BMJ Open

Diagnostic accuracy of the Geriatric Depression Scale-30, Geriatric Depression Scale-15, Geriatric Depression Scale-5 and Geriatric Depression Scale-4 for detecting major depression: protocol for a systematic review and individual participant data meta-analysis
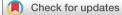
Andrea Benedetti,[1,2,3] Yin Wu,[1,4,5] Brooke Levis,[1,4] Machelle Wilchesky,[4,6] Jill Boruff,[7] John P A Ioannidis,[8,9] Scott B Patten,[10,11] Pim Cuijpers,[12] Ian Shrier,[1,4] Simon Gilbody,[13] Zahinoor Ismail,[10,11,14] Dean McMillan,[13] Nicholas Mitchell,[15] Roy C Ziegelstein,[16] Brett D Thombs[1,3,4,5,17,18]

Correspondence to
Dr Brett D Thombs;
brett.thombs@mcgill.ca

ABSTRACT

Introduction The 30-item Geriatric Depression Scale (GDS-30) and the shorter GDS-15, GDS-5 and GDS-4 are recommended as depression screening tools for elderly individuals. Existing meta-analyses on the diagnostic accuracy of the GDS have not been able to conduct subgroup analyses, have included patients already identified as depressed who would not be screened in practice and have not accounted for possible bias due to selective reporting of results from only better-performing cut-offs in primary studies. Individual participant data meta-analysis (IPDMA), which involves a standard systematic review, then a synthesis of individual participant data, rather than summary results, could address these limitations. The objective of our IPDMA is to generate accuracy estimates to detect major depression for all possible cut-offs of each version of the GDS among studies using different reference standards, separately and among participant subgroups based on age, sex, dementia diagnosis and care settings. In addition, we will use a modelling approach to generate individual participant probabilities for major depression based on GDS scores (rather than a dichotomous cut-off) and participant characteristics (eg, sex, age, dementia status, care setting).

Methods and analysis Individual participant data comparing GDS scores to a major depression diagnosis based on a validated structured or semistructured diagnostic interview will be sought via a systematic review. Data sources will include Medline, Medline In-Process & Other Non-Indexed Citations, PsycINFO and Web of Science. Bivariate random-effects models will be used to estimate diagnostic accuracy parameters for each cut-off of the different versions of the GDS. Prespecified subgroup analyses will be conducted. Risk of bias will be assessed with the Quality Assessment of Diagnostic Accuracy Studies-2 tool.

Ethics and dissemination The findings of this study will be of interest to stakeholders involved in research, clinical practice and policy.

PROSPERO registration number CRD42018104329.

Strengths and limitations of this study

► This study will use individual participant data to estimate diagnostic accuracy for all relevant cut-off scores of the different versions of the Geriatric Depression Scale. Using data from all participants at each cut-off score will overcome limitations related to selective cut-off reporting in primary study publications.

► This study will conduct analyses that exclude patients with current diagnoses of depression or who are undergoing mental health treatment, including antidepressants, at the time of study enrolment, as these patients would not be screened in clinical practice. This will overcome potential bias in primary diagnostic test accuracy studies where these patients are often included.

► This study will include subgroup analyses of diagnostic accuracy across different reference standards and by participant characteristics (eg, sex, age, dementia status, care setting).

► A potential limitation is that the success of the study depends on the ability to obtain the relevant individual participant data and to avoid selective availability of studies with better or worse accuracy results. We do not know the proportion of eligible datasets that will be possible to include in the study.

BACKGROUND

Major depression is present in 5%–10% of the geriatric population internationally.[1 2] Effective treatments for depression are available, but identification is often haphazard. Physicians may fail to recognise up to half of all patients with depression, and most patients with depression do not receive minimally adequate care.[3 4] At the same time, there is a high rate of overdiagnosis and overtreatment, and the majority of patients who are treated do not meet diagnostic criteria.[5–7] Diagnosis of elderly individuals can be particularly difficult for clinicians due to factors such as cognitive impairments, social stigma, medical comorbidity and atypical or vague clinical presentation.[1 8 9]

Some Canadian and international geriatric care organisations recommend screening elderly adults for depression,[10–13] but the Canadian Task Force on Preventive Health Care (CTFPHC), for instance, does not recommend depression screening, including for geriatric individuals.[14] The CTFPHC has expressed concern that published studies may overstate the accuracy of depression screening tools and that screening could lead to high rates of false positive tests, and still not improve depression outcomes.[14]

The 30-item Geriatric Depression Scale (GDS-30) and the GDS-15, GDS-5 and GDS-4, which are 15-item, 5-item and 4-item subsets of the GDS-30, are commonly recommended as depression screening tools for elderly individuals.[15–17] As with other depression screening tools, primary studies on the diagnostic accuracy of the different versions of the GDS have been limited by (1) small samples; (2) the selective reporting of results for cut-offs when they perform well in a given sample, but not when they perform poorly; (3) the inclusion of patients already known by clinicians to have depression and (4) the inability to conduct subgroup analyses (eg, different age groups, dementia diagnosis, care settings) due to small sample sizes. Conventional meta-analyses of the GDS or short versions of the GDS that have synthesised published summary data have not been able to conduct subgroup analyses or exclude already diagnosed patients,[15 16 18] and concerns have been raised about bias in these meta-analyses due to selective cut-off reporting in primary studies that could not be addressed.[18]

Individual participant data meta-analysis (IPDMA), which involves a standard systematic review, followed by synthesis of actual participant data from primary studies, rather than aggregating summary data, can address these problems by including actual participant data from all studies.[19 20] In the context of evaluating the diagnostic accuracy of depression screening tools, IPDMA has three major advantages compared with conventional meta-analyses. First, for the conventional binary screening approach, IPDMA can address bias from the selective publication of diagnostic accuracy results for well-performing cut-offs from small studies since accuracy can be evaluated across all relevant cut-offs for all participants. Second, IPDMA allows the appropriate exclusion of already diagnosed or

already treated patients when primary studies have data on existing diagnoses and treatment. Third, an IPDMA with large numbers of participants and major depression cases would allow subgroup analyses by study-level factors (eg, study setting, risk of bias factors) and individual factors that may influence screening accuracy (eg, age, sex, dementia diagnosis). Finally, a large IPD database would allow the development of a predictive algorithm to generate estimates of the probability of having major depression based on participant characteristics and actual GDS scores, rather than binary classifications of individuals as simply negative or positive based on screening results. This is important because, for instance, an individual with a score of 0 on the GDS-30 may have a lower likelihood of having depression than an individual with a substantially higher, but sub-cut-off, score of 10. Using a dichotomous cut-off method, however, both would be classified as negative screens and assigned the same probability of having depression.

One of the downsides of IPDMAs is that they are resource intensive. In addition, if the primary datasets obtained are not representative of all primary studies, the IPDMA could be biased.[19–22] In a previous IPDMA of the Patient Health Questionnaire-9 (PHQ-9) screening tool, which was the first IPDMA of the diagnostic accuracy of a depression screening tool,[23] we were able to synthesise 58 of 72 eligible primary datasets (17 357 participants, 2312 major depression cases). This suggests that investigators are generally able and willing to provide primary data from studies of the diagnostic accuracy of depression screening tools for use in IPDMA. A preliminary PubMed search for the GDS verified the existence of enough primary studies (more than 100 potentially eligible datasets that appear to have at least 30 000 participants, 4000 cases) to make IPDMA feasible for the GDS.

Thus, the objectives of this IPDMA are to evaluate the diagnostic accuracy of the GDS-30, GDS-15, GDS-5 and GDS-4 among studies using different reference standards, separately; among participant subgroups based on age, sex, dementia diagnosis and care settings; and excluding participants identified as already diagnosed or treated for depression. Furthermore, a prediction model will be generated.

METHODS AND ANALYSIS

This systematic review has been funded by the Canadian Institutes of Health Research (Funding Reference Number PJT-156365).

The IPDMA has been designed and will be conducted in accordance with best-practice standards as elaborated in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (DTA)[24] and other key sources.[19 20 25] Results will be reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of DTA Studies (PRISMA-DTA) statement and the PRISMA-IPD statement.[26 27] The IPDMA protocol does not deviate substantively from previous IPDMA protocols

that we have developed and published for other depression screening tools.[23 28 29]

Sources of evidence

The search strategy was developed by a medical librarian and was adapted from a search strategy developed for a similar systematic review to obtain datasets for IPDMA of the PHQ-9 depression screening tool,[23] which was peer reviewed using the Peer Review of the Electronic Search Strategy standard.[30] The search strategy is also similar to strategies that we have used for systematic reviews and IPDMA of the Hospital Anxiety and Depression Scale and Edinburgh Postnatal Depression Scale.[28 29]

We will search Medline, Medline In-Process & Other Non-Indexed Citations, PsycINFO (OvidSP platform) and Web of Science (Web of Knowledge platform). The Medline search strategy for the GDS was validated by testing against already identified publications from our preliminary search. The strategy was then adapted for PsycINFO and Web of Science. We limited our search strategy to these databases based on research showing that adding other databases (eg, EMBASE) when the Medline search is highly sensitive does not identify additional eligible studies.[31] The Cochrane Handbook for Systematic Reviews of DTA[24] suggests combining concepts of the index test and the target conditions, but this was redundant for depression screening tools as these tests are limited to testing for depression. Thus, the search strategy for electronic databases was composed of two concepts: the index test of interest and studies of screening accuracy. There are no published search hedges designed specifically for mental health screening, but key articles were consulted in developing search terms.[32–34] See online supplementary file 1 for detailed information on searches. To supplement electronic searches, we will search reference lists of included publications and relevant reviews, conduct a related articles search using the PubMed 'related articles' feature, and query authors of included studies for unpublished studies. Search results will be uploaded into the citation management database RefWorks (RefWorks-COS, Bethesda, Maryland, USA), and the RefWorks duplicate check function will be used to identify citations retrieved from multiple sources. Unique citations will then be uploaded into the systematic review programme DistillerSR (Evidence Partners, Ottawa, Canada), and DistillerSR will be used to store and track search results and to track results of the review process.

Selection of eligible studies

To conduct the meta-analysis, we will seek primary datasets that allow us to compare GDS scores to major depression diagnostic status. Datasets from articles in any language will be sought for inclusion if they compare results from any version of the GDS to diagnoses of major depressive disorder (MDD) or major depressive episode (MDE) made with a validated diagnostic interview, administered within 2 weeks of the GDS and based on Diagnostic and Statistical Manual (DSM)

or International Classification of Diseases criteria (ICD), which are similar to DSM criteria and generally used outside of North America.

The 2-week criterion was set because that is the duration of symptoms required for a diagnosis of major depression. Datasets where some participants were administered the screening tool within 2 weeks of the diagnostic interview and some participants were not will be included if the original data allows us to identify and select eligible participants. Most primary studies use MDD as the reference standard, but some may use MDE, which is identical with respect to the symptoms of depression, but does not exclude participants with psychotic disorders or a history of manic episodes. If both are available, we will record both and prioritise DSM over ICD and MDE over MDD in analyses. Data from studies where all participants are known to have psychiatric diagnoses, have been referred for mental health evaluation or are undergoing treatment for depression will be excluded, with the exception of participants treated for substance use disorders, for whom depression screening may be considered. The coding manual for inclusion and exclusion decisions is shown in online supplementary file 2.

Two investigators will review articles independently for eligibility. If either reviewer determines that a study may be eligible based on title or abstract review, a full-text article review will be completed. Disagreement between reviewers after full-text review will be resolved by consensus, including a third investigator as necessary. Translators will be used to evaluate titles/abstracts and articles for languages other than those for which team members are fluent. See online supplementary file 3 for a preliminary PRISMA flow of studies figure.

Transfer of data and data management

Authors of studies containing datasets that meet inclusion criteria will be contacted to invite them to contribute primary data for inclusion. Data will only be used from studies that received ethics approval and all data that are transferred will be properly de-identified prior to transfer. Participant data will be cleaned and coded for uniformity across datasets using an already developed codebook, similar to codebooks used in our previous IPDMAs.[23 28 29] Actual data coding and transfer from original studies into the IPD database will be done by a supervised staff or trainee member of the team. Participant characteristics and screening accuracy results for each study using the cleaned datasets will be compared with those from the original datasets to identify any potential discrepancies.

In addition to obtaining original participant-level data, data will also be extracted from the published articles of included studies. We will cross-check the published data with the original participant-level data obtained from each dataset and any inconsistencies will be discussed with the original authors. Corrections will be made as necessary.

Quality assessment

Two reviewers will independently use the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool[35] to assess risk of bias in primary studies. QUADAS-2 incorporates assessments of risk of bias across four core domains: participant selection, the index test, the reference standard, and the flow and timing of assessments. Two reviewers will independently assess risk of bias, and any discrepancies will be resolved by consensus.

Data analysis

Analyses will estimate sensitivity and specificity separately at each cut-off by bivariate random-effects meta-analysis models as described in Riley *et al*.[36] For each GDS version, we will fit these models, estimated via Gauss-Hermite adaptive quadrature, for the full range of plausible GDS cut-off values.[36] This approach models sensitivity and specificity simultaneously and accounts for the precision of estimates within studies.[36] Data from all included primary studies will be analysed simultaneously with a random-effects model as sensitivity and specificity are assumed to vary across primary studies. We will also construct a pooled receiver operating characteristic curve and identify the optimal cut-off.[36] We will compare results that only include datasets that allow the exclusion of patients diagnosed with depression or receiving depression treatment (including antidepressants with reason unspecified) with results that also include studies where these data are not available. For assessment of each version, we will include studies that report total scores for the specific GDS version or individual GDS item scores from longer versions of the GDS which could be used to calculate total scores for the shorter version. We will consider imputation if a large part of data is missing.

In a previous IPDMA with the PHQ-9,[37] we found that reference standards appeared to perform differently. The Mini International Neuropsychiatric Interview (MINI) is fully structured, but was designed for very rapid administration and described as its authors as being overinclusive as a result. We found that, controlling for depressive symptom scores, the MINI classified approximately twice as many participants with major depression as other fully structured interviews.[38 39] Compared with semistructured interviews, which are intended to be done by experienced diagnosticians and involve some clinical judgement (eg, Structured Clinical Interview for DSM Disorders) fully structured interviews (MINI excluded), diagnosed more participants with low symptom levels as depressed and fewer participants with higher symptom levels. Fully structured interviews can be delivered by lay interviewers and are intended to achieve a high level of standardisation, but may sacrifice accuracy.[40–43] Thus, we will assess possible differences and evaluate sensitivity and specificity separately by reference standard.

In secondary analyses, to the extent that there are sufficient data, we will investigate subgroups according to age, sex, dementia status and severity, dementia subtype, number of medical comorbidities (with specific comorbidities integrated to the extent possible), care setting and risk of bias. QUADAS-2 factors that will be considered include patient selection factors, blinding of reference standard to index test results, and timing between administration of index test and reference standard (eg, 0–7 days, 7–14 days). Additionally, a subgroup analysis will be conducted that includes only data from countries listed as 'very high development' on the United Nation's Human Development Index.[44]

If there is a sufficient number of studies with published diagnostic accuracy data for major depression that are eligible but do not provide data, studies included in the IPDMA will be compared with eligible studies that do not provide data in terms of sensitivity and specificity, using published summary data from the studies that do not provide data. Depending on the number of missing studies, a sensitivity analysis may also be conducted that includes aggregate summary estimates of sensitivity and specificity from the studies that do not provide IPD in the main meta-analysis, along with data from studies that contribute to the IPDMA.[36] If there are a large number of studies that do not contribute primary data, this analysis may become the primary analysis.

Clinical predictive models have not been used previously to generate individualised probabilities that an individual has major depression based on screening tool scores and participant characteristics. There is a rich tradition of using predictive models for risk scores or classifying patients based on diagnostic tests, and our approach will build on those traditions.[45–50] To do this, we will develop binary predictive models that use GDS scores as well as key participant characteristics (eg, sex, age, dementia status, care setting) to estimate the probability and associated 95% CI that an individual has major depression. We will estimate logistic mixed models and then integrate over the distribution of the random effects as described in Pavlou *et al* and Skrondal *et al*.[51 52] Continuous variables (GDS score and age) will be modelled using flexible semiparametric methods (eg, regression splines). We will consider the inclusion of interaction terms. The models will be evaluated in terms of their overall performance (Nagelkerke's R^2, Brier score), calibration (eg, slope of linear predictor; are average, low and high predictions correct) and discrimination (eg, c-statistic; discrimination slope: can we separate subjects with and without major depression).[45 46] Validation with the same subjects used to develop a model results in overly optimistic performance. We will assess internal validation via the bootstrap method, which has been shown to be preferable to split sample validation approaches.[47] Although there are advantages to external validation, given the wide range of study populations that we will be using, it would be unlikely that there would be another comparable dataset large enough for validation. Thus, assessment of internal validity via bootstrapping will allow us to understand how our model may perform in a clinical setting, and by adjusting our regression coefficients for optimism, the performance of our model will be as accurate as possible.

In sensitivity analyses, we will explore including each item from the GDS questionnaires as a separate predictor variable, rather than only the total score.

Patient and public involvement

Patients and members of the public were not involved in the study.

ETHICS AND DISSEMINATION

Only individual studies that obtained ethical clearance and informed consent will be included. Only anonymised data will be provided by the investigators of the original studies.

The main outcomes of the IPDMA reflect knowledge that will influence future research, clinical practice and policy. Strategies for effective dissemination and specific outputs will be based on research showing how to best tailor research outputs to different user groups,[53–58] including research on improving the usefulness of reports of systematic review and meta-analyses for healthcare managers and policy-makers.[56 58] Dissemination will include publication of results in high-impact medical journals with open access, as well as presentations in seminars and symposia to policy-makers, healthcare providers and researchers at national and international conferences.

If the predictive model performs well, a free and easy-to-use online calculation tool will be created to incorporate individual characteristics into accuracy estimates and provide users of our research with probabilities that individual patients have depression based on their GDS score and key characteristics. The calculator will be similar to other successful tools, such as the FRAX Fracture Risk Assessment Tool (http://www.shef.ac.uk/FRAX/index.aspx). The tool that will be made from the results of this study will be modelled on this tool and presented with tablet and app versions.

Author affiliations
[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada
[2]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montreal, Quebec, Canada
[3]Department of Medicine, McGill University, Montreal, QC, Canada
[4]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada
[5]Department of Psychiatry, McGill University, Montreal, QC, Canada
[6]Department of Family Medicine, McGill University, Montreal, Quebec, Canada
[7]Schulich Library of Science and Engineering, McGill University, Montreal, Quebec, Canada
[8]Stanford Prevention Research Center, Department of Medicine, and Department of Health Research and Policy, Stanford School of Medicine, Stanford, California, USA
[9]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, USA
[10]Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada
[11]Department of Psychiatry, University of Calgary, Calgary, Canada
[12]Department of Clinical, Neuro and Developmental Psychology and Amsterdam Public Health research institute, VU University Amsterdam, Amsterdam, The Netherlands
[13]Hull York Medical School and Department of Health Sciences, The University of York, York, UK
[14]Department of Clinical Neurosciences, University of Calgary, Calgary, Alberta, Canada
[15]Department of Psychiatry, University of Alberta, Edmonton, Canada
[16]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, USA
[17]Department of Psychology, McGill University, Montreal, Canada
[18]Department of Educational and Counselling Psychology, McGill University, Montreal, Quebec, Canada

REFERENCES
1. Fiske A, Wetherell JL, Gatz M. Depression in older adults. *Annu Rev Clin Psychol* 2009;5:363–89.
2. Luppa M, Sikorski C, Luck T, *et al*. Age- and gender-specific prevalence of depression in latest-life-systematic review and meta-analysis. *J Affect Disord* 2012;136:212–21.
3. Duhoux A, Fournier L, Nguyen CT, *et al*. Guideline concordance of treatment for depressive disorders in Canada. *Soc Psychiatry Psychiatr Epidemiol* 2009;44:385–92.
4. Duhoux A, Fournier L, Gauvin L, *et al*. What is the association between quality of treatment for depression and patient outcomes? A cohort study of adults consulting in primary care. *J Affect Disord* 2013;151:265–74.
5. Mojtabai R, Olfson M. Proportion of antidepressants prescribed without a psychiatric diagnosis is growing. *Health Aff* 2011;30:1434–42.
6. Mojtabai R. Clinician-identified depression in community settings: concordance with structured-interview diagnoses. *Psychother Psychosom* 2013;82:161–9.
7. Olfson M, Blanco C, Marcus SC. Treatment of adult depression in the United States. *JAMA Intern Med* 2016;176:1482.
8. Jeste DV, Alexopoulos GS, Bartels SJ, *et al*. Consensus statement on the upcoming crisis in geriatric mental health: research agenda for the next 2 decades. *Arch Gen Psychiatry* 1999;56:848–53.
9. Weeks R, Kalucy R, Hill J. Chapter 3: Depression in Elderly Patients. In: Jackson S, Jansen P, Mangoni A, eds. *Prescribing for elderly patients*: John Wiley and Sons, Ltd, 2009.
10. Registered Nurses Association of Ontario. *Registered nurses association of ontario. screening for delirium, dementia, and*

depression in older adults. Toronto, Ontario: Registered Nurses Association of Ontario, 2003.

11. (AMDA). *American Medical Directors Association (AMDA). Depression in the long-term care setting*. Columbia, Maryland: American Medical Directors Association (AMDA), 2011.

12. American Geriatrics SocietyAmerican Association for Geriatric Psychiatry. Consensus statement on improving the quality of mental health care in U.S. nursing homes: management of depression and behavioral symptoms associated with dementia. *J Am Geriatr Soc* 2003;51:1287–98.

13. Buchanan D, Tourigny-Rivard M, Cappeliez P, *et al*. National guidelines for seniors' mental health: the assessment and treatment of depression. *Can J Geriatr* 2006;9:52–8.

14. Joffres M, Jaramillo A, Dickinson J, *et al*. Recommendations on screening for depression in adults. *CMAJ* 2013;185:775–82.

15. Mitchell AJ, Bird V, Rizzo M, *et al*. Which version of the geriatric depression scale is most useful in medical settings and nursing homes? Diagnostic validity meta-analysis. *Am J Geriatr Psychiatry* 2010;18:1066–77.

16. Mitchell AJ, Bird V, Rizzo M, *et al*. Diagnostic validity and added value of the geriatric depression scale for depression in primary care: a meta-analysis of GDS30 and GDS15. *J Affect Disord* 2010;125:10–17.

17. Montorio I, Izal M. The geriatric depression scale: a review of its development and utility. *Int Psychogeriatr* 1996;8:103–12.

18. Pocklington C, Gilbody S, Manea L, *et al*. The diagnostic accuracy of brief versions of the Geriatric Depression Scale: a systematic review and meta-analysis. *Int J Geriatr Psychiatry* 2016;31:837–57.

19. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:c221.

20. Stewart LA, Tierney JF, Clarke M. Chapter 18: Reviews of individual patient data. In: Higgins JPT, Green S, eds. *Cochrane handbook for systematic reviews of interventions. version 5.1.0. chichester*. England: Wiley-Blackwell The Cochrane Collaboration, 2011.

21. Jaspers GJ, Degraeuwe PL. A failed attempt to conduct an individual patient data meta-analysis. *Syst Rev* 2014;3:97.

22. Parish E, Bloom T, Godlee F. Statins for people at low risk. *BMJ* 2015;351:h3908.

23. Thombs BD, Benedetti A, Kloda LA, *et al*. The diagnostic accuracy of the Patient Health Questionnaire-2 (PHQ-2), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-9 (PHQ-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Syst Rev* 2014;3:124.

24. Deeks JJ, Bossuyt PM, Gatsonis C. *Cochrane handbook for systematic reviews of diagnostic Test Accuracy Version 1.0.0*: The Cochrane Collaboration, 2009.

25. Higgins JPT, Green S. *Cochrane handbook for systematic reviews of interventions version 5.1.0*. Chichester, England: Wiley-Blackwell; The Cochrane Collaboration, 2011.

26. McInnes MDF, Moher D, Thombs BD, *et al*. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 2018;319:388–96.

27. Stewart LA, Clarke M, Rovers M, *et al*. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. *JAMA* 2015;313:1657–65.

28. Thombs BD, Benedetti A, Kloda LA, *et al*. Diagnostic accuracy of the Depression subscale of the Hospital Anxiety and Depression Scale (HADS-D) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open* 2016;6:e011913.

29. Thombs BD, Benedetti A, Kloda LA, *et al*. Diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for detecting major depression in pregnant and postnatal women: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open* 2015;5:e009742.

30. McGowan J, Sampson M, Salzwedel DM, *et al*. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol* 2016;75:40–6.

31. Sampson M, Barrowman NJ, Moher D, *et al*. Should meta-analysts search Embase in addition to Medline? *J Clin Epidemiol* 2003;56:943–55.

32. Mann R, Gilbody SM. Should methodological filters for diagnostic test accuracy studies be used in systematic reviews of psychometric instruments? A case study involving screening for postnatal depression. *Syst Rev* 2012;1:9,4053–9.

33. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001;8:391–7.

34. Wong SS, Wilczynski NL, Haynes RB, *et al*. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA Annu Symp Proc* 2003:728–32.

35. Whiting PF, Rutjes AW, Westwood ME, *et al*. QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529–36.

36. Riley RD, Dodd SR, Craig JV, *et al*. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Stat Med* 2008;27:6111–36.

37. Levis B, Benedetti A, Riehm KE, *et al*. Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. *Br J Psychiatry* 2018;212:377–85.

38. Lecrubier Y, Sheehan DV, Weiller E, *et al*. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *European Psychiatry* 1997;12:224–31.

39. Sheehan DV, Lecrubier Y, Harnett Sheehan K, *et al*. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry* 1997;12:232–41.

40. Brugha TS, Jenkins R, Taub N, *et al*. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychol Med* 2001;31:1001–13.

41. Brugha TS, Bebbington PE, Jenkins R. A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. *Psychol Med* 1999;29:1013–20.

42. Nosen E, Woody SR. Chapter 8: Diagnostic assessment in research. McKay D, *Handbook of research methods in abnormal and clinical psychology*: Sage, 2008.

43. Kurdyak PA, Gnam WH. Small signal, big noise: performance of the CIDI depression module. *Can J Psychiatry* 2005;50:851–6.

44. United Nations Development Programme. *Human Development Report 2013*. *The rise of the south: human progress in a diverse world*. New York, NY: United Nations Development Programme, 2013.

45. Steyerberg EW. *Clinical Prediction Models*. New York: Springer, 2009.

46. Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* 2010;21:128–38.

47. Steyerberg EW, Harrell FE, Borsboom GJ, *et al*. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54:774–81.

48. Dalton JE, Kattan MW. Recent advances in evaluating the prognostic value of a marker. *Scand J Clin Lab Invest* 2010;242:59–62.

49. Xh Z, McClish DK, Obuchowski NA. *Statistical methods in diagnostic medicine*. 2nd edn. Hoboken, NJ: Wiley, 2011.

50. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer, 2001.

51. Pavlou M, Ambler G, Seaman S, *et al*. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *BMC Med Res Methodol* 2015;15:59.

52. Skrondal A, Rabe-Hesketh S. Prediction in multilevel generalized linear models. *J R Stat Soc Ser A Stat Soc* 2009;172:659–87.

53. Dobbins M, DeCorby K, Twiddy T. A knowledge transfer strategy for public health decision makers. *Worldviews Evid Based Nurs* 2004;1:120–8.

54. Graham ID, Logan J, Harrison MB, *et al*. Lost in knowledge translation: time for a map? *J Contin Educ Health Prof* 2006;26:13–24.

55. Haines A, Kuruvilla S, Borchert M. Bridging the implementation gap between knowledge and action for health. *Bull World Health Organ* 2004;82:724–31.

56. Dobbins M, Thomas H, O'Brien MA, *et al*. Use of systematic reviews in the development of new provincial public health policies in Ontario. *Int J Technol Assess Health Care* 2004;20:399–404.

57. Dobbins M, Jack S, Thomas H, *et al*. Public health decision-makers' informational needs and preferences for receiving research evidence. *Worldviews Evid Based Nurs* 2007;4:156–63.

58. Lavis J, Davies H, Oxman A, *et al*. Towards systematic reviews that inform health care management and policy-making. *J Health Serv Res Policy* 2005;10:35–48.