Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Clinicians' Practices and Attitudes Toward Depression Assessment: A Cross-Sectional Survey Study

Damon Navandi
damon.navandi@gmail.com
Lund University
Veerle C. Eijsbroek
Lund University
Clara Wiebel
Lund University
Katarina Kjell
Lund University
Marc L. Molendijk
Leiden University
Elizabeth C. Stade
Stanford University
Oscar N. E. Kjell
Lund University

Research Article

Keywords: clinical assessment, depression, structured methods, statistical models, artificial intelligence

Posted Date: March 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8951240/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. O.N.E. Kjell and K. Kjell co-founded a start-up that uses computational language assessments to assess mental health problems. E.C. Stade has received fees for advising a start-up that uses computational language assessment to measure psychological

constructs. The other authors report having no conflicts of interest with respect to the contents, authorship, or publication of this article.

Version of Record: A version of this preprint was published at BMC Psychology on August 19th, 2026. See the published version at https://doi.org/10.1186/s40359-026-05036-w.

Abstract

Accurate psychological evaluation is crucial for early detection, intervention and evaluation of mental health disorders. The validity and reliability of psychological evaluation methods have been systematically researched for many decades. However, assessment methods used by mental health professionals in clinical settings have rarely been empirically documented. This study examines the extent to which empirically supported assessments are used in clinical settings and clinicians' attitudes toward these practices. Clinicians (*N* = 585) from the U.S. (25%), the U.K. (28%), the Netherlands (18%), Sweden (22%), and other countries (7%) completed an online survey reporting their depression assessment methods. Clinicians reported using many different *data collection methods* (e.g., unstructured clinical interviews [83%], rating scales [83%]) in various combinations. A majority (80%) primarily employed clinical methods, as opposed to statistical *data combination methods* (11%). Although they were confident in their assessments, the time to carry out assessments varied widely (*M* = 87 [*SD* = 108] minutes). Seventy-seven percent agreed on the importance of standardized, valid, and reliable assessments, and 43% reported interest in receiving AI decision-support. We did not find many significant cross-cultural differences. A majority of clinicians have not fully adopted standardized empirically supported assessment practices. Predominant reliance on clinical judgment suggests a disconnect between best practices supported by research and everyday clinical practice.

Introduction

Accurate assessment of mental health disorders is crucial for early detection and effective treatment planning. Inaccurate assessment of the type and severity of mental disorders can have significant negative impacts on people's lives, highlighting the importance of precise assessment methods for effective prevention, early detection, and appropriate treatment.[1–3] There are many different methods used to assess mental health problems, including various *data collection methods* such as interviews and rating scales, as well as various *data combination methods*, i.e., approaches for evaluating, integrating, and interpreting different sources of information (Fig. 1).[4,5]

The validity and reliability of depression assessment methods have been extensively studied,[6,7] with some methods being more reliable and valid than others are. However, less is known about the practices of mental health professionals, and the available evidence suggests a discrepancy between guidelines on standards of practice and those on clinical practice.[8] This study aims to gain insights into the extent to which empirically supported assessments are used by mental health professionals, as well as their attitudes toward these methods, which is critical for identifying risks and challenges in current assessment practices and the implementation of new assessment methods. Our study may also add to the discourse on integrating artificial intelligence-based (AI) decision-support systems into assessment processes.[9–11]

Note

The data combination method does not necessarily require data from different data collection methods; for example, responses to items of a single rating scale can be aggregated and interpreted via a statistical model (following an aggregation schema or using an established cut-off point) or interpreted via a clinical method in which individual items are examined via clinical experience. In the statistical method, it is also feasible to include data from a clinical (unstructured) interview in a statistical model (e.g., to what extent the patient responds, their response times, and behaviors).

Data collection methods

Data collection methods range from *unstructured* to *structured*.[12] *Unstructured data collection methods* include *unstructured clinical interviews*, which allow clinicians to explore patients' issues freely and adjust the direction of the assessment on the basis of patient's responses and behaviors without predetermined questions. *Structured data collection methods* are highly organized and follow a predetermined format, ensuring that the same questions are asked of each person. Structured methods aim to yield more reliable and comparable results and are especially well suited for assessing a set of validated criteria for a disorder or clinical phenomenon (e.g., evaluating whether a patient meets a diagnostic manual's criteria for a major depressive episode). Examples include rating scales and structured clinical interviews, as well as semistructured clinical interviews where questions are predetermined, but the interviewer is allowed to ask additional questions or exclude ones to adjust the sourcing of information on a case-to-case basis. The interviews were conducted by clinicians, while for the rating scales, some were clinician-rated, and others were self-reported.

Data combination methods

Two primary approaches for evaluating, integrating, and interpreting different sources of information have been identified for depression assessment: the *clinical* approach and the *statistical* approach.[6,13,14] *The clinical data combination method* is based on human judgment. Clinicians assess various sources of information on the basis of personal judgment, clinical experience, theoretical perspectives as well as surrounding and individual factors (e.g., subtle deviations from a patient's baseline behavior, appearance, overt signs such as direct statements of distress or crying, or general deduction on the basis of the overall information gathered).

In contrast, *the statistical data combination method* makes use of a statistical model that is based on empirically established relationships between the different sources of information and the outcome of interest. This can, for example, involve the clinician entering client data into formulas, actuarial tables, or charts.[13,14] To establish relationships among variables, statistical models may be constructed via relatively simple statistical approaches such as linear or logistic regression. It can also involve more complex statistical techniques, such as AI, which includes techniques such as machine learning and natural language processing. Statistical data combination methods can, for example, be used for screening, comprehensive diagnostic assessment support, treatment personalization, or symptom monitoring.[15,16]

A growing body of research supports the use of AI in mental health care, demonstrating its ability to improve the accuracy and efficiency of diagnostic assessments.[10,16,17] Given its promising applications and increasing capabilities, we focus on AI as a key component of the statistical approach in the following sections.

Final assessment and decision-making

In the data combination phase, both the clinical and the statistical data combination methods can be used at different stages (see Brunswick Lens Model).[18,19,20] For example, a clinician may use a statistical method to integrate a patient's rating scale responses but rely on a clinical method to combine those results with information from an unstructured interview. However, when the results from the two different methods disagree, one cannot follow both of them: When they *disagree*, the final decision will necessarily have to rely on one method over the other.[14,21]

Navigating and integrating different information via the clinical data combination method can be cognitively strenuous and complex,[22,23] especially when the various sources are in conflict with one another. The clinical method has been shown to be less reliable than statistical models.[13,21] However, the implementation of statistical models can be difficult due to insufficient guidelines, education, or resources, such as a lack of validated advanced statistical models available for practical use in psychological assessments.[11] For example, models that enable the statistical combination of data from several rating scales and clinical history are still uncommon.

Since the introduction of the clinical versus statistical controversy in the 1950s, statistical modeling has undergone significant improvements as a result of advancements in the field of AI.[9,10,24] Statistical models, often including AI technology, have become increasingly prevalent in healthcare, especially in medicine.[25,26] Nevertheless, the question remains whether the benefits of such instruments are used for depression assessments in clinical practice and, if so, to what extent, in what way, and how clinicians' attitudes reflect their tendencies to use the statistical data combination methods.

Research support for assessment methods

The validity and reliability of depression assessment methods have been systematically researched for many decades. Semistructured or structured interviews are often regarded as the benchmark standard for depression assessment because they demonstrate greater validity and reliability than unstructured interviews do.[7,27,28] In contrast, unstructured clinical interviews, although widely used in practice, are less reliable and less valid,[12,29] and they are rarely used in research. Self-report questionnaires, while highly reliable, have other shortcomings. For example, rating scales targeting depression: **i)** fail to reliably capture the disorder across scales,[30] **ii)** miss important symptoms,[31] **iii)** differ substantially in

their categorization of depressed patients into severity groups,[32] and **iv)** are multidimensional, thus multiple constructs with inconsistent factor structures across scales are assessed.[33–35]

Additionally, the statistical method for combining information, which involves the use of algorithms or structured approaches, generally results in a 13% increase in accuracy over the clinical method, which relies on the clinician's judgment.[6,13,24,36] Despite the strong evidence supporting the use of structured assessments and statistical methods, many clinicians continue to rely on less empirically supported practices. This disparity highlights a critical gap between research evidence and clinical practice.

Research on depression assessments in practice

With respect to clinical practice, research has examined how mental health professionals use diagnostic classification systems such as the *International Classification of Diseases*[37] or the *Diagnostic and Statistical Manual of Mental Disorders (DSM)*,[38,39] what professionals' attitudes are toward standardized assessment,[40–42] and how clinicians' perceptions of and approaches to the assessment process are related.[43,44]

Study Objectives

How these methods are applied in the practices of clinicians is less known, and the available evidence suggests a discrepancy between guidelines on standards of practice and those of clinical practice.[8] This study aims to gain insights into the extent to which empirically supported assessments are used by clinicians, with a particular focus on the assessment of depression, given that depression is highly prevalent and the leading cause of disability globally.[45,46] Specifically, we investigated (1) collection methods and instruments that clinicians use; (2) the approaches used to integrate and interpret these data (i.e., combination approaches), including clinical and statistical methods; (3) clinicians' attitudes toward standardized assessment tools and AI-based decision support; and (4) clinicians' beliefs about the accuracy of their assessments relative to other methods and other clinicians. These questions were explored across diverse clinical settings, professional roles, and countries, with the goal of understanding the extent to which empirically supported assessment practices are implemented in real-world clinical contexts and the factors influencing their adoption.

Methods

Participants

Clinicians were recruited via convenience- and snowball sampling through several different online platforms such as LinkedIN, Facebook groups, emailing lists, workplace platforms (e.g., Slack channels for clinicians), and Prolific (an online platform where participants can be paid to take part). The inclusion criteria were as follows: clinicians who **1)** are currently working or have past experience in clinical

practice (physically or digitally) and **2)** are assessing or treating depression in patients. Clinicians who had never worked with assessing and/or treating depression were excluded (*n* = 98). The final sample consisted of 585 participating clinicians: 426 (73%) were female, 153 (26%) were male, and six (1%) indicated 'other'. There were 162 (28%) clinicians from the United Kingdom, 148 (25%) from the United States, 127 (22%) from Sweden, 105 (18%) from the Netherlands, and 43 (7%) from other countries. Among the different occupations reported, psychologists (*n* = 195; 33%) and nurses (*n* = 172; 29%) made up the largest part of the sample. Other occupations were physicians (*n* = 88; 15%), psychotherapists (*n* = 86; 15%), psychiatrists (*n* = 22; 4%), and 78 (13%) of the respondents selected 'other' (see Supplementary Table S1 in the Electronic Supplementary Material [ESM] which presents a cross-tabulation of clinician occupations by country). The clinicians reported a mean of 9.08 years (*SD* = 8.2) of experience assessing and/or treating depression. Almost half of the clinicians (*n* = 275; 47%) reported assessing and/or treating depression *daily or several times a week*. The other clinicians reported doing so *once a week or less* (*n* = 118; 20%), *once a month or less* (*n* = 95; 16%), or *having done this previously but not anymore* (*n* = 97; 17%).

Material: Survey Content

The survey was designed to gather insights into clinicians' practices, attitudes, and experiences regarding depression assessment. It included questions on data collection and combination methods, attitudes toward standardized assessment and AI, self-perceived accuracy, clinical experience, and demographics. The six domains within the questionnaire were (1) *Assessment Methods*, where clinicians reported on the data collection methods they use (e.g., unstructured, semi-structured, and structured interviews; rating scales), frequency of use, and specific instruments for both initial assessment and follow-up; (2) *Data Combination Methods*, where clinicians indicated whether they rely on clinical judgment, statistical models, or single data sources to integrate information; (3) *Assessment Attitudes*, where clinicians were asked about their attitudes regarding the importance of standardized methods for diagnostic assessments, as well as their interest in AI decision support coupled with one of the randomly assigned descriptions (1 = no information on the accuracy of AI, 2 = assuming high accuracy of AI, 3 = empirical data on accuracy of AI); (4) *Self-Perceived Accuracy*, where clinicians compared their accuracy against peers using similar or different methods, including AI and combinations of methods; (5) Clinical *Experience and Confidence*, where clinicians reported their experience with depression assessments, their confidence levels, and time spent per assessment; and (6) *Demographics*, including age, gender, country, profession, therapeutic approach, and employment sector. Detailed descriptions of the questions, response options, and randomized AI descriptions are provided on the ESM, along with the full questionnaire.

Procedure

Informed consent

was obtained from all the clinicians at the start of the survey. They were informed about the nature of the study, that their participation was voluntary and anonymous, that they had the right to withdraw at any time without providing a reason, and that their anonymized answers would be shared according to open science practices. Clinicians completed the survey online, following the question sequence outlined in the Materials section. The clinicians were presented with one of the randomized conditions when asked about their attitudes toward AI decision support. However, owing to an issue with the survey tool, the assigned condition was only recorded for a subset of clinicians, specifically those who completed the survey through Prolific. After completing the survey, the clinicians were debriefed and given the opportunity to send feedback. The survey was available in English, Dutch, or Swedish and took, on average, 10.92 minutes (*SD* = 15.89; Med = 7.82) to complete.

Statistical Analyses

The statistical analyses were performed in R (R Core Team, 2022; for details, see ESM). Descriptive statistics (frequencies [Freq], means [*M*], standard deviations [*SD*], medians [Med]) were calculated for each variable. Pearson correlations (*r*) were used to estimate the relationships between the variables, where the cut-off values for the different strengths were as follows: very weak or no correlation (± .0-.2), weak (± .3-.4), moderate (± .5-.6), strong (± .7-.8), and very strong (± .9 – 1). Analysis of variance (ANOVA) was employed for the AI condition comparisons and the cross-cultural comparisons. All ANOVAs were calculated for unbalanced designs (Type III ANOVA). The assumptions of homogeneity of variances were checked via Levene's test and were met. For significant main effects, post-hoc pairwise comparisons were conducted with Bonferroni correction to control for multiple testing. The *alpha*-value was set to 0.05.

Results

Duration of evaluation

The self-reported time it took for clinicians to identify depression in a typical patient ranged from 3 minutes to 15 hours[1], with a median time of 60 minutes (*SD* = 108 minutes, Mean = 87 minutes). Self-reported assessment time did not correlate significantly with the reported use of various data collection methods (see Supplementary Table S2 in the ESM for the correlations between all variables).

Aim 1: Data collection methods

The clinicians reported which data collection method(s) they used (Fig. 1a). Most clinicians reported using unstructured interviews (83%) and/or rating scales (83%), followed by semistructured interviews (65%), structured interviews (40%), and/or 'other' methods (25%).

The clinicians also estimated the proportion of their patients for whom they used these different methods (Fig. 1b). On average, they reported using clinical interviews for 58% of their patients, followed by rating scales for 51% of their patients, and semistructured interviews for 34% of their patients.

Most clinicians (89%) reported using two or more methods across patients (Fig. 1c). Only 11% of the clinicians reported using a single method across all of their patients, of which the unstructured interviews were the most common method. The most commonly reported data collection method combinations were as follows: 1) Unstructured interviews and rating scales (18%); 2) Unstructured interviews, semistructured interviews, structured interviews, and rating scales (18%); and 3) Unstructured interviews, semistructured interviews, and rating scales (*n* = 16%).

Most clinicians (84%) indicated performing follow-up assessments; similar patterns regarding the variety of data collection methods were observed among them. A more detailed description is available in the ESM.

N = 585

Specific data collection scales and (semi-)structured interviews. Clinicians could indicate which standardized tools and instruments they use. A list of the specific instruments and their corresponding figures is detailed in the ESM and the *Mental Health Assessment Dashboard* (see MHAD; https://cwiebel.shinyapps.io/ClinicalPractices/), an interactive online tool that allows for further analysis and custom comparisons of the data.

Aim 2: Data combination methods

The vast majority (80%) of the clinicians reported combining the collected data with clinical experience (i.e., the clinical method; Fig. 2). Eleven percent reported using the statistical data combination method, and 6% reported not using any data combination method since they used only one source of information.

N = 585.

Aim 3: **Attitudes towards assessment methods**

Attitudes toward standardized assessment methods. More than half of the clinicians reported *somewhat* to *strongly agree[ing]* (76%) that it is important to use standardized assessment methods with high validity and reliability (Fig. 3a). A notable number of clinicians reported *somewhat-* to *strongly disagree[ing]* (17%) with this statement. Clinicians' ratings of the importance of using standardized assessments were correlated with their reported frequency of using rating scales ($r = .20$, $p < .001$) and structured interviews ($r = .14$, $p = .002$) but negatively correlated with their reported frequency of using unstructured interviews ($r = -.13$, $p = .002$). Many clinicians have elaborated on their answers. A summary of these elaborations is presented in the SM.

N = 585.

Attitudes toward AI decision support. More clinicians reported to be *probably* or *definitely* interested in receiving AI decision support (43%) than *probably not* or *definitely not* (33%; Fig. 3b). A substantial number reported not knowing or having too little knowledge to take a stand (24%). Greater interest in using AI was correlated with positive agreement on the importance of using standardized assessments

(r = .11, p = .009) and with the reported frequency of using rating scales (r = .17, p < .001). The attitudes toward the use of AI decision support did not significantly differ between the AI information conditions (F[2,324] = 1.60, p = .204, n = 359)[2]. Many clinicians have also elaborated on their answers, which are presented on the ESM.

Aim 4: Beliefs in one's confidence and accuracy

A majority of the clinicians reported feeling *confident* (35%), *very confident* (32%), or *extremely confident* (10%) in identifying depression, whereas few reported feeling *somewhat confident* (19%), *a little confident* (4%), or *not at all confident* (< 1%). Confidence in identifying depression was positively correlated with the reported assessment frequency of depression (r = .33, p < .001), work experience with identifying depression (r = .21, p < .001), and the reported frequency of using rating scales and semistructured interviews as data collection methods (r = .17, $p < .001$ and $r = .11$, $p = .007$). See the ESM for more details.

Clinicians reported believing that their assessments (based on the method[s] they usually use for assessing depression) yield superior accuracy compared to those clinicians who use only one data collection method, but not with those clinicians who use a combination of data collection methods (Fig. 4a). The largest discrepancy between the clinicians' own estimated assessment accuracy compared to others was noted in comparison to those who relied only on rating scales as their single data collection source.

Among the clinicians who reported using more than one method in their assessments (89%), most reported believing their accuracy to be similar to or slightly less than that of other clinicians using multiple methods (Fig. 4c). However, 80% of clinicians rely on clinical experience to interpret results from multiple sources, and these clinicians reported believing their assessment to be more accurate than that of AI or statistical models (Fig. 4b). In summary, while clinicians do not consider themselves more accurate than their peers do when a combination of methods is used, they do view their accuracy as superior to that of statistical models.

The error bars represent the standard error of the mean. N = 585.

Cross-cultural differences

Exploratory analyses comparing clinicians from the U.S., the Netherlands, Sweden, and the U.K. identified few small to moderate cross-cultural differences. Clinicians from the U.S. and the U.K. used unstructured interviews for the smallest percentage of patients, with averages of 51% and 46%, respectively. The finding that the vast majority of clinicians report combining collected data with clinical experience is consistent across countries. For a more exploratory cross-cultural analysis, see the ESM and the *MHAD*. However, these findings should be interpreted with caution due to the heterogeneous nature of the samples, especially with respect to the occupations (e.g., psychologists, psychiatrists, general practitioners, nurses).

Discussion

This study examined the extent to which mental health professionals use empirically supported methods for assessing depression and their attitudes toward different methods. The results show that depression assessments are conducted by different professions with diverse educational backgrounds, using a wide range of methods and reporting large variations in the time it takes to assess, reflecting real-world practices and the diverse experiences patients encounter. The results show that (1) mental health professionals use a wide variety of data collection methods and that (2) they predominantly rely on the clinical method for data combination. (3) A total of 77% of the clinicians agreed on the importance of using standardized, valid, and reliable assessments, and 43% reported being interested in receiving AI decision support. (4) They are generally confident in their assessments but vary substantially in the reported time it takes to carry out a depression assessment. On average, clinicians report believing they are more accurate than AI or other clinicians are when a single data collection method is used. Interestingly, a wide variety of methods and practices were consistently found across all the examined countries, with only small significant cross-cultural differences. This highlights the paradox of consistent inconsistency in depression assessment practices on an international scale.

Data collection methods

Clinicians report using different data collection methods with a wide range of combinations and instruments that vary greatly between individuals. Most clinicians (89%) reported using two or more methods, with unstructured interviews and rating scales being the most common. Although a significant portion of the clinicians (16%) did not conduct follow-up assessments, a similar pattern of diverse methods was observed among those who did. This aligns with previous studies highlighting the lack of use of outcome measures.[47,48] While unstructured interviews have shown low interrater reliability for diagnosing depression and are seldom used in research,[29,49] various rating scales for assessing depression differ significantly (i.e., they capture different symptoms of depression and only partly overlap[30,50]; and are often used interchangeably in research.[30,49–51] This variation in data collection methods may be problematic for providing equal, standardized, high-quality care, especially since research shows that different methods and instruments yield varying results. The absence of a single industry standard for diagnostic manuals further contributes to inconsistencies in assessments and diagnoses.[51]

Data combination methods

While approximately one in ten clinicians reported using the statistical method to form their final assessment, 80% reported interpreting and weighing the collected data on the basis of their clinical experience and judgment. Despite extensive research showing that the statistical method for combining information is as good as or superior to the clinical judgment of an expert clinician,[6,13,24,36] the majority of final assessments still rely on the clinical method. As early as 35 years ago, Dawes et al.[14] argued for the routine use of statistical methods in clinical assessment and decision-making. Grove and Meehl[21]

addressed this issue, concluding that using the less accurate of two assessment methods “is not only unscientific and irrational, it is unethical” (*p*. 323). However, possible explanations for the continued reliance on clinical data combination might be the lack of easily implemented, user-friendly, and validated statistical models that comply with the required regulations (e.g., CE-marking and FDA approval). Clinical validation and readiness for implementation should be weighed heavily before employing statistical models since, for example, flawed algorithms trained on small or biased training data can have adverse effects.[52] This concern is further amplified by widespread apprehensions about the safety and validity of AI models, as highlighted by clinicians in this survey.

Attitudes toward standardized assessment

A sizable group of clinicians, approximately one-fifth, report that they *somewhat*- to *strongly disagree* with the importance of using standardized assessment methods that have been demonstrated in research to have high validity and reliability — which we expected to find a unanimous agreement on. Even though 76% of the clinicians reported that they *somewhat*- to *strongly agree* with the statement, the importance they assigned to standardized assessment methods did not affect the final decision for most clinicians, who still use a combination of clinical data methods. This corresponds to other research examining clinicians’ attitudes, which have shown a mostly, but not completely, positive attitude towards standardized assessment methods: Clinicians seem to be positive about the quality of standardized assessment methods but less positive about their usefulness in practice, and the advantages of standardized assessment over clinical judgment.[40–42] To enable better implementation, the need to overcome specific barriers and issues, such as adequate training and practicality, is highlighted.[40–42,44]

Attitudes toward AI

Clinicians’ interest in AI decision support for depression assessment is mixed: 43% express interest, 33% show disinterest, and 23% remain undecided due to insufficient information. This highlights the need to update clinicians on AI advancements and benefits and involve them in algorithm development and validation. Key concerns, including accountability, transparency, and privacy, must be addressed.[53] Interest in AI support correlates with the use of rating scales and positive attitudes toward standardized methods, suggesting that clinicians familiar with these practices could facilitate AI integration in clinical settings. However, enhanced education on AI and statistical models is essential. For example, a study on psychiatrists’ views of generative AI found widespread use for administrative tasks but skepticism toward its active role in patient care.[54] The lack of differences in attitudes across AI descriptions and diverse responses in open comments further emphasize the need for comprehensive educational initiatives to foster an understanding of AI’s capabilities, limitations, and ethical integration into mental health practices.

Clinicians Assessment Accuracy

Strikingly, a majority of clinicians who rely on a combination of clinical data believe that their assessments are more accurate than those derived from a statistical method or AI, which contradicts

findings from previous research.[6,13,24,36] The clinicians ranked the accuracy of their assessments as superior to those conducted by peers who use only a single method (particularly rating scales alone) but not compared with those using a combination of methods. These findings suggest a general belief among clinicians in the superiority of employing multiple methods for data collection and combining data via a clinical method. The belief in the clinical method could be explained by a more general tendency of individuals to overestimate their cognitive abilities.[55] However, clinicians' confidence and accuracy beliefs have been shown to be poor indicators of their assessment accuracy,[20,23,56] which is troubling given the impact it may have on patients and their course of treatment.

Limitations

The results are constrained by certain limitations, yet they indicate several directions for future research to further our understanding of depression assessment practices and facilitate the adoption of empirically supported assessment methods. Given the study's broad focus on depression assessments in real-world clinical settings, the sample comprises a wide range of occupations, including psychologists, psychotherapists, psychiatrists, and general practitioners. However, this makes it challenging to understand the experiences of specific occupations; therefore, future research could focus on single professional groups to gain deeper insights into their unique practices and challenges. Similarly, the goals and methods of clinicians' assessments can differ significantly both between and within countries, occupations, and clinics, all of which warrant further investigation.

This study included only Western countries, each with distinct guidelines, rules, and regulations governing mental health practices. The distribution of occupations also varies across countries, which necessitates caution when interpreting cross-country differences. Future research should expand to include non-Western countries to gain insights into how mental health practices are carried out in diverse cultural and regulatory contexts.

Additionally, this study specifically examined the assessment of depression, so the findings may not be generalizable to other mental health conditions. The use of a self-report survey format introduces the potential for recall bias and may reflect clinicians' perceptions of best practices rather than their actual behaviors. As a result, the data could overestimate the extent to which empirically supported assessment methods are employed in practice. Finally, online recruitment and sampling may have restricted participation to specific subgroups. The sampling strategy is also likely to favor digitally active clinicians, which should be considered when interpreting the findings. In addition, AI-related randomization data were missing for some participants, which reduced the statistical power. Future research could benefit from random sampling directly from clinical settings or from observational studies of audits of actual practices, etc.

Potential implications and future research

A better understanding of current depression assessment practices can provide clinicians, policymakers, and researchers with a foundation for improving healthcare quality. Understanding current practices can

support clinicians, regulatory bodies, and researchers in joining forces to lead changes and update guidelines and policies.

The Mental Health Assessment Dashboard can inform discussions on the current (*limited*) use of empirically supported assessments and any future integration of AI in clinical settings. Furthermore, there is a need for more research regarding the potential risks of current depression assessment practices and their implications for patients and the challenges for clinicians in embracing newer empirically supported methods.[15,57] There is a lack of research on how clinicians combine assessment methods and the potential shortcomings associated with the clinical combination method.

The variability in assessors and assessment methods reflects real-world practice and highlights the need for clearer guidelines and training to ensure equitable and consistent standardization. In other words, the outcome of a depression assessment for depression should not depend on where, when, or by whom the patient is assessed.

Additionally, many clinicians' concerns about AI should be addressed, including biases and ethical issues, as well as how they can work with AI. Transparent communication and education about the capabilities and limitations of AI, along with evidence from robust studies, can potentially build trust and acceptance among clinicians.[58]

Conclusion

Previous research[7,51] has typically studied depression assessment methods in research settings or focused on attitudes in clinical settings[40–42] rather than actual assessment practices in different clinical practices. Our results highlight heterogeneity in depression assessments: Many different professionals conduct depression assessments in clinical practice (e.g., psychiatrists, psychologists, general practitioners, nurses); employ a wide variety of methods (ranging from unstructured to structured approaches using numerous tools) in different combinations, and report a great difference in the time it takes to carry out the assessment.

The results reveal a pattern of skepticism toward AI, underestimation of statistical models, and overestimation of one's own accuracy, which aligns with findings from previous studies. While clinicians express confidence in their ability to assess depression, the widespread use of unstructured interviews alongside standardized tools, coupled with a predominant reliance on clinical judgment over statistical models, suggests a disconnect between best practices supported by research and everyday clinical practice. Adopting empirically supported practices, including standardized, valid, and reliable data collection and combination assessment methods, is essential to synthesize information systematically and improve the quality of the data used in clinical judgment. This alignment between clinical practice and research evidence would enhance patient care while allowing clinicians to consider individual factors. However, achieving this requires harmonizing clinical judgment with data-driven approaches, as well as aligning manuals and guidelines with empirical research. Future initiatives should focus on

developing accessible, practical, validated tools and training programs that bridge the gap between research and practice, fostering integration and consistency in depression assessment.

Abbreviations

AI
Artificial Intelligence
DSM
Diagnostic and Statistical Manual of Mental Disorders
ESM
Electronic Supplementary Material
ICD
International Classification of Diseases
LLM
Large Language Model *(only include if mentioned earlier; I didn't sit explicitly)*
MHAD
Mental Health Assessment Dashboard
U.K.
United Kingdom
U.S.
United States
WHO
World Health Organization

Declarations

Ethics approval and consent to participate

The study recieved ethical approval from Leiden University (2022-08-10-M.L. Molendijk-V5-4037) and was exempt from ethical approval at the University of Pennsylvania (United States). The study is deemed exempt from requiring ethical approval according to Swedish Law (s§3-4 of the Act [2003:460] on the ethical review of research involving humans in Sweden).

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are available at https://osf.io/pw8mk/. The code for the *Mental Health Assessment Dashboard* is also shared here in case the app no longer can be hosted (https://cwiebel.shinyapps.io/ClinicalPractices/).

Competing interests

O.N.E. Kjell and K. Kjell co-founded a start-up that uses computational language assessments to assess mental health problems. E.C. Stade has received fees for advising a start-up that uses computational language assessment to measure psychological constructs. The other authors report having no conflicts of interest with respect to the contents, authorship, or publication of this article.

Funding

O.N.E. Kjell, K. Kjell and V.C. Eijsbroek were funded by FORTE (2022-01022).

Authors' contributions

Conceptualization: Ideas; formulation or evolution of overarching research goals and aims.

Damon Navandi, Veerle C. Eijsbroek, Clara Wiebel, Katarina Kjell, Marc L. Molendijk, Elizabeth C. Stade, Oscar N. E. Kjell

Data curation: Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.

Damon Navandi, Veerle C. Eijsbroek, Clara Wiebel

Formal analysis: Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data.

Clara Wiebel, Veerle C. Eijsbroek, Damon Navandi, and Oscar N. E. Kjell

Funding acquisition: Acquisition of the financial support for the project leading to this publication.

Katarina Kjell, Oscar N. E. Kjell

Investigation: Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.

Damon Navandi, Veerle C. Eijsbroek, Clara Wiebel, Katarina Kjell, Marc L. Molendijk, Elizabeth C. Stade, Oscar N. E. Kjell

Methodology: Development or design of methodology; creation of models.

Damon Navandi, Veerle C. Eijsbroek, Clara Wiebel, Katarina Kjell, Marc L. Molendijk, Elizabeth C. Stade, Oscar N. E. Kjell

Project administration: Management and coordination responsibility for the research activity planning and execution.

Katarina Kjell, Oscar N. E. Kjell

Resources: Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools.

Damon Navandi, Veerle C. Eijsbroek, Clara Wiebel, Katarina Kjell, Marc L. Molendijk, Elizabeth C. Stade, Oscar N. E. Kjell

Software: Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.

Clara Wiebel

Supervision: Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.

Oscar N. E. Kjell

Validation: Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.

Visualization: Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.

Clara Wiebel, Veerle C. Eijsbroek, Damon Navandi, Oscar N. E. Kjell

Writing—original draft: Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).

Damon Navandi, Veerle C. Eijsbroek, Clara Wiebel, Oscar N. E. Kjell

Writing—review and editing: Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision: including pre- or post-publication stages.

Damon Navandi, Veerle C. Eijsbroek, Clara Wiebel

Acknowledgements

Not applicable

References

1. Shen H, Zhang L, Xu C, Zhu J, Chen M, Fang Y. Analysis of Misdiagnosis of Bipolar Disorder in An Outpatient Setting. Shanghai Archives of Psychiatry [Internet]. 2018;30(2):93–101. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5936046/
2. Stahnke B. A systematic review of misdiagnosis in those with obsessive-compulsive disorder. Journal of Affective Disorders Reports [Internet]. 2021;6(1):100231. Available from: https://www.sciencedirect.com/science/article/pii/S2666915321001578
3. Fried EI, Proppert RKK, Rieble CL. Building an Early Warning System for Depression: Rationale, Objectives, and Methods of the WARN-D Study. Clinical Psychology in Europe [Internet]. 2023;5(3):1–25. Available from: https://cpe.psychopen.eu/index.php/cpe/article/view/10075
4. Eijsbroek VC, Katarina Kjell, Schwartz A, Boehnke JR, Fried EI, Klein DN et al. The LEADING Guideline. Reporting Standards for Expert Panel, Best-Estimate Diagnosis, and Longitudinal Expert All Data (LEAD) Studies. medRxiv (Cold Spring Harbor Laboratory). 2024.
5. Hunsley J, Mash EJ. Evidence-based assessment. Annual review of clinical psychology [Internet]. 2007;3:29–51. Available from: https://www.ncbi.nlm.nih.gov/pubmed/17716047
6. Meehl PE. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press; 1954.
7. Pettersson A, Boström KB, Gustavsson P, Ekselius L. Which instruments to support diagnosis of depression have sufficient accuracy? A systematic review. Nord J Psychiatry. 2015;69(7):497–508.
8. Cook JR, Hausman EM, Jensen-Doss A, Hawley KM. Assessment Practices of Child Clinicians. Assessment. 2016;24(2):210–21.
9. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. Curr Psychiatry Rep. 2019;21(11):116.
10. Oscar NE, Kjell K, Kjell H. Andrew Schwartz. Beyond Rating Scales: With Targeted Evaluation, Language Models are Poised for Psychological Assessment. Psychiatry Res. 2023;333:115667–7.
11. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial Intelligence for Mental Healthcare: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. Biol Psychiatry: Cogn Neurosci Neuroimaging. 2021;6(9):856–64.
12. Mueller AE, Segal DL. Structured versus Semistructured versus Unstructured Interviews. Encyclopedia Clin Psychol. 2015;1(7):1–7.
13. Ægisdóttir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, et al. The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. Couns Psychol. 2006;34(3):341–82.

14. Dawes R, Faust D, Meehl P. Clinical versus actuarial judgment. Science [Internet]. 1989;243(4899):1668–74. Available from: http://meehl.umn.edu/sites/meehl.dl.umn.edu/files/138cstixdawesfaustmeehl.pdf
15. DeRubeis RJ. The history, current status, and possible future of precision mental health. Behav Res Ther. 2019;123:103506.
16. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychological Medicine [Internet]. 2019;49(09):1426–48. Available from: https://www.cambridge.org/core/journals/psychological-medicine/article/abs/machine-learning-in-mental-health-a-scoping-review-of-methods-and-applications/0B70B1C827B3A4604C1C01026049F7D9
17. Bennett CC, Hauser K. Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. Artif Intell Med. 2013;57(1):9–19.
18. Egon Brunswik, Cooksey RW. The Conceptual Framework of Psychology [1952]. 2001;225–37.
19. Hammond KR, Hursch CJ, Todd FJ. Analyzing the components of clinical inference. Psychol Rev. 1964;71(6):438–56.
20. Karelaia N, Hogarth RM. Determinants of linear judgment: A meta-analysis of lens model studies. Psychol Bull. 2008;134(3):404–26.
21. Grove WM, Meehl PE. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. Psychol Public Policy Law. 1996;2(2):293–323.
22. Bowes SM, Ammirati RJ, Costello TH, Basterfield C, Lilienfeld SO. Cognitive biases, heuristics, and logical fallacies in clinical practice: A brief field guide for practicing clinicians and supervisors. Prof Psychology: Res Pract. 2020;51(5):435–45.
23. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inf Decis Mak. 2016;16(1).
24. Garb HN, Wood JM. Methodological advances in statistical prediction. Psychol Assess. 2019;31(12):1456–66.
25. Collins FS, Varmus H. A New Initiative on Precision Medicine. N Engl J Med. 2015;372(9):793–5.
26. Davenport T, Kalakota R. The Potential for Artificial Intelligence in Healthcare. Future Healthcare Journal [Internet]. 2019;6(2):94–8. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6616181/
27. Ali GC, Ryan G, De Silva MJ. Validated Screening Tools for Common Mental Disorders in Low and Middle Income Countries: A Systematic Review. Burns JK, editor. PLOS ONE. 2016;11(6):e0156939.
28. Breedvelt JJF, Zamperoni V, South E, Uphoff EP, Gilbody S, Bockting CLH, et al. A systematic review of mental health measurement scales for evaluating the effects of mental health prevention interventions. Eur J Pub Health. 2020;30(3):510–6.

29. Miller PR, Dasher R, Collins R, Griffiths P, Brown F. Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. Psychiatry Res. 2001;105(3):255–64.
30. Fried EI. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. J Affect Disord. 2017;208(208):191–7.
31. Chevance A, Ravaud P, Tomlinson A, Le Berre C, Teufer B, Touboul S, et al. Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. Lancet Psychiatry. 2020;7(8):692–702.
32. Zimmerman M, Martinez JH, Friedman M, Boerescu DA, Attiullah N, Toba C. How Can We Use Depression Severity to Guide Treatment Selection When Measures of Depression Categorize Patients Differently? J Clin Psychiatry. 2012;73(10):1287–91.
33. Shafer AB. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. J Clin Psychol. 2005;62(1):123–46.
34. van Loo HM, de Jonge P, Romeijn JW, Kessler RC, Schoevers RA. Data-driven subtypes of major depressive disorder: a systematic review. BMC Med. 2012;10(1).
35. Fried EI, van Borkulo CD, Epskamp S, Schoevers RA, Tuerlinckx F, Borsboom D. Measuring depression over time.. Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. Psychol Assess. 2016;28(11):1354–67.
36. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: A meta-analysis. Psychol Assess. 2000;12(1):19–30.
37. ICD-11 [Internet]. icd.who.int. Available from: https://icd.who.int
38. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. 2013;5(5).
39. First MB, Rebello TJ, Keeley JW, Bhargava R, Dai Y, Kulygina M, et al. Do mental health professionals use diagnostic classifications the way we think they do? A global survey. World Psychiatry. 2018;17(2):187–95.
40. Danielson M, Månsdotter A, Fransson E, Dalsgaard S, Larsson J-O. Clinicians' attitudes toward standardized assessment and diagnosis within child and adolescent psychiatry. Child Adolesc Psychiatry Mental Health. 2019;13(1).
41. Jensen-Doss A, Hawley KM. Understanding Barriers to Evidence-Based Assessment: Clinician Attitudes Toward Standardized Assessment Tools. Journal of Clinical Child & Adolescent Psychology [Internet]. 2010;39(6):885–96. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3058768/
42. Jensen-Doss A, Hawley KM. Understanding Clinicians' Diagnostic Practices: Attitudes Toward the Utility of Diagnosis and Standardized Diagnostic Tools. Administration and Policy in Mental Health and Mental Health Services Research [Internet]. 2011;38(6):476–85. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6114089/
43. Bohman B. Clinicians' perceptions and practices of diagnostic assessment in psychiatric services. BMC Psychiatry. 2023;23(1).

44. Cho E, Tugendrajch SK, Marriott BR, Hawley KM. Evidence-Based Assessment in Routine Mental Health Services for Youths. Psychiatric Serv. 2020;72(3):appi.ps.2019005
45. Mathers CD, Loncar D. Projections of Global Mortality and Burden of Disease from 2002 to 2030. Samet J, editor. PLoS Medicine [Internet]. 2006;3(11):e442. Available from: https://pubmed.ncbi.nlm.nih.gov/17132052/
46. Prevalence. and variability of current depressive disorder in 27 European countries: a population-based study. The Lancet Public Health [Internet]. 2021; Available from: https://www.sciencedirect.com/science/article/pii/S2468266721000475#bib9
47. Boswell JF, Kraus DR, Miller SD, Lambert MJ. Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. Psychother Res. 2013;25(1):6–19.
48. Gilbody SM, House AO, Sheldon TA. Psychiatrists in the UK do not use outcomes measures. Br J Psychiatry. 2002;180(2):101–3.
49. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. Am J Psychiatry. 2013;170(1):59–70.
50. Newson JJ, Hunter D, Thiagarajan TC. The heterogeneity of mental health assessment. Frontiers in Psychiatry [Internet]. 2020;11(76):1–24. Available from: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.00076/full
51. Santor DA, Gregus M, Welch A. FOCUS ARTICLE: Eight Decades of Measurement in Depression. Measurement: Interdisciplinary Research & Perspective. 2006;4(3):135–55.
52. Topol EJ. High-performance medicine: the Convergence of Human and Artificial Intelligence. Nature Medicine [Internet]. 2019;25(1):44–56. Available from: https://www.nature.com/articles/s41591-018-0300-7
53. Leslie D, systems in the public sector Dr David Leslie Public Policy Programme. Understanding artificial intelligence ethics and safety A guide for the responsible design and implementation of AI. Understanding artificial intelligence ethics and safety [Internet]. 2019; Available from: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf
54. Blease C, Worthen A, Torous J. Psychiatrists' Experiences and Opinions of Generative Artificial Intelligence in Mental Healthcare: An Online Mixed Methods Survey. Psychiatry Res. 2024;333:115724–4.
55. Heck PR, Simons DJ, Chabris CF. 65% of Americans believe they are above average in intelligence: Results of two nationally representative surveys. van Amelsvoort T, editor. PLOS ONE. 2018;13(7):e0200103.
56. Miller DJ, Spengler ES, Spengler PM. A meta-analysis of confidence and judgment accuracy in clinical decision making. J Couns Psychol. 2015;62(4):553–67.
57. Betancourt TS, Chambers DA. Optimizing an Era of Global Mental Health Implementation Science. JAMA Psychiatry. 2016;73(2):99.

58. Misra R, Keane PA, Hogg HDJ. How should we train clinicians for artificial intelligence in healthcare? Future Healthc J. 2024;11(3):100162.

Footnotes

1. Nine cases with a reported time greater than the mean plus 3 standard deviations were seen as outliers and thus removed from the time analysis. Their reported assessment times ranged from 20 to 60 hours, respectively. Additionally, seven cases reported 0 minutes; these were removed because it is possible that they do not assess depression at all.
2. The lower number of clinicians in this analysis is due to the randomisation condition not being saved for all clinicians.

Figures

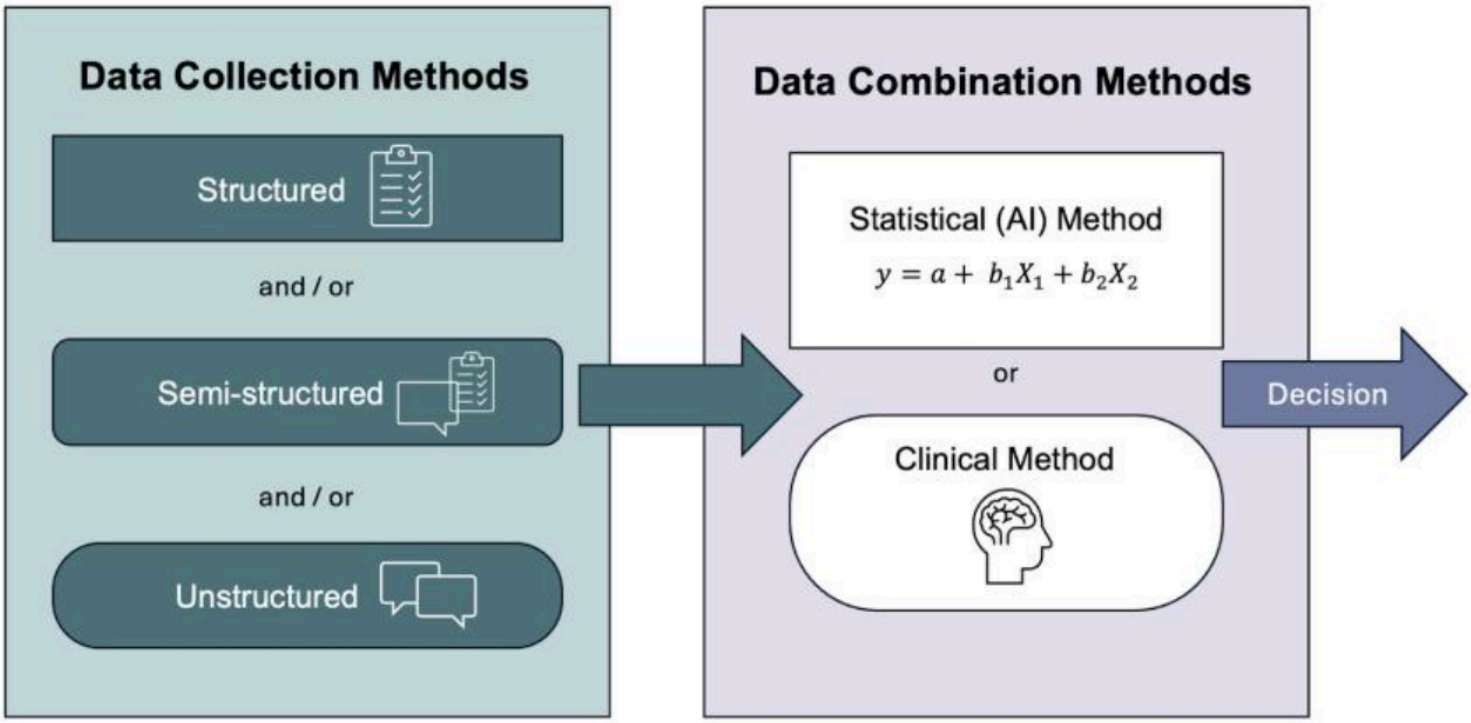


Figure 1

Overview of depression assessment methods, including data collection and combination methods.
Note. The data combination method does not necessarily require data from different data collection methods; for example, responses to items of a single rating scale can be aggregated and interpreted via a statistical model (following an aggregation schema or using an established cut-off point) or interpreted via a clinical method in which individual items are examined via clinical experience. In the statistical method, it is also feasible to include data from a clinical (unstructured) interview in a statistical model (e.g., to what extent the patient responds, their response times, and behaviors).

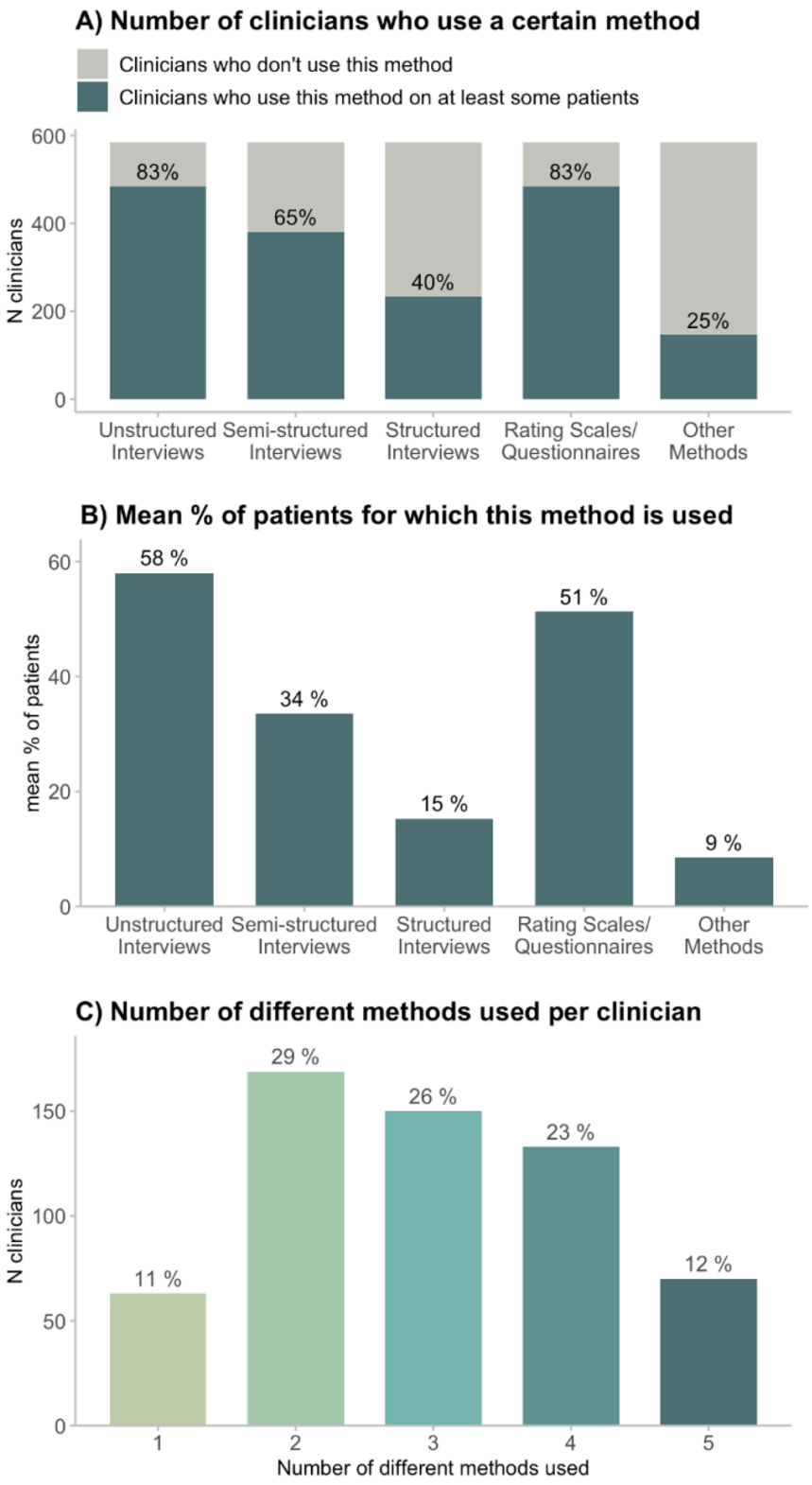


Figure 2

Figure 1 A–C | Clinicians' reported data collection methods: A) number (percentage) of clinicians using each method, B) mean percentage of patients each method is used for and C) the number of methods used per clinician across all patients.

N = 585

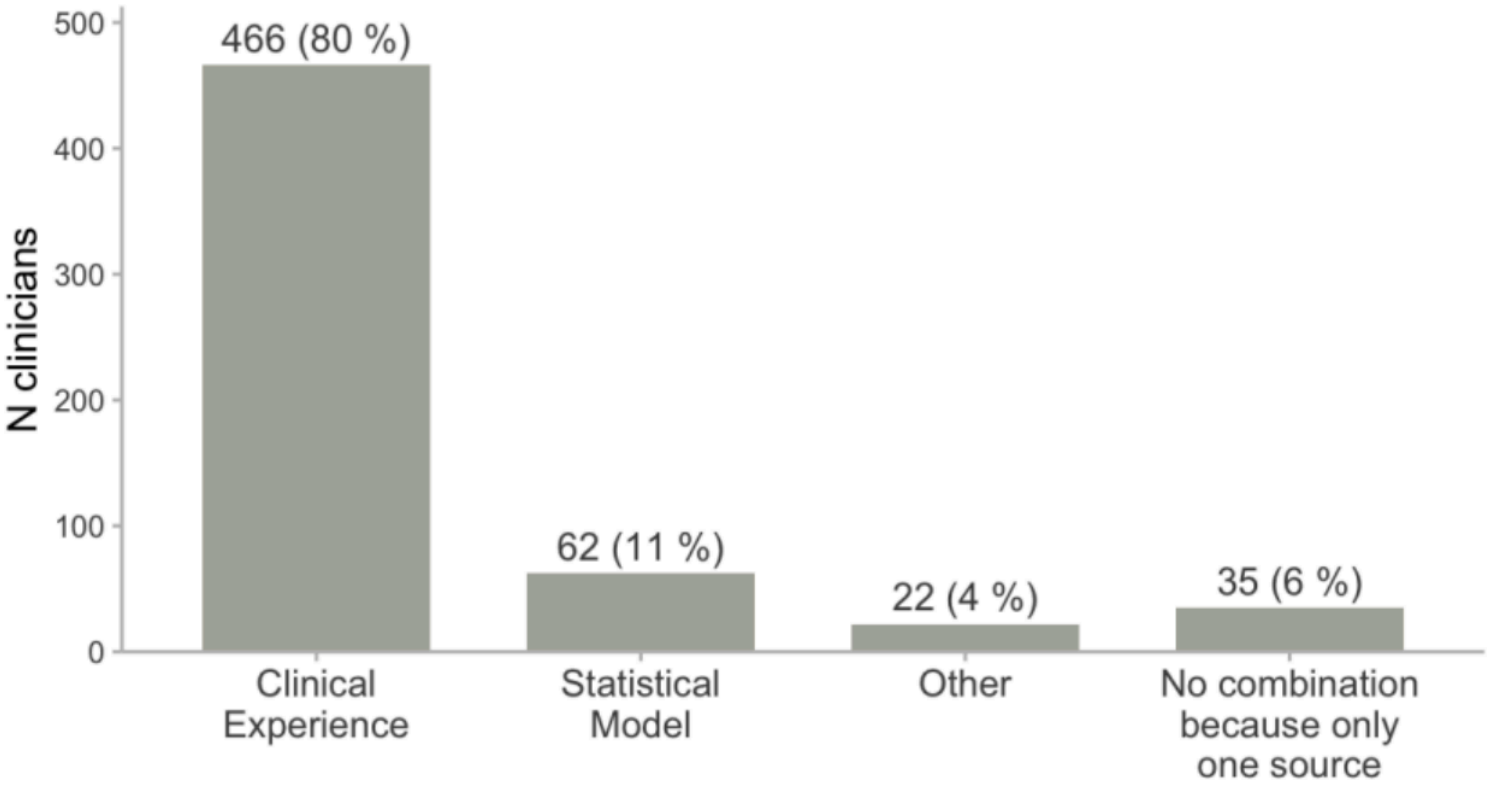


Figure 3

Figure 2 | Clinicians' reported use of different data combination methods

N = 585.

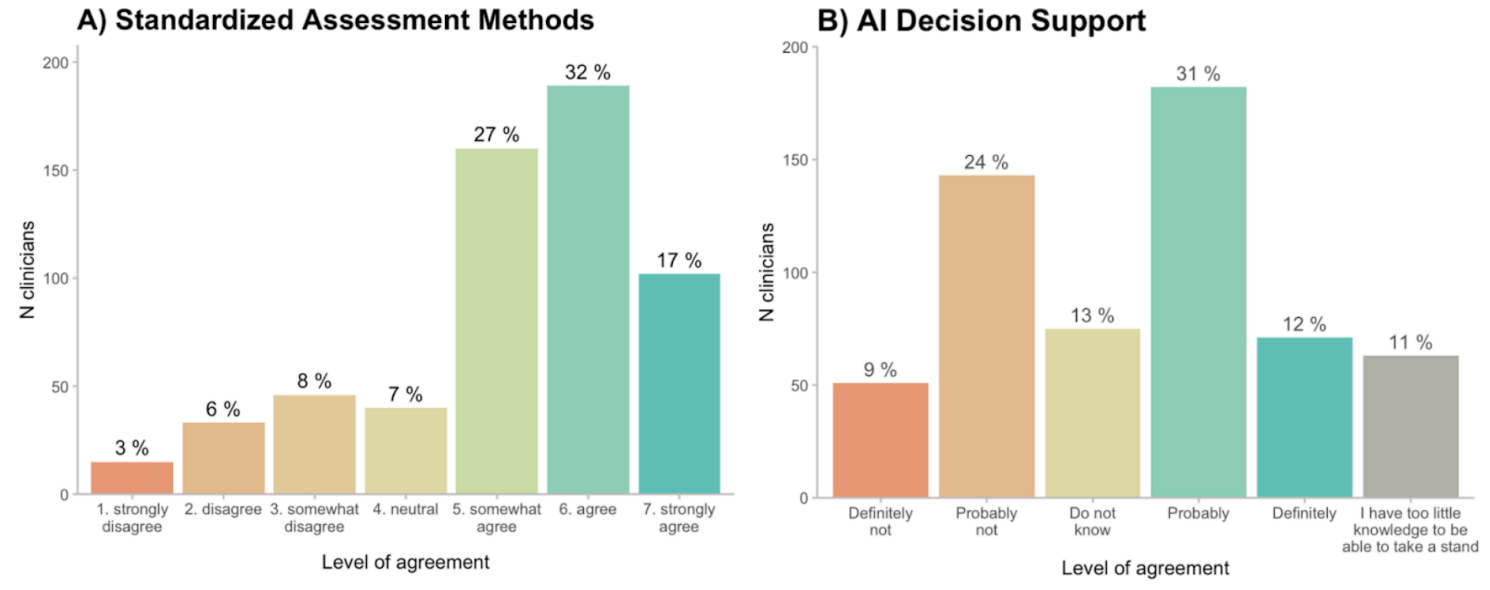


Figure 4

Figure 3 A-B | Clinicians' reported attitudes toward the use of A) standardized assessment methods and B) AI decision support.

N = 585.

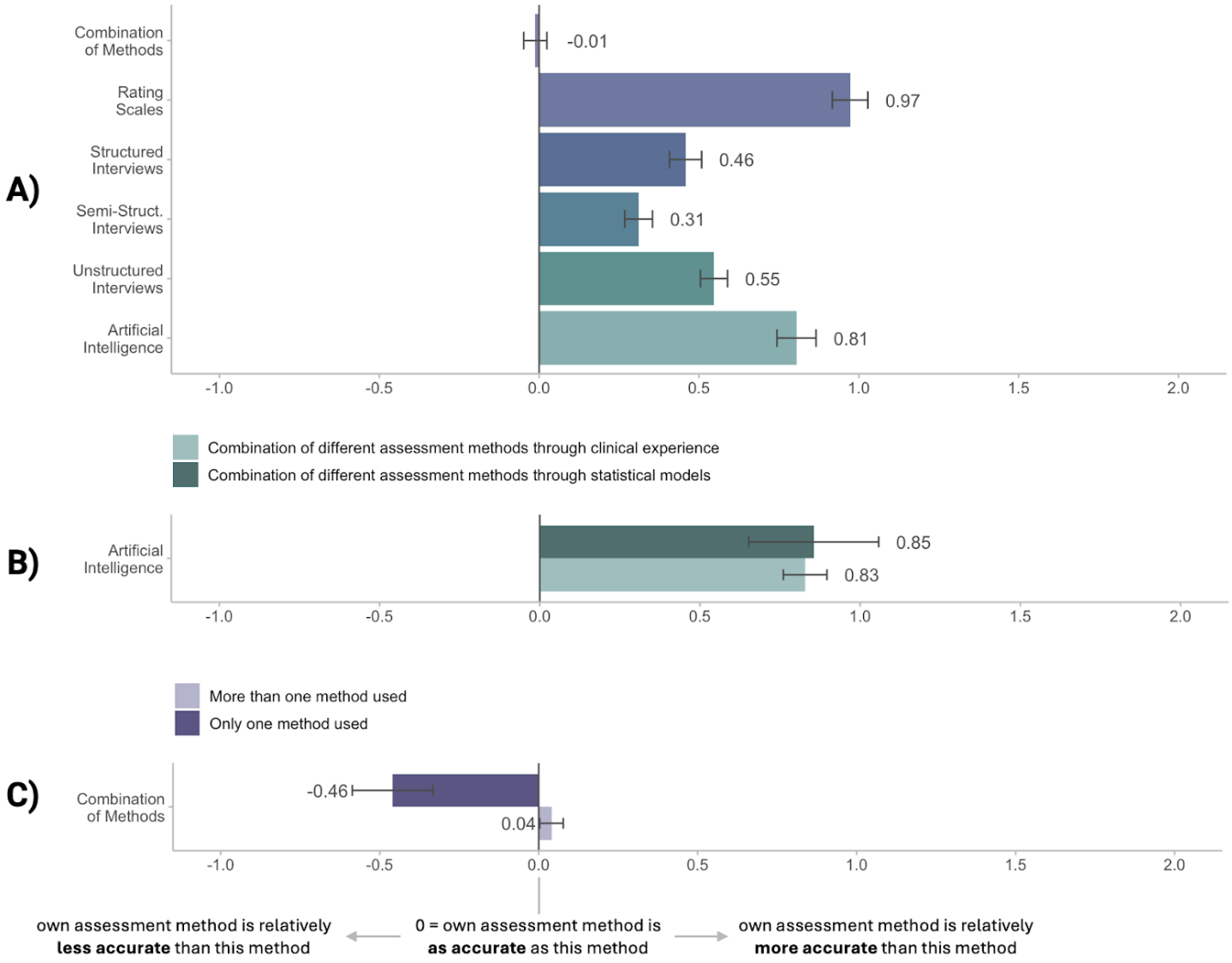


Figure 5

Figure 4 A—C | Clinicians' self-perceived assessment accuracy: A) compared with that of clinicians using other methods, B) compared with that of clinicians using AI, and C) compared with that of clinicians using a combination of methods.
The error bars represent the standard error of the mean. *N* = 585.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx