxiety/fear, happiness, anger, sadness, boredom and disgust) to predict anxiety with good predictive power (accuracy: .84). Salekin and colleagues (50) recruited a sample of low and high socially anxious college students and asked them to give a 3-minute presentation about their experiences at college. The algorithm predicted the category of social anxiety and the participants' ratings of state anxiety with good (F1-measure: .90) and very good predictive power (F1-measure: .93), respectively.

Three quarters of the experiments predicting anxiety (*n* = 15) reported at least acceptable predictive power, among high-quality studies (*n* = 6), it was only 50%. Interestingly, all the high-quality studies induced anxiety in their participants, as opposed to almost a half of all studies (*n* = 7) using actors mimicking anxiety. It is therefore possible that real conditions may be more difficult to predict than artificially created, perhaps exaggerated conditions. Results of 4 studies, including one which externally validated its algorithm, and all studies predicting PTSD, were not interpretable due to insufficient information about numbers of participants.

Only about a quarter of studies was considered high-quality. The most problematic item was sample size criterion, which was met in only one study (4%), showing that recruiting a sufficient sample size for an experiment may be challenging. The samples were balanced in most cases (95%), as was the algorithm validated (87%). However, only one study validated it externally. Two thirds of studies used a ground truth to see whether anxiety was satisfactorily induced in participants, and about a half of these instruments was validated. The validity of the task criterion was crucial for this group of studies, as it demonstrated that the task undergone by the participants to induce anxiety was shown effective in previous research. However, this criterion was met in less than half of the studies (n = 9).

Clinical populations

Most (n = 12, 92%) clinical population studies primarily focused on feature selection or model improvement. Many of the authors collected their own data, while four studies used existing, sometimes public, datasets. For example, Tavabi and colleagues (51) predicted PTSD with good predictive power (F1-measure: .85) from data collected in previous research using a sample of veterans undergoing a virtual reality exposure therapy (52). Like general populations studies, most clinical population studies predicted anxiety or PTSD in case-control designs (n = 10).

The text data (n = 5) was collected as transcripts of face-to-face therapy sessions or on digital therapy platforms (varying from responses to open-ended questions in the intake questionnaire, to patients' e-mail exchanges with the therapists). Audio (n = 4), video (n = 1) recordings, and their combination (n = 3) were, for example, made during clinical interviews with patients. For instance, combination of audio and video recording of a clinical interview data was used by Schultebraucks and colleagues (30) in a sample of 81 patients one month following admission to an emergency department. They found good predictive power in the prediction of PTSD (F1-measure: .83).

All studies included adults who sought help for anxiety or PTSD complaints. However, only slightly above half (n = 8) reported information on participants, e.g., age (19, 28, 30, 51, 53, 54), gender (19, 28, 30, 51, 53-55), ethnicity (30, 51), marital status (51), education (28, 51), employment status (51), self-reported symptoms (19, 53), type of experienced trauma (30, 51, 55), received care (19) or comorbidity (51).

Overall, 92% of studies reported acceptable to very good prediction (for both case-control and cross-sectional studies). This result did not considerably differ per disorder, whether text, audio, or video algorithm was involved, or in high-quality studies. There was only one study which validated the algorithm externally (with good prediction).

Ten studies (77%) were rated as high-quality. Specific criteria were met by around two thirds (sample size, balanced sample) to (almost) all studies (validation, ground truth). Around 80% (both disorders) used a validated measure as ground truth (namely a diagnostic interview, n = 6, or a self-report measure, n = 4), as opposed to annotation by raters (n = 1) or no reported ground truth (n = 2). However, only one study validated the algorithm externally.

Summary

Current research into natural behavior-based ML algorithms for anxiety and PTSD is of a rather fundamental nature, as the focus is mainly on identification of features for prediction or improvement of predictive power of existing models. Yet, it shows promising results, as most studies reach at least acceptable predictive power. Studies conducted in mental-health-care-seeking populations are much fewer than in other settings, but they report higher predictive power and better methodological quality. Social media is a common source of data for general populations studies. Laboratory experiments struggle to recruit sufficient samples compared to studies in other populations, probably due to design requirements, and often recruit actors rather than general population participants or patients.

Discussion

The current study is the first systematic review of ML algorithms using natural behaviors for the prediction of anxiety and PTSD in a narrative accessible for mental health professionals. Most studies were published recently and focused on algorithms using text data, perhaps thanks to the accessibility of free text on social media and progress in natural language processing. Data collection of experimental or clinical data requires more efforts, costs and often an ethical approval. Studies in clinical populations were fewer than other populations, which is in line with the common practice to test new innovations on non-clinical populations first.

Two thirds of all, but also high-quality studies only, reported acceptable to very good predictive power. It could be expected that methodological rigor would lead to higher predictive power, as with better methodology, the aims of the study are more satisfactorily met. Alternatively, we could expect lower predictive power in high-quality studies, as the results are less likely to be unreliable and inflated (56). If a study was deemed high-quality, its results may still be difficult to interpret due to imbalanced samples. Therefore, it is possible that our quality instrument requires improvement to assess the methodological quality in a more fine-grained manner.

Eighteen studies did not confirm the outcome of the prediction by ground truth, meaning a measure providing evidence of anxiety or PTSD. In the studies which used ground truth, only less than half compared their outcome to "gold standard" comparators, meaning validated diagnostic interviews (e.g., SCID-5, or Mini-International Neuropsychiatric Interview; 57, 58) or self-report measures. Only 8 studies (6%) validated the algorithm using a previously unseen dataset, and only one such study was conducted in clinical settings. Generalizability of the findings is thus limited, especially for conclusions in clinical practice. Our findings thus agree with previous reviews on other data sources suggesting that more methodological rigor is needed in clinical prediction models in psychiatry (6).

Given the lack of previous reviews on natural behavior data, the comparability of our results with previous research is limited. Nevertheless, indications can be derived from a previous review by Ramos-Lima and colleagues (37). This review included 49 studies predicting PTSD and acute stress disorder using ML techniques based on various data sources. It included 5 studies using text or audio data which are part of the current review (29, 59-62). The authors reported identical results as we did, however, did not interpret them further. They mentioned complications with comparison of the results of individual studies to each other, since a wide range of performance metrics was reported, and argue for standardization of the reporting. We encountered similar problems and provided solution in the predictive power interpretation that we created. Higher transparency and standardization in methods and reporting is clearly needed.

Our results need to be interpreted considering their limitations. No standardized quality assessment tool for ML-prediction studies using natural behavior was available, we thus created a critical appraisal instrument. Two previous reviews on predictive modelling using various data sources partly overlapped with ours. Sajjadian and colleagues (36) considered all studies with a sufficient sample size and algorithm validation adequate-quality (both items being part of our quality assessment). Ramos-Lima and colleagues (37) developed a 9-item assessment, of which 5 items (sample balance, ground truth, and algorithm validation, appropriateness of the ML algorithm, reporting of relevant performance metrics) were part of our instrument or result interpretation, but did not include 2 additional items we evaluated (sample size, task validity). Their other aspects (i.e., representativeness of the sample, confounders, description of features, and handling missing data) were also originally considered for our review but could not be evaluated due to insufficient reporting in the included studies. Therefore, we recommend that a multidisciplinary international expert team addresses these quality issues, resulting in a consensus statement. The TRIPOD statement (63), providing reporting guidelines for ML prediction studies, was not mentioned by any of the included studies, but should be followed. Furthermore, no meta-analysis of the predictive power of included studies could be performed, given the substantial heterogeneity among these studies, both across and within the algorithm types. It was also impossible to assess the relationship between the predictive power and study quality. However, the current study created a basis for future investigation.

Natural behavior algorithms show promising results: They report at least acceptable performance in most studies, including those focused on social media data, symptom identification in general populations, or in clinical settings. Nevertheless, the evidence is still in its infancy. Future research should focus on the use and validation of ML algorithms in clinical practice, prediction of treatment outcome, translation into routine care, prevention, and implementation. Guidelines should be developed separately for their use in different populations. Acceptance and trust towards their application in health care must be assessed to address potential reluctance of their adoption (64). Furthermore, it is necessary to explore at which stage of the ML-based diagnostic process the clinician should enter (65, 66), to secure adaptability to changing contexts, and provide ethical supervision (67). How the prediction will be influenced by the boom of generative AI and large language models remains also unclear.

In clinical practice, ML algorithms using natural behavior, when based on quality data and methodology, reliable, and valid, could become a part of decision support tool. They could also reveal everyday manifestation of mental states in real-time and in the ecological habitat of the individual (68), as natural expression or onset of the disorder, complementing thus other, both subjective and objective, data sources. In their optimal form, they may also lead to the discovery of additional objective “markers” of mental disorders, e.g., through digital phenotyping, meaning collecting measurable characteristics of an individual’s digital footprint, such as smartphone usage (69, 70).

In the future, integrating these behavioral markers with other biological and clinical data may enhance diagnostic accuracy and treatment outcomes in psychiatry and psychology. It may provide more comprehensive assessments earlier in time and against lower costs. Furthermore, prevention efforts in general populations, for example through symptom identification on social media, may benefit from the use of ML algorithms. However, first, more evidence from clinical settings is necessary.

Acknowledgments: This project was part of the IT4Anxiety project financed by Interreg North-West Europe (NWE 983), European Union. We would like to thank Caroline Planting for conducting the systematic searches for this study, and Leonie Verstraete for the diligent work on screening of the studies and data extraction.

Disclosures: The authors report no conflicts of interest.

References

1. Yamamoto Y, Kanayama N, Nakayama Y, Matsushima N. Current status, issues and future prospects of personalized medicine for each disease. Journal of Personalized Medicine. 2022;12(3):444.

2. Cuijpers P, Ciharova M, Quero S, Miguel C, Driessen E, Harrer M, et al. The contribution of "individual participant data" meta-analyses of psychotherapies for depression to the development of personalized treatments: a systematic review. Journal of Personalized Medicine. 2022;12(1):93.

3. Thompson NC, Greenewald K, Lee K, Manso GF. The computational limits of deep learning. arXiv preprint arXiv:200705558. 2020.

4. Xu Z, Sun J. Model-driven deep-learning. National Science Review. 2018;5(1):22-4.

5. Thieme A, Belgrave D, Doherty G. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. ACM Transactions on Computer-Human Interaction (TOCHI). 2020;27(5):1-53.

6. Meehan AJ, Lewis SJ, Fazel S, Fusar-Poli P, Steyerberg EW, Stahl D, et al. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. Molecular psychiatry. 2022;27(6):2700-8.

7. Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, Irving J, Catalan A, Oliver D, et al. Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. Schizophrenia bulletin. 2021;47(2):284-97.

8. Kim J, Lee J, Park E, Han J. A deep learning model for detecting mental illness from user content on social media. Scientific reports. 2020;10(1):11846.

9. Kim H, Lee S, Lee S, Hong S, Kang H, Kim N. Depression prediction by using ecological momentary assessment, actiwatch data, and machine learning: observational study on older adults living alone. JMIR mHealth and uHealth. 2019;7(10):e14149.

10. Månsson KN, Frick A, Boraxbekk C-J, Marquand A, Williams S, Carlbring P, et al. Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. Translational psychiatry. 2015;5(3):e530-e.

11. Dabek F, Caban JJ. Leveraging big data to model the likelihood of developing psychological conditions after a concussion. Procedia computer science. 2015;53:265-73.

12. van Breda W, Bremer V, Becker D, Hoogendoorn M, Funk B, Ruwaard J, et al. Predicting therapy success for treatment as usual and blended treatment in the domain of depression. Internet interventions. 2018;12:100-4.

13. Fusar-Poli P, Stringer D, MS Durieux A, Rutigliano G, Bonoldi I, De Micheli A, et al. Clinical-learning versus machine-learning for transdiagnostic prediction of psychosis onset in individuals at-risk. Translational Psychiatry. 2019;9(1):259.

14. Tylee DS, Chandler SD, Nievergelt CM, Liu X, Pazol J, Woelk CH, et al. Blood-based gene-expression biomarkers of post-traumatic stress disorder among deployed marines: a pilot study. Psychoneuroendocrinology. 2015;51:472-94.

15. Liu F, Xie B, Wang Y, Guo W, Fouche J-P, Long Z, et al. Characterization of post-traumatic stress disorder using resting-state fMRI with a multi-level parametric classification approach. Brain topography. 2015;28:221-37.

16. Waszkiewicz N. Mentally sick or not—(Bio) Markers of psychiatric disorders needed. MDPI; 2020. p. 2375.

17. van der Tuin S, Booij S, Muller M, van den Berg D, Oldehinkel A, Wigman J. The added value of daily diary data in 1-and 3-year prediction of psychopathology and psychotic experiences in individuals at risk for psychosis. Psychiatry Research. 2023;329:115546.

18. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope investigative otolaryngology. 2020;5(1):96-116.

19. Wiegersma S, Hidajat M, Schrieken B, Veldkamp B, Olff M. Improving web-based treatment intake for multiple mental and substance use disorders by text mining and machine learning: Algorithm development and validation. JMIR mental health. 2022;9(4):e21111.

20. Giannakakis G, Koujan MR, Roussos A, Marias K. Automatic stress analysis from facial videos based on deep facial action units recognition. Pattern Analysis and Applications. 2022:1-15.
21. Giakoumis D, Drosou A, Cipresso P, Tzovaras D, Hassapis G, Gaggioli A, et al. Using activity-related behavioural features towards more effective automatic stress detection. 2012.
22. Aafjes-van Doorn K, Kamsteeg C, Bate J, Aafjes M. A scoping review of machine learning in psychotherapy research. Psychotherapy Research. 2021;31(1):92-116.
23. Kjell K, Johnsson P, Sikström S. Freely generated word responses analyzed with artificial intelligence predict self-reported symptoms of depression, anxiety, and worry. Frontiers in Psychology. 2021;12:602581.
24. Jiang ZP, Levitan SI, Zomick J, Hirschberg J, editors. Detection of mental health from reddit via deep contextualized representations. Proceedings of the 11th international workshop on health text mining and information analysis; 2020.
25. Gu J, Gao B, Chen Y, Jiang L, Gao Z, Ma X, et al. Wearable social sensing: Content-based processing methodology and implementation. IEEE Sensors Journal. 2017;17(21):7167-76.
26. He Q, Veldkamp BP, de Vries T. Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. Psychiatry research. 2012;198(3):441-7.
27. 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(5):e0216591.
28. Burkhardt H, Pullmann M, Hull T, Areán P, Cohen T, editors. Comparing emotion feature extraction approaches for predicting depression and anxiety. Proceedings of the eighth workshop on computational linguistics and clinical psychology; 2022.
29. Marmar CR, Brown AD, Qian M, Laska E, Siegel C, Li M, et al. Speech-based markers for posttraumatic stress disorder in US veterans. Depression and anxiety. 2019;36(7):607-16.
30. Schultebraucks K, Yadav V, Shalev AY, Bonanno GA, Galatzer-Levy IR. Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. Psychological Medicine. 2022;52(5):957-67.
31. Alam MAU, Kapadia D, editors. Laxary: a trustworthy explainable twitter analysis model for post-traumatic stress disorder assessment. 2020 IEEE International Conference on Smart Computing (SMARTCOMP); 2020: IEEE.
32. INTERREG North-West Europe. IT4Anxiety 2020 [updated 11/09/2023. Available from: https://vb.nweurope.eu/projects/project-search/it4anxiety-managing-anxiety-via-innovative-technologies-for-better-mental-health/.
33. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of internal medicine. 2019;170(1):51-8.
34. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. Bmj. 2011;343.
35. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. bmj. 2016;355.
36. Sajjadian M, Lam RW, Milev R, Rotzinger S, Frey BN, Soares CN, et al. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. Psychological Medicine. 2021;51(16):2742-51.
37. Ramos-Lima LF, Waikamp V, Antonelli-Salgado T, Passos IC, Freitas LHM. The use of machine learning techniques in trauma-related disorders: a systematic review. Journal of psychiatric research. 2020;121:159-72.
38. Lemoigne Y, Caner A. Molecular Imaging: Computer Reconstruction and Practice: Springer Science & Business Media; 2008.
39. Hawkins DM. The problem of overfitting. Journal of chemical information and computer sciences. 2004;44(1):1-12.

40. Solanki M, Mittra Y, editors. Using Twitter to measure the impact of immigration by studying people's mood. 2021 8th International Conference on Behavioral and Social Computing (BESC); 2021: IEEE.
41. Sawalha J, Yousefnezhad M, Shah Z, Brown MR, Greenshaw AJ, Greiner R. Detecting presence of PTSD using sentiment analysis from text data. Frontiers in psychiatry. 2022;12:811392.
42. Wang Y, Zhao N. Prediction model of interaction anxiousness based on Weibo data. Frontiers in Public Health. 2022;10:1045605.
43. Leary MR, Kowalski RM. The interaction anxiousness scale: Construct and criterion-related validity. Journal of personality assessment. 1993;61(1):136-46.
44. Buddhitha P, Inkpen D, editors. Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. Proceedings of the tenth international workshop on health text mining and information analysis (LOUHI 2019); 2019.
45. Almeqren MA, Almuqren L, Alhayan F, Cristea AI, Pennington D. Using deep learning to analyze the psychological effects of COVID-19. Frontiers in Psychology. 2023;14:962854.
46. McGinnis EW, Anderau SP, Hruschak J, Gurchiek RD, Lopez-Duran NL, Fitzgerald K, et al. Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood. IEEE journal of biomedical and health informatics. 2019;23(6):2294-301.
47. Nandyal S, Swathi P, editors. Early Childhood Anxiety and Depression Detection Based on Speech Using Machine Learning Analysis. Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces; 2022: Springer.
48. Zbancioc M-D, Feraru M, editors. Recognizing Fear/Anxiety in Relation to Other Emotions. 2020 International Conference on e-Health and Bioengineering (EHB); 2020: IEEE.
49. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B, editors. A database of German emotional speech. Interspeech; 2005.
50. Salekin A, Eberle JW, Glenn JJ, Teachman BA, Stankovic JA. A weakly supervised learning framework for detecting social anxiety and depression. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies. 2018;2(2):1-26.
51. Tavabi L, Poon A, Rizzo AS, Soleymani M, editors. Computer-based PTSD assessment in VR exposure therapy. HCI International 2020–Late Breaking Papers: Virtual and Augmented Reality: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22; 2020: Springer.
52. Loucks L, Yasinski C, Norrholm SD, Maples-Keller J, Post L, Zwiebach L, et al. You can do that?!: Feasibility of virtual reality exposure therapy in the treatment of PTSD due to military sexual trauma. Journal of anxiety disorders. 2019;61:55-63.
53. Demiris G, Corey Magan KL, Parker Oliver D, Washington KT, Chadwick C, Voigt JD, et al. Spoken words as biomarkers: using machine learning to gain insight into communication as a predictor of anxiety. Journal of the American Medical Informatics Association. 2020;27(6):929-33.
54. Dia M, Khodabandelou G, Othmani A, editors. A novel stochastic transformer-based approach for post-traumatic stress disorder detection using audio recording of clinical interviews. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS); 2023: IEEE.
55. Sawadogo MAL, Pala F, Singh G, Selmi I, Puteaux P, Othmani A. PTSD in the wild: a video database for studying post-traumatic stress disorder recognition in unconstrained environments. Multimedia Tools and Applications. 2023:1-23.
56. Cuijpers P, Harrer M, Miguel C, Ciharova M, Karyotaki E. Five decades of research on psychological treatments of depression: A historical and meta-analytic overview. American psychologist. 2023.
57. First MB, Williams JB, Karg RS, Spitzer RL. SCID-5-CV. Intervista Clinica Strutturata per i Disturbi del DSM-5 Versione Per Il Clinico Ed Italiana a cura Di Andrea Fossati e Serena Borroni: Raffaello Cortina Editore Milano. 2017.

58. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. Journal of clinical psychiatry. 1998;59(20):22-33.
59. He Q, Veldkamp BP, Glas CA, de Vries T. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. Assessment. 2017;24(2):157-72.
60. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Scientific reports. 2017;7(1):13006.
61. Vergyri D, Knoth B, Shriberg E, Mitra V, McLaren M, Ferrer L, et al., editors. Speech-based assessment of PTSD in a military population using diverse feature classes. Sixteenth annual conference of the international speech communication association; 2015: Citeseer.
62. Banerjee D, Islam K, Xue K, Mei G, Xiao L, Zhang G, et al. A deep transfer learning approach for improved post-traumatic stress disorder diagnosis. Knowledge and Information Systems. 2019;60:1693-724.
63. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Annals of internal medicine. 2015;162(1):55-63.
64. Chan EY, Saqib NU. Privacy concerns can explain unwillingness to download and use contact tracing apps when COVID-19 concerns are high. Computers in Human Behavior. 2021;119:106718.
65. Jašović-Gašić M, Dunjic-Kostić B, Pantović M, Cvetić T, P Marić N, A Jovanović A. Algorithms in psychiatry: State of the art. Psychiatria Danubina. 2013;25(3):0-283.
66. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. Psychiatric Quarterly. 2022;93(1):249-53.
67. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annual review of clinical psychology. 2018;14:91-118.
68. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ digital medicine. 2019;2(1):1-11.
69. Torous J, Kiang MV, Lorme J, Onnela J-P. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR mental health. 2016;3(2):e5165.
70. Oudin A, Maatoug R, Bourla A, Ferreri F, Bonnot O, Millet B, et al. Digital phenotyping: Data-driven psychiatry to redefine mental health. Journal of Medical Internet Research. 2023;25:e44502.

Table 1

Summative Results per Disorder

	All studies N = 128		General Populations N = 92		Clinical Populations N = 13		Experiments N = 23	
Total Number of Studies	Anxiety N = 87	PTSD N = 41	Anxiety N = 62	PTSD N = 30	Anxiety N = 5	PTSD N = 8	Anxiety N = 20	PTSD N = 3
Modality								
Text	48	24	44	23	4	1	0	0
Audio	17	11	6	5	0	4	11	2
Video	16	2	8	2	1	0	7	0
Multi-modal (audio + video)	6	4	4	0	0	3	2	1
Data-collection								
Existing dataset	20	15	12	11	1	3	7	1
Data collected by the authors	63	24	47	17	4	5	12	2
Both	4	2	3	2	0	0	1	0
Design								
Case-control study	52	35	48	29	4	6	0	0
Cross-sectional study	11	1	11	0	0	1	0	0
Case-control+cross-sectional	4	2	3	1	1	1	0	0
Experiment	20	3	0	0	0	0	20	3
Continent								
Asia	39	5	30	2	1	1	8	2
Australia	2	0	0	0	0	0	2	0
Europe	23	12	14	8	2	3	7	1
North America	20	24	15	20	2	4	3	0
South America	3	0	3	0	0	0	0	0
Type of journal								
Psychology	11	11	9	9	2	2	0	0
Computer science	66	27	44	18	3	6	19	3
General science	10	3	9	3	0	0	1	0

Table 2

Predictive Power per Type of Predicted Disorder and Algorithm Used – Results per Type of Population

	Case-control studies[a]						Cross-sectional studies[b]				
Predictive power[c]	Very good	Good	Acceptable	Poor	Not interpretable (Insufficient information)	Not interpretable (Imbalanced dataset)	No relationship	Weak relationship	Moderate relationship	Strong relationship	**Total**
General populations											
Text-based											
Anxiety	2	7	15	2	12	1	1	2	2		44
PTSD	2	4	6	1	8	1		1			23
Audio-based											
Anxiety			1	1	1		1	1	1		6
PTSD	1	3	1								5
Video-based											
Anxiety	1	3			1	1			1	1	8
PTSD	2										2
Multimodal											
Anxiety		1			1				1	1	4
PTSD											0
Clinical populations											
Text-based											
Anxiety		1	1					1	1		4
PTSD			1								1
Audio-based											
Anxiety											0
PTSD		1	1		1				1		4
Video-based											
Anxiety			1								1
PTSD											0
Multimodal											
Anxiety											0
PTSD	1	2									3
Experiments											
Text-based											
Anxiety											0
PTSD											0
Audio-based											
Anxiety	3	4	1	1	2						11
PTSD					2						2
Video-based											
Anxiety	2	3	1		1						7
PTSD											0
Multimodal											
Anxiety					1					1	2
PTSD					1						1

Note. The best performance per study is reported.
[a] Case-control studies: Studies predicting the disorder of interest categorically, meaning, for example, presence versus absence of the disorder.
[b] Cross-sectional studies: Studies predicting the disorder of interest continuously, meaning, for example, predicting the severity expressed as a score on a continuous instrument.
[c] Predictive power: See Interpretation of Predictive Power of Included Studies, Supplementary material 4.

Table 3

Predictive Power per Type of Predicted Disorder and Algorithm Used – All studies and High-quality Studies Only

	Case-control studies[a]						Cross-sectional studies[b]				
Predictive power[c]	Very good	Good	Acceptable	Poor	Not interpretable (Insufficient information)	Not interpretable (Imbalanced dataset)	No relationship	Weak relationship	Moderate relationship	Strong relationship	**Total**
All studies (*N* = 128)											
Text-based											
Anxiety	2	8	16	2	12	1	1	3	3		48
PTSD	2	4	7	1	8	1		1			24
Audio-based											
Anxiety	3	4	2	2	3		1	1	1		17
PTSD	2	4	2		2				1		11
Video-based											
Anxiety	3	6	2		2	1			1	1	16
PTSD	2										2
Multimodal											
Anxiety		1			2				1	2	6
PTSD	1	2			1						4
Only high quality studies[d] (*N* = 74)											
Text-based											
Anxiety	2	3	8		6		1	3	3		26
PTSD	1	4	5		5	1		1			17
Audio-based											
Anxiety	1	1	1	2	1		1	1	1		9
PTSD	2	3	1						1		7
Video-based											
Anxiety	1	3			1	1			1	1	8
PTSD	2										2
Multimodal											
Anxiety		1							1	1	3
PTSD	1	1									2

Note. The best performance per study is reported.

[a] Case-control studies: Studies predicting the disorder of interest categorically, meaning, for example, presence versus absence of the disorder.

[b] Cross-sectional studies: Studies predicting the disorder of interest continuously, meaning, for example, predicting the severity expressed as a score on a continuous instrument.

[c] Predictive power: See Interpretation of Results of Included Studies, Supplementary material 4.

[d] Only studies meeting at least 4 out of 5 quality criteria.

Table 4

Quality Assessment per Population

Total Number of Studies	All studies *N* = 128	Text		Audio		Video		Multimodal	
		Anxiety *N* = 48	PTSD *N* = 24	Anxiety *N* = 17	PTSD *N* = 11	Anxiety *N* = 16	PTSD *N* = 2	Anxiety *N* = 6	PTSD *N* = 4
	N (%)	*N* (%)	*N* (%)	*N* (%)	*N* (%)	*N* (%)	*N* (%)	*N* (%)	*N* (%)
Number of criteria met									
4-5 (High-quality)	**74 (58)**	**26 (54)**	**17 (71)**	**9 (52)**	**7 (64)**	**8 (50)**	**2 (100)**	**3 (50)**	**2 (50)**
0-3	54 (42)	22 (46)	7 (29)	8 (48)	4 (36)	8 (50)	0 (0)	3 (50)	2 (50)
Individual items positive									
Sample size	62 (48)	25 (52)	17 (71)	6 (35)	3 (27)	6 (38)	1 (50)	2 (33)	2 (50)
Balanced sample	72 (56)	19 (39)	11 (46)	16 (94)	8 (73)	12 (75)	2 (100)	3 (50)	1 (25)
Validation	117 (91)	44 (92)	23 (96)	15 (88)	10 (91)	16 (100)	2 (100)	4 (66)	3 (75)
Externally validated	8 (6)	1 (2)	1 (4)	2 (12)	0 (0)	1 (8)	1 (50)	1 (17)	1 (25)
Ground truth	110 (87)	42 (88)	23 (96)	13 (76)	8 (73)	13 (81)	2 (100)	5 (83)	4 (100)
Type of ground truth									
Diagnostic interview	15 (12)	1 (2)	4 (17)	0 (0)	4 (37)	2 (13)	1 (50)	0 (0)	3 (75)
Self/observer-reported instrument	46 (36)	13 (27)	4 (17)	11 (64)	3 (27)	9 (57)	1 (50)	5 (83)	0 (0)
Self-declared diagnosis	24 (19)	10 (21)	13 (54)	0 (0)	1 (9)	0 (0)	0 (0)	0 (0)	0 (0)
Annotation by raters	15 (12)	9 (18)	1 (4)	2 (12)	0 (0)	2 (13)	0 (0)	0 (0)	1 (25)
Topic (e.g., subreddit)	10 (8)	9 (18)	1 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Validity of the task	114 (89)	48 (100)	24 (100)	11 (65)	9 (82)	11 (69)	2 (100)	6 (100)	3 (75)

Figure Titles

1. **Figure 1 -** *PRISMA Flow Diagram*
2. **Figure 2 -** *Number of Studies per Predictive Power in the Prediction of Anxiety – General Populations, Clinical Populations, and Experiments (Case-control and Cross-sectional Studies)*
3. **Figure 3 -** *Number of Studies per Predictive Power in the Prediction of PTSD – General Populations, and Clinical Populations (Case-control and Cross-sectional Studies)*