Review

Large Language Models for Mental Health Prediction: Scoping Review of Bias and Clinical Utility Documentation in 2019-2024

Clémentine Bleuze[1], MSc; Karën Fort[1], PhD; Vincent P Martin[1], PhD; Aurélie Névéol[2], PhD

[1]Loria, Université de Lorraine, CNRS, Inria, Nancy, France
[2]LISN, Université Paris-Saclay, CNRS, Orsay, France

Corresponding Author:

Clémentine Bleuze, MSc
Loria
Université de Lorraine, CNRS, Inria
Loria Campus Scientifique, BP 239
Nancy F-54000
France
Phone: 33 3-83-59-20-20
Email: clementine.bleuze@univ-lorraine.fr

Abstract

Background: A growing body of literature leverages large language models (LLMs) to make mental health predictions. However, these models are prone to bias, and studies to validate their clinical utility are lacking.

Objective: This scoping review aims to uncover bias and clinical utility limitations stemming from the methodological design of LLM-based mental health predictive systems. In addition, it intends to document the level of self-reflection about bias and clinical challenges reported by authors in their own work.

Methods: This work follows the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines and was registered online. Eligible studies were original research articles in English published between 2019 and 2024, using LLMs to detect mental health conditions in nonsynthetic textual data. The search was conducted in 5 scientific databases (PubMed, Web of Science, IEEE Xplore, ACM Digital Library, and ACL Anthology) with queries associating keywords related to "Mental Health," "Large Language Models," and "Prediction." We extracted both methodological information about the included studies and authors' statements relevant to issues of bias and clinical utility. This extraction was based on a framework screening the entire pipeline of development of LLMs with applications in mental health: research design and selection, data collection, outcome definition, model development, and postdeployment considerations. Statistical description of the retrieved entities, as well as thematic coding, was performed for analysis.

Results: A total of 2472 articles were identified, of which 263 (10.6%) were assessed for eligibility, and 201 (8.1%) were included in the review. Included studies were mostly recent, indicating a growing interest in the use of LLMs for mental health predictions. Our analysis revealed that a majority of studies share similar methodological choices along their development pipeline: most of them focus on depressive disorders identified via processing user texts on social media, mainly with the use of nonspecialist LLMs derived from BERT (Bidirectional Encoder Representations from Transformers). Following previous works on these matters, we highlighted how these choices may hinder the clinical relevance and fairness of the envisioned systems. Similarly, we found that 164 (81.6%) studies mention themes related to bias and clinical utility; however, most of the discussion revolves around data-centered issues. Only 41 (20.4%) articles mention themes associated with at least 3 out of 5 pipeline steps, suggesting a limited appropriation of the notions of bias and clinical utility in such a sensitive context as mental health analysis.

Conclusions: Bias and clinical utility are lightly covered in the field of LLM-based mental health prediction research as of 2019-2024. In-depth approaches involving interdisciplinary teams of clinicians and natural language processing specialists are needed to ensure technical soundness, clinical relevance, and fair outcomes for potential users.

JMIR AI 2026;5:e88082; doi: 10.2196/88082

Keywords: natural language processing; large language models; mental health; bias; clinical utility

Introduction

Background

Biomedical natural language processing (NLP) methods aim to support clinical research and practice, with suggested uses spanning information retrieval, medical documentation generation, and patient data analysis [1,2]. As language has been studied as a marker for conditions such as depression [3], schizophrenia [4,5], and Alzheimer disease [6], mental health–oriented tasks have become popular in NLP. Notably, methods to analyze patient-authored texts or medical records as indicative of a given condition or symptom have been developed, with the underlying hypothesis that this could translate into clinically useful tools for diagnosis or large-scale screening [7-9]. A number of these methods are now based on large language models (LLMs), which recently garnered global attention from the public, industrials, and researchers for their assumed general language capabilities.

However, it is increasingly documented that LLMs can produce unreliable results and potentially harm users by perpetuating and amplifying bias [10-14], which has been defined as "the presence of systematic errors or disparities within decision-making processes that disproportionately affect specific subgroups" [11]. While current medical practice is admittedly not exempt from bias itself [15], automated systems such as LLMs amplify existing stereotypes, which can lead to unjustified differences in simulated diagnoses and treatment plans for vignettes of patients with different ethnicities and genders [13]. Relatedly, it has been reported that male profiles are overgenerated by LLMs (regardless of the true prevalence of the considered condition) both in English and in French [13,14]. Despite substantial research efforts, no reliable mitigation technique has been proved to fully resolve these issues [16,17], which may be intrinsic to LLMs [18].

Moreover, existing evaluation settings for biomedical LLMs are limited [10,19] and real-world impact studies are still lacking in the NLP community [20], leaving many ethical and regulatory challenges unaddressed [21]. As a result, it is still unclear to what extent these LLM-based systems are clinically "[useful] in improving patient outcomes, informing clinical decision-making, and optimizing health care resources" [22] in mental health. In addition, computer scientists and NLP engineers creating such systems may lack the medical expertise needed to make accurate design choices. It seems therefore crucial to methodically audit LLM-based systems both for bias and clinical utility before considering downstream clinical deployments.

Existing reviews have investigated the applications, benefits, and risks of LLMs in health care [23] and mental health [24-26], highlighting general ethical challenges such as data privacy, fairness and bias, clinical integration, and ethical governance. There is, however, a lack of literature that explicitly connects these different issues and systematically assesses them throughout the development steps of the considered systems, taking a distance from indicators of (possibly ill-defined [27]) performance. Notably, although they are crucial to understand downstream risks, existing bias studies in health-related scenarios generally posit bias as a single-dimension construct (eg, gender or racial bias) [17,28], using LLMs as readily available tools whose development is left unquestioned. Yet the decision-making processes from which bias can stem are numerous, which calls for a bias investigation throughout the entire development pipeline of the considered system [29,30]. Similarly, the limitations and barriers to the clinical implementation of NLP technologies [19,31] tend to be listed as post hoc challenges of the systems rather than connected with underlying design choices. In this review, we consider bias and clinical utility not only to be both related and important issues, but to be complementary when it comes to implementing (or envisioning the possible implementation of) any automated system in real-world health care scenarios. Indeed, both these dimensions (How biased will the final system be? How clinically useful?) depend greatly on methodological choices spread throughout the development of the system.

Building on these considerations, we propose to study bias and clinical utility as 2 intricate dimensions rooted in methodological choices, with expected downstream social impact on users. To our knowledge, this is the first such extensive, large-scale joint analysis of bias and clinical utility for predictive mental health applications based on LLMs.

Objectives

The objectives of this work are 2-fold. First, we wish to broadly describe the methodological decisions adopted in publications on LLM-based mental health prediction. This will lead us to analyze possible bias and barriers to clinical utility stemming from these methodological choices, at each step of the system's development pipeline.

Second, we intend to document the awareness of researchers on matters of bias and clinical utility, as indicated by relevant reported considerations in these same publications. Indeed, increased global research attention on these issues does not necessarily translate into actionable plans for other researchers [32]. Furthermore, propensity to bias is still absent from reference evaluation benchmarks used to evaluate these systems, which shows limited appropriation of these issues within the community.

Drawing on these insights, we mean to foster discussions as to the extent to which it is currently possible, safe, and desirable to integrate LLMs into clinical routine and mental health care in particular.

Methods

Registration and Protocol

This scoping review was preregistered in the Open Science Framework registry [33]. We followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and

Meta-Analyses extension for Scoping Reviews) guidelines [34] (refer to Checklist 1).

Eligibility Criteria

We analyzed original research articles published in conferences or journals between January 2019 (to account for the release of BERT [35]) and December 2024. Eligible studies had to (1) assess the presence or severity of a mental health condition; (2) use, for that purpose, an NLP system comprising at least one LLM; and (3) use nonsynthetic textual samples as input data, possibly along with synthetic samples or other modalities. We excluded preprints, reviews, and studies which were not openly available online, as well as papers not written in English.

Importantly, following criterion (1), articles claiming to perform only stress detection or sentiment analysis were not considered eligible. Regarding criterion (2), as it can be underspecified in the literature which models fall under the designation of "LLMs," we followed the recommendation of Rogers and Luccioni [36] to provide an explicit working definition. In accordance with standard considerations [37], we therefore called LLMs NLP tools that (1) model and can generate text, (2) receive large-scale pretraining, (3) can be used for transfer learning, and (4) rely on the Transformer architecture [38]. Notably, this definition includes both encoder models such as BERT [35] and generative decoder models such as GPT [39], in line with existing reviews about applications of LLMs to mental health [24]. Finally, although criterion (3) allows for the inclusion of papers developing systems for audio or imaging data, our analysis was oriented toward the LLM-relevant part of these systems, that is, the one processing text. Besides, even if synthetic datasets are increasingly praised for clinical research [40], we wanted to focus in this review on systems that fulfill at least partially expectations of pending deployment—generally comprising testing in realistic conditions which are, up to now, only imperfectly proxied by generated corpora [41,42]. In addition, our framework comprises a step which focuses on the included data, considering demographics of included participants, which is not applicable to synthetic datasets.

Search Strategy

In order to provide a wide coverage of relevant publications, we searched 5 databases spanning biomedical, NLP, and more general AI-related literature: MEDLINE (PubMed), Web of Science, IEEE Xplore, ACM Digital Library, and the ACL Anthology. All searches were performed on January 20, 2025, with queries following the template (terms related to "MENTAL HEALTH") AND (terms related to "LARGE LANGUAGE MODELS") AND (terms related to "PREDICTION"). The queries were crafted following multiple search rounds and designed to minimize false negatives; hence, the large number of allowed terms in each keyword group (request details are presented in Multimedia Appendix 1). When allowed by the database search engine, we used filters for date, publication language, and publication type according to the eligibility criteria listed above.

Screening Process

The initial set of records identified via database search was loaded into the Rayyan software (Mourad Ouzzani) [43]. A sample of 100 records was first screened based on title and abstract by 3 reviewers (CB, AN, and KF) independently. Interannotator agreement was high, with pairwise Cohen κ [44] scores ranging from 0.654 to 0.807 [45], while resulting discussions helped clarify the screening strategy (see "Eligibility Criteria" section). Subsequently, the first author (CB) screened the remainder of records, as well as the full text of all the eligible studies before conducting extraction. Exclusion reasons and record counts at every step are detailed in the flowchart. Remaining doubts were addressed through discussion between authors.

Data Extraction

In accordance with our research objectives, we extracted 2 types of information in the reviewed papers: methodological information (what do the authors do?), and authors' self-reflections and acknowledgments linked with bias and clinical utility (what do the authors say about what they do?). The full-text PDFs of included studies were loaded into Zotero (Corporation for Digital Scholarship) [46] for reading, alongside an auxiliary spreadsheet document for qualitative entity extraction performed by CB.

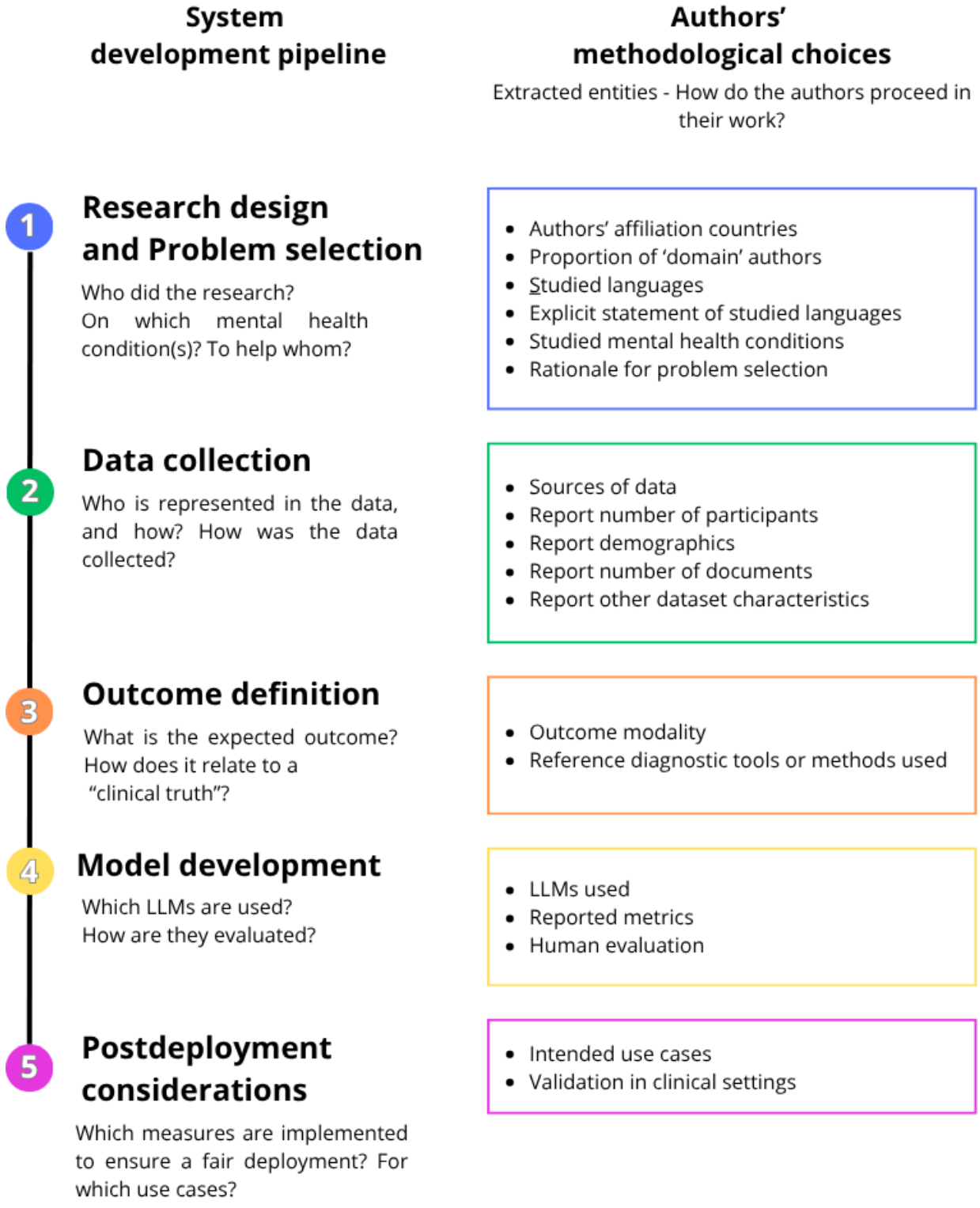
To our knowledge, there is no existing framework specifically designed to study both bias and clinical utility in the conception pipelines for LLMs applied to mental health. However, 2 previous approaches, which we deem complementary to analyze bias and clinical utility at every step of the development pipeline of an LLM-based system for mental health prediction, were proposed.

On the one hand, Hovy and Prabhumoye [29] identified 5 sources of bias in NLP systems: the data, the annotations made on the data, the input representations fed to NLP models, NLP models themselves, and larger research design choices such as the treated language. On the other hand, Chen et al [30] sketch a 5-step pipeline of ethical stakes for machine learning (ML) in health care, considering (1) the unequal distribution of funding and research attention between health issues with regards to the share of the world population that they affect (problem selection), (2) nonrepresentative data (data collection), (3) imperfect medical proxies for target outcomes (outcome definition), (4) ML models optimization parameters potentially exacerbating bias (algorithmic development), and (5) blindness to issues such as generalizability to various clinical settings, downstream impact assessment or regulatory compliance before real-life system deployment (postdeployment considerations).

Inspired by these 2 previous approaches, we propose a 5-step framework (Figure 1) to study both bias and clinical utility within the selected articles. We identify key arguments of each work, and define qualitative entities to extract from papers for future analysis. For instance, Chen et al [30] discuss the importance of measuring performance metrics across demographic groups to ensure fair treatment among patients, which we operationalize as a binary entity

"per-group performance" to be set to "true" when authors report disaggregated results (eg, against sex or gender, race or ethnicity, or other sensitive attributes). This entity relates indeed both to bias (as some groups may be discriminated against if the system performs poorly for them) and to clinical utility (as clinical outcomes are expected to be fair among patients). We redirect the reader to Multimedia Appendix 2 for more details on our analysis framework design and the operationalization of entities.

Figure 1. Left: our proposed 5-step pipeline for analyzing bias and clinical utility in large language model–based mental health prediction systems. Right: entities extracted in papers, in accordance with each pipeline step. LLM: large language model.



We also collected sentences from the included articles that refer explicitly to any sort of bias or to elements related to the clinical utility of the developed system, as a proxy for self-reflection of researchers regarding these matters. We extracted these statements independently from the arguments they defend (eg, recognizing the presence of bias in presented results vs stating that bias has been mitigated), within our extraction framework.

Data Analysis

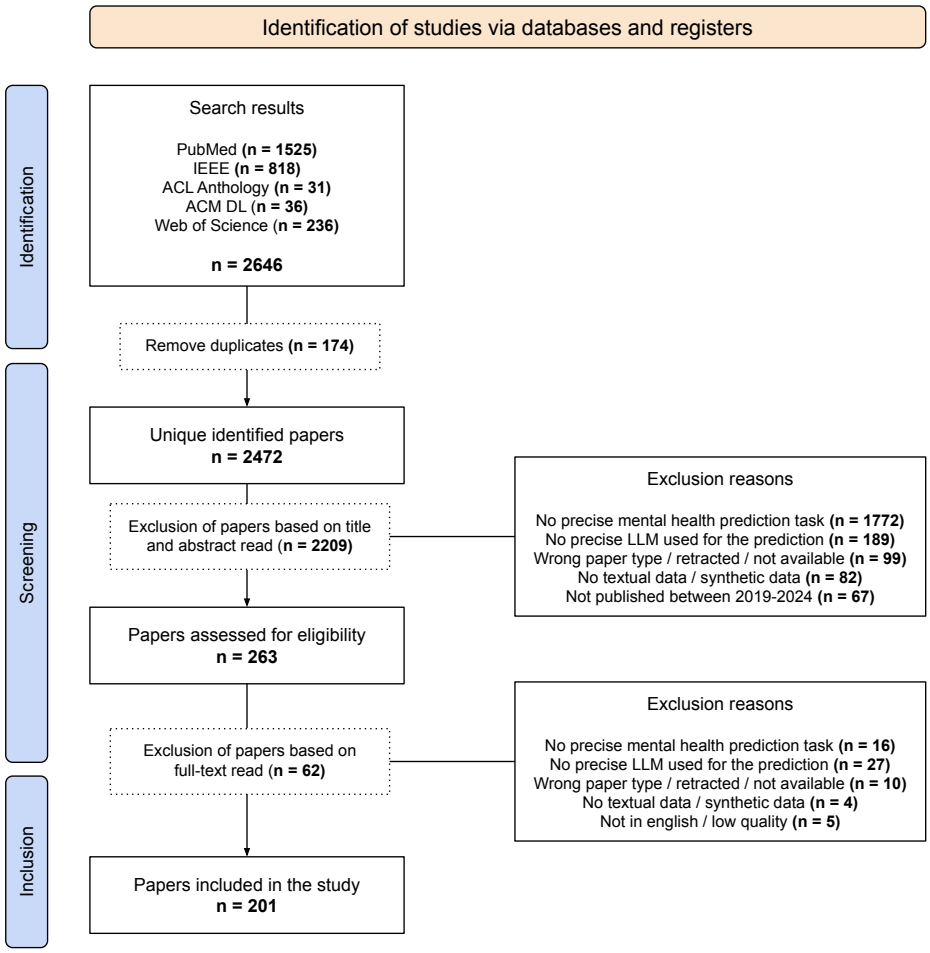
We performed the analysis of the extracted data in a Jupyter Notebook (Project Jupyter) environment, using Python libraries (Guido van Rossum). When applicable, usual descriptive statistics (distribution of values, mean value, range, and so on) were computed. In some cases when necessary conditions were met, we modeled the effect of categorical variables (eg, the effect of medical affiliation on the studied conditions) using chi-squared contingency tests (χ^2) computed with the *SciPy* Python package (without correction). The entities for which full sentences or noisy outputs were extracted were further analyzed by clustering the extracted data into meaningful categories, in an iterative and inductive aggregation process inspired by grounded theory [47]. For authors' self-reflections about bias and clinical utility, we used the same approach to map the highlighted sentences to themes. The final results (extracted entities or categories obtained after coding for each paper) following data analysis are presented as a spreadsheet in Multimedia Appendix 3.

Results

Study Selection

We initially identified 2646 candidate records in databases (search performed on January 20, 2025). After deduplication, 2472 unique papers were screened for eligibility based on title and abstract. The remaining 263 papers were then screened based on full-text read. Finally, 201 (8.1%) studies were included in this review. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of the process is detailed in Figure 2, and the full list of included papers is available in Multimedia Appendix 4.

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.



Included studies (N=201) are mostly recent: 149 (74.1%) are published in 2023-2024, 46 (22.9%) in 2021-2022, and only 6 (3%) in 2019-2020. Most of the retrieved studies are sourced from IEEE Xplore (124/201, 61.7%) and PubMed (57/201, 28.4%). The ACL Anthology, the Web of Science,

and the ACM Digital Library, respectively, raised 10 (5%), 7 (3.5%), and 3 (1.5%) papers.

Global Trends in Bias and Clinical Utility Report by the Authors

We found mentions of bias and clinical utility-related themes in 164/201 (81.6%) papers. The pipeline step that triggered the most discussion is that of data collection (107/201, 53.2%), followed by model development (69/201, 34.3%), and problem selection and research design (60/201, 29.8%). Themes associated with outcome definition and postdeployment considerations were only evoked in, respectively, 45/201 (22.4%) and 31/201 (15.4%) papers. In 77/201 (38.3%) papers, the discussion was restricted to a single step of the pipeline. However, 46/201 (22.9%), 26/201 (12.9%), and 10/201 (5%) papers evoked themes associated with 2, 3, and 4 of these steps. A total of 5 (2.5%) papers mentioned considerations from all 5 steps of the pipeline [48-52].

Bias and Clinical Utility Along the System Development Pipeline

Overview

In this section, we describe entities extracted from the included studies, as illustrated in Figure 1. For each of the 5 steps of pipeline design potentially subject to biases and clinical utility limitations, we both report the extracted entities and the relevant themes explicitly mentioned by authors about their work under a paragraph entitled "Mentions of Bias and Clinical Utility–Related Themes."

Step 1: Research Design and Problem Selection

Overview

The step of research design and problem selection raises the following questions: who did the research? On which mental health conditions? To help whom?

Affiliation Countries of Authors

Forty-five distinct countries are represented in author affiliations, spanning all 6 continents. There is, however, a concentration of publications with author affiliations in China (43/201, 21.4%), the United States (42/201, 20.9%), and India (32/201, 15.9%). International collaborations are present in 54 papers (26.9%), with up to 4 different countries of affiliation in a single publication.

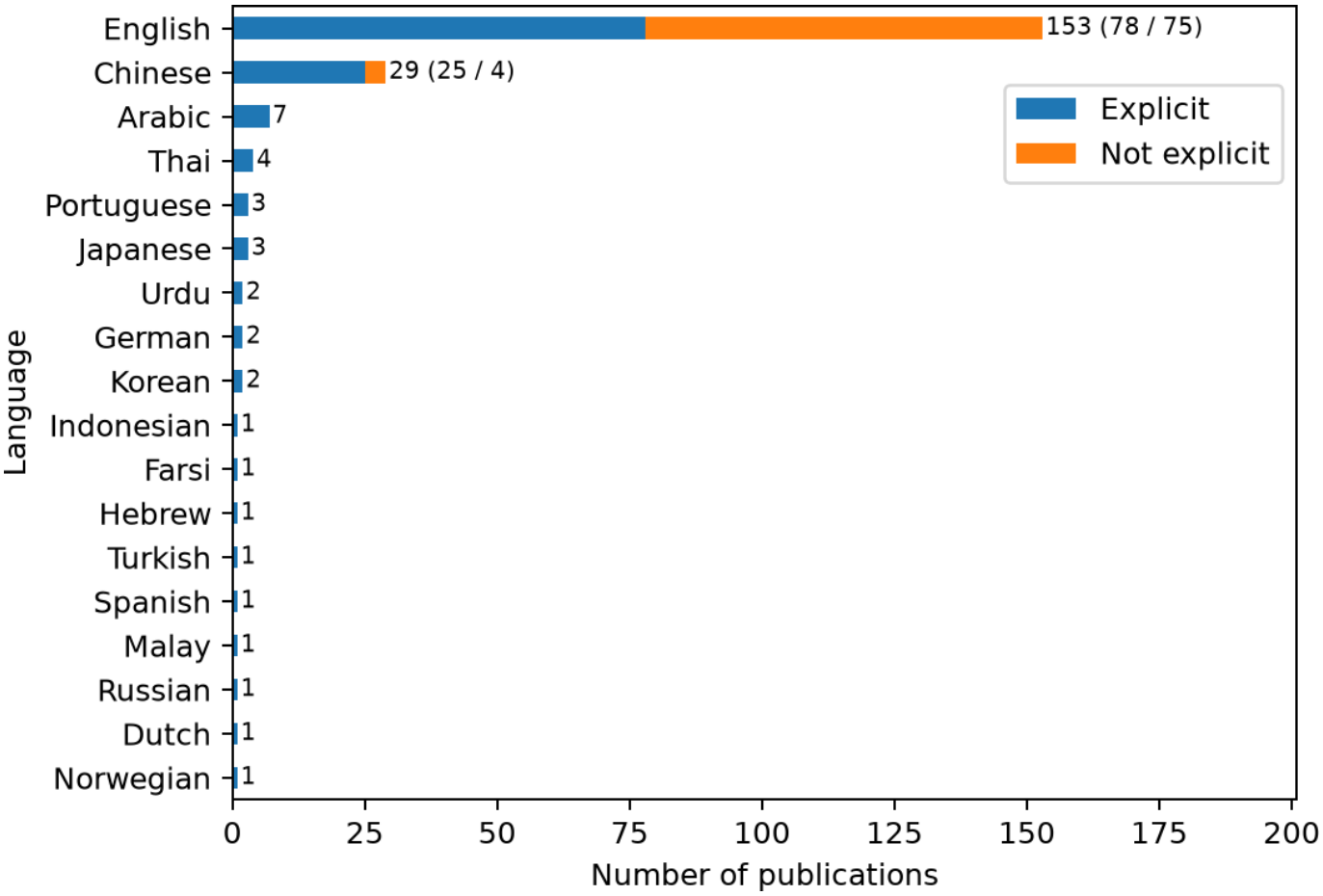
Domain Affiliation of Authors

A total of 76 (37.8%) papers comprise at least one author declaring an affiliation related to the medical domain (domain author). The average ratio of domain authors over all authors within a paper is 0.22, with a median value of 0.0 (IQR 0.33) and an SD of 0.34. A total of 20 (10%) papers have a domain-to-all ratio of 1.0, indicating that they are written only by domain-affiliated authors.

Languages

As for the languages the authors work with (ie, those of the processed data), 18 distinct languages are identified (Figure 3). While English is overtly predominant (153/201, 76.1%), other treated languages include Chinese (29/201, 14.4%), Arabic (7/201, 3.5%), Thai (4/201, 2%), Portuguese (3/201, 1.5%), Japanese (3/201, 1.5%), and others. Of the publications on English and Chinese, respectively, 75/153 (49%) and 4/29 (13.8%) omit to state the treated language explicitly; therefore, this information has to be deduced based on the models or resources used.

Figure 3. Distribution of the languages studied in the included papers (ie, the languages of the data the systems make predictions about). For each language, it is also indicated whether it is explicitly mentioned by the authors in the publications.



Studied Mental Health Conditions

Fifteen mental health disorder groups (mapped to condition groups found in the *DSM-5* [*Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*] [53], as well as a generic category of "Suicidal Risk" for papers studying suicidal ideation, behavior, or attempts) are studied in the selected articles (Table 1). A majority of the reviewed papers work on predicting depressive disorders (148/201, 73.6%), significantly more so when the papers comprise no domain author (χ^2_1=10.95, P<.001). Other prevalent selected conditions include suicidal risk (47/201, 22.9%) and anxiety disorders (17/201, 8.5%), with a variety of other condition groups marginally represented. Besides, while 175 (87.1%) papers focus on a single condition, 26 (12.9%) papers study at least 2 of them, with a maximum coverage of 10 conditions in a single paper [54].

Table 1. Distribution of mental health disorder groups among studies (some studies include multiple disorder groups).

Mental health disorder group	Paper count, n (%)
Depressive disorders	148 (73.6)
Suicidal risk	47 (23.4)
Anxiety disorders	17 (8.5)
Trauma and stressor-related disorders	14 (7)
Bipolar and related disorders	13 (6.5)
Schizophrenia spectrum and other psychiatric disorders	11 (5.5)
Attention-deficit hyperactivity disorder	10 (5)
Personality disorders	8 (4)
Feeding and eating disorders	7 (3.5)
Major and mild neurocognitive disorders	5 (2.5)
Autism spectrum disorders	5 (2.5)
Intellectual disabilities	3 (1.5)

Mental health disorder group	Paper count, n (%)
Obsessive-compulsive and related disorders	2 (1)
Dissociative disorders	1 (0.5)
Other (perinatal psychiatry)	1 (0.5)

Rationale for Problem Selection

Three major themes are identified which relate to the conditions under study or to mental health in general: negative impacts on individuals (165/201, 82.1%; eg, "Mental illness is [...] significantly impacting the lives of individuals across diverse communities" [54]), a high or increasing prevalence (170/201, 84.6%; eg, "Suicide is one of the main causes of death in the world" [55]), and shortcomings of the "traditional" medical system to efficiently tackle the problem (127/201, 63.2%; eg, "Constrained clinician time and a strong focus on anticancer treatment may contribute to the insufficient identification of patients at risk for depression" [56]). Besides, 9/201 (4.5%) papers explain that their study takes place in the context of shared tasks organized by the scientific community. These tasks are hosted by labs or workshops at conferences such as CLEF, CLPsych, ACL, and IEEE BigData.

Mentions of Bias and Clinical Utility–Related Themes

Overall, 60/201 (29.8%) papers mention themes related to the step of research design and problem selection, and 38/201 (18.9%) mention language and culture as a potential bias source related to the step of problem selection. This comprises both acknowledgments that results on English may not generalize to other languages, and cases where authors highlight that they work on an underserved language. Others (22/201, 10.9%) explicate how population disparities related to the condition under study oriented the design of their work toward more vulnerable groups. For instance, some researchers study mental health issues of pregnant or postpartum women [49], US service members and veterans [57], children or adolescents [58], patients with cancer [59], and so on. As for the choice of a condition, 4 (6.7%) papers claim to tackle a research gap, such as Juhng et al [52] who state that "much less work in the NLP community has focused on detecting anxiety disorders as has been done for depressive disorders." Finally, 1 (0.5%) paper explicitly acknowledges the "absence of direct involvement with mental health experts," which we map to the theme of workforce diversity.

Step 2: Data Collection

Overview

The step of data collection raises the following questions: who is represented in the data, and how? How was the data collected?

Data Sources

We identify 3 major data provenances for the datasets used in the reviewed papers. First, posts from social media and online forums are used in 133/201 (66.2%) papers, under the frequent rationale that it consists of large-scale, spontaneous, and anonymous testimonies of users about their mental health. This data source is significantly more adopted in papers comprising no domain authors (χ^2_1=23.57, P<.001). The most prevalent sources within this category are Reddit (Reddit, Inc; 72/201, 35.8%), Twitter (X Corp; 44/201, 21.9%), and Sina Weibo (Weibo Corporation; 11/201, 5.5%). Data collections such as those provided in the CLEF conference for eRisk labs constitute a popular example of such social media datasets (see Textbox 1). Second, 62/201 (30.8%) papers make use of data collected from participants during clinical or semiclinical interactions. This comprises interview transcripts (eg, Patient Health Questionnaire (PHQ)–led interviews in the distressed analysis interview corpus and extended-distressed analysis interview corpus [60], refer to Textbox 1), narrative transcripts provided by participants about their medical experiences, texts entered on mental health monitoring apps, and therapy transcripts. Third, electronic health records of patients coming from different sites are used in 10/201 (4.9%) papers. Among them, one paper uses publicly available ScAN [61] (from MIMIC-III [62]), while the others rely on private datasets accessed through their institutional affiliations. A total of 2 (1%) papers gather multisite records, while the others are from a single source.

Textbox 1. Erisk and the Distressed Analysis Interview Corpus (DAIC): 2 common datasets for automated mental health prediction.

Erisk labs have been held at CLEF since 2017. Each edition comprises one or more tasks designed to "explore issues of evaluation methodology, effectiveness metrics, and other processes related to early risk detection" [63]. As of 2025, while multiple tasks have been proposed (anorexia and related eating disorders, self-harm, pathological gambling), the major focus has been on depression detection. Notably:

- A first dataset for the early detection of depression was used and extended throughout 2017 (task 1), 2018 (task 1), and 2022 (task 2) editions [63-65]. It initially consisted of 531,000 Reddit (Reddit, Inc) posts [66] collected from 892 Reddit users (137 depressed and 755 control) following a template-based heuristic and manual check. "Depression" posts are those of users who unambiguously stated their diagnosis (eg, "I was diagnosed with depression," but not

"I think I have depression"), while "Control" posts were those of random users as well as some posting in the r/Depression forum without explicit diagnosis statements.
- A second dataset was constituted for measuring the severity of the signs of depression in 2019 (task 3) and extended for 2020 (task 2) and 2021 (task 3) [67-69]. It contains the entire Reddit posting history of 20 users (up to 90 in 2021) who agreed to fill BDI-based questionnaires as ground-truth data. Lab participants thus have to predict the users' scores for each item, as well as an overall depression assessment score, eventually mapped to 1 of 4 coarse-grained categories (minimal, mild, moderate, and severe depression).

The DAIC is a collection of 621 clinical interviews designed to "support the diagnosis of psychological distress conditions such as anxiety, depression, and post traumatic stress disorder" [60]. It comprises 4 subcorpora corresponding to distinct interview setups: face-to-face, teleconference, Wizard-of-Oz (WOZ, ie, interviews are led by a human-controlled virtual interviewer), and automated (ie, interviews are led by a fully automated virtual agent). The corpus data consist of audio and video recordings of interviews, partial transcriptions, and both verbal and nonverbal annotations. Interviewees consisted of veterans of the US armed forces and voluntary participants recruited via online ads in California, 397/621 (63.9%) of whom had their interview tagged as "distressed." An extended version (E-DAIC) has since been constituted for the AVEC workshop at ACM MM 2019, although documentation is still lacking [70].

Reporting of Participants Count and Demographic Characteristics

We find that 122/201 (60.7%) papers omit to provide the number of participants included in the datasets they use in their work. When that count is disclosed, we find both very small (eg, 22 participants in the study by Li et al [71]) and very large datasets (eg, more than 43,000 patients in the study by Meng et al [72]). Demographic information about participants is reported in 43/201 (21.4%) papers. More precisely, information can be found about their age (35/201, 17.4%), their sex or gender (32/201, 15.9%), their race or ethnicity or cultural background (11/201, 5.5%), their education level or academic status (9/201, 4.5%), and other attributes such as marital status, household income, or additional medical information (9/201, 4.5%).

Reporting of Documents Count and Other Dataset Characteristics

Document counts (ie, the data samples which are fed to the model, such as social media posts or clinical notes) are reported in 172/201 (85.6%) papers. We note that this document count ranges from dozens of documents (eg, the study by Hayati et al [73]) to several hundred thousand documents (eg, the study by Feng et al [74]). Details about the time of data collection, class size (eg, the number of data samples tagged with or without depression), or outcome-related statistics (eg, the average anxiety score in a questionnaire-based dataset) are reported by the authors in 127/201 (63.2%) papers.

Mentions of Bias and Clinical Utility–Related Themes

We find mentions related to the step of data collection in 107/201 (53.2%) papers, making it the most discussed pipeline step with regard to bias and clinical utility. A total of 60/201 (29.8%) studies discuss the challenge of class imbalance (ie, when target labels are unevenly distributed in the data), which can lead to biased results. This imbalance is typically mitigated by authors using sampling or weighting techniques to avoid negative impacts on model performance. Representation bias mentions (ie, when the attributes of the people represented in a dataset mismatch those of the target population) are found in 52/201 (25.9%) papers and linked with diverse methodological (eg, site of collection and sampling method) and sociodemographic features (eg, gender, race and ethnicity, and social status). Other themes relate to the challenges of dealing with a small dataset (22/201, 10.9%), preexisting bias (3/201, 1.5%) in the data (eg, Hutto et al [51] acknowledge as a limitation that "bias may be introduced by the author of the [medical] note" and static datasets (3/201, 2.8%) which can only account for a person's condition at a given point in time, whereas clinical decisions are usually made based on a person's medical history.

Step 3: Outcome Definition

Overview

The step of outcome definition raises the following questions: what is the expected outcome? How does it translate to clinically actionable information?

Outcome Modalities

Different outcome modalities are found in the reviewed studies, with sometimes several ones in the same paper. In 146/201 (72.6%) papers, simple binary labels are assigned to data samples for the condition under study (eg, "depression" vs "no depression"). In 46/201 (22.8%) papers, severity degrees are instead estimated on ordinal, qualitative scales typically comprised 3 to 4 values (eg, "no depression," "low depression," "moderate depression," and "severe depression"). Continuous, quantitative scales, on the other hand, are used in 19/201 (9.5%) papers when scores are computed based on reference questionnaires (see paragraph below). Finally, 9/201 (4.5%) papers provide symptom-level predictions by focusing on individual items of aforementioned questionnaires, emotions, or other behavioral clues.

Reference Diagnostic Tools or Methods Used

Only 64/201 (31.8%) articles explicitly mention their reliance on standard psychiatric tools and classifications for condition assessment of the included participants. A total of 10/201 (5%) studies refer to the *DSM* (*Diagnostic and Statistical*

Manual of Mental Disorders), and 11/201 (5.5%) to the *ICD* (*International Classification of Diseases*). A range of questionnaires are also mentioned, including the PHQ (30/201, 14.9%) with 8 (PHQ-8) or 9 items (PHQ-9) [75], the Hamilton Depression Scale [76] (6/201, 3%) and the Beck Depression Inventory [77] (6/201, 3%) for depression, as well as the posttraumatic stress disorder checklist for civilians [78] (3/201, 1.5%), or the Mini-Mental State Examination [79] (3/201, 1.5%) for cognitive impairment. It should be noted that the authors use such questionnaires in various ways: sometimes, scores obtained from participants' self-administration are used as readily available labels, while in other cases the questionnaire is used to structure interviews whose transcripts will be used as input to the prediction system. In 137/201 (68.2%) studies, the authors rely on other custom methods for condition identification, or do not specify the means by which their data were annotated. These custom methods can involve asking annotators to classify texts based on the presence of certain keywords or on social media–based heuristics (eg, when texts are extracted from specific discussion forums). In 17/201 (8.5%) studies, the authors specify that labels were provided by trained clinicians.

Mentions of Bias and Clinical Utility–Related Themes

Mentions related to the step of outcome definition are found in 45/201 (22.4%) papers. The validity of the proxy used as a marker of a given mental health condition is questioned in 38/201 (18.9%) papers, which stress the limitations of relying on binary labels or self-report statements expressed on social media. For instance, Ohse et al [80] note that the absence of a differential diagnosis may limit the validity of their findings (derived from self-report measures), as their ground truth may have been overly inclusive. Annotation bias is also mentioned in 8/201 (4%) papers, which comprises acknowledgments that low interannotator agreement or stereotypical bias from the annotators may hinder the validity of the labels used.

Step 4: Model Development

Overview

The step of model development raises the following questions: which LLMs are used? How are they evaluated?

LLMs Used

A total of 112 distinct LLMs are identified in our corpus, with an average count of 2.5 LLMs used per study (range: 1-14). As detailed in Table 2, most studies rely on encoder-only models mainly derived from BERT [35] and related models such as RoBERTa [81], DistilBERT [82], and others. These models are typically used to produce contextual representations of input sequences, which are later fed to classification layers or to other models for mental health prediction. Decoder-based models derived from GPT [39] and XLNet [83] as well as encoder-decoder models are used to a lesser extent; researchers then tend to use them for prediction in a few-shot setting or, more rarely, for data augmentation or explanation generation. In 38/201 (18.9%) papers, at least one LLM which was previously trained on clinical or mental health–related content is used. We note, however, that in the remaining 163/201 (81.1%) papers, general models are preferred, which can be used off-the-shelf and optionally trained on domain data. When working on languages other than English, authors explicitly state the need to use multilingual or specialized monolingual models such as BERT derivatives for the Chinese or Arabic language.

Table 2. List of frequently used models in our corpus (with cutoff frequency of 4 papers)[a].

Model	Clinical or mental health training data	Paper count, n (%)
BERT[b] [35]	—[c]	124 (61.7)
RoBERTa [81]	—	49 (24.4)
DistilBERT [82]	—	25 (12.4)
MentalBERT [84]	Reddit mental health posts	21 (10.4)
ALBERT [85]	—	15 (7.5)
GPT-3.5	—	15 (7.5)
XLNet [83]	—	15 (7.5)
GPT-4 [86]	—	10 (5)
GPT or ChatGPT	—	10 (5)
XLM-RoBERTa [87]	—	9 (4.5)
BERT-chinese [35]	—	9 (4.5)
MentalRoBERTa [84]	Reddit mental health posts	7 (3.5)
SBERT [88]	—	6 (3)
mBERT [35]	—	6 (3)
BioClinicalBERT [89]	Clinical notes (from MIMIC-III)	5 (2.5)

Model	Clinical or mental health training data	Paper count, n (%)
MPNET [90]	—	5 (2.5)
ELECTRA [91]	—	5 (2.5)
DeBERTa [92]	—	5 (2.5)
AraBERT [93]	—	4 (2)
BioBERT [94]	Biomedical literature (PubMed, PMC)	4 (2)
DepRoBERTa [95]	Reddit mental health posts	4 (2)

[a]Indicates cases where the precise version of the GPT model used is not disclosed.
[b]BERT: Bidirectional Encoder Representations from Transformers.
[c]No clinical or mental health training data have been performed for these models.

Reported Metrics and Human Evaluation

Standard classification metrics such as accuracy, precision, specificity, recall (sensitivity), F-measure, area under the receiver operating characteristic curve, or raw confusion matrices are reported in 187/201 (93%) papers. Similarly, standard regression metrics such as mean squared error, mean absolute error, or correlation coefficients (eg, Pearson *r*) are reported in 25/201 (12.4%) papers. A range of more specific metrics used to fit the constraints of particular prediction setups was also identified: for instance, the early risk detection error [67] is designed to take into account both the correctness of a system's binary decision and the delay needed to make that decision (measured by the number of text items seen before providing an answer). On the other hand, human-led, qualitative evaluation is very rare in the corpus. As an example, Wang et al [96] ask 50 medical interns specializing in mental illness to rate the safety, usability, and fluency of their depression detection system on a 1-10 scale.

Mentions of Bias and Clinical Utility Related Themes

Mentions related to the step of model development are found in 69/201 (34.3%) papers. The main reported theme is that of the technical limitations of models (56/201, 27.8%), which can have an effect on their downstream clinical utility, although this consequence is hardly ever made explicit by the authors. These limitations include small context window size (making the processing of long documents challenging), nondomain training, an elevated computational and time cost, the need for large training datasets, etc. In addition, 13/201 (6.5%) papers explicitly mention the notion of model bias, that is, bias intrinsically encoded and amplified by LLMs. Group fairness is evoked in 5/201 (2.5%) papers, some of which report disaggregated results based on sociodemographic features (eg, age and gender [97] and sex, race, and ethnicity [48,56]).

Step 5: Postdeployment Considerations

Overview

The step of postdeployment considerations raises the following questions: which measures are implemented to ensure a fair deployment? For which use cases?

Intended Use Cases

Many papers do not propose an explicit, concrete use case for their work. Instead, some express the general ambition to facilitate the early detection of mental health issues in individuals (57/201, 28.4%). As for more precise applications, some researchers propose to use clues in patients' medical data to anticipate the possible onset of mental health issues (eg, depression in patients with cancer [48,56]). Others evoke the use of NLP methods to monitor global mental health trends in social media (eg, postpartum depression [98] and suicide ideation [99]), sometimes even suggesting that users at risk should be recommended to mental health specialists [59] or identified for real-world interventions [100]. Some proposals are also designed to fit more directly into the patient-caregiver relationship. Notably, Diniz et al [55] developed a web app for doctors to monitor suicidal ideations of patients estimated from their smartphone keyboard data, while Shimamoto et al [101] proposed to automatically estimate depression severity based on oral responses to an automated version of the Montgomery-Åsberg Depression Rating Scale.

Validation in Clinical Settings Procedures

The corpus contains no description of rigorous, systematic validation procedures in realistic clinical settings of the reviewed automated systems for mental health prediction. This could be explained by the fact that a large number of studies are retrospective, that is, they use existing data from patients the authors did not interact directly with. In that case, there is consequently no downstream impact on the concerned stakeholders, and the clinical utility is minimal. To our knowledge, none of the reviewed systems has been deployed in routine mental health care.

Mentions of Bias and Clinical Utility–Related Themes

Mentions related to the step of postdeployment considerations are found in 31/201 (15.4%) papers. A total of 26/201 (12.9%) papers evoke the performance of their system in the context of a future possible deployment; both positive and negative judgments are emitted. As an illustration, Matero et al [102] explicitly warn that "at this time [they] do not suggest [their] model(s) be used in practice to label mental health states," whereas Bartal et al [50] are more confident that their model "has the potential to fit seamlessly

into routine obstetric care." Others anticipate whether the generalizability (4/201, 2%) of their system is sufficient or whether it may be confronted with integration challenges (4/201, 2.0%). Interestingly, Khalil et al [103] suggest that federated learning (ie, "an alternative that leaves the training data distributed on the mobile devices, and learns a shared model by aggregating locally-computed updates" [104]) could address issues of data privacy and regulation disparities between countries when making mental health predictions in a multilingual setting.

Discussion

Principal Findings

With this study, we introduced a framework to analyze methodological bias and clinical utility throughout the development pipeline of LLM-based mental health prediction systems. Our review notably sketched the prototypical profile of such a prediction system: it focuses on depressive disorders, leverages English data collected from social media, and uses BERT-based classifiers in the absence of a clinically useful outcome for the included participants. In what follows, we argue that this is problematic for multiple reasons.

LLM-Based Mental Health Prediction Suffers From Bias All Along the Development Pipeline

First, the focus on depressive disorders leaves nearly untouched other condition families, echoing the observations of Wang et al [26]. Although it is true that they account for a large part of the world's mental health burden, millions of individuals are also affected by anxiety disorders, bipolar disorder, schizophrenia, or eating disorders, to name a few [105]. Besides, it seems that communication disorders (affecting individuals' language abilities) could also particularly benefit from NLP inputs [106]. In addition, despite a relative diversity among affiliation countries, papers almost exclusively work on English data, which exacerbates the existing bias toward English-based systems in NLP [29]. This bias often goes unnoticed, with nearly half of papers working on English omitting to mention it explicitly, a phenomenon that has been otherwise measured in NLP conferences [107,108]. In order to benefit a wider range of individuals, non-English or multilingual approaches should be considered.

Second, the data used in the studies largely come from social media (which was also observed by Wang et al [26]), even more so when no domain authors are involved. Some authors highlight the ease of access and collection of such data (as opposed to "traditional" clinical data), its alleged relevance to the task of mental health detection, and its wide accessibility (eg, "social media's ubiquity presents a platform for individuals to express their feelings, instead of traditional, formal clinical settings, with 8 out of 10 people disclosing their suicidal thoughts and plans." [109]). However, social media users do not make up representative samples of individuals, and it is difficult to ensure the quality of social media posts [30]. This is furthermore problematic as demographic information is generally absent from these datasets, and users are rarely able to provide consent for the collection of their data. In order to avoid harming users, social media research should comply with ethical guidelines as regards consent of participants, data deidentification, and sharing [110].

Third, the proxies used to account for participants' alleged mental health status lack clinical grounding. It is unlikely that heuristics based on keywords within texts or self-administered questionnaire scores would be considered reasonable indicators by clinicians to validate diagnostic labels. This leads us to question the maturity of the field for clinical integration.

Fourth, authors tend to overlook practical requirements when adopting a model. We note that popular, nonspecialist models (and, increasingly, generative LLMs) are generally preferred, yet these choices are rarely clinically motivated. It is currently debated whether specialist models can actually be competitive with bigger, more recent generalist models [111] on specialized downstream tasks. However, the size and eventual proprietary status of the latter hinder nonnegotiable needs of data privacy and control over computational cost in clinical environments with limited resources, along with issues of environmental impact and models' propensity to bias [112,113].

Finally, there is an overall lack of awareness of the ethical issues implied by a possible clinical deployment of the considered systems. Importantly, some works seem not to aim for such clinical use cases and deny that their systems should be used as diagnostic tools already. Yet this leads to questioning the clinical utility of tasks confined to computer science laboratories, whose results may be mistakenly interpreted as clinical conclusions. Overall, authors mainly initiate discussion themes that focus on generic ML issues related to model constraints at the model development step (eg, input length constraints, computational cost seen as a practical barrier and not an environmental problem), while revealing a data-centric view of bias around the data collection step (eg, calling for bigger, class-balanced datasets). This technical approach to bias, which calls for technical solutions, has been previously identified as the engineering ethos [114]. However, such complex issues as biases cannot be solved with a sole technique [12], if at all [18]. In what follows, we argue that an ethical approach to LLMs applied to the domain of mental health would benefit from more interdisciplinarity.

Interdisciplinarity Is Needed to Increase Clinical Validity and Utility

In addition to the biases identified above, our observations raise the question of interdisciplinarity and its influence on the clinical utility of proposed solutions. Indeed, while we

stressed global shortcomings, it should be acknowledged that studies were more robust in terms of disorder diversity, specificity of disorder definition, and quality of data sources (clinical interviews and electronic health records) when at least one author had a medical affiliation. This aligns with previous findings that constituting interdisciplinary teams (including clinicians and computer scientists) improves results validity in ML problems [115].

For instance, clinical expertise is needed to adequately define what is referred to as "depression" in studies. Some papers ambiguously assimilate the concepts of "depression" and "negative mood/sentiment" (eg, "the task is to discover the mood of the user" [116]), or "depression" and "suicidal risk" (eg, Wang et al [117] guidelines for depression level estimation are actually based on reports of suicidal ideation and plan), or do not seem to make a conceptual difference between self-diagnosed depression and clinical diagnoses produced by health care professionals. The task of disambiguating what is designated by "depression" is all the more complex because according to MeSH definitions, it can cover a sign or symptom ("Depression" [118]) or a disorder with different levels of severity ("Depressive Disorder" [119] or "Major Depressive Disorder" [120]). These differences are rarely described or implemented in data annotation protocols in studies, even though they play a crucial role in diagnostic reasoning and patient care. Consequently, training a system to automatically detect "depression" and claiming it is ready for use in clinical practice is suboptimal at best (what kind of "depression" does the system detect?), and misleading if users are left to make their own assumption of what concept of "depression" is operationalized. The close collaboration between computer scientists and medical doctors is thus required to ensure and to propagate the clinical value of the targeted outcome.

Interdisciplinary collaborations between diverse fields can also improve the robustness of systems' evaluation. Indeed, while some studies warrant caution before real-world deployment, others present their system as deployment-ready (even in the absence of such validation), while a majority do not address the question at all. Notably, no included study presented a rigorous clinical validation for an LLM-based prediction system to be applied in practice, for example, through a randomized controlled trial [20].

Extending even beyond sole clinical validation, interdisciplinary frameworks for the validation of digital devices in medicine have recently been published. For instance, the certification of digital devices—and particularly those embarking LLMs—by the US Food and Drug Administration or the European Medicines Agency requires them to comply with the criteria of the V3 framework [121]. This framework requires rigorous clinical validation (ie, demonstrating that the metrics computed by the device correlate with meaningful physiological or clinical dimensions), but also rigorous hardware verification (validation of the sensor, if there is one) and analytical validation (ie, demonstrating that the algorithm indeed measures a behavior or a physiological marker).

This validation, however, does not ensure that the device, in its context of use, will be useful. Indeed, for most of the articles included in this review, the motivation for the research often remains vague, although untested potential use cases are described, such as large-scale mental health screening or patient monitoring. For others, very few contextual elements help grasp the medical stakes, which we believe have to be mentioned in such research. Is this to say that mental health prediction could become yet another LLM-equipped task, ensuring reasonable chances of getting a paper published?

Interestingly, an updated version of the V3 framework has been recently introduced: the V3+ framework, completing the 3 previous validations by a criterion of usability validation [122]. Although it differs from clinical utility, the addition of utility validation incorporates considerations regarding the projected use, the target population, and the intended context into the validation of digital health devices. Clarifying these elements, which requires interdisciplinary consultation between computer scientists (regarding technical validity) and clinicians (regarding usage), would thus prevent misunderstandings between the two communities, thereby avoiding empty promises fueled by each community's overclaims regarding the capabilities of the other—particularly regarding LLMs' ability in clinical context [123].

Limitations

While this scoping review focuses on 201 papers published between 2019 and 2024, this period covers both the introduction of masked (eg, BERT) and autoregressive (eg, GPT) language models, which provides an opportunity to observe how they were used for mental health prediction early on. In addition, we believe that the large number of included studies allows us to provide a representative snapshot of the field in this 6-year window, through an original joint analysis of bias and clinical utility. While more recent articles could have been included, we assume that they would only have had a marginal effect on the trends we evidenced. Notably, Reiter [20] recently reported that as of March 2025, only 0.1% of papers from the ACL Anthology provided any form of downstream impact evaluation, echoing our findings in the case of mental health prediction.

Our approach is that of a scoping review, aiming at presenting a representative view of what researchers or clinicians may encounter when looking for LLM-based mental health prediction works. As such, we did not perform quality assessment of the articles at the screening stage [124]. In addition, due to the large number of included studies, we could not ensure an independent extraction by multiple reviewers, which may introduce bias in the presented results: as for other phases of the review, this was mitigated with regular discussions and methodological refinements. Since we considered a large number of qualitative entities, methodological choices were necessary to extract them into quantitative metrics. For instance, authors were labeled as "domain authors" whenever they reported being affiliated with a medical institution, a medical university department or school, or a company providing medical devices or services.

While we acknowledge this to be an imperfect proxy for medical expertise, our data are freely available in Multimedia Appendices 1–4 for interested researchers to replicate our results with different methodological choices.

Finally, this study was conducted by a group of French, White, NLP researchers, including some with a background in medical informatics or psychiatry. Although we have aimed at adopting an open and interdisciplinary vision, taking into account issues related to bias and clinical utility, we may have missed some relevant aspects to these questions, starting with the phase of article identification. Because this is the main language used for research dissemination, we reviewed only articles written in English in major scientific databases, which nonetheless allowed us to retrieve studies working on other languages. We intentionally crafted broad queries to minimize false negatives, and considered multiple literature sources relevant to NLP and medical sciences. Yet, different requests and additional sources may retrieve slightly different results. Also, information about eventual clinical deployments of the considered systems outside of what is reported in the papers may have been missed.

Broader Implications

This review focused solely on LLM-based mental health prediction. We did not consider interventional applications (eg, chatbots for mental health support), nor other medical domains, and some of our observations may not generalize. Nevertheless, we believe that some of our findings are part of a broader picture. Our study suggests that over the past 6 years, the perspective of downstream usage (and users) has seldom been considered in the development of LLM-based, predictive mental health applications. This echoes trends that have been reported in NLP [20]. The overemphasis on algorithm development [125] and quantitative, benchmark-based competition for performance [126] poses a risk of harm by assuming medical applications are "yet another task" to tackle despite increased ethical risks. Finally, it would be interesting to study more deeply the "promises" and narratives around AI and NLP for (mental) health care (eg, the promise to assist health care workers without replacing them, the promise to be cost-effective, and so on), which is left for future research.

Conclusions

LLMs are increasingly used in mental health prediction tasks. Following a 5-step pipeline as a framework to analyze NLP proposals, we described an emerging field that exhibits some marked tendencies, notably toward the detection of depressive disorders and the use of social media data in English. We showed that researchers are generally aware of some bias and clinical utility–related issues, but that reports in papers are sparse and not systematic. In particular, postdeployment considerations and impact studies are lacking, which leads us to question the idea that LLMs have the potential to revolutionize mental health care. When dealing with NLP-assisted mental health research, we advocate for a combined view of bias and clinical utility, implying that no biased system can be clinically useful, and vice versa. This requires interdisciplinarity and careful efforts all along the research pipeline. We strongly believe that the goal to contribute to mental health research and, ultimately, to benefit patients should remain central in NLP efforts towards mental health.

Acknowledgments

The authors thank Gaétan Kerdelhué from DéSaN (Digital Health Department, University Hospital of Rouen, France) for his guidance in crafting MEDLINE queries and selecting an appropriate review management platform. The authors attest that no AI was used to assist the creation of the manuscript, nor the underlying research work.

Funding

This work has received support from the French government (Agence Nationale pour la Recherche) under grant agreement ANR-23-IAS1-0004 (InExtenso).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization: CB, KF, AN
Data curation: CB, KF, AN
Formal analysis: CB
Funding acquisition: KF
Investigation: CB, VPM
Methodology: CB, KF, AN
Supervision: KF, VPM, AN
Visualization: CB
Writing - Original Draft: CB, KF, VPM, AN
Writing - Review & Editing: CB, KF, VPM, AN

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategy for studies inclusion with full string requests.
[DOCX File (Microsoft Word File), 11 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Extraction guide detailing the entities extracted, the rationale for inclusion, and operationalization.
[PDF File (Adobe File), 76 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Table of extracted and analyzed entities from the reviewed papers.
[XLSX File (Microsoft Excel File), 47 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Complete list of included studies with metadata.
[XLSX File (Microsoft Excel File), 28 KB-Multimedia Appendix 4]

Checklist 1

PRISMA-ScR checklist.
[PDF File (Adobe File), 36175 KB-Checklist 1]

References

1. He K, Mao R, Lin Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Inform Fusion. Jun 2025;118:102963. [doi: 10.1016/j.inffus.2025.102963]
2. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics. Sep 2024;11(3):57. [doi: 10.3390/informatics11030057]
3. Tølbøll K. Linguistic features in depression: a meta-analysis. J Lang Works. Dec 16, 2019;4(2):39. URL: https://tidsskrift.dk/lwo/article/view/117798 [Accessed 2026-07-27]
4. Marini A, Spoletini I, Rubino IA, et al. The language of schizophrenia: an analysis of micro and macrolinguistic abilities and their neuropsychological correlates. Schizophr Res. Oct 2008;105(1-3):144-155. [doi: 10.1016/j.schres.2008.07.011] [Medline: 18768300]
5. Amblard M, Musiol M, Rebuschi M. Discourse coherence - from psychology to linguistics and back again. In: Amblard M, Musiol M, Rebuschi M, editors. Coherence of Discourse - Formal and Conceptual Issues of Language. Springer; 2021:1-17. [doi: 10.1007/978-3-030-71434-5_1]
6. Appell J, Kertesz A, Fisman M. A study of language functioning in Alzheimer patients. Brain Lang. Sep 1982;17(1):73-91. [doi: 10.1016/0093-934x(82)90006-2] [Medline: 7139272]
7. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. Oct 2009;42(5):760-772. [doi: 10.1016/j.jbi.2009.08.007] [Medline: 19683066]
8. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimers Dis. 2016;49(2):407-422. [doi: 10.3233/JAD-150520] [Medline: 26484921]
9. Leroy G, Gu Y, Pettygrove S, Galindo MK, Arora A, Kurzius-Spencer M. Automated extraction of diagnostic criteria from electronic health records for autism spectrum disorders: development, evaluation, and application. J Med Internet Res. Nov 7, 2018;20(11):e10497. [doi: 10.2196/10497] [Medline: 30404767]
10. Hiebel N, Ferret O, Fort K, Névéol A. Clinical text generation: are we there yet? Annu Rev Biomed Data Sci. Aug 2025;8(1):173-198. [doi: 10.1146/annurev-biodatasci-103123-095202] [Medline: 40101215]
11. Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. J Biomed Inform. Jun 2024;154:104646. [doi: 10.1016/j.jbi.2024.104646] [Medline: 38677633]
12. Hofmann V, Kalluri PR, Jurafsky D, King S. AI generates covertly racist decisions about people based on their dialect. Nature. Sep 2024;633(8028):147-154. [doi: 10.1038/s41586-024-07856-5] [Medline: 39198640]
13. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. Jan 2024;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
14. Ducel F, Hiebel N, Ferret O, Fort K, Névéol A. "Women do not have heart attacks!" gender biases in automatically generated clinical cases in French. Presented at: Findings of the Association for Computational Linguistics; Apr 29 to May 4, 2025; Albuquerque, NM. URL: https://aclanthology.org/2025.findings-naacl [Accessed 2026-08-05]
15. P Goddu A, O'Conor KJ, Lanzkron S, et al. Do words matter? Stigmatizing language and the transmission of bias in the medical record. J Gen Intern Med. May 2018;33(5):685-691. [doi: 10.1007/s11606-017-4289-2] [Medline: 29374357]

16. Sun L, Mao C, Hofmann V, Bai X. Aligned but blind: alignment increases implicit bias by reducing awareness of race. In: Che W, Nabende J, Shutova E, Pilehvar MT, editors. Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 27 to Aug 1, 2025:22167-22184; Vienna, Austria. [doi: 10.18653/v1/2025.acl-long.1078]
17. Omar M, Soffer S, Agbareia R, et al. Sociodemographic biases in medical decision making by large language models. Nat Med. Jun 2025;31(6):1873-1881. [doi: 10.1038/s41591-025-03626-6] [Medline: 40195448]
18. Resnik P. Large language models are biased because they are large language models. Comput Linguist. Sep 1, 2025;51(3):885-906. [doi: 10.1162/coli_a_00558]
19. Wang L, Bhanushali T, Huang Z, Yang J, Badami S, Hightow-Weidman L. Evaluating generative AI in mental health: systematic review of capabilities and limitations. JMIR Ment Health. May 15, 2025;12:e70014. [doi: 10.2196/70014] [Medline: 40373033]
20. Reiter E. We should evaluate real-world impact. Comput Linguist. Dec 1, 2025;51(4):1419-1431. [doi: 10.1162/COLI.a.18]
21. Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. Jun 2024;6(6):e428-e432. [doi: 10.1016/S2589-7500(24)00061-X] [Medline: 38658283]
22. Badrick T, Bowling F. Clinical utility - information about the usefulness of tests. Clin Biochem. Nov 2023;121-122:110656. [doi: 10.1016/j.clinbiochem.2023.110656] [Medline: 37802380]
23. Moulaei K, Yadegari A, Baharestani M, Farzanbakhsh S, Sabet B, Reza Afrash M. Generative artificial intelligence in healthcare: a scoping review on benefits, challenges and applications. Int J Med Inform. Aug 2024;188:105474. [doi: 10.1016/j.ijmedinf.2024.105474] [Medline: 38733640]
24. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. Oct 18, 2024;11(1):e57400. [doi: 10.2196/57400] [Medline: 39423368]
25. Jin Y, Liu J, Li P, et al. The applications of large language models in mental health: scoping review. J Med Internet Res. May 5, 2025;27:e69284. [doi: 10.2196/69284] [Medline: 40324177]
26. Wang X, Zhou Y, Zhou G. The application and ethical implication of generative AI in mental health: systematic review. JMIR Ment Health. Jun 27, 2025;12:e70610. [doi: 10.2196/70610] [Medline: 40577783]
27. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
28. Schnepper R, Roemmel N, Schaefert R, Lambrecht-Walzinger L, Meinlschmidt G. Exploring biases of large language models in the field of mental health: comparative questionnaire study of the effect of gender and sexual orientation in anorexia nervosa and bulimia nervosa case vignettes. JMIR Ment Health. Mar 20, 2025;12(1):e57986. [doi: 10.2196/57986] [Medline: 40111287]
29. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. Lang Linguist Compass. Aug 2021;15(8):e12432. [doi: 10.1111/lnc3.12432] [Medline: 35864931]
30. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. Jul 2021;4:123-144. [doi: 10.1146/annurev-biodatasci-092820-114757] [Medline: 34396058]
31. Wenderott K, Krups J, Weigl M, Wooldridge AR. Facilitators and barriers to implementing AI in routine medical imaging: systematic review and qualitative analysis. J Med Internet Res. Jul 21, 2025;27:e63649. [doi: 10.2196/63649] [Medline: 40690758]
32. Ghosh S, Wilson K. Bias Is a math problem, AI bias is a technical problem: 10-year literature review of AI/LLM bias research reveals narrow [gender-centric] conceptions of 'Bias', and academia-industry gap. AIES. Oct 15, 2025;8(2):1091-1106. [doi: 10.1609/aies.v8i2.36613]
33. Large language models for mental health diagnosis: a scoping review of biases and applicability concerns. OSF. 2025. URL: https://osf.io/ygknh/overview?view_only=aac52b70d01f4bcb994ab1533c0bbc74 [Accessed 2025-11-14]
34. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 2, 2018;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
35. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Jun 2-9, 2019:4171-4186; Minneapolis, MN. [doi: 10.18653/v1/N19-1423]
36. Rogers A, Luccioni S. Position: key claims in LLM research have a long tail of footnotes. Presented at: Proceedings of the Forty-first International Conference on Machine Learning; Jul 21-27, 2024:42647-42466; Vienna, Austria. [doi: 10.5555/3692070.3693805]
37. Jurafsky DH, Martin J. Speech and Language Processing. 3rd ed. Stanford University; 2026. URL: https://web.stanford.edu/~jurafsky/slp3/ [Accessed 2026-07-27]

38. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: Annual Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, CA. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [Accessed 2026-07-27]
39. Radford A, Narasimhan K. Improving language understanding by generative pre-training. Semantic Scholar. 2018. URL: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035 [Accessed 2026-07-27]
40. Smolyak D, Bjarnadóttir MV, Crowley K, Agarwal R. Large language models and synthetic health data: progress and prospects. JAMIA Open. Dec 2024;7(4):ooae114. [doi: 10.1093/jamiaopen/ooae114] [Medline: 39464796]
41. Sharoff S, Baker J, Hunt DDF, Simpson A. Almost clinical: linguistic properties of synthetic electronic health records. In: Danilova V, Kurfalı M, Söderfeldt Y, Reed J, Burchell A, editors. Presented at: Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026); Mar 28, 2026:115-126; Rabat, Morocco. [doi: 10.18653/v1/2026.healing-1.10]
42. Kapania S, Ballard S, Kessler A, Vaughan JW. Examining the expanding role of synthetic data throughout the AI development pipeline. Presented at: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency; Jun 23-26, 2025:45-60; Athens Greece. Jun 23, 2025.[doi: 10.1145/3715275.3732005]
43. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. Dec 5, 2016;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
44. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. Apr 1960;20(1):37-46. [doi: 10.1177/001316446002000104]
45. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
46. Zotero. URL: https://www.zotero.org/ [Accessed 2026-07-27]
47. Smith JM, Smith DCP, Hsiao DK. Database abstractions: aggregation and generalization. ACM Trans Database Syst. 1977;2(2):105-133. [doi: 10.1145/320544.320546]
48. van Buchem MM, de Hond AAH, Fanconi C, et al. Applying natural language processing to patient messages to identify depression concerns in cancer patients. J Am Med Inform Assoc. Oct 1, 2024;31(10):2255-2262. [doi: 10.1093/jamia/ocae188] [Medline: 39018490]
49. Bartal A, Jagodnik KM, Chan SJ, Babu MS, Dekel S. Identifying women with postdelivery posttraumatic stress disorder using natural language processing of personal childbirth narratives. Am J Obstet Gynecol MFM. Mar 2023;5(3):100834. [doi: 10.1016/j.ajogmf.2022.100834] [Medline: 36509356]
50. Bartal A, Jagodnik KM, Chan SJ, Dekel S. AI and narrative embeddings detect PTSD following childbirth via birth stories. Sci Rep. Apr 11, 2024;14(1):8336. [doi: 10.1038/s41598-024-54242-2] [Medline: 38605073]
51. Hutto A, Zikry TM, Bohac B, et al. Using a natural language processing toolkit to classify electronic health records by psychiatric diagnosis. Health Informatics J. 2024;30(4):14604582241296411. [doi: 10.1177/14604582241296411] [Medline: 39466373]
52. Juhng S, Matero M, Varadarajan V, Eichstaedt J, V Ganesan A, Schwartz HA. Discourse-level representations can improve prediction of degree of anxiety. Presented at: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); Jul 9-14, 2023:1500-1511; Toronto, Canada. [doi: 10.18653/v1/2023.acl-short.128]
53. Diagnostic and Statistical Manual of Mental Disorders: DSM-5TM. 5th ed. American Psychiatric Association Publishing; 2013. [doi: 10.1176/appi.books.9780890425596]
54. Adel S, Elmadany N, Sharkas M. AI - driven mental disorders categorization from social media: a deep learning pre-screening framework. Presented at: 2024 International Conference on Machine Intelligence and Smart Innovation (ICMISI); May 12-14, 2024:238-243; Alexandria, Egypt. [doi: 10.1109/ICMISI61517.2024.10580665]
55. Diniz EJS, Fontenele JE, de Oliveira AC, et al. Boamente: a natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. Healthcare (Basel). Apr 8, 2022;10(4):698. [doi: 10.3390/healthcare10040698] [Medline: 35455874]
56. de Hond A, van Buchem M, Fanconi C, et al. Predicting depression risk in patients with cancer using multimodal data: algorithm development study. JMIR Med Inform. Jan 18, 2024;12:e51925. [doi: 10.2196/51925] [Medline: 38236635]
57. Zuromski KL, Low DM, Jones NC, et al. Detecting suicide risk among U.S. servicemembers and veterans: a deep learning approach using social media data. Psychol Med. Sep 2024;54(12):3379-3388. [doi: 10.1017/S0033291724001557] [Medline: 39245902]
58. Zhang Q, Meng W, Hao J, et al. BERT- BiLSTM-Caps Language Model for Screening of Children's Severe Mental Retardation. Presented at: 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS); Dec 20-22, 2021:296-301; London, United Kingdom. [doi: 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00056]

59. Podina IR, Bucur AM, Todea D, et al. Mental health at different stages of cancer survival: a natural language processing study of Reddit posts. Front Psychol. 2023;14:1150227. [doi: 10.3389/fpsyg.2023.1150227] [Medline: 37425170]
60. Gratch J, Artstein R, Lucas G, et al. The distress analysis interview corpus of human and computer interviews. In: Calzolari N, Choukri K, Declerck T, editors. Presented at: Ninth International Conference on Language Resources and Evaluation; May 26-31, 2014:3123-3128; Reykjavik, Iceland. [doi: 10.63317/3o7bccg9xequ]
61. Rawat BPS, Kovaly S, Yu H, Pigeon W. ScAN: suicide attempt and ideation events dataset. In: Carpuat M, Marneffe MC, Ruiz M IV, editors. Presented at: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics; Jul 10-15, 2022:1029-1040; Seattle, WA. [doi: 10.18653/v1/2022.naacl-main.75]
62. Johnson A, Pollard T, Mark R. MIMIC-III clinical database. PhysioNet. URL: https://physionet.org/content/mimiciii/1.4/ [Accessed 2026-07-27]
63. Losada DE, Crestani F, Parapar J. ERISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. Presented at: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017). Sep 11-14, 2017:Springer International Publishing. 346-360; Dublin, Ireland. 2017.[doi: 10.1007/978-3-319-65813-1_30]
64. Losada DE, Crestani F, Parapar J, et al. Overview of erisk: early risk prediction on the internet. In: Bellot P, Trabelsi C, Mothe J, editors. Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2018). Springer International Publishing; 2018:343-361. [doi: 10.1007/978-3-319-98932-7_30]
65. Parapar J, Martín-Rodilla P, Losada DE, et al. Overview of erisk 2022: early risk prediction on the internet. In: Barrón-Cedeño A, Da San Martino G, Degli Esposti M, editors. Presented at: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2018); Sep 5-8, 2022:233-256; Bologna, Italy. 2022.[doi: 10.1007/978-3-031-13643-6_18]
66. Losada DE, Crestani F. A test collection for research on depression and language use. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N, editors. Presented at: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association (CLEF 2016); Sep 5-8, 2016:28-39; Évora, Portugal. [doi: 10.1007/978-3-319-44564-9_3]
67. Losada DE, Crestani F, Parapar J. Overview of erisk 2020: early risk prediction on the internet. In: Arampatzis A, Kanoulas E, Tsikrika T, et al, editors. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer International Publishing; 2020:272-287. [doi: 10.1007/978-3-030-58219-7_20]
68. Losada DE, Crestani F, Parapar J. Overview of erisk 2019 early risk prediction on the internet. In: Crestani F, Braschler M, Savoy J, et al, editors. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer International Publishing; 2019:340-357. [doi: 10.1007/978-3-030-28577-7_27]
69. Parapar J, Martín-Rodilla P, Losada DE, Crestani F. Overview of erisk at CLEF 2021: early risk prediction on the internet (extended overview). In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer International Publishing; 2021:324-344. [doi: 10.1007/978-3-030-85251-1_22]
70. DAIC-WOZ database & extended DAIC database. University of Southern California. URL: https://dcapswoz.ict.usc.edu/ [Accessed 2025-10-21]
71. Li S, Nair R, Naqvi SM. Acoustic and text features analysis for adult ADHD screening: a data-driven approach utilizing DIVA interview. IEEE J Transl Eng Health Med. 2024;12:359-370. [doi: 10.1109/JTEHM.2024.3369764] [Medline: 38606391]
72. Meng Y, Speier W, Ong MK, Arnold CW. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE J Biomed Health Inform. Aug 2021;25(8):3121-3129. [doi: 10.1109/JBHI.2021.3063721] [Medline: 33661740]
73. Hayati MFM, Ali M, Rosli A. Depression detection on malay dialects using GPT-3. Presented at: 2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES); Dec 7-9, 2022:360-364; Kuala Lumpur, Malaysia. [doi: 10.1109/IECBES54088.2022.10079554]
74. Feng W, Wu H, Ma H, et al. Applying contrastive pre-training for depression and anxiety risk prediction in type 2 diabetes patients based on heterogeneous electronic health records: a primary healthcare case study. J Am Med Inform Assoc. Jan 18, 2024;31(2):445-455. [doi: 10.1093/jamia/ocad228] [Medline: 38062850]
75. Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. Psychol Med. Jun 2020;50(8):1368-1380. [doi: 10.1017/S0033291719001314] [Medline: 31298180]
76. Hamilton M, Sartorius N, Ban TA. The Hamilton Rating Scale for Depression. In: Sartorius N, Ban TA, editors. Assessment of Depression. Springer; 1986:143-152. URL: https://doi.org/10.1007/978-3-642-70486-4_14 [Accessed 2025-10-24] [doi: 10.1007/978-3-642-70486-4_14]
77. Beck AT, Steer RA, Brown G. Beck Depression Inventory–II. APA PsycNet. 2011. URL: https://doi.apa.org/doi/10.1037/t00742-000 [Accessed 2026-07-27]

78. Weathers FW, Litz B, Herman D, Juska J, Keane T. PTSD checklist—civilian version. APA PsycNet. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft02622-000 [Accessed 2026-04-23]
79. Tombaugh TN, McIntyre NJ. The Mini-Mental State Examination: a comprehensive review. J Am Geriatr Soc. Sep 1992;40(9):922-935. [doi: 10.1111/j.1532-5415.1992.tb01992.x] [Medline: 1512391]
80. Ohse J, Hadžić B, Mohammed P, et al. GPT-4 shows potential for identifying social anxiety from clinical interview data. Sci Rep. Dec 16, 2024;14(1):30498. [doi: 10.1038/s41598-024-82192-2] [Medline: 39681627]
81. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. URL: http://arxiv.org/abs/1907.11692 [Accessed 2024-06-17] [doi: 10.48550/arXiv.1907.11692]
82. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. Preprint posted online on Mar 1, 2020. URL: http://arxiv.org/abs/1910.01108 [Accessed 2025-09-10] [doi: 10.48550/arXiv.1910.01108]
83. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. Presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019:5753-5763; Red Hook, NY. [doi: 10.5555/3454287.3454804]
84. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E, et al. MentalBERT: publicly available pretrained language models for mental healthcare. In: Calzolari N, Béchet F, Blache P, editors. Presented at: Thirteenth Language Resources and Evaluation Conference; Jun 20-25, 2022:7184-7190; Marseille, France. [doi: 10.63317/2d7umxifj8wz]
85. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. Preprint posted online on Feb 9, 2020. URL: http://arxiv.org/abs/1909.11942 [Accessed 2025-09-10] [doi: 10.48550/arXiv.1909.11942]
86. OpenAI, Achiam J, Adler S, Agarwalrwa S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 4, 2024. URL: http://arxiv.org/abs/2303.08774 [Accessed 2024-07-18]
87. Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale [Webinar]. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020:8440-8451; Online. [doi: 10.18653/v1/2020.acl-main.747]
88. Reimers N, Gurevych I, Inui K, Jiang J, Ng V, Wan X. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X, editors. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019:3982-3992; Hong Kong, China. [doi: 10.18653/v1/D19-1410]
89. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical BERT embeddings. In: Rumshisky A, Roberts K, Bethard S, Naumann T, editors. Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; Jun 7, 2019:72-78; Minneapolis, MN. [doi: 10.18653/v1/W19-1909]
90. Song K, Tan X, Qin T, Lu J, Liu TY. MPNet: masked and permuted pre-training for language understanding. Presented at: Proceedings of the 34th International Conference on Neural Information Processing Systems; Dec 6, 2020:16857-16867; Red Hook, NY, United States. URL: https://dl.acm.org/doi/10.5555/3495724.3497138 [Accessed 2026-07-30]
91. Clark K, Luong MT, Le QV, et al. ELECTRA: pre-training text encoders as discriminators rather than generators. Presented at: 8th International Conference on Learning Representations; Apr 26-30, 2020; Online. URL: https://nlp.stanford.edu/pubs/clark2020electra.pdf [Accessed 2026-07-30]
92. He P, Liu X, Gao J, Chen W. Deberta: decoding-enhanced bert with disentangled attention. Presented at: 9th International Conference on Learning Representations; May 3-7, 2021; Online. URL: https://iclr.cc/virtual/2021/poster/2562 [Accessed 2026-07-30]
93. Antoun W, Baly F, Hajj H, et al. Transformer-based model for arabic language understanding. In: Al-Khalifa H, Magdy W, Darwish K, Elsayed T, Mubarak H, editors. Presented at: 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection; May 12, 2020; Marseille, France. URL: https://www.researchgate.net/publication/353371883_AraBERT_Transformer-based_Model_for_Arabic_Language_Understanding [Accessed 2026-07-30]
94. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15, 2020;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
95. Poświata R, Perełkiewicz M, Chakravarthi BR. OPI@LT-EDI-ACL2022: detecting signs of depression from social media text using roberta pre-trained language models. In: Chakravarthi BR, Bharathi B, McCrae JP, Zarrouk M, Bali K, Buitelaar P, editors. Presented at: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion; May 27, 2022:276-282; Dublin, Ireland. [doi: 10.18653/v1/2022.ltedi-1.40]

96. Wang X, Liu K, Wang C. Knowledge-enhanced pre-training large language model for depression diagnosis and treatment. Presented at: 2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS); Aug 12-13, 2023:532-536; Dali, China. [doi: 10.1109/CCIS59572.2023.10263217]
97. Huang G, Shen W, Lu H, Hu F, Li J, Liu H. Multimodal depression detection based on factorized representation [Webinar]. Presented at: 2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS); Dec 10-11, 2022:190-196; [doi: 10.1109/HDIS56859.2022.9991717]
98. Dhankar A, Katz A. Tracking pregnant women's mental health through social media: an analysis of Reddit posts. JAMIA Open. Dec 2023;6(4):ooad094. [doi: 10.1093/jamiaopen/ooad094] [Medline: 38033783]
99. Boonyarat P, Liew DJ, Chang YC. Leveraging enhanced BERT models for detecting suicidal ideation in Thai social media content amidst COVID-19. Inf Process Manag. Jul 2024;61(4):103706. [doi: 10.1016/j.ipm.2024.103706]
100. Wu EL, Wu CY, Lee MB, Chu KC, Huang MS. Development of internet suicide message identification and the Monitoring-Tracking-Rescuing model in Taiwan. J Affect Disord. Jan 1, 2023;320:37-41. [doi: 10.1016/j.jad.2022.09.090] [Medline: 36162682]
101. Shimamoto M, Ishizuka K, Ohtani K, et al. Machine learning algorithm-based estimation model for the severity of depression assessed using Montgomery-Asberg depression rating scale. Neuropsychopharmacol Rep. Mar 2024;44(1):115-120. [doi: 10.1002/npr2.12404] [Medline: 38115795]
102. Matero M, Hung A, Schwartz HA. Evaluating contextual embeddings and their extraction layers for depression assessment. In: Barnes J, Clercq O, Barriere V, editors. Presented at: Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis; May 26, 2022:89-94; Dublin, Ireland. [doi: 10.18653/v1/2022.wassa-1.9]
103. Khalil SS, Tawfik NS, Spruit M. Federated learning for privacy-preserving depression detection with multilingual language models in social media posts. Patterns (N Y). Jul 12, 2024;5(7):100990. [doi: 10.1016/j.patter.2024.100990] [Medline: 39081573]
104. McMahan B, Moore E, Ramage D, Hampson S, Arcas BY. Communication-efficient learning of deep networks from decentralized data. Presented at: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics; Apr 20-22, 2017; Fort Lauderdale, Florida, USA. URL: https://research.google/pubs/communication-efficient-learning-of-deep-networks-from-decentralized-data/ [Accessed 2026-07-30]
105. Mental disorders. World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [Accessed 2025-11-17]
106. Bååth R, Sikström S, Kalnak N, Hansson K, Sahlén B. Latent semantic analysis discriminates children with developmental language disorder (DLD) from children with typical language development. J Psycholinguist Res. Jun 2019;48(3):683-697. [doi: 10.1007/s10936-018-09625-8] [Medline: 30684119]
107. Bender EM. The #BenderRule: on naming the languages we study and why it matters. The Gradient. 2019. URL: https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/ [Accessed 2025-10-24]
108. Ducel F, Fort K, Lejeune G, Lepage Y. Do we name the languages we study? the #benderrule in LREC and ACL articles. Presented at: Thirteenth Language Resources and Evaluation Conference; Jun 20-25, 2022:564-573; Marseille, France. [doi: 10.63317/3mqaq2qsx5mv]
109. Sawhney R, Joshi H, Gandhi S, Jin D, Shah RR. Robust suicide risk assessment on social media via deep adversarial learning. J Am Med Inform Assoc. Jul 14, 2021;28(7):1497-1506. [doi: 10.1093/jamia/ocab031] [Medline: 33779728]
110. Benton A, Coppersmith G, Dredze M. Ethical research protocols for social media health research. In: Hovy D, Spruit S, Mitchell M, Bender EM, Strube M, Wallach H, editors. Presented at: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing; Apr 4, 2017:94-102; Valencia, Spain. [doi: 10.18653/v1/W17-1612]
111. Vishwanath K, Alyakin A, Ghosh M, et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. Nat Med. Jul 2026;32(7):2405-2409. [doi: 10.1038/s41591-026-04431-5] [Medline: 42286322]
112. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus. Oct 2023;15(10):e46454. [doi: 10.7759/cureus.46454] [Medline: 37927664]
113. Morand C, Ligozat AL, Névéol A. MLCA: a tool for machine learning life cycle assessment. Presented at: 2024 10th International Conference on ICT for Sustainability (ICT4S); Jun 24-28, 2024:227-238; Stockholm, Sweden. [doi: 10.1109/ICT4S64576.2024.00031]
114. Forsythe DE. Engineering knowledge: the construction of knowledge in artificial intelligence. Soc Stud Sci. Aug 1993;23(3):445-477. [doi: 10.1177/0306312793023003002]
115. Littmann M, Selig K, Cohen-Lavi L, et al. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. Nat Mach Intell. Jan 2020;2(1):18-24. [doi: 10.1038/s42256-019-0139-8]
116. Esackimuthu S, Hariprasad S, Sivanaiah R, S A, Rajendram SM, T T M. SSN_MLRG3 @lt-edi-acl2022-depression detection system from social media text using transformer models. Presented at: Proceedings of the Second Workshop

on Language Technology for Equality, Diversity and Inclusion; May 27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.ltedi-1.26]
117. Wang X, Chen S, Li T, et al. Assessing depression risk in chinese microblogs: a corpus and machine learning methods. Presented at: 2019 IEEE International Conference on Healthcare Informatics (ICHI); Jun 10-13, 2019:1-5; Xi'an, China. [doi: 10.1109/ICHI.2019.8904506] [Medline: 32537571]
118. Depression MeSH descriptor data 2026. National Library of Medicine. URL: https://meshb.nlm.nih.gov/record/ui?ui=D003863 [Accessed 2025-10-17]
119. Major depressive disorder MeSH descriptor data 2026. National Library of Medicine. 2025. URL: https://meshb.nlm.nih.gov/record/ui?ui=D003865 [Accessed 2025-10-17]
120. Depressive disorder MeSH descriptor data 2026. National Library of Medicine. URL: https://meshb.nlm.nih.gov/record/ui?ui=D003866 [Accessed 2025-10-17]
121. Goldsack JC, Coravos A, Bakker JP, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). NPJ Digit Med. 2020;3:55. [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
122. Bakker JP, Barge R, Centra J, et al. V3+ extends the V3 framework to ensure user-centricity and scalability of sensor-based digital health technologies. NPJ Digit Med. Jan 24, 2025;8(1):51. [doi: 10.1038/s41746-024-01322-2] [Medline: 39856145]
123. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med. Jul 29, 2023;6(1):135. [doi: 10.1038/s41746-023-00879-8] [Medline: 37516790]
124. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libraries J. Jun 2009;26(2):91-108. [doi: 10.1111/j.1471-1842.2009.00848.x]
125. Wagstaff KL. Machine learning that matters. arXiv. Preprint posted online on Jun 18, 2012. [doi: 10.48550/arXiv.1206.4656]
126. Raji ID, Bender EM, Paullada A, Denton E, Hanna A. AI and the everything in the whole wide world benchmark. Presented at: 35th Conference on Neural Information Processing Systems (NeurIPS 2021); Dec 6-14, 2021; Online. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html [Accessed 2026-07-30]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
***DSM*:** *Diagnostic and Statistical Manual of Mental Disorders*
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
***ICD*:** *International Classification of Diseases*
LLM: large language model
ML: machine learning
NLP: natural language processing
PHQ: Patient Health Questionnaire
PHQ-8: Patient Health Questionnaire-8
PHQ-9: Patient Health Questionnaire-9
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by Bradley Malin; peer-reviewed by Shihua Cao, Songbo Hu; submitted 19.Nov.2025; final revised version received 19.Jun.2026; accepted 22.Jun.2026; published 13.Aug.2026

Please cite as:
Bleuze C, Fort K, Martin VP, Névéol A
Large Language Models for Mental Health Prediction: Scoping Review of Bias and Clinical Utility Documentation in 2019-2024
JMIR AI 2026;5:e88082
URL: https://ai.jmir.org/2026/1/e88082
doi: 10.2196/88082