Vol. 10, No. 4, 2025, pp. 267-283
DOI: https://doi.org/10.29210/30036999000

Contents lists available at [Journal ELORA]

JRTI (Jurnal Riset Tindakan Indonesia)
ISSN: 2502-079X (Print) ISSN: 2503-1619 (Electronic)

Journal homepage: https://jrti.eloracenter.org/jrti

Beyond what people say: a systematic review of artificial intelligence-based detection of psychological non-disclosure and hidden mental states in counseling and mental health care

Ifdil Ifdil[1*)], Nur Adila Wafiqoh Zulvi[1], M. Fahli Zatrahadi[2]
[1] Universitas Negeri Padang, Padang, Indonesia
[2] Universitas Islam Negeri Sultan Syarif Kasim Riau, Riau, Indonesia

Article Info

Article history:

Received Sep 16th, 2025
Revised Nov 30th, 2025
Accepted Dec 16th, 2025

Keyword:

Artificial intelligence
Digital phenotyping
Hidden mental states
Mental health counseling
Psychological non-disclosure

ABSTRACT

Not revealing psychologically distressing thoughts, symptoms, and experiences, that is, psychological non-disclosure, contributes to the failure of detection and treatment in many cases of mental illness. Since diagnostic methods mostly depend on self-reporting, such states often remain unrecognized. This article is a literature review and it summarizes the research on how AI and computational methods can be applied to detect psychological non-disclosure and secret states. According to PRISMA 2020, literature was searched using three databases: Scopus, PubMed, and ScienceDirect; only eligible sources were selected and the Mixed Methods Appraisal Tool was used for evaluating. After excluding duplicates only 791 records remained with 20 of those (2020-2025) being included. Six methods are mentioned: language and text analysis; speech and acoustic biomarkers; facial micro-expressions and physiological signals; large language model conversational systems; predictive machine learning; and disclosure augmentation. The study shows that multimodal AI can identify hidden distress, self-injury, and suicidal states although the knowledge is still not fully developed, validation is not enough, and safeguards against misuse remain weak. A framework mapping the different modalities to possible targets was suggested as well the article highlights the need for more thorough validation, transparent reporting, and privacy-preserving use of AI.

Corresponding Author:

Ifdil Ifdil,
Universitas Negeri Padang, Padang, Indonesia
Email: ifdil@konselor.org

Introduction

Mental health disorders are one of the major global public health issues, and there is still a big gap between the number of people in need and who actually get help. Chronic shortage of service providers, limited access at different places of residence, and a pervasive stigma altogether act as limiting factors for early intervention, and hence lead to the rapid development of self-help and digital methods across the world (Mo et al., 2025; Knippler et al., 2024; Balcombe & De Leo, 2022). In contrast, help-seeking is often postponed or not done until quite a large part of the population who suffers from psychological disorders gets no formal assistance at all. Hence, recognizing why distress isn't revealed and how to detect it early has become the core issues in mental health research today and in designing care pathways that are both scalable and fair.

The big challenge that is characteristic of today's environment is psychological non-disclosure, a phenomenon referring to people concealing, minimizing or not even talking to their physicians or the clinicians at all, the symptom, the experience or behavior they have. The reason for this concealment is that in clinical settings and the rest of the time, the client's contribution is what determines the assessment. Therefore, concealment opens the possibility of major mistakes both in diagnosis and follow-up. The factors that hinder truthful expression are shame, fear of others' perception, and anticipation of negative repercussions. In some cases, disorders are also the reason that one does not know the internal state. (Miqdadi et al., 2025; Kosyluk et al., 2024). So we are always having a discrepancy between what an individual says and his or her real feelings, and in cases of suicidal thoughts or self-harm such a discrepancy may lead to very serious consequences.

There are quite a few studies that have aimed at supplementing self-report with signs that are not self-reflective or psychological. For instance, digital platforms for psychological help, conversation agents, and measurement-based care systems nowadays collect data on behavior and language that could reveal signs of psychological distress not through the client's self-reporting directly (Aschbacher et al., 2023; Alemu et al., 2024; Leung et al., 2021). Also, it was shown that when the use of such multimodal digital service has been directed at help-seekers and, for example, to children and teenagers, they have found that interaction data and use patterns were able to convey clinical information (Alvarez-Jimenez et al., 2020; Alemu et al., 2023). In a sense, with the aid of these studies, the problem of recognizing the hidden state becomes a matter of computing: as a person's hidden thoughts can be reflected in his/her language, voice, facial expressions, or physique, the machine can detect them, and thus, the limitations of human perception and questionnaires will be compensated to some extent.

Recent methodological advances have pushed this direction forward quite effectively. NLP, deep learning, and large language models can nowadays automatically analyze text, speech, and conversation at a broad scale, whereas ecological momentary and passive-sensing data collection allow a better understanding of the client's state through observations in the daily life situation (Zainal et al., 2025; Lazaridou et al., 2025; Han et al., 2024). Talking to computers, such as conversational agents and other automated systems, can not only help but also collect information through the interaction, or produce reports based on data collected. This can be very important for the development and application of mental health services, as it helps to cover many people at a low cost and collect a lot of details on psychological problems (Jiang et al., 2025; Keyan et al., 2025; Li et al., 2025; Kim et al., 2025). These features make it feasible for us to create mental health models which identify very minute, subtle, distributed hints that someone might be distressed.

Although it is moving forward, the evidentiary base is quite scattered across various disciplines, disorders, diagnostic approaches and modalities, and as of yet no one has synthesized how specifically AI is being used to detect psychological non-disclosure and hidden mental states. There is the problem of a highly fragmented literature most papers can be roughly categorized as coming either from computer science or a form of psychology, which is often either psychiatry or counseling. Moreover, different authors use vastly different metrics or results to present the findings and almost never there is convergence in terms of language of the underlying concept of concealing one's true self, which is the subject of concealment. It follows that we don't have any information to determine how certain modalities correspond to which hidden states, how effective the detection is, and where the field's center of gravity, i.e. main research activity or methodologies, is located. So, this lack is of a type wherein there is a need for a more comprehensive framework by which various detection methods can be linked to the phenomena of concealment they aim to uncover.

There is another lack and it is theoretical and methodological in nature. A lot of the work done so far has focused on the technical aspect, like a high accuracy benchmarked against standard datasets, but has left many details on the clinical validity, cross-cultural adaptability aspects, and the ethical issues of inferring what someone has chosen to remain silent about (Fritz et al., 2025; Albikawi et al., 2025) uncovered. On the other hand AI-aided care can be a double-edged sword since people's mental health and their views significantly influence their acceptance leading to the trustworthiness, consent, and risk of surveillance aspects being questioned (Fritz et al., 2025; Savir et al., 2025). Only limited efforts exist to prospectively authenticate detection methods in real life clinical settings or to deal with privacy issues when implementing algorithms and hence leaving a chasm between the idealistic view of what the algorithm does versus the responsible use of it.

Bridging these gaps is quite a challenging task but of course the stakes couldn't be higher. Once AI systems become a part of the routine screening, triage and monitoring processes they have the capacity to reveal someone's hidden distress without being the person themselves and as a consequence who gets the assistance, who is referred and who misses out or gets misclassified altogether will be decided at least partly by such inference mechanisms (Peng et al., 2025; Pisani et al., 2024; Wen et al., 2023). A well-developed overview of the available evidence and a discussion of its strengths and weaknesses is crucial in helping clinical experts, software developers and regulators find the way towards valid, ethical deployment only after they fully appreciate the potential and pitfalls at stake from AI. To that end this review paper presents an analysis of research papers on AI-based psychological non-disclosure detection and categorizes them based on the method of detection and hidden condition targeted in the study, evaluates them for quality and suggests what next steps, for the field to align with clinical practice and build AI systems one can trust, would be most reasonable.

Method

Research Design and Framework

A systematic literature review (SLR) was adopted to identify, appraise, and synthesize empirical evidence in a transparent and reproducible manner. The SLR method is appropriate where a body of research is dispersed and methodologically heterogeneous, enabling rigorous aggregation while limiting reviewer bias (Tranfield et al., 2003; Liberati et al., 2009). Reporting followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement, which specifies standards for search documentation, screening, and flow reporting (Page et al., 2021; Moher et al., 2009).

Search Strategy

Prior to constructing the search string, the research questions were decomposed using the PICOS framework (Population, Intervention/Interest, Comparison, Outcome, Study design) to ensure systematic identification of key concepts (Table 1). Each PICOS element was translated into a cluster of synonyms and combined with Boolean operators.

Table 1. <PICOS framework guiding the search strategy>

Element	Definition in this review
Population (P)	Individuals engaging with counseling or mental health services and communities, across clinical and non-clinical settings and age groups.
Intervention/Interest (I)	AI, machine learning, and computational methods (NLP and text mining, speech/acoustic analysis, facial micro-expression, physiological/EEG signals, large language models, predictive ML) used to detect non-disclosed, concealed, latent, or under-reported psychological states.
Comparison (C)	Human clinician judgment, standard self-report, or usual care where reported; a comparator was not required for inclusion.
Outcome (O)	Detection/prediction performance, identification of concealed or latent states, disclosure or help-seeking behavior, and clinical utility.
Study design (S)	Empirical peer-reviewed studies (quantitative, qualitative, mixed methods) published 2020–2025 in English.

The Boolean search string combined three concept clusters-concealment, artificial intelligence, and mental health context-using field codes restricting matches to title, abstract, and keywords (TITLE-ABS-KEY in Scopus, with equivalent field tags in the other databases):

("non-disclosure" OR "nondisclosure" OR conceal OR hidden OR latent OR "under-report*" OR alexithymi* OR "self-disclosure" OR "help-seeking") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "natural language processing" OR "large language model*" OR "neural network*" OR NLP OR chatbot OR "digital phenotyp*") AND ("mental health" OR counsel* OR psychiatr* OR psycholog* OR depress* OR anxiet* OR suicid* OR "self-injur*" OR distress)*

Truncation (e.g., conceal*) captured morphological variants such as conceal, concealed, and concealment, while wildcards broadened coverage of spelling variants. Limiters restricted results to peer-reviewed articles, English language, and the 2020–2025 window to reflect the recent maturation of large language models and deep-learning methods. The string was piloted and refined iteratively: an initial version yielded excessive intervention-only records, prompting the addition of concealment-specific terms to sharpen relevance.

Database and Information Sources

Three databases were searched as information sources: Scopus (primary), PubMed, and ScienceDirect, selected for complementary coverage of multidisciplinary, biomedical, and engineering literature respectively. The Scopus export served as the authoritative bibliographic evidence base for screening and data extraction; PubMed and ScienceDirect provided supplementary identification. Searches were executed and de-duplicated prior to screening.

Eligibility Criteria

Table 2. <Inclusion and exclusion criteria>

Criterion	Inclusion	Exclusion
Language	English	Non-English
Document type	Peer-reviewed journal article	Conference paper, book chapter, editorial, preprint
Publication period	2020–2025	Before 2020
Focus	Uses AI/computational methods to detect non-disclosed, concealed, latent, or under-reported psychological states	No detection component, or intervention-only without detection focus
Population/context	Counseling or mental health context (clinical or community)	Unrelated disciplines or non-psychological outcomes
Accessibility	Full text retrievable	Full text not retrievable

Study Selection Process

Selection proceeded in three stages. First, titles and abstracts of de-duplicated records were screened against the eligibility criteria. Second, full texts of potentially eligible reports were retrieved and assessed in detail. Third, studies meeting all criteria were retained for synthesis. Records excluded at the full-text stage were documented with reasons. Discrepancies during screening were resolved through discussion and re-examination of the source record to reach consensus.

Quality Assessment

Methodological quality was appraised using the Mixed Methods Appraisal Tool (MMAT), version 2018, which accommodates the heterogeneous designs anticipated in this literature-quantitative descriptive, quantitative non-randomized, quantitative randomized, qualitative, and mixed methods (Hong et al., 2018). Each study was first classified by design; the corresponding five-item MMAT checklist was then applied (rated Yes = 1, Can't tell/Partial = 0.5, No = 0). Studies scoring ≥ 60% (≥ 3 of 5 criteria) were regarded as methodologically adequate. Consistent with MMAT guidance, quality scores were not used as an automatic exclusion cutoff; lower-confidence studies were retained but weighted accordingly in synthesis. A summary appears in Table 3.

Table 3. <MMAT quality assessment summary of included studies>

No	Author(s), Year	Study design	Q1	Q2	Q3	Q4	Q5	Total	Category
1	Magnani et al. (2023)	Quant. descriptive	Y	Y	Y	CT	Y	4.5	High
2	Singh & Malik (2025)	Review (JBI)	Y	Y	CT	Y	Y	4.5	High
3	Kim & Lee (2025)	Quant. descriptive	Y	Y	Y	Y	Y	5.0	High
4	Meier (2025)	Mixed methods	Y	CT	Y	CT	Y	4.0	High
5	Zeiler et al. (2025)	Quant. non-randomized	Y	Y	Y	Y	Y	5.0	High
6	Zhang & Zhang (2025)	Quant. descriptive	Y	Y	CT	Y	Y	4.5	High
7	Zhou et al. (2024)	Quant. descriptive	Y	Y	Y	Y	Y	5.0	High
8	Surber et al. (2024)	Quant. non-randomized	Y	Y	Y	Y	CT	4.5	High
9	Nagaoka et al. (2024)	Quant. descriptive	Y	Y	Y	Y	Y	5.0	High
10	Ronneberg et al. (2024)	Quant. randomized	Y	Y	CT	Y	Y	4.5	High
11	Reyner-Fuentes et al. (2025)	Quant. descriptive	Y	Y	Y	CT	Y	4.5	High
12	Taiwo & Al-Bander (2025)	Quant. descriptive	Y	CT	Y	CT	Y	4.0	High
13	Fennig et al. (2025)	Quant. descriptive	Y	Y	Y	Y	Y	5.0	High
14	Warner et al. (2021)	Quant. randomized	Y	Y	Y	Y	Y	5.0	High

No	Author(s), Year	Study design	Q1	Q2	Q3	Q4	Q5	Total	Category
15	Feldhege et al. (2021)	Quant. descriptive	Y	Y	Y	CT	Y	4.5	High
16	Oppenheimer et al. (2021)	Quant. descriptive	Y	CT	Y	Y	CT	4.0	High
17	Yang et al. (2021)	Mixed methods	Y	Y	CT	Y	CT	4.0	High
18	Maes (2022)	Review (JBI)	Y	CT	Y	CT	Y	4.0	High
19	Jungilligens et al. (2020)	Quant. non-randomized	Y	Y	Y	Y	Y	5.0	High
20	Scholich et al. (2025)	Mixed methods	Y	Y	Y	CT	Y	4.5	High

Note. Y = Yes (1.0); CT = Can't tell/Partial (0.5); N = No (0). Total = score out of 5. Category: High ≥ 4.0; Medium 3.0–3.5; Low < 3.0. Reviews were appraised with the equivalent JBI checklist and mapped onto the five-item scale.

Data Extraction Procedure

A structured extraction form captured, for each study: author(s) and year, country, study design, sample and data source, detection modality and technology, target hidden/latent state, outcome measures, and key findings. Extraction was performed directly from the source records to preserve fidelity, and entries were cross-checked against the original abstracts.

Network and Bibliometric Analysis Methodology

Descriptive bibliometric characterization complemented the thematic synthesis. Included studies were profiled by publication year (Figure 2), detection modality (Figure 3), and targeted concealed or latent state (Figure 4). These distributions were derived solely from the extracted study characteristics and are used to describe the composition of the evidence base rather than to make inferential claims.

Data Analysis and Synthesis

Findings were integrated using thematic synthesis, which supports the aggregation of heterogeneous quantitative, qualitative, and mixed-methods evidence (Thomas & Harden, 2008). Coding proceeded inductively from the extracted findings, generating descriptive themes that were then organized into analytical clusters aligned with the research questions. Themes were reviewed iteratively against the primary studies to ensure they were grounded in the source evidence.

Reporting and Documentation

The review was conducted and reported in accordance with the PRISMA 2020 checklist and flow requirements (Page et al., 2021). The complete selection pathway, including counts and exclusion reasons, is documented in the PRISMA 2020 flow diagram (Figure 1).

Results and Discussions

Study Selection Results

The database searches identified 791 records (Scopus n = 755; PubMed n = 22; ScienceDirect n = 14). After removing 15 duplicate records, 776 records were screened at the title and abstract level, of which 712 were excluded as irrelevant to AI-based detection of non-disclosed states. The remaining 64 reports were assessed in full text, and 44 were excluded for the following reasons: no AI or computational detection method (n = 17); not targeting non-disclosed, hidden, or latent states (n = 13); intervention-only without a detection focus (n = 9); and full text not retrievable (n = 5). Twenty studies met all eligibility criteria and were included in the synthesis. The complete selection pathway is depicted in Figure 1.

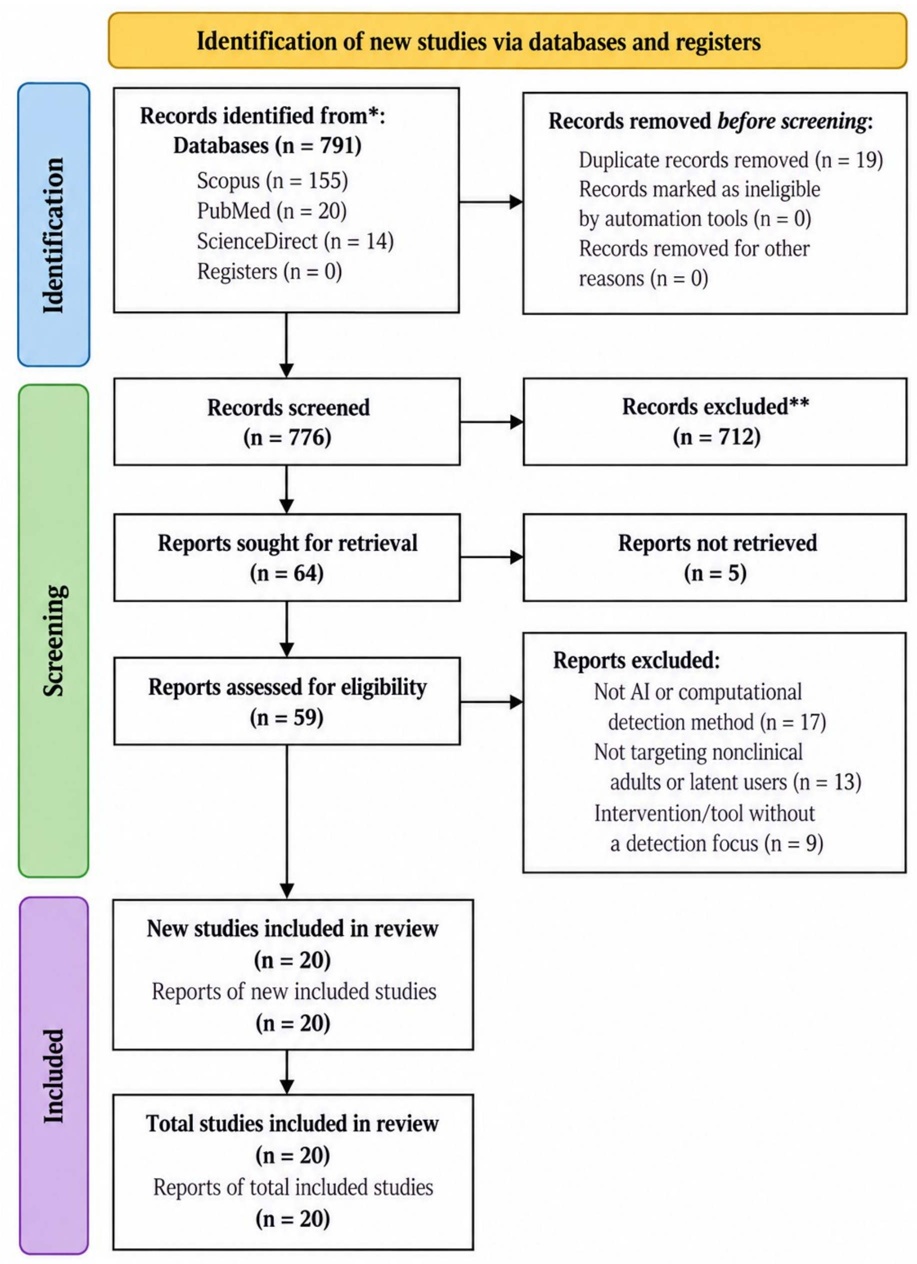

Figure 1 <PRISMA 2020 flow diagram of the study selection process (adapted from Page et al., 2021).>

Descriptive Characteristics

The 20 included studies were published between 2020 and 2025, with a marked concentration in 2024–2025 (13 of 20 studies), reflecting the recent acceleration of deep-learning and large language model research (Figure 2). Study designs spanned quantitative descriptive, quantitative non-randomized, randomized, mixed-methods, and review formats, and all met the MMAT adequacy threshold. Full characteristics of each study are presented in Table 4, and their thematic and methodological classification in Table 5.

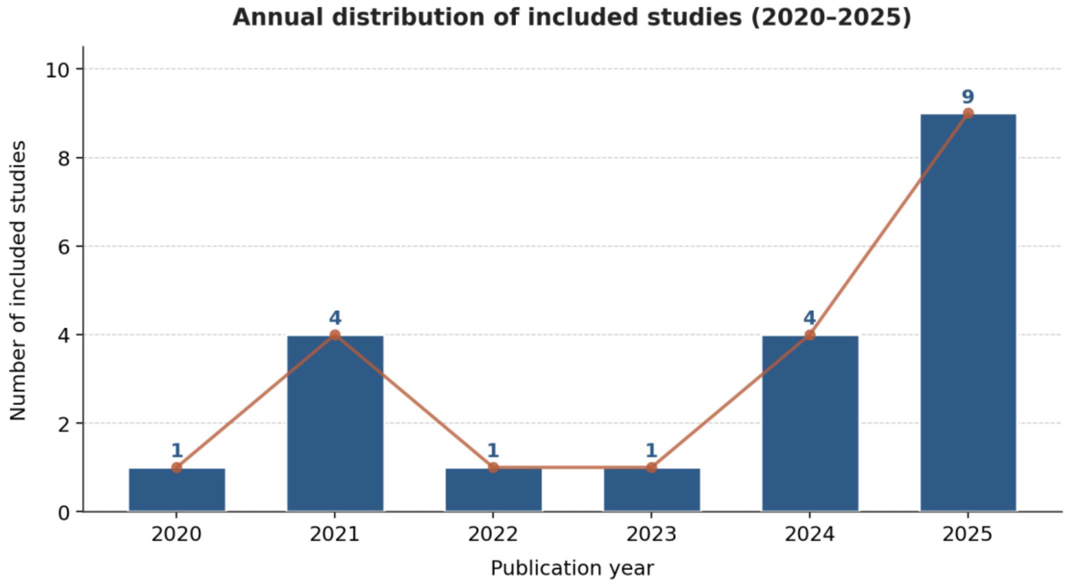

Figure 2 <Annual distribution of the 20 included studies (2020–2025).>

Table 4. <Summary of included studies>

No	Author(s)	Year	Country	Method / modality	Full title	Key findings
1	Magnani et al.	2023	Italy	NLP / automated linguistic profiling	Linguistic profile automated characterisation in pluripotential clinical high-risk mental state (CHARMS) conditions: methodology of a multicentre observational study	Establishes an NLP-based methodology to characterise language markers of clinical high-risk mental states, enabling automated detection of early, often unrecognised psychopathology.
2	Singh & Malik	2025	India	Facial micro-expression recognition (review)	Unveiling hidden emotions: a review of microexpression recognition, classification, and datasets	Synthesises micro-expression recognition methods that reveal genuine emotions people attempt to mask, and identifies gaps in datasets, robustness, and multimodal integration.
3	Kim & Lee	2025	South Korea	Topic modelling + machine learning (text)	Development and validation of an automated machine for self-injury assessment via young Koreans' natural writings	The K-SITR tool analyses natural online writing to assess self-injury severity, addressing concealment among youth who hide their self-injurious identity.
4	Meier	2025	USA	Large language model (ChatGPT)	Initial evaluation of an AI-augmented progress monitoring and outcome assessment	ChatGPT analysis of unstructured session notes generated clinically relevant progress themes, offering a low-burden route to outcome monitoring that clinicians often omit.
5	Zeiler et al.	2025	Austria	Emotion-recognition assessment (longitudinal)	Alexithymia and Emotion Recognition Over the Treatment Course in Adolescents and Emerging Adults With Anorexia Nervosa	Alexithymia and facial emotion recognition changed over treatment and independently predicted eating-disorder outcome, marking difficulty in identifying and disclosing emotion.

No	Author(s)	Year	Country	Method / modality	Full title	Key findings
6	Zhang & Zhang	2025	China	EEG-based prediction model	Emotion dysregulation as a marker in adolescent mental health with EEG-based prediction model	An EEG-based model integrated with psychosocial risk indicators detected emotional dysregulation, providing a physiological marker of latent adolescent mental-health risk.
7	Zhou et al.	2024	China	Regression + random forest (ML)	Prediction of non-suicidal self-injury in adolescents at the family level using regression methods and machine learning	Machine-learning models predicted adolescent non-suicidal self-injury from family-level factors, improving identification of risk not evident from direct report.
8	Surber et al.	2024	Germany	Eye-tracking / attention paradigm	Deployment of attention to facial expressions varies as a function of emotional quality-but not in alexithymic individuals	Alexithymic individuals showed altered spontaneous attention to emotional faces, revealing measurable processing signatures of impaired emotional awareness.
9	Nagaoka et al.	2024	Japan	Deep clustering (deep learning)	Identify adolescents' help-seeking intention on suicide through self- and caregiver's assessments of psychobehavioral problems: deep clustering of the Tokyo TEEN Cohort study	Deep clustering of self- and caregiver reports identified adolescent subgroups differing in suicidal help-seeking intention, surfacing concealed risk profiles.
10	Ronneberg et al.	2024	USA	Voice-enabled AI counselor	Study of a PST-trained voice-enabled artificial intelligence counselor for adults with emotional distress (SPEAC-2): Design and methods	A voice-based AI coach (Lumen) delivering problem-solving therapy targets emotional distress at scale; a pilot showed changes in cognitive-control neuroimaging markers.
11	Reyner-Fuentes et al.	2025	Spain	Speech ML + machine unlearning	Machine Unlearning for Speaker-Agnostic Detection of Gender-Based Violence Condition in Speech	A speaker-agnostic model extracts paralinguistic trauma biomarkers to detect gender-based-violence condition from speech, enabling privacy-preserving, non-invasive screening.
12	Taiwo & Al-Bander	2025	UK	BERT + GPT chatbot	Emotion-aware psychological first aid: Integrating BERT-based emotional distress detection with Psychological First Aid-Generative Pre-Trained Transformer chatbot for mental health support	A BERT distress-detection module coupled to a PFA-GPT chatbot delivers emotion-aware first aid, recognising distress that generic chatbots miss.
13	Fennig et al.	2025	Israel	Large language models (topic/affect mining)	Bridging the conversational gap in epilepsy: Using large language models to reveal insights into patient behavior and concerns from online discussions	LLM analysis of online posts surfaced concerns, emotional distress, and suicidal ideation rarely raised in clinical visits, revealing an unspoken patient agenda.
14	Warner et al.	2021	Canada	Risk-prediction algorithm (RCT)	Self-help behaviors partially mediate the relationship	Disclosing algorithm-derived personalized depression risk

No	Author(s)	Year	Country	Method / modality	Full title	Key findings
					between personalized depression risk disclosure and psychological distress: A mediation analysis using data from a randomized controlled trial	influenced distress partly through self-help behaviours, showing how risk feedback shapes disclosure and coping.
15	Feldhege et al.	2021	Germany	NLP / linguistic style analysis	Changes in language style and topics in an online eating disorder community at the beginning of the COVID-19 Pandemic: Observational Study	Linguistic-style shifts in an online eating-disorder community tracked rising distress during COVID-19, demonstrating language as a passive marker of concealed symptoms.
16	Oppenheimer et al.	2021	USA	LASSO regression (ML) + neuroimaging	Informing the study of suicidal thoughts and behaviors in distressed young adults: The use of a machine learning approach to identify neuroimaging, psychiatric, behavioral, and demographic correlates	Machine learning identified combinations of neuroimaging, psychiatric, and behavioural correlates of suicidal thoughts, improving detection of concealed suicidality.
17	Yang et al.	2021	South Korea	Machine-learning text analysis (mixed)	Factors to improve distress and fatigue in Cancer survivorship; further understanding through text analysis of interviews by machine learning	ML text analysis of survivor interviews extracted under-reported distress and fatigue drivers, complementing survey scores with narrative-derived needs.
18	Maes	2022	Thailand	Precision-medicine ML modelling (review)	Precision Nomothetic Medicine in Depression Research: A New Depression Model, and New Endophenotype Classes and Pathway Phenotypes, and A Digital Self	Proposes ML-derived endophenotype classes and a “digital self” to model latent depression subtypes beyond overt symptom report, guiding precision detection.
19	Jungilligens et al.	2020	Germany	Neuropsychological assessment battery	Impaired emotional and behavioural awareness and control in patients with dissociative seizures	Patients with dissociative seizures showed impaired emotional awareness, interoception, and agency, objectively indexing under-recognised affective processing deficits.
20	Scholich et al.	2025	USA	Large language model chatbots	A Comparison of Responses from Human Therapists and Large Language Model-Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study	Comparing chatbot and therapist responses exposed chatbots’ strengths and safety limitations in recognising and responding to disclosed and implied distress.

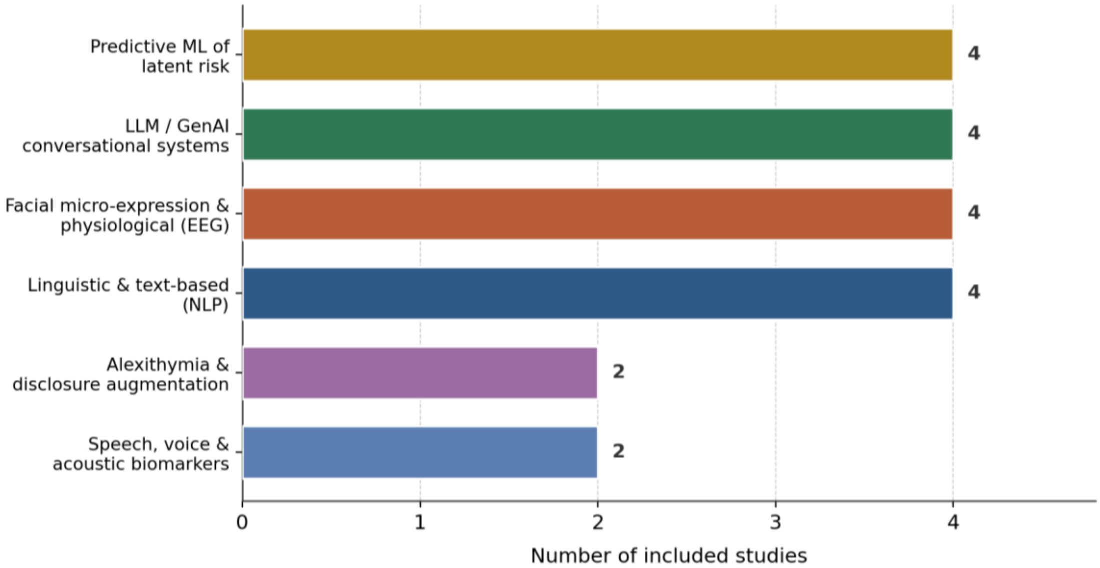

Figure 3. <Distribution of included studies by AI detection modality.>

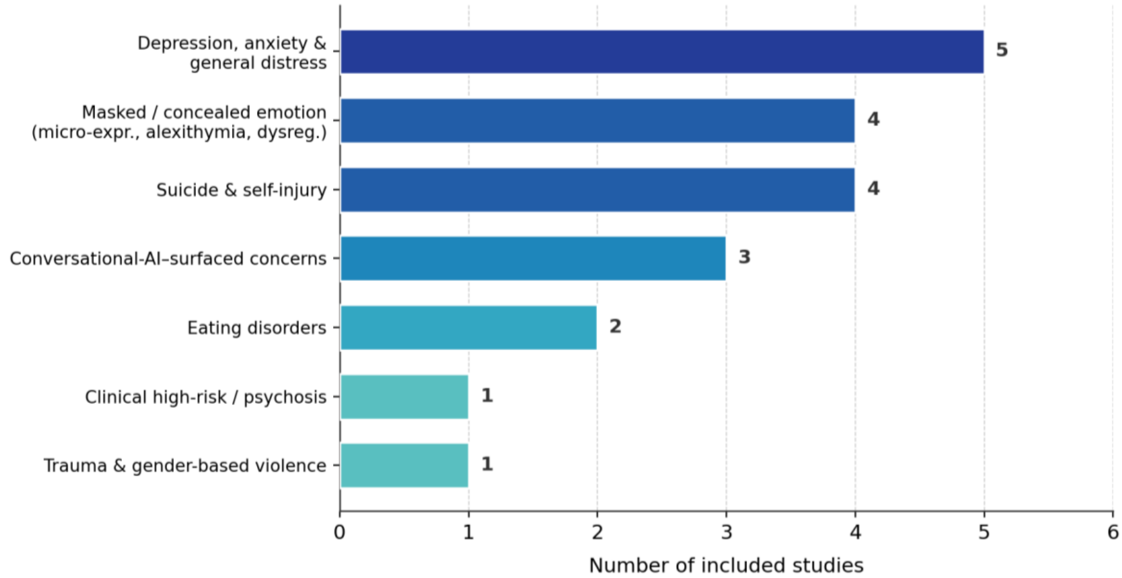

Figure 4. <Distribution of included studies by targeted non-disclosed or latent psychological state.>

Table 5. <Study classification by theme, method, and outcome>

No	Author(s)	Year	Country	Research design	Detection theme	Technology	Outcome/target
1	Magnani et al.	2023	Italy	Quant. descriptive	Linguistic & text-based (NLP)	NLP linguistic profiling	Detection of clinical high-risk state
2	Kim & Lee	2025	South Korea	Quant. descriptive	Linguistic & text-based (NLP)	LDA + ML (K-SITR)	Self-injury severity from writing
3	Feldhege et al.	2021	Germany	Quant. descriptive	Linguistic & text-based (NLP)	Linguistic-style analysis	Eating-disorder distress markers
4	Yang et al.	2021	South Korea	Mixed methods	Linguistic & text-based (NLP)	ML text analysis	Under-reported survivor distress
5	Ronneberg et	2024	USA	Quant.	Speech, voice &	Voice-enabled AI	Distress reduction

No	Author(s)	Year	Country	Research design	Detection theme	Technology	Outcome/target
	al.			randomized	acoustic	(PST)	/ detection
6	Reyner-Fuentes et al.	2025	Spain	Quant. descriptive	Speech, voice & acoustic	Speech ML + unlearning	GBV condition from speech
7	Singh & Malik	2025	India	Review	Facial & physiological	Micro-expression recognition	Masked-emotion detection
8	Surber et al.	2024	Germany	Quant. non-randomized	Facial & physiological	Eye-tracking attention	Alexithymic processing signature
9	Zhang & Zhang	2025	China	Quant. descriptive	Facial & physiological	EEG prediction model	Emotion-dysregulation marker
10	Jungilligens et al.	2020	Germany	Quant. non-randomized	Facial & physiological	Neuropsych. battery	Impaired emotional awareness
11	Meier	2025	USA	Mixed methods	LLM / GenAI conversational	ChatGPT note analysis	Progress/outcome surfacing
12	Taiwo & Al-Bander	2025	UK	Quant. descriptive	LLM / GenAI conversational	BERT + PFA-GPT	Emotion-aware distress detection
13	Fennig et al.	2025	Israel	Quant. descriptive	LLM / GenAI conversational	LLM topic/affect mining	Unspoken patient concerns
14	Scholich et al.	2025	USA	Mixed methods	LLM / GenAI conversational	LLM chatbot comparison	Therapeutic-response quality
15	Zhou et al.	2024	China	Quant. descriptive	Predictive ML of latent risk	Regression + random forest	NSSI risk prediction
16	Nagaoka et al.	2024	Japan	Quant. descriptive	Predictive ML of latent risk	Deep clustering	Suicidal help-seeking profiles
17	Oppenheimer et al.	2021	USA	Quant. descriptive	Predictive ML of latent risk	LASSO regression	Suicidality correlates
18	Maes	2022	Thailand	Review	Predictive ML of latent risk	Precision-medicine ML	Latent depression subtypes
19	Zeiler et al.	2025	Austria	Quant. non-randomized	Alexithymia & disclosure	Emotion-recognition assessment	Alexithymia & ED outcome
20	Warner et al.	2021	Canada	Quant. randomized	Alexithymia & disclosure	Risk-prediction algorithm	Risk disclosure & self-help

Thematic Synthesis

Findings for RQ1: AI modalities and methods for detecting non-disclosure

The included studies employed six distinct detection modalities (Figure 3). Linguistic and text-based analysis was the most common, using natural language processing to extract markers of concealed distress from clinical language, online writing, and interview transcripts (Magnani et al., 2023; Kim & Lee, 2025; Feldhege et al., 2021; Yang et al., 2021). These approaches share a premise that language carries involuntary traces of psychological state even when disclosure is guarded: automated linguistic profiling flagged clinical high-risk conditions (Magnani et al., 2023), topic modelling quantified self-injury severity from natural writing (Kim & Lee, 2025), and shifts in linguistic style tracked rising eating-disorder distress during the pandemic (Feldhege et al., 2021).

Beyond text, four further modalities were evident. Speech and acoustic methods detected trauma-related and emotional-distress signatures from paralinguistic features (Reyner-Fuentes et al., 2025; Ronneberg et al., 2024). Facial micro-expression and physiological approaches targeted affect that individuals attempt to mask, using micro-expression recognition, eye-tracking, EEG, and neuropsychological assessment (Singh & Malik, 2025; Surber et al., 2024; Zhang & Zhang, 2025; Jungilligens et al., 2020). Large language model and generative systems mined conversation and notes to surface concerns not explicitly raised (Fennig et al., 2025;

Meier, 2025; Taiwo & Al-Bander, 2025; Scholich et al., 2025). Finally, predictive machine learning modelled latent risk-particularly suicidality and self-injury-by integrating heterogeneous predictors (Zhou et al., 2024; Nagaoka et al., 2024; Oppenheimer et al., 2021; Maes, 2022). Across modalities, deep learning and transformer architectures predominated in the most recent studies, signalling a methodological shift toward multimodal and representation-learning techniques.

Findings for RQ2: Targeted states and detection performance

The most frequently targeted states were depression, anxiety, and general distress, followed by suicide and self-injury, and masked or concealed emotion (Figure 4). Suicidality and self-injury attracted particular attention because concealment in these domains carries the gravest consequences: models predicted non-suicidal self-injury from family-level data (Zhou et al., 2024), identified adolescent subgroups by suicidal help-seeking intention (Nagaoka et al., 2024), and isolated multimodal correlates of suicidal thoughts in distressed young adults (Oppenheimer et al., 2021). Concealed and difficult-to-articulate emotion was targeted through micro-expression, physiological, and alexithymia-focused work (Singh & Malik, 2025; Zhang & Zhang, 2025; Zeiler et al., 2025; Surber et al., 2024).

Where quantitative performance was reported, results were promising but heterogeneous. Text- and topic-based tools achieved workable classification of self-injury severity (Kim & Lee, 2025), speech models extracted robust trauma biomarkers under speaker-agnostic constraints (Reyner-Fuentes et al., 2025), and LLM analyses recovered clinically salient themes-including suicidal ideation-rarely surfaced in consultations (Fennig et al., 2025). However, outcome metrics varied widely, from accuracy and consistency figures to predictive correlates and neuroimaging change, and few studies benchmarked against clinician judgment or validated prospectively. Clinical utility was therefore demonstrated more often in principle than in deployed practice, with several contributions still at the design, protocol, or feasibility stage (Ronneberg et al., 2024; Meier, 2025).

Findings for RQ3: Clinical, ethical, and implementation implications

The reviewed evidence points to a clear clinical opportunity: AI can act as a complementary sensor that surfaces distress patients do not, or cannot, disclose, extending screening into online spaces, routine notes, and daily interaction (Fennig et al., 2025; Meier, 2025; Taiwo & Al-Bander, 2025). Privacy-preserving designs-such as speaker-agnostic speech modelling-illustrate how sensitive detection might be made safer for vulnerable groups (Reyner-Fuentes et al., 2025). Yet the same capabilities raise ethical tensions. Inferring what a person has deliberately withheld implicates consent, autonomy, and the risk of surveillance, and comparisons of chatbots with therapists exposed safety limitations in how automated systems handle disclosed and implied risk (Scholich et al., 2025). Implementation is further constrained by fragmented validation, heterogeneous outcomes, and limited integration into clinical workflows, underscoring that responsible adoption depends as much on governance as on model accuracy.

Comparative and Critical Analysis

Methodologically, the corpus was dominated by quantitative descriptive designs, with fewer randomized or prospective clinical evaluations (Table 5). This distribution reveals a field advancing rapidly on technical feasibility while lagging on clinical validation. The temporal pattern is instructive: earlier studies (2020–2022) relied on regression, neuropsychological, and conventional machine-learning approaches (Jungilligens et al., 2020; Oppenheimer et al., 2021; Maes, 2022), whereas recent studies increasingly adopt transformers, deep clustering, and large language models (Nagaoka et al., 2024; Fennig et al., 2025; Taiwo & Al-Bander, 2025). Multimodal integration remains uncommon; most studies exploit a single channel, leaving the complementary value of combining language, voice, face, and physiology largely untested. Comparability is further limited by inconsistent reporting of performance and the scarcity of shared benchmarks for concealment specifically, as opposed to overt symptom classification.

Interpreted together, these findings indicate that psychological non-disclosure is computationally tractable: concealed and latent states leave detectable traces across multiple channels, and contemporary AI can recover a meaningful portion of them. This reframes a long-standing clinical problem-the dependence on self-report-as an opportunity for augmentation rather than replacement, positioning AI as a second, unobtrusive source of evidence alongside the therapeutic relationship (Aschbacher et al., 2023; Alemu et al., 2024).

Theoretically, the results extend models of help-seeking and disclosure by showing that the barriers documented in the wider literature-stigma, label avoidance, and impaired emotional awareness-have measurable behavioral and linguistic correlates (Miqdadi et al., 2025; Kosyluk et al., 2024; Olatunde et al., 2025). Detection thus operationalizes constructs that were previously accessible only through self-report, and the emergence of endophenotype and "digital self" framings suggests a move toward latent, data-driven characterization of distress (Maes, 2022; Han et al., 2024). Practically, the evidence supports embedding detection within measurement-based and stepped-care systems, where algorithmic flags can trigger clinician review, adapt care level, or prompt outreach (Wen et al., 2023; Pisani et al., 2024; Balcombe & De Leo, 2022).

Situated against prior reviews of digital mental health and conversational agents, which have emphasized intervention efficacy and engagement (Jiang et al., 2025; Keyan et al., 2025; Alvarez-Jimenez et al., 2020), this review foregrounds detection of concealment as a distinct and under-synthesized objective. Contradictions in the literature are notable: some work celebrates the reach of generative chatbots (Li et al., 2025; Kim et al., 2025), while direct comparisons reveal safety and quality gaps relative to human clinicians (Scholich et al., 2025). Acceptance is similarly conditional, shaped by users' mental health status, attitudes, and capacity to engage (Fritz et al., 2025; Savir et al., 2025; Albikawi et al., 2025). These tensions caution against uncritical deployment.

Three research gaps stand out. First, multimodal fusion for concealment detection is largely unexplored, despite evidence that single channels each capture only part of the signal. Second, prospective clinical validation against expert judgment is rare, leaving diagnostic and predictive value uncertain in real settings (Leung et al., 2021; Zainal et al., 2025). Third, ethical and governance frameworks specific to inferring withheld information are underdeveloped, including consent, transparency, and safeguards against misclassification and surveillance (Fritz et al., 2025; Peng et al., 2025).

This review has limitations. The evidence base was restricted to English-language, peer-reviewed articles from 2020–2025 in three databases, which may omit relevant grey literature and non-English work. The construct of non-disclosure is heterogeneous, and mapping studies onto modality and target categories required interpretive judgment. Finally, heterogeneous outcomes precluded quantitative meta-analysis, so synthesis remains narrative. Future research should prioritize multimodal, prospectively validated detection integrated into clinical workflows; standardized reporting and shared benchmarks for concealment; privacy-preserving and explainable models; and co-designed governance that centers patient autonomy (Knippler et al., 2024; Mo et al., 2025).

In summary, and in direct answer to the research questions: AI detects non-disclosure through six converging modalities (RQ1); it most often targets depression, suicidality/self-injury, and masked emotion, with promising but incompletely validated performance (RQ2); and while it offers a powerful complement to self-report, its responsible use depends on validation, transparency, and ethical safeguards (RQ3).

Conclusions

This systematic review synthesized 20 studies on artificial intelligence for detecting psychological non-disclosure and hidden mental states in counseling and mental health care. In answer to the research questions, AI achieves detection through six modalities-linguistic, speech, facial/physiological, large language model, predictive machine learning, and disclosure augmentation (RQ1); it most frequently targets depression and general distress, suicidality and self-injury, and masked emotion, with promising though heterogeneously validated performance (RQ2); and it offers a valuable complement to self-report while raising substantive ethical and implementation challenges (RQ3). The core contribution is an integrative modality-by-target framework that clarifies a fragmented field. Practically, detection can strengthen measurement-based and stepped care when embedded responsibly. Key limitations are the scarcity of prospective clinical validation and multimodal integration, and immature governance. Future work should pursue validated, explainable, privacy-preserving, and ethically co-designed detection systems.

References

Albikawi, Z., Abuadas, M., & Rayani, A. M. (2025). Nursing Students' Perceptions of AI-Driven Mental Health Support and Its Relationship with Anxiety, Depression, and Seeking Professional Psychological Help: Transitioning from Traditional Counseling to Digital Support. Healthcare (Switzerland), 13(9), Article 1089. https://doi.org/10.3390/healthcare13091089

Alemu, Y., Chen, H., Duan, C., Caulley, D., Arriaga, R. I., & Sezgin, E. (2023). Detecting Clinically Relevant Emotional Distress and Functional Impairment in Children and Adolescents: Protocol for an Automated Speech Analysis Algorithm Development Study. JMIR Research Protocols, 12, Article e46970. https://doi.org/10.2196/46970

Alemu, Y., Bautista, E. C., Vinson, S., Ohiomoba, P., Melecio, N., Amare, H., Vivia, J., Karippara, H., & Harris, C. (2024). Implementing And Analyzing The Advantages Of Voice Ai As Measurement-Based Care To Address Behavioral Health Treatment Disparities Among Youth In Economically

Disadvantaged Communities. Telehealth and Medicine Today, 9(4), Article 506. https://doi.org/10.30953/thmt.v9.506

Alvarez-Jimenez, M., Rice, S., D'Alfonso, S., Leicester, S., Bendall, S., Pryor, I., Russon, P., McEnery, C., Santesteban-Echarri, O., Da Costa, G., Gilbertson, T., Valentine, L., Solves, L., Ratheesh, A., McGorry, P. D., & Gleeson, J. (2020). A novel multimodal digital service (moderated online social therapy+) for help-seeking young people experiencing mental ill-health: Pilot evaluation within a national youth e-mental health service. Journal of Medical Internet Research, 22(8), Article e17155. https://doi.org/10.2196/17155

Aschbacher, K., Rivera, L. M., Hornstein, S., Nelson, B. W., Forman-Hoffman, V. L., & Peiper, N. C. (2023). Longitudinal Patterns of Engagement and Clinical Outcomes: Results From a Therapist-Supported Digital Mental Health Intervention. Psychosomatic Medicine, 85(7), 651–658. https://doi.org/10.1097/PSY.0000000000001230

Balcombe, L., & De Leo, D. (2022). Linking music streaming platform advertisements with a digital mental health assessment and interventions. Frontiers in Digital Health, 4, Article 964251. https://doi.org/10.3389/fdgth.2022.964251

Feldhege, J., Moessner, M., Wolf, M., & Bauer, S. (2021). Changes in language style and topics in an online eating disorder community at the beginning of the COVID-19 Pandemic: Observational Study. Journal of Medical Internet Research, 23(7), Article e28346. https://doi.org/10.2196/28346

Fennig, U., Yom-Tov, E., Savitzky, L., Nissan, J., Altman, K., Loebenstein, R., Boxer, M., Weinberg, N., Gofrit, S. G., & Maggio, N. (2025). Bridging the conversational gap in epilepsy: Using large language models to reveal insights into patient behavior and concerns from online discussions. Epilepsia, 66(3), 686–699. https://doi.org/10.1111/epi.18226

Fritz, B., Eppelmann, L., Edelmann, A., Rohrmann, S., & Wessa, M. (2025). How mental health status and attitudes toward mental health shape AI Acceptance in psychosocial care: a cross-sectional analysis. BMC Psychology, 13(1), Article 617. https://doi.org/10.1186/s40359-025-02954-z

Han, N., Wen, Y., Wang, B., Huang, F., Liu, X., Li, L., & Zhu, T. (2024). Developing a machine learning-based instrument for subjective well-being assessment on Weibo and its psychological significance: An evaluative and interpretive research. Applied Psychology: Health and Well-Being, 16(4), 2246–2265. https://doi.org/10.1111/aphw.12590

Hong, Q. N., Pluye, P., Bujold, M., & Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada.

Jiang, X.-H., Yuan, X.-H., Zhao, H., & Peng, J.-S. (2025). The Effectiveness and Feasibility of Conversational Agents in Supporting Care for Patients With Cancer: Systematic Review and Meta-Analysis. Journal of Medical Internet Research, 27, Article e76968. https://doi.org/10.2196/76968

Jungilligens, J., Wellmer, J., Schlegel, U., Kessler, H., Axmacher, N., & Popkirov, S. (2020). Impaired emotional and behavioural awareness and control in patients with dissociative seizures. Psychological Medicine, 50(16), 2731–2739. https://doi.org/10.1017/S0033291719002861

Keyan, D., Hall, J., Jordan, S., Watts, S., Au, T., Dawson, K. S., Sway, R. A., Crawford, J., Sorsdahl, K., Luitel, N. P., de Graaff, A. M., Ghalayini, H., Habashneh, R., El-Dardery, H., Fanatseh, S., Malik, A., Servili, C., Faroun, M., Abualhaija, A., … Carswell, K. (2025). The development of a World Health Organization transdiagnostic chatbot intervention for distressed adolescents and young adults. Frontiers in Digital Health, 7, Article 1528580. https://doi.org/10.3389/fdgth.2025.1528580

Kim, S., & Lee, D.-G. (2025). Development and validation of an automated machine for self-injury assessment via young Koreans' natural writings. PLoS ONE, 20(1), Article e0316619. https://doi.org/10.1371/journal.pone.0316619

Kim, N., Lee, J., Park, S. H., On, Y., Lee, J., Keum, M., Oh, S., Song, Y., Lee, J., Won, G. H., Shin, J. S., Lho, S. K., Hwang, Y. J., & Kim, T.-S. (2025). GPT-4 generated psychological reports in psychodynamic perspective: a pilot study on quality, risk of hallucination and client satisfaction. Frontiers in Psychiatry, 16, Article 1473614. https://doi.org/10.3389/fpsyt.2025.1473614

Knippler, E. T., Martinez, A. J., Amiri, I., Madundo, K., Mmbaga, B. T., Goldston, D. B., Relf, M. V., & Knettel, B. A. (2024). Challenges and opportunities for improving mental health care and preventing suicide among people living with HIV: Perspectives of mental health professionals in Tanzania. PLOS

Global Public Health, 4(2 February), Article e0002762. https://doi.org/10.1371/journal.pgph.0002762

Kosyluk, K., Baeder, T., Greene, K. Y., Tran, J. T., Bolton, C., Loecher, N., DiEva, D., & Galea, J. T. (2024). Mental distress, label avoidance, and use of a mental health chatbot: Results from a us survey. JMIR Formative Research, 8, Article e45959. https://doi.org/10.2196/45959

Lazaridou, A., Sivakumar, S., Biefer, H. R. C., Weilenmann, S., Princip, M., Zuccarella-Hackl, C., Petzschner, F. H., Heinzle, J., Stephan, K. E., Dzemali, O., & von Känel, R. (2025). Predictive modelling of clinically significant depressive symptoms after coronary artery bypass graft surgery: protocol for a multicentre observational study in two Swiss hospitals (the PsyCor study). BMJ Open, 15(9), Article e108061. https://doi.org/10.1136/bmjopen-2025-108061

Leung, Y. W., Wouterloot, E., Adikari, A., Hirst, G., de Silva, D., Wong, J., Bender, J. L., Gancarz, M., Gratzer, D., Alahakoon, D., & Esplen, M. J. (2021). Natural language processing-based virtual cofacilitator for online cancer support groups: Protocol for an algorithm development and validation study. JMIR Research Protocols, 10(1), Article e21453. https://doi.org/10.2196/21453

Li, J., Zhang, T., Cheung, J. P. Y., Meng, N., Wu, Z., Ma, D. C. F., Chan, E. A., Yu, L., Cao, Y., & Li, Y. (2025). Chatbot-delivered structured psychological intervention (SPI-Bot) for teenagers with adolescent idiopathic scoliosis in Hong Kong: protocol for a pilot randomised controlled trial. BMJ Open, 15(7), Article e098734. https://doi.org/10.1136/bmjopen-2025-098734

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions. PLoS Medicine, 6(7), Article e1000100. https://doi.org/10.1371/journal.pmed.1000100

Maes, M. (2022). Precision Nomothetic Medicine in Depression Research: A New Depression Model, and New Endophenotype Classes and Pathway Phenotypes, and A Digital Self. Journal of Personalized Medicine, 12(3), Article 403. https://doi.org/10.3390/jpm12030403

Magnani, L., Carmisciano, L., Dell'orletta, F., Bettinardi, O., Chiesa, S., Imbesi, M., Limonta, G., Montagna, E., Turone, I., Martinasso, D., Aguglia, A., Serafini, G., Amore, M., Amerio, A., Costanza, A., Sibilla, F., Calcagno, P., Patti, S., Molino, G., … Saginario, A. (2023). Linguistic profile automated characterisation in pluripotential clinical high-risk mental state (CHARMS) conditions: methodology of a multicentre observational study. BMJ Open, 13(3), Article e066642. https://doi.org/10.1136/bmjopen-2022-066642

Meier, S. T. (2025). Initial evaluation of an AI-augmented progress monitoring and outcome assessment. Counselling and Psychotherapy Research, 25(1), Article e12869. https://doi.org/10.1002/capr.12869

Miqdadi, A. I., Alhalabi, M. N., Alhadidi, M., Sinnokrot, S., Abdallah, F. M., AbuQamar, Q. A., Aldaieflih, M. M., Mahmoud, K. F., Abu-Shennar, J. A., Younis, M., Alhusamiah, B. K., & Aldiqs, M. A. (2025). Attitudes Toward Seeking Professional Psychological Help, Self-Stigma of seeking help, and Preferences for AI Mental Health Chatbots vs. Human Counsellors: A Mixed Methods Study. International Journal of Mental Health and Addiction. https://doi.org/10.1007/s11469-025-01595-y

Mo, L., Li, H., Zhang, Y., Li, A., Xiong, Z., Cun, P., & Zhu, T. (2025). Self-help psychological intervention for young individuals during the post-COVID-19 era: development of a PST chatbot using GPT-4. Frontiers in Digital Health, 7, Article 1627268. https://doi.org/10.3389/fdgth.2025.1627268

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Annals of Internal Medicine, 151(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Nagaoka, D., Uno, A., Usami, S., Tanaka, R., Minami, R., Sawai, Y., Okuma, A., Yamasaki, S., Miyashita, M., Nishida, A., Kasai, K., & Ando, S. (2024). Identify adolescents' help-seeking intention on suicide through self- and caregiver's assessments of psychobehavioral problems: deep clustering of the Tokyo TEEN Cohort study. The Lancet Regional Health - Western Pacific, 43, Article 100979. https://doi.org/10.1016/j.lanwpc.2023.100979

Olatunde, O. E., Richards, D., Dobbs, P. D., Nahar, V. K., Sharma, M., & Davis, R. E. (2025). Psychological Help-Seeking Among College Students: Applying the Multi-Theory Model of Health Behavior Change in Assessing Telehealth Use for Psychological Help-Seeking. American Journal of Lifestyle Medicine, 19(3), 450–462. https://doi.org/10.1177/15598276221116561

Oppenheimer, C. W., Bertocci, M., Greenberg, T., Chase, H. W., Stiffler, R., Aslam, H. A., Lockovich, J., Graur, S., Bebko, G., & Phillips, M. L. (2021). Informing the study of suicidal thoughts and behaviors in distressed young adults: The use of a machine learning approach to identify neuroimaging, psychiatric, behavioral, and demographic correlates. Psychiatry Research - Neuroimaging, 317, Article 111386. https://doi.org/10.1016/j.pscychresns.2021.111386

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71

Peng, Z., Su, X., & Hou, Y. (2025). How do GenAI and social media provide psychosocial information support for cancer survivors and professionals? An analysis using structural topic modeling for exploration and comparison. European Journal of Oncology Nursing, 79, Article 103018. https://doi.org/10.1016/j.ejon.2025.103018

Pisani, A. R., Wyman, P. A., Cero, I., Kelberman, C., Gurditta, K., Judd, E., Schmeelk-Cone, K., Mohr, D., Goldston, D., & Ertefaie, A. (2024). Text Messaging to Extend School-Based Suicide Prevention: Pilot Randomized Controlled Trial. JMIR Mental Health, 11, Article e56407. https://doi.org/10.2196/56407

Reyner-Fuentes, E., Rituerto-González, E., & Peláez-Moreno, C. (2025). Machine Unlearning for Speaker-Agnostic Detection of Gender-Based Violence Condition in Speech. Applied Sciences (Switzerland), 15(22), Article 12270. https://doi.org/10.3390/app152212270

Ronneberg, C. R., Lv, N., Ajilore, O. A., Kannampallil, T., Smyth, J., Kumar, V., Barve, A., Garcia, C., Dosala, S., Wittels, N., Xiao, L., Aborisade, G., Zhang, A., Tang, Z., Johnson, J., & Ma, J. (2024). Study of a PST-trained voice-enabled artificial intelligence counselor for adults with emotional distress (SPEAC-2): Design and methods. Contemporary Clinical Trials, 142, Article 107574. https://doi.org/10.1016/j.cct.2024.107574

Savir, T., & Baumel, A. (2025). Capacity to Invest Effort as a Predictor of Preference for Digital Mental Health Interventions Over Psychotherapy: Cross-Sectional Study Using an Ecological Digital Screening Tool. Journal of Medical Internet Research, 27(1), Article e77802. https://doi.org/10.2196/77802

Scholich, T., Barr, M., Stirman, S. W., & Raj, S. (2025). A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. JMIR Mental Health, 12, Article e69709. https://doi.org/10.2196/69709

Singh, J., & Malik, P. (2025). Unveiling hidden emotions: a review of microexpression recognition, classification, and datasets. Multimedia Tools and Applications, 84(42), 50063–50118. https://doi.org/10.1007/s11042-025-21100-w

Surber, C., Hoepfel, D., Günther, V., Kersting, A., Rufer, M., Suslow, T., & Bodenschatz, C. M. (2024). Deployment of attention to facial expressions varies as a function of emotional quality-but not in alexithymic individuals. Frontiers in Psychiatry, 15, Article 1338194. https://doi.org/10.3389/fpsyt.2024.1338194

Taiwo, O., & Al-Bander, B. (2025). Emotion-aware psychological first aid: Integrating BERT-based emotional distress detection with Psychological First Aid-Generative Pre-Trained Transformer chatbot for mental health support. Cognitive Computation and Systems, 7(1), Article e12116. https://doi.org/10.1049/ccs2.12116

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, Article 45. https://doi.org/10.1186/1471-2288-8-45

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management, 14(3), 207–222. https://doi.org/10.1111/1467-8551.00375

Warner, E., Nannarone, M., Manuel, D., Lashewicz, B., Patten, S., Schmitz, N., & Wang, J. (2021). Self-help behaviors partially mediate the relationship between personalized depression risk disclosure and psychological distress: A mediation analysis using data from a randomized controlled trial. Journal of Psychiatric Research, 140, 7–14. https://doi.org/10.1016/j.jpsychires.2021.05.047

Wen, A., Wolitzky-Taylor, K., Gibbons, R. D., & Craske, M. (2023). A randomized controlled trial on using predictive algorithm to adapt level of psychological care for community college students: STAND triaging and adapting to level of care study protocol. Trials, 24(1), Article 508. https://doi.org/10.1186/s13063-023-07441-7

Yang, K., Kim, J., Chun, M., Ahn, M. S., Chon, E., Park, J., & Jung, M. (2021). Factors to improve distress and fatigue in Cancer survivorship; further understanding through text analysis of interviews by machine learning. BMC Cancer, 21(1), Article 741. https://doi.org/10.1186/s12885-021-08438-8

Zainal, N. H., Tan, H. H., Hong, R. Y. S., & Newman, M. G. (2025). Prescriptive Predictors of Mindfulness Ecological Momentary Intervention for Social Anxiety Disorder: Machine Learning Analysis of Randomized Controlled Trial Data. JMIR Mental Health, 12, Article e67210. https://doi.org/10.2196/67210

Zeiler, M., Wittek, T., Truttmann, S., Klang, J., Kopp, K., Krauss, H., Auer-Welsbach, E., Ohmann, S., Sackl-Pammer, P., Werneck-Rohrer, S., Schmidt, U., Karwautz, A., & Wagner, G. (2025). Alexithymia and Emotion Recognition Over the Treatment Course in Adolescents and Emerging Adults With Anorexia Nervosa. European Eating Disorders Review, 33(6), 1265–1275. https://doi.org/10.1002/erv.70005

Zhang, Z., & Zhang, L. (2025). Emotion dysregulation as a marker in adolescent mental health with EEG-based prediction model. Scientific Reports, 15(1), Article 38178. https://doi.org/10.1038/s41598-025-22067-2

Zhou, S. C., Zhou, Z., Tang, Q., Yu, P., Zou, H., Liu, Q., Wang, X. Q., Jiang, J., Zhou, Y., Liu, L., Yang, B. X., & Luo, D. (2024). Prediction of non-suicidal self-injury in adolescents at the family level using regression methods and machine learning. Journal of Affective Disorders, 352, 67–75. https://doi.org/10.1016/j.jad.2024.02.039