DOI(Journal): 10.31703/gssr
DOI(Volume): 10.31703/gssr.2025(X)
DOI(Issue): 10.31703/gssr.2025(X.IV)

p-ISSN: 2520-0348

e-ISSN: 2616-793X

www.gssrjournal.com
Global Social Sciences Review
exploring humanity

GSSR

GLOBAL SOCIAL SCIENCES REVIEW

HEC-RECOGNIZED CATEGORY-Y

Volum X, ISSUE IV FALL (DECEMBER-2025)

HumanityPublications
sharing research
www.humanitypublications.com
Karachi | Islamabad | Lahore

Double-blind Peer-review Journal
www.gssrjournal.com
© Global Social Sciences Review

Humanity Publications(HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

Article Title

Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study Across Demographic Groups

Abstract

Depression screening models may perform differently across demographic groups despite satisfactory overall accuracy. This protocol describes an independent, paired evaluation of a locked depression screening model and the Patient Health Questionnaire-9 (PHQ-9) across age, race and ethnicity, gender, and preferred-language groups. Consecutive eligible adults attending participating primary care sites will complete both tests during the same clinical episode. All consenting participants, regardless of screening results, will be offered a blinded clinician-administered diagnostic interview for current major depressive disorder within 14 days. The model, threshold, subgroup definitions, and sample size will be finalized and registered before recruitment. The primary outcome is between-group differences in sensitivity at the prespecified threshold. Secondary analyses will assess specificity, calibration, missed cases, and referral workload. Missing data and interview noncompletion will undergo sensitivity analyses. This manuscript presents a planned study; no participant data, diagnostic accuracy, fairness, or clinical-benefit findings are reported.

Keywords: Depression Screening, Diagnostic Accuracy, Algorithmic Bias, Health Equity; PHQ-9, Prospective Validation, Study Protocol

Authors:

Mazharul Ahosan Fahim: (Corresponding Author)
Department of Sociology, East Tennessee State University, Johnson City, Tennessee, USA.
(Email: fahimmazharulahosan@gmail.com)
ORCID: https://orcid.org/0009-0002-2123-2735

Ashrafa Hossain: Department of Psychology, Chittagong College, National University of Bangladesh, Chittagong, Bangladesh.
ORCID: https://orcid.org/0009-0006-0627-9745

Khadija Jahan Mukta: Department of Psychology, Goldsmiths University of London, London, UK.
ORCID: https://orcid.org/0009-0006-4991-450X

Pages: 230-243
DOI: 10.31703/gssr.2025(X-IV).19
DOI link: https://dx.doi.org/10.31703/gssr.2025(X-IV).18
Article link: http://www.gssrjournal.com/article/algorithmic-bias-in-depression-screening-protocol-for-a-prospective-multicenter-diagnostic-accuracy-study-across-demographic-groups
Full-text Link: https://gssrjournal.com/article/algorithmic-bias-in-depression-screening-protocol-for-a-prospective-multicenter-diagnostic-accuracy-study-across-demographic-groups
Pdf link: https://www.gssrjournal.com/jadmin/Auther/31rvIolA2.pdf

Global Social Sciences Review

p-ISSN: 2520-0348 **e-ISSN:** 2616-793x
DOI(journal):10.31703/gssr
Volume: X (2025)
DOI (volume):10.31703/gssr.2025(X)
Issue: IV Fall (December-2025)
DOI(Issue):10.31703/gssr.2025(X-IV)

Home Page
www.gssrjournal.com

Volume: X (2025)
https://www.gssrjournal.com/Current-issue

Issue: IV-Fall (December 2025)
https://www.gssrjournal.com/issue/10/4/2025

Scope
https://www.gssrjournal.com/about-us/scope

Submission
https://humaglobe.com/index.php/gssr/submissions

Google scholar

Visit Us

Humanity Publications (HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

Citing this Article

19	Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study Across Demographic Groups		
Authors	Mazharul Ahosan Fahim Ashrafa Hossain Khadija Jahan Mukta	DOI	10.31703/gssr.2025(X-IV).19
		Pages	230-243
		Year	2025
		Volume	X
		Issue	IV

Referencing & Citing Styles	
APA	Fahim, M. A., Hossain, A., & Mukta, K. J. (2025). Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study Across Demographic Groups. *Global Social Sciences Review*, *10*(4), 230-243. https://doi.org/10.31703/gssr.2025(X-IV).19
CHICAGO	Fahim, Mazharul Ahosan, Ashrafa Hossain, and Khadija Jahan Mukta. 2025. "Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study Across Demographic Groups." *Global Social Sciences Review* 10 (4):230-243. doi: 10.31703/gssr.2025(X-IV).19.
HARVARD	FAHIM, M. A., HOSSAIN, A. & MUKTA, K. J. 2025. Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study Across Demographic Groups. *Global Social Sciences Review,* 10, 230-243.
MHRA	Fahim, Mazharul Ahosan, Ashrafa Hossain, and Khadija Jahan Mukta. 2025. 'Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study Across Demographic Groups', *Global Social Sciences Review*, 10: 230-43.
MLA	Fahim, Mazharul Ahosan, Ashrafa Hossain, and Khadija Jahan Mukta. "Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study across Demographic Groups." *Global Social Sciences Review* 10.4 (2025): 230-43. Print.
OXFORD	Fahim, Mazharul Ahosan, Hossain, Ashrafa, and Mukta, Khadija Jahan (2025), 'Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study Across Demographic Groups', *Global Social Sciences Review,* 10 (4), 230-43.
TURABIAN	Fahim, Mazharul Ahosan, Ashrafa Hossain, and Khadija Jahan Mukta. "Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study across Demographic Groups." *Global Social Sciences Review* 10, no. 4 (2025): 230-43. https://dx.doi.org/10.31703/gssr.2025(X-IV).19.

©2025 The Author(s). This work is licensed under a Creative Commons Attribution-Non-Commercial-No Derivatives 4.0 International License.

e-ISSN:2616-793x Volume: X (2025) Issue: IV-Fall (December 2025) p-ISSN:2520-0348

Global Social Sciences Review
www.gssrjournal.com
DOI:http://dx.doi.org/10.31703/gssr

Pages: 230-243 URL: https://doi.org/10.31703/gssr.2025(X-IV).19 Doi: 10.31703/gssr.2025(X-IV).19

Title

Algorithmic Bias in Depression Screening Protocol for a Prospective Multicenter Diagnostic Accuracy Study Across Demographic Groups

Authors:

Mazharul Ahosan Fahim: (Corresponding Author)
Department of Sociology, East Tennessee State University, Johnson City, Tennessee, USA.
(Email: fahimmazharulahosan@gmail.com)
ORCID: https://orcid.org/0009-0002-2123-2735

Ashrafa Hossain: Department of Psychology, Chittagong College, National University of Bangladesh, Chittagong, Bangladesh.
ORCID: https://orcid.org/0009-0006-0627-9745

Khadija Jahan Mukta: Department of Psychology, Goldsmiths University of London, London, UK.
ORCID: https://orcid.org/0009-0006-4991-450X

Contents

- Introduction
- Methods
- Design and reporting framework
- Population and recruitment
- Index test and comparator
- Reference standard and blinding
- Demographic attributes and data governance
- Outcomes and estimands
- Statistical analysis
- Bias, reproducibility, and patient safety
- Requirements before recruitment
- Planned reporting and interpretation
- Discussion
- Conclusion
- References

Abstract

Depression screening models may perform differently across demographic groups despite satisfactory overall accuracy. This protocol describes an independent, paired evaluation of a locked depression screening model and the Patient Health Questionnaire-9 (PHQ-9) across age, race and ethnicity, gender, and preferred-language groups. Consecutive eligible adults attending participating primary care sites will complete both tests during the same clinical episode. All consenting participants, regardless of screening results, will be offered a blinded clinician-administered diagnostic interview for current major depressive disorder within 14 days. The model, threshold, subgroup definitions, and sample size will be finalized and registered before recruitment. The primary outcome is between-group differences in sensitivity at the prespecified threshold. Secondary analyses will assess specificity, calibration, missed cases, and referral workload. Missing data and interview noncompletion will undergo sensitivity analyses. This manuscript presents a planned study; no participant data, diagnostic accuracy, fairness, or clinical-benefit findings are reported.

Keywords:

Depression Screening, Diagnostic Accuracy, Algorithmic Bias, Health Equity; PHQ-9, Prospective Validation, Study Protocol

Introduction

Depression screening is a "household" entry point to assessment and treatment. The outcome of a screening does not make the diagnosis of major depressive disorder; however, it can be a signal for a subsequent discussion with someone or for delaying a discussion with someone. There is consensus by the US Preventive Services Task Force that depression screening should be implemented in adults when systems in place can enable an evaluation of the depression diagnosis and evidence-based interventions following a positive screening result (USPSTF, 2023). The closer the

©2025 The Author(s). This work is licensed under a Creative Commons Attribution-Non-Commercial-No Derivatives 4.0 International License.

score is to the perfect score, the better the quality of that doorway, and as more clinics are implementing scores based on questionnaires or electronic records, that doorway is hardly average.

If a model is able to reasonably rank patients in the overall population, it may incorrectly leave out a higher proportion of patients in one group, and it may incorrectly identify a higher proportion of patients as having a higher risk in another group or it may counsel more patients to go for an interview than are necessary. These are likely to be an artifact of how the model has been trained, but also could represent different ways in which language and symptoms of conditions are expressed, inequitable access to care, or inconsistencies in clinical labels for training and testing the tool. There is a social context to race and ethnicity, and disparities between groups serve as indicators and not proof of either biological differences in depression or biological differences in groups.

The PHQ-9 is a practical comparator because its original validation and later participant-level meta-analysis established a well studied score threshold, while also showing that the choice of diagnostic interview affects estimated sensitivity (Kroenke et al., 2001; Levis et al., 2019). Evidence from NHANES supported broad measurement invariance of the PHQ-9 across major US sex, race and ethnicity, and education groups, but that finding does not prove every translated administration or local clinical workflow is equivalent (Patel et al., 2019). Existing studies have also examined fairness in depression prediction across populations and smartphone measures (Dang et al., 2024; Adler et al., 2024). The proposed study narrows the question to clinically meaningful diagnostic accuracy at a referral threshold, verified against an interview and reported separately across groups. AI tools are beginning to be explored in the field of psychiatric care recently, but there are questions about algorithmic bias. (Samiun et al., 2025). Thus, it is crucial to evaluate the construct of mental health screening algorithm's diagnostic validity among different demographics. The literature testing the PHQ-9 also identifies differences and qualifications with regard to cultural, racial and demographic groups (Baas et al., 2011, Chae et al., 2025, Harry et al., 2023, Reich et al., 2018). Studies in speech-based depression detection also have reported gender related performance distortion (Bailey & Plumbley, 2021).

The proposed study questions include whether there are differences across age, prespecified race/ethnicity, gender, and preferred-language groups with regard to sensitivity when using the locked model threshold, and whether there are differences in specificity and expected referral workload across these groups as well; and how missing interviews, language, and site mix influence interpretation. This is a provisional threshold for clinical concern for the purposes of sample size planning and interpretation, motivated by an absolute difference in sensitivity of 0.10, rather than an observed difference or evidence of unfairness. The final threshold group comparisons and primary analysis family needs to be approved by clinical and patient steering group prior to registration.

One question point of importance is that the question is clinically important - it acts at a single encounter, depression, symptom reporting and access to an interview are serious clinical issues and happen over time. A numerical score is only beneficial if understood at this meeting in terms of meaning and implications. A patient with a negative result can be comforted and a patient with a positive result may require a timely clinical evaluation. If there are differences in how the system is applied, by language or social position, these may exacerbate any inequity in access. The comparison thus relates to a “real” decision of whether to provide additional evaluation at this time. It does not consider a higher AUC to necessarily be good evidence that the screen leads to better care.

Evaluations published in the literature may vary with regard to its condition predicted. Some refer to a cutoff on the questionnaire some to an EHR diagnosis that comes months later and some to a clinical interview (as a means of identification) for a current disorder. There are several questions these labels address. An algorithm can be taught with past diagnoses and learn who had an opportunity to be diagnosed including trends of appointments, insurance or documentation. A model that is able to duplicate one of these labels may look like it's on track on the chart, but it may be missing people who weren't reported on the chart. We chose primary outcome of currently BDII

interview-positive MDD as the condition of primary interest to anchor the evaluation to a clinically interpretable condition. The approach doesn't mean the interview is somehow like a complete assessment of distress or that any particular mental health outcome is necessarily the most important one to take into account. The difference between a symptom score and an interview defined outcome/input is important in other depression models, where the models differ in the measurement properties of their outcome or input (Vu et al., 2025; Xu et al., 2025).

The next issue has to do with denominators. A smaller number of positive screens for a sub-group may give the impression of a less burdened sub-group when in fact a larger proportion of the true number of positive screens may not have been identified. The pathway that leads to the denominator – the number of cases recognized is the sensitivity, it makes no sense without understanding that pathway. Specificity measures the percentage of people who are not afflicted with the disorder who do not receive a positive test result and a high specificity does not necessarily correlate with high sensitivity. Predictive values will vary from tool to tool based on condition prevalence and hence follow different trends as a tool moves between two clinics. These measures are essential for a fair analysis but also the screening and diagnostics assessment, who was the target population offered the screening and who was the population that completed a diagnostic assessment. Meta-analysis conducted on the PHQ-9 reveals why it is not possible to calculate a portable accuracy estimate that comes from combining values for setting, reference interviews and frequency of occurrence (Levis et al., 2020; López Malo Vázquez de Lara et al., 2023).

The term bias has two similar yet different meanings. Statistical bias: Distortion(s) in an estimate arising from one or more of recruitment, verification, measurement or analysis. Algorithmic bias is an analysis of the difference between the model's errors and/or effects that is systematic across groups and potentially has ethical dimensions. Attending to the difference is a beginning to inquiry. This observed difference could be the result of a number of different factors such as nonrepresentative training data, the questionnaire being interpreted differently, the quality of the translation, missing inputs, differential follow up, and selection of an average-population threshold. To propose a technical (i.e., non-biological) correction, it is important that the effect(s) due to these mechanisms are distinguished from one another. Threshold adjustment can't fix imbalanced access to an interview or treatment. However, it is also important to consider the context of the application of fairness interventions and the error tradeoffs they introduce in clinical risk prediction and detection of depression (Pfohl et al., 2021).

A prospective same-patient comparison of a locked model (using the PHQ-9) and an interview PHQ independently of the two screens is planned. Previous research also shows that there are age-group differences in depression prediction (Adler et al., 2024, Dang et al., 2024, Harry et al., 2023) and in questionnaires. This protocol builds out on that literature by delineating screening verification across screening outcomes as well as the translation of screening error to missed cases or potential referrals per 1,000 adults screened. The result of the design will give new empirical information depending on the completed multicenter data and the model chosen for validation.

Methods:

The conceptual target population would be adults who visit for routine primary care visits and who may have either one of the index tests performed. The study is not a case-control sample, put together from known depressed and non depressed patients. This enriched sample can give a good estimate of some types of conditional accuracy measures if they can assume it is representative, but it is not an accurate representation of the clinical prevalence and the number of referrals that a screening policy would generate. The unit of enrollment is the "person at the eligible encounter. There were 2-3 encounters within the unit of the principal prediction. The recruitment time period (or the part of it), weekday/evening or interruption of the recruitment period will be noted in the study log indicating selection in the validation sample.

All the above will be approved by the steering group, consisting of primary care and mental health clinicians, statisticians, data engineers and patient representatives, before recruitment begins. Amendments will be made to a protocol log that is dated. Instruments which have been translated

prior to enrollment will be changes once different personnel have been recruited and interviews have been conducted and their impact on comparability will be communicated.

Design and reporting framework

This is a future multicenter (prospective) diagnostic accuracy protocol and not a diagnostic study. Clinics that participate will be identified and their pathways of providing screening pre-ethics submission. The consecutive recruitment process will be recorded in the predefined clinic sessions in an eligibility log. Prior to the enrollment of the first participant the protocol, the statistical analysis plan, model version, threshold derivation and primary subgroup contrasts will be registered. The eventual results report will follow STARD 2015 and STARD-AI and include relevant TRIPOD+AI items for model provenance/validation, Risk of bias assessment is presented in the format of PROBAST+AI (Bossuyt et al., 2015; Sounderajah et al., 2025; Collins et al., 2024; Moons et al., 2025).

Population and recruitment

The proposed intended use is in adults age 18–>65 for routine primary care in the detection of major depressive disorder. The potentially main group will not include individuals who are actively depressed from a documented major depression as well as those who require immediate emergency treatment, and that accommodations will be provided and that language will be confirmed prior to making the determination that an individual is not able to complete interview. These criteria need to be considered and agreed prior to registration, in reference to the chosen model. Failure to complete a demographic response will not disqualify a participant, the non-demographic responses must be missing as well. Eligibility, approaches, consent, test/completion of interview, refusal, and reason for lack of assessment (by site) will all be documented in the screening log.

To avoid overweighting frequent users, and to make use of the first time period the patient was eligible for screening, the first screening encounter per patient will be used in the primary analysis. If the model is going to be used more than once, possibly a secondary clustered analysis can be done which in that scenario would include later encounters. The language of each questionnaire/interpretation of interview will be noted on site and whether translation/interpretation was required.

Before this protocol can be activated an identified and accessible model with version must be available. It will have all the above information on its model card: the developer, training population, intended use, the complete predictor list, coding, look back windows, missing value rules, transformations, calibration, software version or hash, and the time predictions are available. A blind audit will generate re-sampled feature vectors from the source records to re-sense the leakage of these post screening. If a suitable model is not available, the model development should be done following a development plan that is separately specified and under an independent validation cohort, which the present protocol cannot freeze a developed model over the validation cohort.

The decision from the screening encounter (the exact model probability, locked decision, and PHQ-9 decision/complete PHQ-9 score) will be captured prior to thresholding. TWO tests will be correlated on a patient level. The team will record their time taken and any overlap of item in the PHQ-9 and the model predictor. No inputs to the model may come from a validation diagnosis, a post encounter code, treatment, referral event or an interview.

Index test and comparator

This is because the selected model's probability of having current major depressive disorder will be the index test. The complete specification of it and threshold setting dataset will be sent together with the registered protocol. Only existing development data shall be used to select the threshold; a pre-specified rule for referring to interview workload shall be used to choose the threshold(s) amongst qualifying thresholds with the approval of the steering group for a minimum acceptable sensitivity and feasible interview capacity respectively. This clinical review is an interim one to evaluate previously proposed sensitivity target of 0.85. If none of the cut-off values are achieved within the approved target, there is something with the protocol and its use which should be reconsidered; the model will not be silently returned with the validation data.

Model threshold, preprocessing and software will be fixed prior to enrolment and will not be changed, and in the primary analysis. The sum of the nine items, and cutoff of ≥10, will be used for the PHQ-9 comparator. The instrument used will have a prespecified "instrument rule" with missing items: an incomplete instrument screening will not be considered a negative screening. There is no difference in safety assessment for any response to self harm, irrespective of research scores and usual clinical practice is followed. If a threshold or recalibration (post outcome) has been chosen other than the standard model threshold, the evaluation will be marked as exploratory.

Reference standard and blinding

The standard to be suggested is a face-to-face SCID-5 interview by a trained clinician within seven days and at most 14 days after both screens for current major depression. The interview version, validated translations, procedures of the interpreter, training of the interviewers and adjudication of cases of uncertain diagnosis/interpretation will be determined prior to registration. The model probability, model decision, and PHQ-9 total will be masked from interviewers, as will be the study referral classification. The date of the interview, interview language, and the chronicity of the date of the screening, and relevant symptom change will be recorded, and analyses will consider them within the 7 day window, testing timing sensitivity.

Each participant that consents, with both the model and PHQ-9, will be presented with the same reference interview, irrespective of the polarity of both model and PHQ-9. There is no intention to do any result based sampling of interviews. Failure to schedule or refusal, illness, language barriers and loss to follow-up will be recorded and the participants will be contacted by a uniform protocol. Non-completion as reported below will be broken out by site, index result and demographic group. Case estimates will be complemented by inverse-probability ‑ weighting relying on all the variables measured for all the screened individuals and bounds-of-sensitivity analyses for outcomes that may be missing not ‑ at ‑ random.

Demographic categories must be selected to align with the local context, along with the clinical question – and should not be copied from a training database. We'll keep the participants' exact answers in a secure source field and develop analysis categories using documented mapping which can be audited. The multiple race selections, a specific field for ethnicity, description of gender, and a language besides the predominant one on the site should not be eliminated in the cleaning process. While public tables may suppress very small cells to protect privacy, the internal analysis would process and keep the cells and indicate an impact on disparity due to suppression. Decisions on group names for publications should be made with patients. In the case of cross-cultural validation studies, it is important for the language of the questionnaire as well as the behaviour of the items to be recorded and assessed (Carroll et al., 2020; Reich et al., 2018).

There are a lot of seemingly neutral variables that can affect prediction based upon a protected attribute. Information about resources and opportunities to access care can be captured in zip code, digital engagement, missed visits and in prior diagnosis. A simple 'ageing at a rate' approach to removing race or gender from the inputs to a model is therefore not sufficient to make the process fair. On the other hand recording an attribute on an audit has no requirement to use it to change a person's prediction. Our preset primary audit is based on self-reported attributes of the observed model behavior. If someone proposes an additional attribute for a recalibration model, an independent clinical, legal and patient-centred argument establishing the balance of humanity and harms needs to be made.

Demographic attributes and data governance

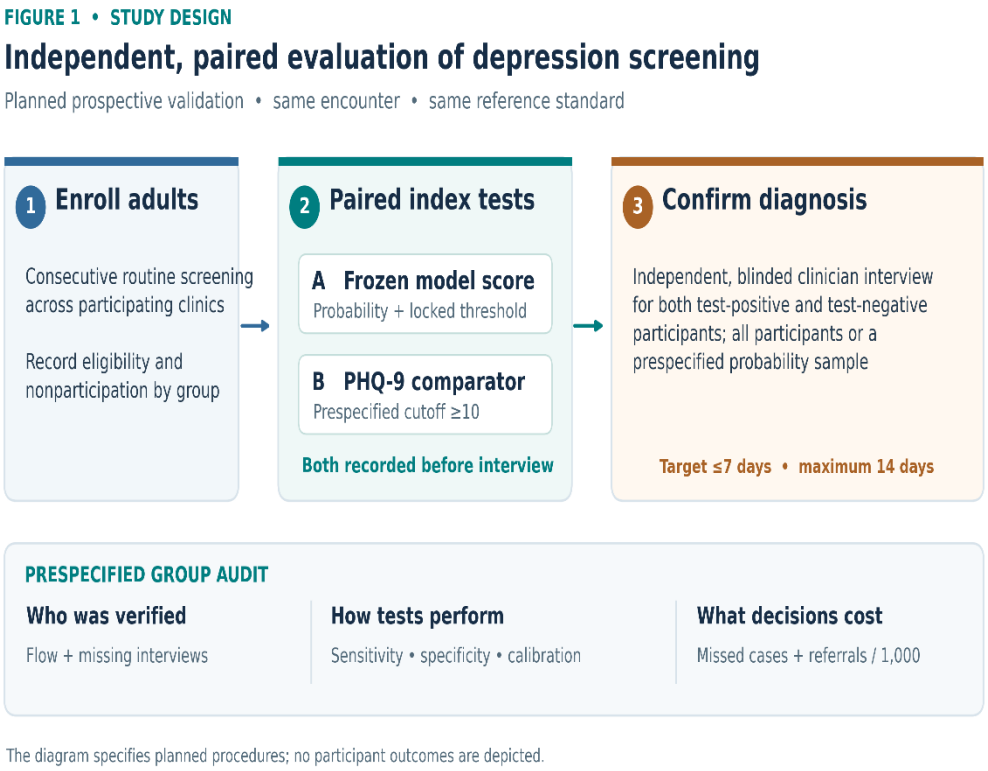
Age, race, ethnicity, gender identity and preferred language will be self-reported with questions, which have been approved locally, allowing for multiple answers and none at all. The collection of sex assigned at birth will only be for an approved purpose. The registration will designate a name for each basic category, comparison, and disclosure rule based on sites and model participants and patient representatives without any post hoc rules. There will be separate numbers for missing, unknown and declined responses. The counts for small groups will be made available and displayed with the uncertainty (where possible, specify the

need of privacy). However the imprecise estimates do not represent equal performance.

The main model assessment will be to look at those factors which are 'protected' as audit variables. The appropriateness of their inclusion into the model should depend on the design idea of the intended use and needs to be communicated. Access to identifiable data will be role limited and the public version of replication files shall not contain data for identifiable patients unless there is consent and governance to release deidentified patient data, only codes, data definitions and synthetic examples shall be released.

Figure 1

Prospective diagnostic accuracy workflow. Both index tests precede the blinded interview. Boxes describe planned procedures, not observed participants or findings.



The main target population are the responding adults that meet all the criteria in the pre-defined workflow for first detection in the participating sites. The main estimand is the absolute difference in sensitivity in the case of registered groups at a single locked threshold, by applying the same definition of current MDD based on the same interview. Interviews for all eliminate the need for planned verification by test positivity, but some who might not participate or be interviewed may bias observed accuracy. Estimates will be reported on for the screened population and cannot be generalized to those attending clinics but not screened, and those who did not attend the clinics.

Absolute differences will be given more emphasis than ratios. The change from 0.90 to 0.80 sensitivity, no matter expressed as a ratio, means 10 more "missed" cases per 100 people with the condition. However, the number of missed cases per 1,000 screened also is dependent on prevalence. There are two presentations that we will discuss. A simple complete-verification scenario with estimated prevalence of 12% (using the interview as the criterion) would suggest a sensitivity of 0.80 which means that about 24 cases would have been missed out of 1,000 examined; this is a mathematical occurrence and is not observed in the study results. An appropriate balance also

requires the amount of interviews produced by false positives and the capacity of the local referral service.

Calibrating is the process we'll examine; the whole story will be told by no single graphical curve. An intercept is an average overprediction or underprediction (corresponding to an average poor skill score); a slope is an indication of whether the range of predictions is too extreme or too mild (corresponding to an average poor skill score). These risk distribution plots will be used to demonstrate the number of patients that fall into each range of probability to try to avoid getting confused by a large number of patients with a seemingly high deviation that are actually in an almost-empty tail of the distribution. Confidence bands will be computed and displayed for subgroup curves along with the denominators. While a score on discrimination and/or calibration may be added to these displays, it should not be used as a substitute for an explanation of either of the two. The fairness claims will be linked with the point at which clinicians would act with regard to the referral threshold.

Outcomes and estimands

The sensitivity for each group g at each model threshold t is TPg/(TPg + FNg), specificity is TNg/(TNg + FPg), and PPV is TPg/(TPg + FPg). The main analysis will focus on estimating absolute differences in sensitivity for the pre-defined group pairs, 95% confidence intervals for the difference, absolute raw numbers of interviews confirmed cases and the percentage of interviews completed in each group. The categories for the comparison and reference group (if one has been used), must appear in the study registration at the start of the study and not in any of the validation labels. The clinical concern threshold of the absolute difference of 0.10 is proposed and is not a statistical or ethical cut-off.

Secondary endpoints are specificity, false-positive differences, positive predictive values with observed prevalence, AUROCs,_area under precision recall curves', Brier score, intercept of calibration curve, slope of calibration curve and clinically relevant calibration plots. Consequences at the locked threshold will be provided in the context of the estimation and will be given as missed cases at that threshold confirmed by interviews and positive cases who screened per 1,000 screened with uncertainty intervals. These are anticipated screening outcomes; since the number of actual clinical referrals, treatment and patient outcomes cannot be measured as results of model implementation, these are anticipated instead.

A group-related threshold analysis will be an exploratory approach. Mechanical matching can decrease mechanically sensitivity and can obscure the reasons for differences and cause a decrease in specificity. Any recommended threshold changes should need only clinical review, evidence of benefit and the input of stakeholders. The study will not take as its equivalence statistical parity and ethical fairness.

Measure	Numerator / definition	Clinical reading
Sensitivity	True positives / all interview-confirmed cases	Share of cases flagged for assessment
Specificity	True negatives / all interview-negative people	Share spared an unnecessary referral
Positive predictive value	True positives / all model-positive people	Yield of positive screens; varies with prevalence
Calibration	Observed case frequency against predicted risk	Whether estimated risks can be trusted
Missed cases per 1,000	False negatives / all screened × 1,000	Potential cases not prompted for assessment

This data freeze will occur prior to the last model run. A flow diagram, missingness report, cross-tabulations of the index test with the interview and subgroup estimates of the analysis file are generated with a reproducible script. Any discrepancies will be arbitrated between the screening log and model output, and against a dated audit trail. The analysis will be separately done for interview-positive and interview-negative participants overall and by site to show the distribution of scores. This enables a reader to determine if a group difference indicates score separation, which means differential scores resulting from differential levels or differential places a threshold on a specific score distribution.

The same individuals, both with the algorithm and with the PHQ-9, will be paired and have an completed PHQ-9 interview. The number of cases identified by both, algorithm only, PHQ-9 only and neither will be calculated. The 4-way accounting of the same population of interview-negatives will be described to describe what their referral workload is. Uncertainty for differences in sensitivity and specificity will be estimated by a paired bootstrap, which will maintain the dependence between tests. The differences in AUROC alone, will be secondary in providing the immediate referral decision. If the PHQ-9 items are used in the algorithm, it will be specified that the tests are not independent technologies.

Sparse intersectional cells can lead ordinary asymptotic tests to give false confidence or may be inadequate due to the lack of an event in a cell. Exact counts and appropriate intervals will be reported and in a preformed hierarchical sensitivity analysis, partial pooling will be discussed. Partial pooling can improve precision of estimates, but can lead to an underestimate of a real local issue, and toward the overall pattern. For this reason, "pooled" estimates will be provided next to the "observed proportions" that are unsmoothed, and no interval crossing o will be interpreted as being "safe. Limitation and the widest possible aggregate is reported if the cell value was not able to be released.

The frozen version of the model will be assessed in every site and for the early and late recruitment events at the same threshold. Leave one site out retraining will not be carried out in the primary validation it would alter the validation test being assessed. If there are multiple recalibration exercises, the ones that are performed will be separate from the reported test set and will be marked as exploratory. The apparent pooled adequacy will not compensate for inadequate denominators at sites and/or in underrepresented groups.

Emphasis will be on the size and confidence interval of each difference, number of people impacted at realistic clinic volume and clinical tradeoffs. The registered primary contrast family will be adjusted for multiplicity using Holm adjustment and all the estimates and unadjusted intervals for the subgroups will still be visible. Any estimates that are exploratory intersectional will be distinguished as such. In a small group failing to reject a difference will not be said to be evidence of equitable performance.

Statistical analysis

A locked script will be provided to compute the full cohort flow each site-by-group, the elements of the confusion matrix, absolute differences in sensitivities, sensitivity, specificity of the model and PHQ-9. 95% confianza intervals of the differences between paired tests and between group differences will be estimated at the individual patient level using a bootstrap approach to resampling, grouped by site. The first eligible encounter, if not multiple encounters for the patient, will be considered as the primary analysis and the multiple encounters, if collected, will be added as the sensitivity analysis in clustered analysis. Reference groups and statistical plan will be kept prior to recruitment and then implemented post data freeze.

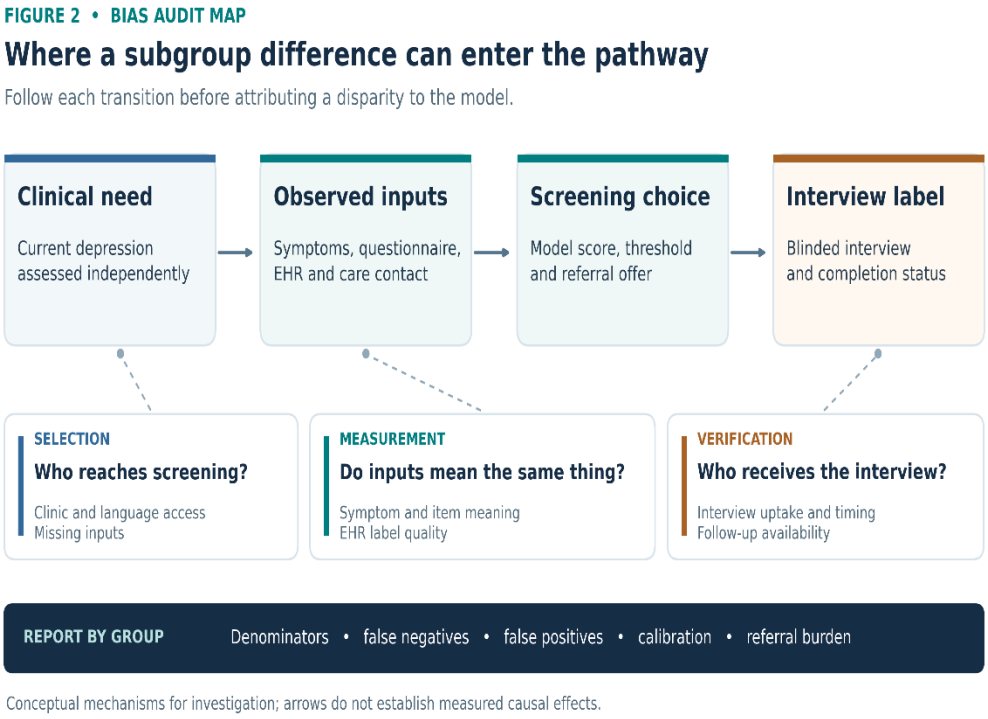
Pairwise classification of interview identified cases and noncases will be used to determine those who were identified by both screens, by the model alone, by PHQ-9 alone and by neither. Probability distributions and calibration (by group) will be shown in a plot with denominators and uncertainty bands. In the locked threshold screen and in exploratory threshold curves sensitivity will be shown as # of referrals projected per 1000 screened. Site-adjustment, adjustment for too short or too long an interval, and adjustment for recorded case-mix, will be of a descriptive nature only and will not be given as evidence that a disparity measured is harmless.

Per missing predictor variable(s), the percentages of the questionnaires, missing demographic response(s), and uncompleted interviews per site and per group will be reported. The development version of imputation rules will be "disparmed" prior to validation. Inverse-probability weights will use prespecified variables for the individuals that were screened (showing weight distributions, positivity, and extreme weights for missed interviews). A weighted estimate and complete case estimate plus a prespecified pattern-mixture (or tipping-point) analysis will be reported separately and together. Reference Diagnoses will NEVER be checked off as "observed".

Target final enrollment is determined prior to Ethics submission and Registration. We will simulate minimum 10,000 multicenter cohorts run on documented estimates of prevalence of depression, the number in each primary population, variation due to site, the number of interviews that will be completed, sensitivity when the null hypothesis is true, and the difference of 0.10, as well as planned Holm adjusted comparisons between primary populations. Feasibility will be controlled by the smallest-planned group. Advantage is that the prespecified decision criterion (e.g., requiring at least 80% simulated power for the clinically important contrast) is accompanied by the expected 95% interval width, and assumptions, code, the sample size specification to be targeted and the maximum number of participants that can be approached need to be added to the registered protocol prior to approaching any participant. In the absence of on-the-ground site and group estimates, an honest number of participants in the enrolment cannot be given here.

Figure 2

Conceptual pathways that can produce apparent subgroup differences. The diagram distinguishes symptoms, observed inputs, selection for screening or interview, and decisions; it does not claim a causal effect has been measured.



It is important to note the study team will be performing a directed audit of participants who do not have an interview. Where applicable, reasons will differentiate between refusal, scheduling failure, inability to contact, language access, illness and death. The above reasons are not mutually exclusive and may be similar to depression as well as demographic group. This probability of

verification will be modeled using available data for all the participants in the screening (site and index result) prior to the generation of inverse probability weights. Extremes weights will be shown, and will be truncated if their weight is found to be very sensitive in a documented sensitivity analysis. A tipping-point type of exercise will ask how much more or less that the number of unverified cases need to be from the verified cases in order to switch the underlying conclusion of the primary discrepancy.

During the diagnostic validation the prospective score will be confidential from the clinical staff to minimise incorporation and feedback bias. Normal assessment and self harm precautions are maintained. Note that the present study is a purely observational study and estimates only test performance and not whether model use leads to an increase in completed assessment, treatment initiation, symptoms, or harm. These outcomes will need to be implemented in a separately designed implementation study or pragmatic trial under prospective monitoring.

Bias, reproducibility, and patient safety

These include selective attendance and recruitment; the noncompletion of the interviews to get to a referent; differences in translation of measures; temporal leakage; and drift between training and the intended situation. A sample of feature time stamps and interview records will be audited by the blinded reviewer. The protocol deviations, versions of software, model card, an aggregate participant flow, and code (as allowed by privacy protections) will be reported in the completed study. STARD-AI and TRIPOD+AI will eventually be used as guidance for the empirical report and not considered to be ongoing checklists for this protocol.

Observational algorithm during algorithm validation period. They should not make predictions which supersede a clinician's sense of proportion. The algorithm will not replace existing urgent assessment procedures for self-harm content–the wider context of this case will be assessed as appropriate. This paper does not discuss the prediction of suicidal behaviour, but rather the screening for depression. Any identified difference between the algorithms should be reviewed by a patient and/or clinician advisory group independent of the groups creating the algorithms to ascertain the impacts on referral flows and acceptability of proposed changes.

Requirements before recruitment

The study team will need to identify the model and its existing development track, name all of the study's settings and screening periods, approve the population of the study and the language(s) that were validated for use by the interview process, lock the model and threshold, define all primary demographic categories and contrasts, run the enrollment simulation, seek ethics approval, and enter the protocol and analysis plan. This manuscript is not for either enrolment nor for validation study. Any changes in these decisions must be reported, as well as the date of the amendment, and whether or not validation labels were available.

Planned reporting and interpretation

This manuscript resists making any statements on data from the participants or on any data coming out of the model. The results reported in this document are not for observed outcomes—the recruited cohort, interviewed and not interviewed cohort, prespecified number for each cohort and accuracy estimate with uncertainty will be reported on in a separate results document. Final report will display model and PHQ-9 response data, separately for each site and demographic registered, with the same interview diagnosis.

The final report will describe the number of participants that progressed from being eligible for each test and the blinded interview as well as providing a description of reasons for not progressing in the tests and the blinded interview. Prior to percentages, it will present the cells in the confusion matrix, as well as interview-confirmed cases and noncases, and missing scores. These counts are needed to determine whether the difference in the rate of errors by the sub-groups is due to a subgroup error or whether the estimate is simply not likely to have enough precision to be meaningful.

By specifying sensitivity and specificity, the results display can include calibration, score distribution, and the number of cases projected to be missed and referrals for more investigation for a

positive screen (in the population of screened adults). Operational tradeoff will be able to be seen because of paired discordance level between the model and PHQ-9 and the denominators of groups. Each listing (a threshold curve or post hoc recalibration) will be kept in a separate listing from the locked primary listing (comparison).

If there are differences in sub-groups the interpretation will focus on noncompletion of interviews, instruments (versions of), site mix, and predictors (availability of) before moving on to proposing a mechanism. Both weighted and unweighted results will be reported, as well as the sensitivity bound for the missed interviews. Numerical differences or a post hoc change in threshold does not provide causal explanation nor does it provide solution to inequitable access to health care.

Since statements about the likelihood of a positive are hidden behind the curtain during validation, it is not possible to say if a positive model statement is a determining factor in an attended interview or better treatment outcomes. A future implementation study may encompass the offers of screening, positive screening results, scheduled and completed screening assessments, diagnosis, and treatment options as well as a patient reported experience.

Patient and clinician representatives will be involved in determining which of the missed cases, incorrect/future referrals, and privacy issues are most important to the interpretation of finished results. They will be asked to participate, give decisions, and the decisions and any protocol modifications will be recorded prior to analysis if possible.

Discussion

This protocol evaluates a clinically actionable threshold for the performance of the subgroups in comparison with an independent interview. The main methodological contribution of the study is the verification not related to screen results and paired evaluation with PHQ-9 in the same individuals. Despite the existing literature on demographic bias or in the prediction of depression, there is no literature to provide the results of the test (Adler et al., 2024; Dang et al., 2024; Huang et al., 2024). Any potential difference would only come with the completion of a multi-center analysis, which would determine the difference and its significance and accuracy.

Independent interviews may help to mitigate historical dependency on EHR label diagnoses but have a number of drawbacks that may impact the outcome: language, interviewer interpretation, timing of interviews and noncompletion. The PHQ-9 is a screen, and not a gold standard for diagnosis. These conditions should be brought to light and analyzed for the final report and it cannot be assumed that the comparator is impartial.

Imbalances in the information of differences between the groups may indicate unequal access, measurement, training data and/or clinical context. The existence of a measure gap would indicate investigation but it would not show that this gap, as a result of demographic identity, brought about an error. Model recalibration and group specific thresholds would be interventions that would have to be separately clinically, ethically and externally validated.

At this stage this draft can not be submitted as a registered, finalised protocol. The version of the model, the sites that participate, the nature of the interviews and translations, exact comparisons of demographic characteristics, simulation of enrollment, declaration of ethics, and registration number are not provided. These facts to be fed in must only be from the real sites and investigators and not myopically deduced from the literature. The study will be protocol till completion of data gathering and data analysis.

Validation to be done is observational, and a "blind" validation, or model scores will be hidden during validation. It can provide an approximation of how accurately the diagnostic can be achieved and can estimate the workload for providers who utilize the algorithm for referral, but cannot determine if the algorithm improves or worsens patients' outcomes, access to the assessment or safety. Interpretation may be challenging even after recruitment with small number of intersectional groups and unverified participants. The limitations should be indicated next to the estimates in the final report.

The results that follow should be interpreted with the patients and clinicians from each center with which the collaborating centers are familiar. Race and ethnicity will be regarded as social categories, rather than biological explanations. When privacy issues do not exist, findings will be

disaggregated and descriptions of local screening, language access and follow up procedures will be provided

Conclusion

The manuscript presents a proposed framework for the assessment of the diagnostic accuracy of subgroups in depression screening. A locked model and a clearly described 'first detection population' is required for the design, along with independent interviews on both screening results and on prespecified group comparisons with full denominator and uncertainty. It does not commit itself to any empirical assertions about the accuracy, bias, biaslessness, or usefulness in application of any model. Study-specific and governance aspects are outstanding that need to be done first before recruiting in order to prepare an original-results article after the study has been done and analyzed.

References

Adler, D. A., Stamatis, C. A., Meyerhoff, J., Mohr, D. C., Wang, F., Aranovich, G. J., Sen, S., & Choudhury, T. (2024). Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data. *npj Mental Health Research*, *3*, Article 17. https://doi.org/10.1038/s44184-024-00057-y

Google Scholar Worldcat Fulltext

Baas, K. D., Cramer, A. O. J., Koeter, M. W. J., van de Lisdonk, E. H., van Weert, H. C., & Schene, A. H. (2011). Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). *Journal of Affective Disorders*, *129*(1–3), 229–235. https://doi.org/10.1016/j.jad.2010.08.026

Google Scholar Worldcat Fulltext

Bailey, A., & Plumbley, M. D. (2021). Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)* (pp. 596–600). IEEE. https://doi.org/10.23919/EUSIPCO54536.2021.9615933

Google Scholar Worldcat Fulltext

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., & Cohen, J. F. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, *351*, Article h5527. https://doi.org/10.1136/bmj.h5527

Google Scholar Worldcat Fulltext

Carroll, H. A., Hook, K., Rojas Perez, O. F., Denckla, C., Vince, C. C., Ghebrehiwet, S., Ando, K., Touma, M., Borba, C. P. C., Fricchione, G. L., & Henderson, D. C. (2020). Establishing reliability and validity for mental health screening instruments in resource-constrained settings: Systematic review of the PHQ-9 and key recommendations. *Psychiatry Research*, *291*, Article 113236. https://doi.org/10.1016/j.psychres.2020.113236

Google Scholar Worldcat Fulltext

Chae, D., Lee, J., & Lee, E. (2025). Internal structure of the Patient Health Questionnaire-9: A systematic review and meta-analysis. *Asian Nursing Research*, *19*(1), 1–12. https://doi.org/10.1016/j.anr.2024.12.005

Google Scholar Worldcat Fulltext

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, Article e078378. https://doi.org/10.1136/bmj-2023-078378

Google Scholar Worldcat Fulltext

Costantini, L., Pasquarella, C., Odone, A., Colucci, M. E., Costanza, A., Serafini, G., Aguglia, A., Belvederi Murri, M., Brakoulias, V., Amore, M., Ghaemi, S. N., & Amerio, A. (2021). Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): A systematic review. *Journal of Affective Disorders*, *279*, 473–483. https://doi.org/10.1016/j.jad.2020.09.131

Google Scholar Worldcat Fulltext

Dang, V. N., Cascarano, A., Mulder, R. H., Cecil, C., Zuluaga, M. A., Hernández-González, J., & Lekadir, K. (2024). Fairness and bias correction in machine learning for depression prediction across four study populations. *Scientific Reports*, *14*, Article 7848. https://doi.org/10.1038/s41598-024-58427-7

Google Scholar Worldcat Fulltext

Harry, M. L., Sanchez, K., Ahmedani, B. K., Beck, A. L., Coleman, K. J., Coley, R. Y., Daida, Y. G., Lynch, F. L., Rossom, R. C., Waring, S. C., & Simon, G. E. (2023). Assessing the differential item functioning of PHQ-9 items for diverse racial and ethnic adults with mental health and/or substance use disorder diagnoses: A retrospective cohort study. *Journal of Affective Disorders*, *338*, 402–413. https://doi.org/10.1016/j.jad.2023.04.091

Google Scholar Worldcat Fulltext

Huang, Y., Alvernaz, S., Kim, S. J., Maki, P., Dai, Y., & Peñalver Bernabé, B. (2024). Predicting prenatal depression and assessing model bias using machine learning models. *Biological Psychiatry Global Open Science*, *4*(6), Article 100376. https://doi.org/10.1016/j.bpsgos.2024.100376

Google Scholar Worldcat Fulltext

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Google Scholar Worldcat Fulltext

Levis, B., Benedetti, A., Ioannidis, J. P. A., Sun, Y., Negeri, Z., He, C., Wu, Y., Krishnan, A., Bhandari, P. M., Neupane, D., Imran, M., Rice, D. B., Riehm, K. E., Saadat, N., Azar, M., Boruff, J., Cuijpers, P., Gilbody, S., Kloda, L. A., … Thombs, B. D. (2020). Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: Individual participant data meta-analysis. *Journal of Clinical Epidemiology*,

122, 115–128.e1. https://doi.org/10.1016/j.jclinepi.2020.02.002

Google Scholar Worldcat Fulltext

Levis, B., Benedetti, A., Thombs, B. D., & the DEPRESSD Collaboration. (2019). Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. *BMJ*, *365*, Article l1476. https://doi.org/10.1136/bmj.l1476

Google Scholar Worldcat Fulltext

López Malo Vázquez de Lara, A., Bhandari, P. M., Wu, Y., Levis, B., Thombs, B., Benedetti, A., Sun, Y., He, C., Krishnan, A., Neupane, D., Negeri, Z., Imran, M., Rice, D. B., Riehm, K. E., Saadat, N., Azar, M., Boruff, J., Cuijpers, P., Gilbody, S., … Zhang, Y. (2023). A case study of an individual participant data meta-analysis of diagnostic accuracy showed that prediction regions represented heterogeneity well. *Scientific Reports*, *13*(1), Article 9275. https://doi.org/10.1038/s41598-023-36129-w

Google Scholar Worldcat Fulltext

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., Maier-Hein, L., Liu, X., Logullo, P., McCradden, M. D., Liu, N., … van Smeden, M. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, *388*, Article e082505. https://doi.org/10.1136/bmj-2024-082505

Google Scholar Worldcat Fulltext

Patel, J. S., Oh, Y., Rand, K. L., Wu, W., Cyders, M. A., Kroenke, K., & Stewart, J. C. (2019). Measurement invariance of the Patient Health Questionnaire-9 depression screener in U.S. adults across sex, race/ethnicity, and education level: NHANES 2005–2016. *Depression and Anxiety*, *36*(9), 813–823. https://doi.org/10.1002/da.22940

Google Scholar Worldcat Fulltext

Perez-Toro, P. A., Dineley, J., Iniesta, R., Zhang, Y., Matcham, F., Siddi, S., Lamers, F., Haro, J. M., Penninx, B. W. J. H., Folarin, A. A., Arias-Vergara, T., Orozco-Arroyave, J. R., Nöth, E., Maier, A., Wykes, T., Vairavan, S., Dobson, R., Narayan, V. A., Hotopf, M., & Cummins, N. (2025). Exploring biases related to the use of large language models in a multilingual depression corpus. *Scientific Reports*, *15*(1), Article 36197. https://doi.org/10.1038/s41598-025-19980-x

Google Scholar Worldcat Fulltext

Pfohl, S. R., Foryciarz, A., & Shah, N. H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, *113*, Article 103621. https://doi.org/10.1016/j.jbi.2020.103621

Google Scholar Worldcat Fulltext

Reich, H., Rief, W., Brähler, E., & Mewes, R. (2018). Cross-cultural validation of the German and Turkish versions of the PHQ-9: An IRT approach. *BMC Psychology*, *6*(1), Article 26. https://doi.org/10.1186/s40359-018-0238-z

Google Scholar Worldcat Fulltext

Samiun, M., Rony, M. K. K., Yeasmin, S., Manik, M. M. T. G., Debnath, A., Aziz, M. B., Tasnim, A. F., & Nilima, S. I. (2025). The role of artificial intelligence in managing hospitalized patients with mental illness: A scoping review. *Discover Public Health*, *22*, Article 421. https://doi.org/10.1186/s12982-025-00814-0

Google Scholar Worldcat Fulltext

Sounderajah, V., Guni, A., Liu, X., Collins, G. S., Karthikesalingam, A., Markar, S. R., Golub, R. M., Denniston, A. K., Shetty, S., Moher, D., Bossuyt, P. M., Darzi, A., Ashrafian, H., & STARD-AI Steering Committee. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine*, *31*(1), 3283–3289. https://doi.org/10.1038/s41591-025-03953-8

Google Scholar Worldcat Fulltext

U.S. Preventive Services Task Force. (2023). *Depression and suicide risk in adults: Screening*. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/screening-depression-suicide-risk-adults

Google Scholar Worldcat Fulltext

Vu, T., Dawadi, R., Yamamoto, M., Tay, J. T., Watanabe, N., Kuriya, Y., Oya, A., Tran, P. N. H., & Araki, M. (2025). Prediction of depressive disorder using machine learning approaches: Findings from the NHANES. *BMC Medical Informatics and Decision Making*, *25*(1), Article 83. https://doi.org/10.1186/s12911-025-02903-1

Google Scholar Worldcat Fulltext

Xu, R., Liu, Z., Ouyang, S., Dong, Q., Yan, W., & Xu, D. (2025). Machine learning-driven development of a stratified CES-D screening system: Optimizing depression assessment through adaptive item selection. *BMC Psychiatry*, *25*(1), Article 286. https://doi.org/10.1186/s12888-025-06693-8