

RESEARCH

Risk of bias from inclusion of patients who already have diagnosis of or are undergoing treatment for depression in diagnostic accuracy studies of screening tools for depression: systematic review

Brett D Thombs *associate professor of psychiatry*¹, Erin Arthurs *research assistant*¹, Ghassan El-Baalbaki *postdoctoral fellow*¹, Anna Meijer *doctoral student*², Roy C Ziegelstein *professor of medicine*³, Russell J Steele *associate professor of mathematics and statistics*⁴

¹Lady Davis Institute for Medical Research, Jewish General Hospital and McGill University, Montreal, Quebec, Canada H3T 1E4; ²Interdisciplinary Centre for Psychiatric Epidemiology, University Medical Centre Groningen, University of Groningen, 9713 GZ, Netherlands; ³Johns Hopkins University School of Medicine, Baltimore, Maryland 21224, USA; ⁴Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Ouest, Montreal, Quebec, H3A 2K6

Abstract

Objectives To investigate the proportion of original studies included in systematic reviews and meta-analyses on the diagnostic accuracy of screening tools for depression that appropriately exclude patients who already have a diagnosis of or are receiving treatment for depression and to determine whether these systematic reviews and meta-analyses evaluate possible bias from the inclusion of such patients.

Design Systematic review.

Data sources Medline, PsycINFO, CINAHL, Embase, ISI, SCOPUS, and Cochrane databases were searched from 1 January 2005 to 29 October 2009.

Eligibility criteria for selecting studies Systematic reviews and meta-analyses in any language that reported on the diagnostic accuracy of screening tools for depression.

Results Only eight of 197 (4%) unique publications from 17 systematic reviews and meta-analyses specifically excluded patients who already had a diagnosis of or were receiving treatment for depression. No systematic reviews or meta-analyses commented on possible bias from the inclusion of such patients, even though 10 reviews used quality assessment tools with items to rate risk of bias from composition of the sample of patients.

Conclusions Studies of the accuracy of screening tools for depression rarely exclude patients who already have a diagnosis of or are receiving treatment for depression, a potential bias that is not evaluated in systematic reviews and meta-analyses. This could result in inflated estimates of accuracy on which clinical practice and preventive care guidelines are often based, a problem that takes on greater importance

as the rate of diagnosed and treated depression in the population increases.

Introduction

Depression is a common and disabling condition,¹ and improving care has been prioritised. Routine screening for depression is one solution that has been proposed. Depression screening involves the use of screening tools to identify patients who might have depression but who are not seeking treatment for symptoms and whose depression is not otherwise recognised by their physicians so that they can be further assessed and, if appropriate, treated.²⁻³ Screening for depression has been recommended in several medical settings, including cardiovascular care,⁴ perinatal care,⁵⁻⁷ oncological care,⁸ and primary care,⁹ although no clinical trial has found better depression outcomes for screened versus unscreened patients when the same treatment and care resources are potentially available to both groups.¹⁰⁻¹¹ Screening for depression can identify patients with depression who might otherwise go undetected, but it can also lead to misdiagnosis, the identification of patients as being depressed who are not, and overdiagnosis, which occurs when some patients with mild conditions are identified as depressed and exposed to the risk of labelling and treatment, even when the condition might not cause measurable morbidity or mortality. Recently, a report from the National Institute for Health and Clinical Excellence (NICE)¹¹ noted a lack of evidence for benefit from depression screening and, rather than routine screening, recommended case identification

Correspondence to: B D Thombs brett.thombs@mcgill.ca

Extra material supplied by the author (see <http://www.bmj.com/content/343/bmj.d4825/suppl/DC1>)

Appendix 1: Characteristics of diagnostic accuracy studies included in systematic reviews and meta-analyses

Appendix 2: Estimate of screening results including and excluding patients who already had diagnosis of depression

strategies to identify depression among high risk groups of patients or patients otherwise identified by physicians as possibly having depression.

A great deal of research has been conducted to determine the diagnostic accuracy of depression screening tests in different clinical settings. Based on data from such studies, expert panels have considered the risks and benefits of screening and issued recommendations to screen for depression in various settings.^{9 11} Diagnostic or screening tests, however, are useful only to the extent that they distinguish between disordered and non-disordered states that are not otherwise obvious to clinicians¹² and if they are accurate across the spectrum of patients who will be assessed in clinical practice.¹²⁻¹⁸

The term “spectrum effect” has been used to describe variations in test performance that sometimes occur across subgroups of patients that differ in demographic or clinical features. Spectrum effects raise questions about the generalisability of study results to specific populations of patients that might differ in important ways from study samples.¹⁹ The term “spectrum bias” is related and also describes situations in which the accuracy of a test is heterogeneous across subgroups of patients. Spectrum bias is said to be present when a study samples preferentially from certain portions of the patient spectrum but provides a global estimate of accuracy that could misrepresent what would be experienced in actual practice.¹²⁻¹⁹ Estimates of diagnostic accuracy that are based on case-control designs and whose samples include only obvious cases and healthy controls, for instance, have been shown to substantially overestimate diagnostic accuracy.^{13 14 18}

Self reported depression questionnaires are used for various purposes (such as screening for unidentified cases, tracking severity of symptoms, detecting relapse). For the purpose of screening, which involves the identification of cases not previously recognised, if individuals who already have a diagnosis of depression are not specifically excluded from studies assessing the diagnostic accuracy of depression screening tools, examined cohorts will have a greater prevalence and severity of depression than if only individuals without clinically recognised depression were screened. Not excluding patients who already have a diagnosis would, in turn, lead to determinations of screening accuracy and new case yield that are inflated compared with what would be achieved if the instrument were used to screen patients in clinical practice.¹²⁻¹⁸

Systematic reviews and meta-analyses are highly cited and are prioritised in grading evidence for practice guidelines.^{20 21} If studies of the diagnostic accuracy of depression screening tools that include patients who already have a diagnosis or are receiving treatment are included in systematic reviews and meta-analyses without adjustment for potential bias, these reviews could provide misleading accuracy estimates, thereby misleading calculations of risk-benefit by expert panels and, thus, clinicians.

We evaluated the proportion of studies included in systematic reviews and meta-analyses of the diagnostic accuracy of depression screening tools that excluded patients who already had a diagnosis of or were receiving treatment for depression. We also assessed whether authors of systematic reviews and meta-analyses noted the possibility of spectrum bias from the inclusion of such patients in the original research studies they reviewed. We hypothesised that few studies of depression screening tools would exclude such patients and that systematic reviews and meta-analyses would not consider spectrum bias from their inclusion.

Methods

Selection of systematic reviews and meta-analyses

We searched Medline, PsycINFO, CINAHL, Embase, ISI, SCOPUS, and Cochrane databases from 1 January 2005 to 29 October 2009 for systematic reviews and meta-analyses of the diagnostic accuracy of depression screening tools. We restricted the search to this period to obtain recent systematic reviews and meta-analyses that reflect relatively current practice. The search terms used were ((systematic review OR meta-analysis) AND (screening OR sensitivity OR specificity) AND depression). Eligible articles included systematic reviews and meta-analyses in any language published in final form or on the internet before final publication that reviewed the accuracy of screening tools for depression compared with a diagnosis of depression. Depression screening tools included any self report measure used to attempt to identify patients with depression. We included systematic reviews and meta-analyses that reviewed diagnostic accuracy and other psychometric characteristics of depression questionnaires (such as validity and reliability) but extracted data only on diagnostic accuracy. We excluded systematic reviews and meta-analyses that compared scores only on self report screening tools with classifications of depression based on cut offs from other self report screening tools but not a diagnosis of depression. Two investigators reviewed systematic reviews and meta-analyses for eligibility independently. If either reviewer deemed a systematic review or meta-analysis potentially eligible based on a review of the title and abstract, we carried out a full text review of the systematic review or meta-analysis. Any disagreement between reviewers after full text review was resolved by consensus after consultation with an independent third reviewer. Chance corrected agreement between reviewers was assessed with Cohen's κ .

Data extraction

Two investigators independently extracted and entered on a standardised spreadsheet data items from the systematic reviews and meta-analyses, as well as from the original studies included in the reviews, with discrepancies resolved by consensus. For each systematic review or meta-analysis, they recorded whether or not original studies mentioned possible bias because of the inclusion of patients who already had a diagnosis of or were receiving treatment for depression. Investigators also determined whether or not each systematic review or meta-analysis included an assessment of the quality of included diagnostic accuracy studies. If so, they recorded the tool that was used to do this and whether or not the tool included an evaluation of the risk of spectrum bias. Investigators also recorded the impact factor of the journal in which each systematic review or meta-analysis was published, using the impact factor for the year of publication.²² In addition, they reviewed the introduction and discussion sections and recorded the described purpose for which accuracy of the screening tool was being assessed (such as screening or identification of new cases, monitoring progress of treatment, detection of relapse).

Original diagnostic accuracy studies included in the systematic reviews and meta-analyses were classified as having excluded patients who already had a diagnosis of or were receiving treatment for depression if the authors of the study specifically indicated this in the exclusion criteria. If studies did not specifically indicate that such patients were excluded they were classified as having included them.

For each systematic review or meta-analysis, and overall, we determined the number of unique publications on the diagnostic

accuracy of depression screening tools, as well as the number of unique cohorts of patients. We assessed the number of publications and the number of cohorts because, in some cases, there were multiple publications from the same cohort. This occurred, for instance, when different publications reported results from different screening tools or criterion standards with the same group of patients, when one or more publications reported on a subset of the sample from another publication, or when the same patients were assessed at different time points (such as during pregnancy and after delivery). Identification of different publications from the same cohort was done by cross referencing authors and coauthors, characteristics of patients, and countries in which the research was conducted. Verification was done by comparing information in the publications. Cohort status was coded conservatively in that publications that seemed to be from the same cohort were coded as such, even if this could not be confirmed with 100% certainty.

We did not publish or register a review protocol for this study. All methods were determined a priori with the exception of reviewing the introduction and discussion sections to record the described purpose for which the accuracy of depression screening tool was being assessed. This additional step was added to the study methods after data extraction and tabulation of results to clarify whether the intention of the included systematic reviews and meta-analyses was to assess diagnostic accuracy for identification of new cases versus other possible uses of depression symptom questionnaires.

Results

Search results

The electronic database search yielded 1216 unique titles and abstracts for review. Of these, 1160 were excluded after review of titles and abstracts because they did not report results from a systematic review or meta-analysis or because they reported data from a systematic review or meta-analysis that was not related to the diagnostic accuracy of a depression screening tool. Of the 56 articles that underwent full text review, we excluded 39, leaving 17 eligible systematic reviews and meta-analyses (figure). Chance corrected agreement on inclusion and exclusion decisions between reviewers, as assessed with the Cohen's κ , was 0.95.

Table 1 shows the characteristics of selected systematic reviews and meta-analyses. Of the 17 systematic reviews and meta-analyses included, 10 were systematic reviews,²³⁻³² and seven were meta-analyses.³³⁻³⁹ The systematic reviews and meta-analyses included between two and 63 original studies and were published in a wide range of journals in terms of impact factor. Two meta-analyses assessed the nine item depression scale of the patient health questionnaire (PHQ-9)^{33,39}; one systematic review²³ and two meta-analyses^{37,38} evaluated the geriatric depression scale; seven systematic reviews^{24,26,27,29-32} and one meta-analysis³⁶ assessed depression screening tools, generally, in defined medical populations; two systematic reviews assessed specific screening tools, other than the patient health questionnaire or geriatric depression scale, in defined patient populations^{25,28}; and two meta-analyses assessed brief screening tools (for example, fewer than five items) in primary care³⁴ and palliative care.³⁵ All 17 systematic reviews and meta-analyses described the purpose of the review as related to determining diagnostic accuracy for new case detection by screening, and none discussed how their results might apply to other uses of depression screening tools (such as monitoring progress of treatment, detection of relapse).

Inclusion or exclusion of patients who already had a diagnosis or were receiving treatment

The 17 systematic reviews and meta-analyses included a total of 197 unique publications on the diagnostic accuracy of screening tools for depression in 170 unique cohorts of patients. The diagnostic accuracy studies examined more than 25 different screening tools in a wide range of patients (see appendix 1 on bmj.com). Only eight of 197 unique publications (4%) and eight of 170 cohorts (5%) specifically excluded patients who already had a diagnosis of or were receiving treatment for depression (see appendix 1). As shown in table 1, 11^{23,26,27,30-33,35,37-39} of the 17 systematic reviews or meta-analyses did not examine a single cohort of patients that specifically excluded those who already had a diagnosis of or were receiving treatment for depression.

Table 2 shows that only four⁴⁰⁻⁴³ of the eight studies that excluded such patients reported the number of patients who were excluded because of pre-existing mental health treatment. The proportion of patients excluded for this reason was 22% in a Veteran's Affairs primary care setting in the United States (published in 2004)⁴³; 10% in a 2003 study of patients in general practice from New Zealand⁴²; 2% in a 2004 study of postpartum women from Turkey⁴⁰; and 0.2% in a 1996 study of postpartum women from Sweden.⁴¹

Treatment of spectrum bias in systematic reviews and meta-analyses

As shown in table 1, 13^{23-25,27,30-36,38,39} of the 17 systematic reviews and meta-analyses conducted some form of quality assessment of included studies, including two meta-analyses^{36,39} that used the quality assessment for diagnostic accuracy studies (QUADAS) tool⁴⁴; one systematic review²⁷ that used the diagnostic test studies evaluation tool⁴⁵; one meta-analysis³⁴ that used the Newcastle-Ottawa scale⁴⁶; two systematic reviews^{30,32} that used methods developed by the US Preventive Services Task Force (USPSTF)^{47,48}; one systematic review³¹ that based quality review on guidelines from the American Academy of Neurology⁴⁹; one systematic review²⁵ that evaluated quality items based on a system from the York Centre for Reviews and Dissemination⁵⁰; one systematic review²⁴ that used a study specific tool based on criteria identified by the Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests⁵¹; one meta-analysis³⁵ that based quality ratings on a published article by Pai et al⁵²; and one systematic review²³ and two meta-analyses^{33,38} that used ad hoc procedures, such as extracting data on one to two items related to study quality.

Of these, 10 systematic reviews or meta-analyses^{24,25,27,30-32,34-36,39} used quality assessment methods that included an assessment of spectrum bias. The authors of one of these systematic reviews²⁴ noted study limitations from the lack of non-white patients, and the authors of another³² reported that younger children were poorly represented in studies of children and adolescents. The authors of one meta-analysis reported that half of studies reviewed did not include representative samples but did not provide a rationale for this conclusion.³⁶ The authors of another noted the possibility of a "disease progression bias" in one study of patients after stroke and indicated that none of the other 11 studies reviewed had limitations related to composition of patients.³⁹ In one systematic review, one of four included studies was downgraded because of the description of the sample, but an explanation was not provided.²⁷ The authors of the five other systematic reviews or meta-analyses that used quality assessment methods that included an assessment of

spectrum bias did not comment specifically on quality ratings related to possible spectrum bias.^{21 26 27 30 31}

Overall, none of the 17 systematic reviews or meta-analyses commented on possible spectrum bias from the inclusion in studies of patients who already had a diagnosis of or were receiving treatment for depression.

Discussion

We found that less than 5% of studies on the diagnostic accuracy of depression screening tools appropriately excluded patients who already had a diagnosis of or were receiving treatment for depression. The importance of this finding relates to the potential effect on assessments of the accuracy of depression screening instruments and the number of new cases they will uncover and, therefore, on their utility in clinical practice. The diagnostic accuracy of a screening test is often considered a fixed characteristic of a test, but it can vary substantially in populations with different clinical features.¹⁶ Studies that have examined accuracy of diagnostic tests consistently show that increased prevalence or severity of disease in the cohort of patients being examined inflates the reported sensitivity of the test being assessed.¹⁴ If the accuracy of screening tools for depression was studied in a group of patients, some of whom had already received a diagnosis for the condition, the assessments would be biased by the inclusion of individuals with a greater prevalence and severity of depression than if the instruments were used in clinical practice to screen patients without clinically recognised depression. This would, in turn, lead to inflated, and potentially misleading, estimates of accuracy on which clinical practice and preventive care guidelines are generally based.

Potential magnitude of problem

The potential magnitude of this problem grows as the prevalence of already diagnosed and treated depression in the population increases.^{53 54} Estimates of the prevalence of depression in primary care range from 5% to 13%, including 6% to 9% among adults aged 55 or older.⁵⁵ Rates are somewhat higher in patients with chronic physical illness.¹ Among adults aged 35 and older in the US, rates of antidepressant use increased from 8% to 14% from 1996 to 2005, with a third to a half of prescriptions specifically for psychiatric problems.⁵³ Rates of prescriptions for antidepressants might be even higher among patients with chronic physical disease. Based on provincial data from Ontario, Canada, for instance, the rate of antidepressant prescriptions within six months of an acute myocardial infarction doubled from 8% in 1993 to 16% in 2002 among patients aged 65 and older.⁵⁶ In a more recent cohort of more than 1200 outpatients with stable cardiovascular disease, just under 20% were treated with an antidepressant at the time of enrolment in the study.^{57 58} In addition to patients who receive treatment with antidepressants, a relatively small percentage of people receive psychotherapy for depression without drug treatment,⁵⁹ and some people are recognised by their physicians as depressed but choose not to undergo treatment.

A recent meta-analysis found that general practitioners correctly identify about 50% of patients with depression without the assistance of a screening tool.⁶⁰ Dichotomising a doctor's identification or non-identification of depressive disorders, however, could underestimate the degree to which they recognise depression. A study of over 700 patients in primary care from the US and the Netherlands, for instance, found that complete disagreement between physicians' assessments and a diagnostic interview for depression was much less common

than is often thought.⁶¹ In that study, only 27% of false negative cases based on physician assessments were true false negatives. In most cases of false negatives, physicians recognised symptoms of depression but underestimated severity compared with the diagnostic interview (40%) or gave another psychiatric diagnosis (33%). Thus, in many settings, a substantial proportion of depressed patients are recognised as depressed without screening, either because they seek treatment for their depression or because a healthcare professional otherwise recognises their symptoms. Based on reported rates of prescriptions for antidepressants and estimates of physicians' ability to recognise depression, it could be that as many as half or more of patients who are detected as cases in studies assessing the diagnostic accuracy of screening tools would not even be screened in clinical practice.

Data are not available that would allow a precise calculation of the degree by which studies that fail to exclude patients who already have a diagnosis of or are receiving treatment for depression might overestimate diagnostic accuracy and the number of new patients who would be identified through depression screening. Two reviews, however, have reported that studies of other types of diagnostic tests that have used case-control designs¹³ or case-control designs that compared severely affected patients and healthy controls¹⁸ substantially overestimate diagnostic accuracy (relative diagnostic odds ratios 3.0¹³ and 4.9,¹⁸ respectively).

Even a relatively small increase in reported diagnostic accuracy resulting from the inclusion of patients who already have a diagnosis or are receiving treatment would result in a substantial overestimate of the positive predictive value and new case yield from depression screening compared with what would be expected in clinical practice. A systematic review of the diagnostic accuracy of depression screening tools in primary care found a median sensitivity of 85% and median specificity of 74%.⁶² Based on this, in a primary care setting with a prevalence rate of 10%,⁵⁵ 32% of all patients would screen positive for depression, of whom 27% would be true positive cases, equivalent to 9% of all patients screened. If existing studies overestimated the sensitivity by even 10% because of the inclusion of patients with a diagnosis or being treated (relative diagnostic odds ratio 1.9), and it is conservatively assumed that physicians recognise 50% of depressed patients without screening, the rate of screening with positive results would decrease only slightly, from 32% to 27%. Only 14% of these, however, would be true positives, and, overall, less than 4% of patients screened would be newly identified cases of depression (see appendix 2 on bmj.com).

We know of only one study, which was not included in any of the systematic reviews or meta-analyses that we reviewed, that assessed the yield of screening for depression with and without excluding patients with psychiatric disorders already treated with psychotropic drugs.⁶³ In that study of 113 women with breast cancer, the true positive rate of screening for depression fell from 21% to 7% after exclusion of patients who were already receiving treatment for depression before screening.

Our results should be considered in the context of studies that have assessed whether screening for depression benefits patients. There are at least 11 trials in primary care,¹⁰ as well as trials in perinatal care,^{64 65} and cancer care,⁶⁶ that have tested whether screening and referral for depression treatment improves depression outcomes, and all have had negative results. Reflecting this, the US Preventive Services Task Force recommends screening for depression only when it is supported by integrated staff assisted depression management programmes.⁹ To our knowledge, only one published research

study has documented an attempt to screen and provide collaborative care, as recommended by the task force, in a clinical setting.⁶⁷ In that study, from the Netherlands, 1687 high risk patients were invited to enrol in a screening trial, 780 participated, and 71 cases of major depression were detected. Of the 71 patients identified, 36 were already receiving treatment for depression and 18 additional patients refused treatment or did not attend their scheduled appointment. Thus, only 17 people of 1687 potentially screened started treatment for depression.

Strengths and limitations of review

One possible limitation of the current study is that we searched for systematic reviews and meta-analyses, rather than for original studies, and there are probably many original studies on the diagnostic accuracy of depression screening tools that were not included. Our purpose, however, was to assess whether original studies appropriately excluded patients who already had a diagnosis or were receiving treatment and to determine whether systematic reviews and meta-analyses reflected potential bias from the failure to do this, which required a review of reviews. It is unlikely that including additional studies that were not listed in recent systematic reviews or meta-analyses would have substantively altered the results.

Another potential limitation is that the proportion of patients who already had a diagnosis of or were receiving treatment for depression who were inappropriately included in the diagnostic accuracy studies reviewed is unknown. Only four of the studies that excluded such patients reported the proportion excluded for this reason, and this varied widely depending on the setting and the time period of the study. It was less than 2% in studies that collected data from 10 years ago in Turkey⁴⁰ and more than 15 years ago in Sweden,⁴¹ but about 10% in a 2003 study of patients in general practice from New Zealand⁴² and just over 20% in a 2004 study of primary care patients treated in a US Veteran's Affairs setting.⁴³ In addition, the small number and substantial heterogeneity of studies that excluded patients who already had a diagnosis or were receiving treatment did not allow for an assessment of the effect of inclusion and exclusion decisions on diagnostic accuracy estimates. On the other hand, numerous studies have found that the inclusion of established cases among examined cohorts consistently inflates assessments of the accuracy of a diagnostic test,¹⁴ and it is likely that this would also be the case in studies of depression screening tools.

Conclusions and policy implications

The importance of our findings relates to the use of depression questionnaires for screening, a procedure conducted to identify previously unrecognised cases.^{2,3} In clinical practice, depression questionnaires are sometimes used for purposes other than screening, including monitoring the severity of symptoms in patients who already have a diagnosis of depression and assessing patients for recurrence of symptoms while they are being treated. The introduction and discussion sections of the 17 systematic reviews and meta-analyses we reviewed indicate that all were intended to assess the diagnostic accuracy and utility of depression questionnaires for the purpose of screening—that is, for identification of new cases. None discussed how findings might apply to other possible uses for the questionnaires (such as monitoring progress of treatment or detection of relapse). In addition, the recommendations that have been issued by expert panels regarding depression screening in various settings discuss the use of screening instruments as a means of identifying new cases.

Screening for depression is somewhat different from many other types of screening in that a history or interview might not necessarily be part of the evaluation before a screening tool is administered. To illustrate, the US Preventive Services Task Force recommends screening for cervical cancer in women who have been sexually active and have a cervix.⁶⁸ On the other hand, such screening is not recommended for women older than 65 or for women who have recently had a normal result on a smear test. This approach to screening is predicated on some “filtering” to determine the appropriate individuals or groups to be screened. On the other hand, the task force's recommendations regarding depression screening⁹ focus on issues in healthcare systems, such as the availability of staff assisted depression care, rather than on any upstream evaluation of patients before screening. In clinical settings, screening tools for depression might be routinely administered to all patients in the waiting room of a hospital, physician's office, or clinic, as has been recommended by expert panels.⁴ Regardless of whether these screening tools are used with or without upstream “filtering” in clinical practice, accurate determinations of test characteristics that reflect the ability to detect previously unrecognised cases can be obtained only if this upstream “filtering” is done in studies to exclude patients who already have a diagnosis of depression. Our findings show that this is rarely done, and, as a result, existing evidence on the accuracy and case yield of depression screening tools could substantially overestimate their utility in clinical practice. Well designed studies that exclude patients who already have a diagnosis of or are receiving treatment for depression are needed to generate realistic determinations of the accuracy of depression screening tools in clinical settings to inform decisions about risks and benefits with screening.

We thank Allison Leavens, Lisa R Jewett, and Brooke Levis, all of the Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada, for verification of referencing and study counts and proofreading the manuscript. They were not compensated for their contributions.

Contributors: BDT was responsible for the study concept and design, wrote the review protocol, supervised and carried out the data extraction, and drafted the manuscript with the input of the other authors. EA reviewed articles for inclusion, carried out the data extraction, contributed to the analysis, interpretation, and presentation of data, and conducted a critical revision of the manuscript. GE-B and AM participated in the design of the study, reviewed articles for inclusion, carried out the data extraction, and contributed a critical revision of the manuscript. RCZ contributed to the study design and contributed a critical revision of the manuscript. RJS contributed to the study design and analysis and interpretation of the data and contributed a critical revision of the manuscript. All authors had full access to all of the data (including statistical reports and tables) and take responsibility for the integrity of the data and the accuracy of the data analysis. BDT is guarantor.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. BDT is supported by a New Investigator Award from the Canadian Institutes of Health Research and an Établissement de Jeunes Chercheurs award from the Fonds de la Recherche en Santé Québec. RCZ is supported by the National Center for Complementary and Alternative Medicine (grant No R24AT004641) and the Miller Family Scholar Program of the Johns Hopkins Center for Innovative Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Complementary and Alternative Medicine or the National Institutes of Health.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on

What is already known on this topic

The results of studies on the accuracy of screening tools for depression are routinely used by expert panels to make decisions about the potential benefits of depression screening

What this study adds

Studies of the accuracy of screening tools for depression rarely exclude patients who already have a diagnosis or are receiving treatment, a potential bias that is not evaluated in systematic reviews and meta-analyses

This can result in inflated accuracy and estimates of the yield of new cases on which clinical practice and preventive care guidelines are often based, a problem that takes on greater importance as the rate of diagnosed and treated depression in the population increases

request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: No additional data available.

- 1 Evans DL, Charney DS, Lewis L, Golden RN, Gorman JM, Krishnan KR, et al. Mood disorders in the medically ill: scientific review and recommendations. *Biol Psychiatry* 2005;58:175-89.
- 2 UK National Screening Committee. Second report of the UK National Screening Committee. Departments of Health for England, Scotland, Northern Ireland and Wales, 2000.
- 3 Raffle A, Gray M. Screening: evidence and practice. Oxford University Press, 2007.
- 4 Lichtman JH, Bigger JT Jr, Blumenthal JA, Frasure-Smith N, Kaufmann PG, Lesperance F, et al. Depression and coronary heart disease: recommendations for screening, referral, and treatment: a science advisory from the American Heart Association Prevention Committee of the Council on Cardiovascular Nursing, Council on Clinical Cardiology, Council on Epidemiology and Prevention, and Interdisciplinary Council on Quality of Care and Outcomes Research. *Circulation* 2008;118:1768-75.
- 5 Scottish Intercollegiate Guidelines Network. SIGN 60: postnatal depression and perinatal psychosis. SIGN, 2002.
- 6 Dell DL. Depression in women. In: Clinical updates in women's health. Vol 1. American College of Obstetricians and Gynecologists, 2002.
- 7 National Institute for Health and Clinical Excellence. Antenatal and postnatal mental health: the NICE guideline on clinical management and service guidance. NICE, 2007.
- 8 National Comprehensive Cancer Network. Distress management. NCCN clinical practice guidelines in oncology. 2011. www.nccn.org/professionals/physician_gls/PDF/distress.pdf.
- 9 US Preventive Services Task Force. Screening for depression in adults: US Preventive Services Task Force recommendation statement. *Ann Intern Med* 2009;151:784-92.
- 10 Gilbody SD, Sheldon TD, House AD. Screening and case-finding instruments for depression: a meta-analysis. *CMAJ* 2008;178:997-1003.
- 11 National Collaborating Centre for Mental Health. The NICE guideline on the management and treatment of depression in adults (updated edition). National Institute for Health and Clinical Excellence, 2010.
- 12 Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-based Medicine Working Group. *JAMA* 1994;271:389-91.
- 13 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
- 14 Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189-202.
- 15 Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926-30.
- 16 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;274:645-51.
- 17 Willis BH. Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies. *Fam Pract* 2008;25:390-6.
- 18 Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469-76.
- 19 Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med* 2002;137:598-602.
- 20 Patsopoulos NA, Analatos AA, Ioannidis JP. Relative citation impact of various study designs in the health sciences. *JAMA* 2005;293:2362-6.
- 21 Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. *BMJ* 2001;323:334-6.
- 22 Institute for Scientific Information. Journal citation reports. 2011. www.isiknowledge.com/jcr/.
- 23 Allen J, Annells M. A literature review of the application of the geriatric depression scale, depression anxiety stress scales and post-traumatic stress disorder checklist to community nursing cohorts. *J Clin Nurs* 2009;18:949-59.
- 24 Gaynes BN, Gavin N, Meltzer-Brody S, Lohr KN, Swinson T, Gartlehner G, et al. Perinatal depression: prevalence, screening accuracy, and screening outcomes. *Evid Rep Technol Assess (Summ)* 2005;119:1-8.
- 25 Gibson J, McKenzie-McHarg K, Shakespeare J, Price J, Gray R. A systematic review of studies validating the Edinburgh postnatal depression scale in antepartum and postpartum women. *Acta Psychiatr Scand* 2009;119:350-64.
- 26 Kalpakjian CZ, Bombardier CH, Schomer K, Brown PA, Johnson KL. Measuring depression in persons with spinal cord injury: a systematic review. *J Spinal Cord Med* 2009;32:6-24.
- 27 Mirkhil S, Kent PM. The diagnostic accuracy of brief screening questions for psychosocial risk factors of poor outcome from an episode of pain: a systematic review. *Clin J Pain* 2009;25:340-8.
- 28 Morse R, Kendall K, Barton S. Screening for depression in people with cancer: the accuracy of the hospital anxiety and depression scale. *Clin Eff Nurs* 2006;9:188-96.
- 29 Thekkumpurath P, Venkateswaran C, Kumar M, Bennett MI. Screening for psychological distress in palliative care: a systematic review. *J Pain Symptom Manage* 2008;36:520-8.
- 30 Thoms BD, de Jonge P, Coyne JC, Whooley MA, Frasure-Smith N, Mitchell AJ, et al. Depression screening and patient outcomes in cardiovascular care: a systematic review. *JAMA* 2008;300:2161-71.
- 31 Thoms BD, Magyar-Russell G, Bass EB, Stewart KJ, Tsilidis KK, Bush DE, et al. Performance characteristics of depression screening instruments in survivors of acute myocardial infarction: review of the evidence. *Psychosomatics* 2007;48:185-94.
- 32 Williams SB, O'Connor EA, Eder M, Whitlock EP. Screening for child and adolescent depression in primary care settings: a systematic evidence review for the US Preventive Services Task Force. *Pediatrics* 2009;123:e716-35.
- 33 Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the patient health questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007;22:1596-602.
- 34 Mitchell AJ, Coyne JC. Do ultra-short screening instruments accurately detect depression in primary care? A pooled analysis and meta-analysis of 22 studies. *Br J Gen Pract* 2007;57:144-51.
- 35 Mitchell AJ. Are one or two simple questions sufficient to detect depression in cancer and palliative care? A Bayesian meta-analysis. *Br J Cancer* 2008;98:1934-43.
- 36 Hewitt C, Gilbody S, Brealey S, Paudyal M, Palmer S, Mann R, et al. Methods to identify postnatal depression in primary care: an integrated evidence synthesis and value of information analysis. *Health Technol Assess* 2009;13:1-145,147-230.
- 37 Mitchell AJ, Bird V, Rizzo M, Meader N. 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;125:10-7.
- 38 Wancata G, Alexandrowicz R, Marquart B, Weiss M, Friedrich F. The criterion validity of the geriatric depression scale: a systematic review. *Acta Psychiatr Scand* 2006;114:398-410.
- 39 Wittkamp KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the patient health questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007;29:388-95.
- 40 Aydin N, Inandi T, Yigit A, Hodoglugil NN. Validation of the Turkish version of the Edinburgh postnatal depression scale among women within their first postpartum year. *Soc Psychiatry Psychiatr Epidemiol* 2004;39:483-6.
- 41 Wickberg B, Hwang CP. The Edinburgh postnatal depression scale: validation on a Swedish community sample. *Acta Psychiatr Scand* 1996;94:181-4.
- 42 Arroll B, Khin N, Kerse N. Screening for depression in primary care with two verbally asked questions: cross sectional study. *BMJ* 2003;327:1144-6.
- 43 Corson K, Gerrity MS, Dobscha SK. Screening for depression and suicidality in a VA primary care setting: 2 items are better than 1 item. *Am J Manag Care* 2004;10:839-45.
- 44 Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.
- 45 NHS Public Health Resource Unit. 12 questions to help you make sense of a diagnostic test study. 2011. www.sph.nhs.uk/.
- 46 Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses. 2011. www.ohri.ca/programs/clinical_epidemiology/oxford.htm.
- 47 Pignone M, Gaynes BN, Rushton JL, Mulrow CD, Orleans CT, Whitener BL, et al. Screening for depression. Systematic evidence review No 6. Agency for Healthcare Research and Quality, 2001.
- 48 Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med* 2001;20:21-35.
- 49 Edlund W, Gronseth G, Yuen S, Franklin G, eds. American Academy of Neurology clinical practice guideline process manual. American Academy of Neurology, 2004.
- 50 Khan KS, ter Riet G, Popay J, Nixon J, Kleijnen J. Study quality assessment. In: Khan KS, ter Riet G, Glanville J, Snowdon AJ, Kleijnen J, eds. Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews. 2nd ed. CRD Report 4. University of York. 2001. www.york.ac.uk/inst/crd/crdreports.htm.
- 51 Cochrane Methods Working Group on Systematic Reviews of Screening and Diagnostic Tests. Screening and diagnostic tests: recommended methods. Cochrane Methods Working Group, 1996.
- 52 Pai M, McCulloch M, Enanoria W, Colford JM. Systematic reviews of diagnostic test evaluations: what's behind the scenes? *Evid Based Med* 2004;9:101-3.
- 53 Olsson M, Marcus SC. National patterns in antidepressant medication treatment. *Arch Gen Psychiatry* 2009;66:848-56.
- 54 Marcus SC, Olsson M. National trends in the treatment for depression from 1998 to 2007. *Arch Gen Psychiatry* 2010;67:1265-73.
- 55 O'Connor EA, Whitlock EP, Beil TL, Gaynes BN. Screening for depression in adult patients in primary care settings: a systematic evidence review. *Ann Intern Med* 2009;151:793-803.

- 56 Benazon NR, Mamdani MM, Coyne JC. Trends in the prescribing of antidepressants following acute myocardial infarction, 1993-2002. *Psychosom Med* 2005;67:916-20.
- 57 Gehi A, Haas D, Pipkin S, Whooley MA. Depression and medication adherence in outpatients with coronary heart disease: findings from the Heart and Soul Study. *Arch Intern Med* 2005;165:2508-13.
- 58 Whooley MA, de Jonge P, Vittinghoff E, Otte C, Moos R, Carney RM, et al. Depressive symptoms, health behaviors, and risk of cardiovascular events in patients with coronary heart disease. *JAMA* 2008;300:2379-88.
- 59 Olfson M, Marcus SC. National trends in outpatient psychotherapy. *Am J Psychiatry* 2010;167:1456-63.
- 60 Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet* 2009;374:609-19.
- 61 Tiemens BG, von Korff M, Lin EHB. Diagnosis of depression by primary care physicians versus a structured diagnostic interview: understanding discordance. *Gen Hosp Psychiatry* 1999;21:87-96.
- 62 Williams JW Jr, Pignone M, Ramirez G, Perez Stellato C. Identifying depression in primary care: a literature synthesis of case-finding instruments. *Gen Hosp Psychiatry* 2002;24:225-37.
- 63 Coyne JC, Palmer SC, Shapiro PJ, Thompson R, DeMichele A. Distress, psychiatric morbidity, and prescriptions for psychotropic medication in a breast cancer waiting room sample. *Gen Hosp Psychiatry* 2004;26:121-8.
- 64 Webster J, Linnane J, Roberts J, Starrenburg S, Hinson J, Dibley L. Identify, educate and alert (IDEA) trial: an intervention to reduce postnatal depression. *BJOG* 2003;110:842-6.
- 65 Yonkers KA, Smith MV, Lin H, Howell HB, Shao L, Rosenheck RA. Depression screening of perinatal women: an evaluation of the healthy start depression initiative. *Psychiatr Serv* 2009;60:322-8.
- 66 Maunsell E, Brisson J, Deschenes L, Frasure-Smith N. Randomized trial of a psychological distress screening program after breast cancer: effects on quality of life. *J Clin Oncol* 1996;14:2747-55.
- 67 Baas KD, Wittkamp KA, van Weert HC, Lucassen P, Huyser J, van den Hoogen H, et al. Screening for depression in high-risk groups: prospective cohort study in general practice. *Br J Psychiatry* 2009;194:399-403.
- 68 US Preventive Services Task Force. Screening for cervical cancer: recommendations and rationale. *Am J Nurs* 2003;103:101-2,105-6,108-9.
- 69 Arroll B, Goodyear-Smith F, Kerse N, Fishman T, Gunn J. Effect of the addition of a "help" question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study. *BMJ* 2005;331:884.
- 70 Beck CT, Gable RK. Screening performance of the postpartum depression screening scale-Spanish version. *J Transcult Nurs* 2005;16:331-8.
- 71 Lloyd-Williams M, Friedman T, Rudd N. Criterion validation of the Edinburgh postnatal depression scale as a screening tool for depression in patients with advanced metastatic cancer. *J Pain Symptom Manage* 2000;20:259-65.
- 72 Lloyd-Williams M, Friedman T, Rudd N. An analysis of the validity of the hospital anxiety and depression scale as a screening tool in patients with advanced metastatic cancer. *J Pain Symptom Manage* 2001;22:990-6.
- 73 Vittayanont A, Liabsuetrakul T, Pitanupong J. Development of postpartum depression screening scale (PDSS): a Thai version for screening postpartum depression. *J Med Assoc Thai* 2006;89:1-7.

Accepted: 23 May 2011

Cite this as: *BMJ* 2011;343:d4825

Tables

Table 1 | Systematic reviews (SR) and meta-analyses (MA) of diagnostic accuracy of depression screening tools

Study	Journal impact factor*	Screening tool, patients/setting	Review type	Publications reviewed	Cohorts reviewed	Cohorts that excluded diagnosed or treated patients	Method of quality assessment	Quality assessment included spectrum bias†	Inclusion or exclusion of diagnosed or treated patients noted
Gaynes, 2005 ²⁴	NA	Depression screening tools in perinatal care	SR	23	20	1 (5.0%)	Based on criteria from Cochrane working group	Yes	No‡
Morse, 2006 ²⁸	NA	HADS in patients with cancer	SR	10	10	1 (10.0%)	No	NA	No
Wancata, 2006 ³⁸	3.9	GDS in elderly patients	MA	42	37	0 (0%)	Ad hoc§	No	No
Gilbody, 2007 ³³	2.9	PHQ-9 in primary care and hospital settings	MA	18	15	0 (0%)	Ad hoc§	No	No
Mitchell, 2007 ³⁴	2.2	Short (<5 items) screening tools in primary care patients	MA	12	10	3 (30.0%)	Newcastle-Ottawa scale	Yes	No
Thombs, 2007 ³¹	2.2	Depression screening tools in acute myocardial infarction patients	SR	2	2	0 (0%)	Based on AAN review guidelines	Yes	No
Wittkamp, 2007 ³⁹	2.1	PHQ-9 in primary care and hospital settings	MA	12	9	0 (0%)	QUADAS	Yes	No
Mitchell, 2008 ³⁵	4.8	1-2 questions in cancer and palliative care	MA	10	10	0 (0%)	Based on Pai et al ⁴⁸	Yes	No
Thekkumpurath, 2008 ²⁹	2.7	Depression screening tools in palliative care	SR	8	8	1 (12.5%)	No	NA	No
Thombs, 2008 ³⁰	31.7	Depression screening tools in cardiovascular care	SR	11	11	0 (0%)	USPSTF	Yes	No
Allen, 2009 ²³	1.2	GDS in older adults or veterans in outpatient settings	SR	4	4	0 (0%)	Ad hoc§	No	No
Gibson, 2009 ²⁵	3.7	EPDS in perinatal care	SR	37	35	2 (5.7%)	Based on York Centre for Reviews and Dissemination system	Yes	No
Hewitt, 2009 ³⁶	6.9	Depression screening tools in perinatal care	MA	63	56	4 (7.1%)	QUADAS	Yes	No
Kalpakjian, 2009 ²⁶	1.4	Depression screening tools in spinal cord injury patients	SR	4	4	0 (0%)	No	NA	No
Mirkhil, 2009 ²⁷	3.0	Depression screening tools in patients with pain episode	SR	4	4	0 (0%)	Diagnostic test studies evaluation tool	Yes	No
Williams, 2009 ³²	4.7	Depression screening tools in children and adolescents	SR	9	9	0 (0%)	USPSTF	Yes	No
Mitchell, 2010¶ ³⁷	3.8	GDS in older primary care patients	MA	13	12	0 (0%)	No	NA	No

AAN=American Academy of Neurology; EPDS=Edinburgh postnatal depression scale; GDS=geriatric depression scale; HADS=hospital anxiety and depression scale; NA=not applicable; PHQ-9=patient health questionnaire-9; QUADAS=quality assessment for diagnostic accuracy studies; USPSTF=US Preventive Services Task Force.

*Impact factor from year systematic review or meta-analysis was published.

†Includes quality items related to "representativeness" of samples.

‡In methods authors wrote "We excluded studies that included patients with a known current depressive illness (for whom a screen would not provide new information)." Of 23 studies included in systematic review, however, 22 did not exclude patients who were already recognised as depressed or treated for depression. Authors of review did not comment on inclusion or exclusion of such patients in results or discussion.

Table 1 (continued)

Study	Journal impact factor*	Screening tool, patients/setting	Review type	Publications reviewed	Cohorts reviewed	Cohorts that excluded diagnosed or treated patients	Method of quality assessment	Quality assessment included spectrum bias†	Inclusion or exclusion of diagnosed or treated patients noted
-------	------------------------	----------------------------------	-------------	-----------------------	------------------	---	------------------------------	--	---

§Reported extraction of one to two items related to study quality (for example, blinding).

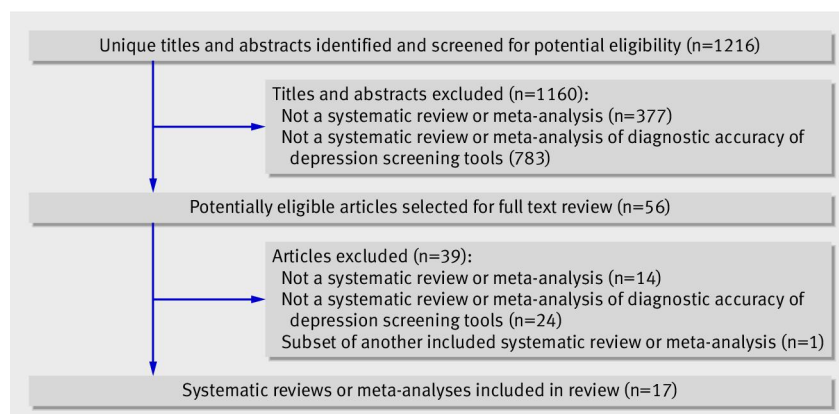
¶Article was epublication ahead of print at time of our search and was subsequently published in 2010.

Table 2| Cohorts of diagnostic accuracy studies that excluded patients who already had diagnosis of or were receiving treatment for depression

Study (review/s included in)	Country of original study	Population	Year(s) data collected	No (%) excluded	Exclusion criterion
Arroll, 2003 ⁴² (Mitchell ³⁴)	New Zealand	General practice patients	NR	47/476 (10%)	Taking psychotropic drugs
Arroll, 2005 ⁶⁹ (Mitchell ³⁴)	New Zealand	General practice patients	NR	NR	Receiving psychotropic drugs
Aydin, 2004 ⁴⁰ (Gibson, ²⁵ Hewitt ³⁶)	Turkey	Postpartum women	2001	6/347 (2%)	Psychiatric treatment history
Beck, 2005 ⁷⁰ (Hewitt ³⁶)	US	Postpartum women	NR	NR	Diagnosis of depression during current pregnancy
Corson, 2004 ⁴³ (Mitchell ³⁴)	US	Veteran's Affairs primary care patients	2002-3	762/3466 (22%)	Mental health appointment in chart within past 6 months
Lloyd-Williams, 2000, 2001 ^{71 72} (Morse, ²⁸ Thekkumpurath ²⁹)	UK	Cancer patients in palliative care	NR	NR	Currently prescribed antidepressant medication
Vittayanont, 2006 ⁷³ (Hewitt ³⁶)	Thailand	Women 6-8 weeks postpartum	2003-4	NR	Current diagnosis of and receiving treatment for psychiatric disorder
Wickberg, 1996 ⁴¹ (Gaynes, ²⁴ Gibson, ²⁵ Hewitt ³⁶)	Sweden	Women 2-3 months postpartum	NR	4/1655 (0.2%)	Already in contact with general practitioner or psychiatrist

NR=not reported.

Figure



Selection of systematic reviews and meta-analyses of diagnostic accuracy of screening tools for depression