**Selective Cutoff Reporting in Studies of Diagnostic Test Accuracy: A Comparison of Conventional and Individual Patient Data Meta-Analyses of the Patient Health Questionnaire-9 Depression Screening Tool**

**Short Title: Selective Cutoff Reporting**

Brooke Levis, Andrea Benedetti, Alexander W Levis, John PA Ioannidis, Ian Shrier, Pim Cuijpers, Simon Gilbody, Lorie A Kloda, Dean McMillan, Scott B Patten, Russell J Steele, Roy C Ziegelstein, Charles H Bombardier, Flavia de Lima Osório, Jesse R Fann, Dwenda Gjerdingen, Femke Lamers, Manote Lotrakul, Sonia R Loureiro, Bernd Löwe, Juwita Shaaban, Lesley Stafford, Henk CPM van Weert, Mary A Whooley, Linda S Williams, Karin A Wittkampf, Albert S Yeung, and Brett D Thombs.

**Correspondence to:**

Dr. Brett D Thombs, PhD

Jewish General Hospital
4333 Cote Ste Catherine Road
Montreal, Quebec H3T 1E4
Tel (514) 340-8222 ext. 5112
Fax (514) 340-8124
E-mail: brett.thombs@mcgill.ca

**Word count:** 3,878

**ABSTRACT**

In diagnostic test accuracy studies, authors sometimes only report results for a range of cutoffs around data-driven “optimal” cutoffs. We assessed selective cutoff reporting in studies of the diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9) depression screening tool. We compared conventional meta-analysis of published results only to individual patient data meta-analysis of results from all cutoffs, using data from 13 of 16 studies published from 2004-2009 included in a published conventional meta-analysis. For the “standard” cutoff of 10, 11 studies published accuracy results. For all other relevant cutoffs, 3-6 studies published results. For all cutoffs examined, specificity estimates in conventional and individual patient data meta-analyses were within 1%. Sensitivity estimates were similar for cutoff 10, but differed 5-15% for other cutoffs. In samples where the PHQ-9 was poorly sensitive at the standard cutoff, authors tended to report results for lower cutoffs that yielded optimal results. When the PHQ-9 was highly sensitive, authors more often reported results for higher cutoffs. Consequently, in the conventional meta-analysis, sensitivity increased as cutoff severity increased across part of the cutoff range, an impossibility if all data are analyzed. In sum, selectively reporting well-performing cutoffs can bias accuracy estimates in meta-analyses using published results only.

**Abstract word count**: 200

**Key words:** Bias, depression, diagnostic test accuracy, individual patient data meta-analysis, screening, selective cutoff reporting

**Abbreviations:** IPD = individual patient data; MDD = major depressive disorder; PHQ-9 = Patient Health Questionnaire-9.

Depression screening is controversial (1-3), and guidelines and policies vary substantially. In the US, the US Preventive Services Task Force recommends depression screening for adults in primary care settings (4). Depression screening is necessary for accreditation for many healthcare providers (5), and it is a required component of Medicare’s Annual Wellness Visit (6). In the UK, on the other hand, neither the National Institute for Health and Care Excellence (7) nor the UK National Screening Committee (8) recommends routine depression screening. The UK Quality and Outcome Framework incentivized depression screening in primary care from 2006 to 2013, but discontinued the program due to disappointing outcomes (9, 10). In Canada, the Canadian Task Force on Preventative Health Care previously recommended depression screening in primary care (11), but recommended against it in their most recent guideline (12). In their recommendation, the Canadian Task Force raised the concern that existing research may exaggerate the diagnostic accuracy of depression screening tools (12).

To date, there are no clinical trials that have randomized untreated patients prior to screening for depression and then provided the same depression treatment to patients identified as depressed in both unscreened and screened trial arms (13). If published evidence concerning the diagnostic accuracy of depression screening tools is indeed exaggerated, the potential effectiveness of screening programs that are currently in place might be dramatically different than expected. Exaggerated estimates of accuracy could lead to inappropriate clinical decisions, including a potentially high rate of overdiagnosis and, subsequently, over-treatment (14, 15).

One mechanism by which publications of research studies could overstate diagnostic accuracy is through selective reporting. Selective reporting occurs when multiple analyses are conducted in a study, and the decision of which to report depends on the results (16). The selective reporting of statistically significant, favourable outcomes and non-reporting of non-significant outcomes is well-established as an important source of bias in clinical trials (16-21), but has received less attention in diagnostic test accuracy studies. Selective reporting may occur in these studies when different cutoff thresholds can be used to classify test results as positive versus negative. Often, only results from a single cutoff, selected from multiple possible cutoffs, are reported in published studies. If reported cutoffs are selected based on maximizing accuracy estimates, then resulting accuracy estimates will tend to overestimate what would occur in practice. The same problem may arise when results for a small range of cutoffs around the optimal cutoff are reported.

Generally accepted “standard” cutoffs have been published for many depression screening tools. These are often based, however, on early studies that included too few depression cases to be confident that the best cutoff was identified (22, 23). Meta-analyses have the potential to overcome limitations due to small samples in primary studies. Results, however, are likely to be biased if primary studies selectively publish results from cutoffs that optimize accuracy. Selective cutoff reporting may be common in depression screening tool studies. A meta-analysis of the depression subscale of the Hospital Anxiety and Depression Scale (24), for instance, excluded 16 of 41 (39%) otherwise eligible studies because they only reported results from study-specific optimal cutoffs and not from standard cutoffs, which were meta-analyzed. In another example, a Patient Health Questionnaire-9 (PHQ-9) meta-analysis (25) evaluated sensitivity and specificity for cutoffs 7-15, including the standard cutoff of 10 (23, 26-29). Sensitivity and specificity were estimated at each cutoff by analyzing data from all primary studies that reported results for the cutoff. Due to incomplete reporting, however, estimates of sensitivity actually improved as cutoffs increased from 9 (less severe symptoms) to 11 (more severe symptoms). This would be mathematically impossible if complete data were analyzed.

No studies have examined patterns of selective cutoff reporting in studies of the diagnostic accuracy of depression screening tools or how selective cutoff reporting may influence results. Thus, in the present study, we re-analyzed studies included in the most recently published meta-analysis of the PHQ-9 (25), which included only published results. Our objective was to dissect how bias operates by comparing results of a conventional meta-analysis of published results to results from an individual patient data (IPD) meta-analysis of both published and unpublished results. For every potentially relevant cutoff, we assessed how reporting choices were associated with perceived sensitivity, specificity, overall diagnostic performance and heterogeneity.

METHODS

Data source

The PHQ-9 (23, 28, 29) is a 9-item measure of depressive symptoms that is increasingly used in medical populations (26, 27). The maximum score is 27, and higher scores represent more severe symptoms. The standard cutoff to identify possible depression is 10 (23, 26-29), which originates from the first PHQ-9 study, which included only 41 depression cases (total N = 580) (23, 29).

We used the most recently published PHQ-9 meta-analysis (25) as a framework to explore bias related to selective cutoff reporting. Studies were eligible for the original meta-analysis (25) if they (1) defined major depressive disorder (MDD) according to standard classification systems; (2) used a validated diagnostic interview for MDD as the reference standard; and (3) provided sufficient data to calculate 2x2 contingency tables. The meta-analysis included 18 primary studies with 17 unique patient samples. One study, however, did not publish accuracy results and was included by integrating individual patient data. Thus, there were 16 unique studies with published results available to assess selective cutoff reporting patterns. We contacted the authors of all 16 studies and invited them to contribute de-identified primary data. The Research Ethics Committee of the Jewish General Hospital in Montreal approved the study.

Data Preparation

From the publications associated with each primary study, we extracted study country, setting, reference standard, and the cutoffs for which accuracy results were reported. From each primary dataset, we extracted PHQ-9 scores and MDD diagnostic status for each patient, as well as information pertaining to weighting. For studies where sampling procedures merited weighting, but the original study did not weight (30, 31), we constructed appropriate weights using inverse selection probabilities. This occurred, for instance, when all patients with positive screens, but only a random subset of patients with negative screens, were administered a diagnostic interview.

We compared patient characteristics and published diagnostic accuracy results with results obtained using the raw datasets. When primary data and original publications were discrepant, we resolved discrepancies in consultation with the original investigators. When 2x2 tables in study publications could not be reproduced using the data due to errors in the publications, we corrected the published results using the raw data and confirmed with the authors (see Web Table 1). We used corrected results because our objective was to evaluate differences in estimates when including all versus only some cutoffs, not the extent or influence of unintentional error.

Data Analysis

First we performed a conventional meta-analysis, where for each cutoff from 7-15, we included data from the studies that reported results for the cutoff in the original publications. For instance, if a study published results for cutoffs 9-13, study data were included in our meta-analyses of cutoffs 9-13, but not our meta-analyses of cutoffs 7, 8, 14 or 15. Second, we performed an IPD meta-analysis, where for each cutoff from 7-15, we included data from all studies.

For both meta-analyses, bivariate random-effects models were estimated via Gauss-Hermite adaptive quadrature, as described in Riley et al. (32), for cutoffs 7-15 separately. This approach models sensitivity and specificity at the same time, accounting for the inherent correlation between them and for the precision of estimates within studies. A random-effects model was used so that sensitivity and specificity were assumed to vary across primary studies. For each cutoff, for conventional and IPD meta-analyses separately, this model provided an estimate of overall pooled sensitivity and specificity. In order to examine differences in results produced by conventional and IPD meta-analyses, we used the results of the random-effects meta-analyses at each cutoff to construct separate pooled receiver operator characteristic curves for each method.

Differences in Sensitivity and Specificity Estimates using Published vs. All Data

We compared sensitivity and specificity estimates based on published results to estimates using all data. To do this, we calculated differences in sensitivity and specificity estimates between the two methods at each cutoff. Confidence intervals for the differences were constructed via the “cases bootstrap” approach, as described by van der Leeden et al. (33, 34), resampling at the study and subject level.

Reporting Patterns

We assessed whether primary studies tended to preferentially report low or high cutoffs depending on their optimal cutoff. In each primary study, we defined the optimal cutoff as the cutoff maximizing Youden’s J statistic (sensitivity + specificity – 1) (35). Youden’s J does not necessarily reflect decision-maker preferences for the use of tests (36), but is commonly used to select optimal cutoffs in depression screening tool accuracy studies. For each study, we plotted the optimal cutoff, along with all other cutoffs for which results were published. We noted whether the reported cutoffs tended to be low or high.

Comparison of Results from Standard and Optimal Cutoffs

We compared results from our IPD meta-analysis of the standard cutoff of 10 to an additional IPD meta-analysis of results from each primary study’s optimal cutoff (i.e., the cutoff that maximized Youden’s J). We hypothesized that heterogeneity in sensitivity when all studies used the standard cutoff would be reduced when results from optimal cutoffs were combined, but that substantially different optimal cutoffs would be identified across studies. We compared forest plots of sensitivities and specificities using first a cutoff of 10 for each study, then, the optimal cutoff for each study. We also quantified heterogeneity by reporting the estimated variances of the random effects for sensitivity and specificity (2 and estimating R, which is the ratio of the estimated standard deviation of the pooled sensitivity (or specificity) from the random-effects model to the estimated standard deviation of the pooled sensitivity (or specificity) from the fixed-effects model (37). Additionally, we compared results from optimal versus the standard cutoff for studies that reported results from low-value cutoffs, high-value cutoffs, or only the standard cutoff.

RESULTS

We successfully obtained 13 of 16 eligible datasets. The 3 studies (23, 38, 39) for which data were not obtained were published between 2001 and 2006, and the authors indicated that data were no longer retrievable. Of the 5743 total patients and 1103 depression cases in the 16 eligible studies, we obtained data for 4589 patients (80%) and 1037 depression cases (94%). Characteristics of included studies are shown in Table 1. Sensitivity and specificity estimates for each cutoff, for each primary study, are shown in Web Table 1.

As shown in Table 2, 11 of 13 studies (83% of patients; 70% of MDD cases) published accuracy results for the standard cutoff of 10, whereas for other cutoffs, only 3-6 studies published results (21-46% of patients; 14-53% of MDD cases per cutoff).

As shown in Figure 1, the receiver operator characteristic curve based on published data only (Figure 1A) illustrates the distortions that occurred due to reporting results from some cutoffs but not others in primary studies. The receiver operator characteristic curve based on all data (Figure 1B), on the other hand, presents a plausible pattern of joint sensitivity and specificity across cutoffs. Youden’s J was maximized at cutoff 11 based on published data and at the standard cutoff of 10 based on all data.

Differences in Sensitivity and Specificity Estimates using Published vs. All Data

Table 3 summarizes discrepancies in sensitivity and specificity for each cutoff based on published data versus all data. For the cutoff of 10, where results were available from 11 of 13 studies, estimates of sensitivity and specificity were similar. For all other cutoffs, where less than half of patient data were available, sensitivity estimates deviated by 5-15%. For all cutoffs, specificity estimates based on published data and all data were within 1%.

Reporting Patterns

Cutoff reporting patterns from the 13 primary studies are shown in Figure 2. In studies where the PHQ-9 was poorly sensitive at the standard cutoff of 10 (sensitivity 0.47-0.75) (30, 40-43), optimal cutoffs were less than 10. In these studies, accuracy results tended to be reported for low cutoff values up to the standard of 10. In studies where the PHQ-9 was highly sensitive at the standard cutoff (sensitivity 0.88-1.00) (31, 44-47), optimal cutoffs were > 10 with one exception where cutoffs 10 and 11 performed equivalently (44). In these studies, results tended to be reported for the standard cutoff of 10 and above. Three studies reported only the standard cutoff of 10 (48-50). Of these, two (49, 50) had good sensitivity (0.91 and 0.93) and specificity (0.89 and 0.85), whereas the third (48) had low sensitivity (0.75) and good specificity (0.91).

As shown in Table 3, this pattern of reporting results from cutoffs close to the optimal cutoff in each study led to the underestimation of sensitivity in published data for cutoffs < 10, and overestimation for cutoffs > 10. All primary studies published results from the study-specific optimal cutoff, and 7 of the 13 studies published results from only 1-3 cutoffs in the cutoff range of 4-16.

Comparison of Results from Standard and Optimal Cutoffs

Forest plots of sensitivity estimates using the standard cutoff of 10 for all studies (Figure 3A) versus the optimal cutoff for each study (Figure 3B) are shown in Figure 3, while those for specificity estimates are shown in Figures 4A and 4B, respectively. There was relatively low heterogeneity in specificity estimates (Figure 4A, cutoff of 10, 2 = 0.18, R = 2.68; Figure 4B, optimal cutoff, 2 = 0.46, R = 4.71). However, there was a high degree of heterogeneity in sensitivity estimates using cutoff 10 for all studies (Figure 3A, 2 = 1.95, R = 5.63), but much less with the optimal cutoff for each study (Figure 3B, 2 = 0.68, R = 2.51) (Figure 3). Optimal cutoffs, however, were highly disparate (5 to 15) (Figure 2).

For all 13 studies, Youden’s J for with a cutoff of 10 (Table 2; Youden’s J = 0.75) and based on optimal cutoffs was similar (sensitivity = 0.92, 95% CI = 0.86, 0.95; specificity = 0.85, 95% CI = 0.79, 0.89; Youden’s J = 0.77). When the 5 studies that reported optimal cutoffs < 10 (30, 40-43) were analyzed separately, Youden’s J was 0.10 greater for optimal cutoffs (sensitivity = 0.88, 95% CI = 0.80, 0.93; specificity = 0.74, 95% CI = 0.71, 0.77; Youden’s J = 0.62) than for the standard cutoff of 10 (sensitivity = 0.63, 95% CI = 0.51, 0.73; specificity = 0.89, 95% CI = 0.87, 0.91; Youden’s J = 0.52). For the 4 studies that reported optimal cutoffs > 10 (not counting one study (44) where cutoffs of 10 and 11 performed equivalently) (31, 45-47), Youden’s J was 0.05 higher for the optimal cutoffs (sensitivity = 0.94, 95% CI = 0.87, 0.98; specificity = 0.87, 95% CI = 0.79, 0.92; Youden’s J = 0.81) compared to the standard cutoff (sensitivity = 0.95, 95% CI = 0.88, 0.98; specificity = 0.82, 95% CI = 0.75, 0.87; Youden’s J = 0.77). For the 3 studies that reported results only for the standard cutoff of 10 (48-50), and the 1 study where cutoffs 10 and 11 performed equivalently (44), estimated sensitivity was 0.94 (95% CI = 0.77, 0.99) and estimated specificity was 0.92 (95% CI = 0.84, 0.96). Sensitivity, specificity, and Youden’s J for each primary study, based on the standard cutoff and the optimal cutoff, are shown in Web Table 2.

DISCUSSION

This is the first study, to our knowledge, that has used IPD meta-analysis to examine selective cutoff reporting patterns in diagnostic test accuracy studies. Using studies from the most recently published meta-analysis of the PHQ-9 (25), we compared published results from primary studies, most of which reported results from only a subset of potentially relevant cutoffs, to results using IPD meta-analysis, where results from all relevant cutoffs for all studies were included.

For the standard PHQ-9 cutoff score of 10, 11 of 13 included studies published accuracy results. Sensitivity and specificity estimates were similar using published results and using all data. For all other cutoffs evaluated, however, fewer than half of the studies published accuracy results. For those cutoffs, specificity estimates were similar across the two meta-analytic methods, but sensitivity estimates differed by 5-15%. Studies where the PHQ-9 was poorly sensitive at the standard cutoff of 10 identified optimal cutoffs < 10 and tended to publish results from only lower cutoff scores up to cutoff 10. On the other hand, studies where the PHQ-9 was highly sensitive at cutoff 10 identified optimal cutoffs > 10 and tended to publish results for cutoffs 10 and above. When only published data were considered, sensitivity was underestimated for cutoffs below 10, but overestimated for cutoffs above 10. As a result, instead of seeing reductions in sensitivity as cutoff severity increased, the published data suggested an implausible increase in sensitivity estimates.

When we compared all primary studies using the same cutoff score of 10, there was substantial heterogeneity in sensitivity estimates. When we evaluated each study using the optimal cutoff, sensitivity estimates were generally homogeneous, but optimal cutoffs across studies ranged from 5-15, a range far too wide to meaningfully help clinicians determine how to use the PHQ-9 in practice.

The standard cutoff score of 10 for the PHQ-9 was based on only 41 cases from the first PHQ-9 study (23, 29). Subsequent studies did not report accuracy results from other cutoffs consistently enough to compare accuracy across cutoffs, but the IPD meta-analysis results suggest that a score of 10 may maximize combined sensitivity and specificity. The purpose of this study, however, was not to identify the cutoff that maximizes sensitivity and specificity. It is possible that when additional data are accumulated, this result may not replicate. Other depression screening tools have standard cutoffs that were also determined using small numbers of patients and have not been revised as evidence has accrued. For instance, the standard cutoff of 11 for “definite” cases of depression using the Hospital Anxiety and Depression Scale, which is the most commonly used depression screening tool in medically ill patients (51, 52), was based on 12 cases, while the cutoff of 8 or higher for “probable” cases was based on 22 cases (22). One Hospital Anxiety and Depression Scale meta-analysis excluded 16 of 41 (39%) otherwise eligible studies because they reported only optimal cutoffs, but not results from either of the standard cutoffs (8 and 11) (24). Other Hospital Anxiety and Depression Scale meta-analyses integrated results from different cutoffs into the same meta-analysis as if all studies used the same cutoff, including many studies that only published non-standard optimal cutoffs, making it impossible to compare accuracy across cutoffs (53, 54).

Outside of mental health, simulation studies have demonstrated that using data-driven cutoff thresholds to estimate diagnostic accuracy generates overly optimistic estimates, particularly when sample sizes are small (55, 56). We were able to identify only one meta-analysis that attempted to compare sensitivity and specificity from studies that used data-driven optimal cutoffs versus using pre-determined cutoffs. In that study, which evaluated D-dimer for suspected deep vein thrombosis, estimates of sensitivity, but not specificity, were significantly higher in studies that used data-driven cutoffs (57). However, the determination of whether a primary study cutoff was pre-determined versus data-driven was based on author report in the primary studies, and the wide range of cutoffs described as “a priori” suggests that some may not actually have been established as target cutoffs in advance. Among 91 analyses using an enzyme-linked immunoassay test, the 76 cutoffs described as a priori ranged from 40-1000 ng/mL, while the 11 described as post-hoc ranged from 50-2000 ng/mL. In 4 analyses, there was no indication of how the cutoff was determined.

Riley et al., have developed models for dealing with missing threshold data in meta-analyses of diagnostic test accuracy (58). These models, however, assume that missing data are missing at random. Thus, they cannot be used in situations where selective cutoff reporting is apparent, as the missingness within the studies is not random. The Quality Assessment of Diagnostic Accuracy Studies-II tool (59) includes an item assessing whether reported cutoffs were chosen a priori or post-hoc; however, this classification is based solely on how authors *report* that they chose their cutoffs. In studies where index test data are ordinal, as is the case with depression screening tools, one potential solution to overcome selective reporting is to require all authors to provide diagnostic accuracy data (e.g., sensitivity, specificity, positive predictive value and negative predictive value; or, alternatively, 2x2 raw tables of true positives, false positives, true negatives and false negatives) for all potentially relevant cutoffs in an appendix. In studies where the index test is continuous, a potential solution is to require authors to make their data publicly available upon publication, which is already expected in some journals. IPD meta-analyses can also be used, but they are time- and resource-intensive, and their quality depends on the ability to obtain representative data (60).

A potential limitation of our study is that we were unable to acquire data for 3 eligible studies included in the original meta-analysis that we replicated. However, we were able to include data for 80% of patients and 94% of cases from the original meta-analysis. Our IPD meta-analysis included data from over 4500 patients, including more than 1000 depression cases. More primary studies using the PHQ-9 have been conducted subsequently, however, and ideally our findings will be replicated in larger samples and with other screening tools. Additionally, it should be noted that while we preformed our study under the assumption that authors selected cutoffs to maximize Youden’s J, it may be the case that other processes were used in primary studies, such as selecting the highest possible threshold that would preserve some minimal level of sensitivity.

The objective of our study was to replicate an existing meta-analysis that included only published data and to examine how selective reporting of results from some cutoffs and not others influences diagnostic accuracy estimates. The intent was not to evaluate the accuracy of the PHQ-9 or to examine potential sources of heterogeneity in accuracy estimates. Thus, consistent with the original meta-analysis, we did not evaluate the quality of primary studies, conduct moderator analyses, or seek to explain inter-study heterogeneity. Currently, work has begun on identifying and gathering primary data to conduct a full IPD meta-analysis of the PHQ-9 (61), which will address these considerations.

In summary, this study demonstrates that in the presence of selective cutoff reporting, when there is a cutoff considered “standard”, estimates of sensitivity may be systematically underestimated for cutoffs below the standard and overestimated for cutoffs above the standard. Because standard cutoffs often originate from preliminary studies with few cases, selective reporting in subsequent studies may impede our ability to use conventional evidence synthesis methods to identify optimal cutoff scores. More work is needed to understand the influence of selective cutoff reporting in studies of the diagnostic test accuracy of depression screening tools and in studies of other types of tests. Users of meta-analyses should be aware of the possibility that meta-analyses based on published data alone may not accurately portray diagnostic accuracy. Similar to movements towards open data access in other fields of research, researchers should routinely report diagnostic test accuracy results for all relevant cutoffs, or make data accessible upon publication.

ACKNOWLEDGEMENTS

1. Author affiliations

Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada (Brooke Levis, Alexander W Levis, Ian Shrier, Russell J Steele, Brett D Thombs); Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada (Brooke Levis, Andrea Benedetti, Alexander W Levis, Ian Shrier, Brett D Thombs); Department of Medicine, McGill University, Montréal, Québec, Canada (Andrea Benedetti, Brett D Thombs); Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada (Andrea Benedetti); Department of Medicine, Department of Health Research and Policy, Department of Statistics, Stanford University, Stanford, California, USA (John PA Ioannidis); Department of Clinical, Neuro and Developmental Psychology, Vrije Universiteit (VU) University, Amsterdam, The Netherlands (Pim Cuijpers); Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, United Kingdom (Simon Gilbody, Dean McMillan); Libraries, Concordia University, Montréal, Québec, Canada (Lorie A Kloda); Departments of Community Health Sciences and Psychiatry, University of Calgary, Calgary, Alberta, Canada (Scott B Patten); Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada (Russell J Steele); Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA (Roy C Ziegelstein); Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA (Charles H Bombardier); Department of Neuroscience and Behavior, Faculty of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil (Flavia de Lima Osório, Sonia R Loureiro); Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA (Jesse R Fann); Department of Family Medicine and Community Health, University of Minnesota, Minneapolis, Minnesota, USA (Dwenda Gjerdingen); Department of Psychiatry, EMGO Institute, VU University Medical Center, Amsterdam, The Netherlands (Femke Lamers); Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (Manote Lotrakul); Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf and Schön Klinik Hamburg Eilbek, Hamburg, Germany (Bernd Löwe); Department of Family Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia (Juwita Shaaban); Centre for Women’s Mental Health, Royal Women’s Hospital, Parkville, Victoria, Australia (Lesley Stafford); Department of General Practice, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands (Henk CPM van Weert); Department of Veterans Affairs Medical Center, San Francisco, California, USA (Mary A Whooley); Richard L Roudebush Veteran Affairs Medical Center, Health Services Research and Development, Indiana University School of Medicine, Indianapolis, Indiana, USA (Linda S Williams); Department of Psychiatry and Department of General Practice, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands (Karin A Wittkampf); Depression Clinical and Research Program, Massachussets General Hospital, Boston, Massachussets, USA (Albert S Yeung); Departments of Psychiatry, Educational and Counselling Psychology, and Psychology, and School of Nursing, McGill University, Montréal, Québec, Canada (Brett D Thombs).

2. Author contributions

BLevis, AB, JPAI, IS, PC, SG, LAK, DM, SBP, RJS, RCZ and BDT were responsible for the study conception and design. SG, CHB, FLO, JRF, DG, FL, SRL, BLöwe, ML, JS, LS, HCPMvW, MAH, LSW, KAW and ASY were responsible for collection of primary patient data included in this study. BLevis and BDT contributed to data extraction and coding for the meta-analysis. BLevis, AB, AWL and BDT contributed to the data analysis. BLevis, AB and BDT contributed to drafting the manuscript. All authors provided a critical review and approved the final manuscript. BDT is the guarantor.

3. Grants and/or financial support

Ms. Levis was supported by a Fonds du recherche du Québec – Santé (FRQS) doctoral award. Dr. Benedetti was supported by an FRQS researcher salary award. Mr. Levis was supported by a Canadian Institutes of Health Research (CIHR) Frederick Banting and Charles Best Canadian Graduate Scholarships – Master’s Award. Collection of data for the primary study by Fann et al. was supported by grant RO1 HD39415 from the US National Center for Medical Rehabilitation Research. Collection of data for the primary study by Gjerdingen et al. was supported by grants from the National Institutes of Mental Health (R34 MH072925, K02 MH65919, P30 DK50456). Dr. Stafford received PhD scholarship funding from the University of Melbourne. Dr. Thombs was supported by an Investigator Award from the Arthritis Society.

4. Conflicts of interest

No funders had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Dr. Patten declares that he has received a grant, outside the submitted work, from the Hotchkiss Brain Institute. The grant was jointly funded by the Institute and Pfizer. Pfizer was the original sponsor of the development of the PHQ-9, which is now in the public domain. All other authors declare no competing interests.

REFERENCES

1. Palmer SC, Coyne JC. Screening for depression in medical care: Pitfalls, alternatives, and revised priorities. *J Psychosom Res* 2003; 54 (4):279-287.
2. Gilbody S, Sheldon T, Wessely S. Should we screen for depression? *BMJ* 2006; 332 (7548):1027-1030.
3. Thombs BD, Coyne JC, Cuijpers P, et al. Rethinking recommendations for screening for depression in primary care. *CMAJ* 2012; 184 (4):413-418.
4. Siu AL, and the US Preventive Services Task Force (USPSTF). Screening for Depression in Adults: US Preventive Services Task Force Recommendation Statement. *JAMA* 2016; 315 (4):380-387.
5. NCQA level 3 PCMH recognition requirements compared to 2011 Joint Commission standards and EPs. http://[www.jointcommission.org/assets/1/18/PCMH-NCQA\_crosswalk-final\_June\_2011.pdf](http://www.jointcommission.org/assets/1/18/PCMH-NCQA_crosswalk-final_June_2011.pdf). Accessed August 21, 2012.
6. Centers for Medicare and Medicaid Services. Medicare Program; Payment policies under the physician fee schedule and other revisions to part B for CY 2011 November 29, 2010. http://[www.federalregister.gov/articles/2010/11/29/2010-27969/medicare-program-payment-policies-under-the-physician-fee-schedule-and-otherrevisions-to-part-b-for#h-117](http://www.federalregister.gov/articles/2010/11/29/2010-27969/medicare-program-payment-policies-under-the-physician-fee-schedule-and-otherrevisions-to-part-b-for#h-117). Accessed August 21, 2015.
7. National Collaborating Center for Mental Health. The NICE guideline on the management and treatment of depression in adults (Updated edition). London, United Kingdom: National Institute for Health and Clinical Excellence; 2010.
8. Allaby M. Screening for depression: A report for the UK National Screening Committee (Revised report). London, United Kingdom: UK National Screening Committee; 2010.
9. Thombs BD, Ziegelstein RC. Does depression screening improve depression outcomes in primary care? *BMJ* 2014; 348:g1253.
10. Burton C, Simpson C, Anderson N. Diagnosis and treatment of depression following routine screening in patients with coronary heart disease or diabetes: a database cohort study. *Psychol Med* 2013; 43 (3):529-537.
11. Joffres M, Jaramillo A, Dickinson J, et al. Recommendations on screening for depression in adults. *CMAJ* 2013; 185 (9):775-782.
12. Thombs BD, Arthurs E, El-Baalbaki G, et al. Risk of bias from inclusion of already diagnosed or treated patients in diagnostic accuracy studies of depression screening tools: A systematic review. *BMJ* 2011; 343:d4825.
13. Thombs BD, Ziegelstein RC, Roseman M, et al. There are no randomized controlled trials that support the United States Preventive Services Task Force Guideline on screening for depression in primary care: a systematic review. *BMC Med* 2014; 12:13.
14. Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. *BMJ* 2012; 344:e3502.
15. Jerant A, Kravitz RL, Fernandez Y, et al. Potential antidepressant overtreatment associated with office use of brief depression symptom measures. *J Am Board Fam Med* 2014; 27 (5):611-620.
16. Kirkham JJ, Dwan KM, Altman DG, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010; 340:c365.
17. Dwan K, Altman DG, Cresswell L, et al. Comparison of protocols and registry entries to published reports for randomised controlled trials. *Cochrane Database Syst Rev* 2011; (1):MR000031.
18. Sterne JAC, Egger M, Moher D. Chapter 10: Addressing reporting biases. In: Higgins JPT, editor. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. London, United Kingdom: The Cochrane Collaboration; 2011.
19. Dwan K, Gamble C, Williamson PR, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLOS ONE* 2013; 8 (7):e66844.
20. DeAngelis CD, Drazen JM, Frizelle FA, et al. Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *JAMA* 2004; 292 (11):1363-1364.
21. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0. London, United Kingdom: The Cochrane Collaboration; 2011.
22. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand* 1983; 67 (6):361-370.
23. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; 16 (9):606-613.
24. Brennan C, Worrall-Davies A, McMillan D, et al. The Hospital Anxiety and Depression Scale: a diagnostic meta-analysis of case-finding ability. *J Psychosom Res* 2010; 69 (4):371-378.
25. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ* 2012; 184 (3):E191-6.
26. Wittkampf KA, Naeije L, Schene AH, et al. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007; 29 (5):388-395.
27. Gilbody S, Richards D, Brealey S, et al. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007; 22 (11):1596-1602.
28. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann* 2002; 32 (9):1-7.
29. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA* 1999; 282 (18):1737-1744.
30. Azah N, Shah M, Juwita S, et al. Validation of the Malay version brief Patient Health Questionnaire (PHQ-9) among adult attending family medicine clinics. *International Medical Journal* 2005; 12 (4):259-263.
31. Yeung A, Fung F, Yu SC, et al. Validation of the Patient Health Questionnaire-9 for depression screening among Chinese Americans. *Compr Psychiatry* 2008; 49 (2):211-217.
32. 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 (29):6111-6136.
33. van der Leeden R, Busing FMTA, Meijer E. Bootstrap methods for two-level models. Technical Report PRM 97-04, Leiden University, Department of Psychology, Leiden, The Netherlands, 1997.
34. van der Leeden R, Meijer E, Busing FMTA. Chapter 11: Resampling multilevel models. In: Leeuw J, Meijer E, eds. *Handbook of multilevel analysis* New York, NY: Springer; 2008:401–433.
35. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3 (1):32-35.
36. Smits N. A note on Youden's J and its cost ratio. *BMC Med Res Methodol* 2010; 10:89.
37. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; 21 (11):1539-1558.
38. Watnick S, Wang P, Demadura T, et al. Validation of 2 depression screening tools in dialysis patients. *Am J Kidney Dis* 2005; 46 (5):919-24.
39. Adewuya AO, Ola BA, Afolabi OO. Validity of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *J Affect Disord* 2006; 96 (1-2):89-93.
40. Stafford L, Berk M, Jackson H. Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease. *Gen Hosp Psychiatry* 2007; 29 (5):417-24.
41. Thombs BD, Ziegelstein RC, Whooley MA. Optimizing detection of major depression among patients with coronary artery disease using the Patient Health Questionnaire: data from the Heart and Soul Study. *J Gen Intern Med* 2008; 23 (12):2014-7.
42. Lamers F, Jonkers CCM, Bosma H, et al. Summed score of the Patient Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients. *J Clin Epidemiol* 2008; 61 (7):679-87.
43. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry* 2008; 8:46.
44. De Lima Osório F, Vilela Mendes A, Crippa J, et al. Study of the discriminative validity of the PHQ-9 and PHQ-2 in a sample of Brazilian women in the context of primary health care. *Psychiatr Care* 2009; 45 (3):216-227.
45. Gräfe K, Zipfel S, Herzog W, et al. Screening for psychiatric disorders with the Patient Health Questionnaire (PHQ). Results from the German validation study. *Diagnostica* 2004; 50 (4):171-181.
46. Fann JR, Bombardier C, Dikmen S, et al. Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. *J Head Trauma Rehabil* 2005; 20 (6):501-511.
47. Gilbody S, Richards D, Barkham M. Diagnosing depression in primary care using self-completed instruments: UK validation of PHQ-9 and CORE-OM. *Br J Gen Pract* 2007; 57 (541):650-2.
48. Gjerdingen D, Crow S, McGovern P, et al. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *Ann Fam Med* 2009; 7 (1):63-70.
49. Williams LS, Brizendine E, Plue L, et al. Performance of the PHQ-9 as a screening tool for depression after stroke. *Stroke* 2005; 36 (3):635-8.
50. Wittkampf K, van Ravesteijn H, Baas K, et al. The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. *Gen Hosp Psychiatry* 2009; 31 (5):451-9.
51. Meader N, Mitchell AJ, Chew-Graham C, et al. Case identification of depression in patients with chronic physical health problems: a diagnostic accuracy meta-analysis of 113 studies. *Br J Gen Pract* 2011; 61 (593):e808-820.
52. Meader N, Moe-Byrne T, Llewellyn A, et al. Screening for poststroke major depression: a meta-analysis of diagnostic validity studies. *J Neurol Neurosurg Psychiatry* 2014; 85 (2):198-206.
53. Mitchell AJ, Meader N, Davies E, et al. Meta-analysis of screening and case finding tools for depression in cancer: evidence based recommendations for clinical practice on behalf of the Depression in Cancer Care consensus group. *J Affect Disord* 2012; 140 (2):149-160.
54. Mitchell AJ, Meader N, Symonds P. Diagnostic validity of the Hospital Anxiety and Depression Scale (HADS) in cancer and palliative settings: a meta-analysis. *J Affect Disord* 2010; 126 (3):335-348.
55. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol* 2006; 59 (8):798-801.
56. Leeflang MM, Moons KG, Reitsma JB, et al. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clin Chem* 2008; 54 (4):729-737.
57. Goodacre S, Sampson FC, Sutton AJ, et al. Variation in the diagnostic performance of D-dimer for suspected deep vein thrombosis. *QJM* 2005; 98 (7):513-527.
58. Riley RD, Ahmed I, Ensor J, et al. Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds. *Syst Rev* 2015; 4:12.
59. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; 155 (8):529-536.
60. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010; 340:c221.
61. 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.

**Table 1.** Characteristics of Included Primary Studies of the Diagnostic Accuracy of the PHQ-9 (2004 – 2009)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **First Author, Year**  | **Country** | **Population** | **Diagnostic Interview** | **Classification System** | **Published Cutoffs** | **Total N** | **Major Depressive Disorder** |
| **N** | **%** |
| **Azah, 2005** (30) | Malaysia | Family medicine clinic patients | CIDI | ICD-10 | 5 to 12 | 180 | 30 | 17 |
| **De Lima Osório, 2009** (44) | Brazil | Females in primary care | SCID | DSM-IV | 10 to 21 | 177 | 60 | 34 |
| **Fann, 2005** (46) | United States | Adults within one year of traumatic brain injury | SCID | DSM-IV | 10, 12 | 135 | 45 | 33 |
| **Gilbody, 2007** (47) | United Kingdom | Primary care patients | SCID | DSM-III-R | 9 to 13 | 96 | 36 | 38 |
| **Gjerdingen, 2009** (48) | United States | Mothers registering their newborns for well-child visits, medical or paediatric clinics | SCID | DSM-IV | 10 | 438 | 20 | 5 |
| **Gräfe, 2004** (45) | Germany | Psychosomatic patients and patients at walk-in clinics and family practices | SCID | DSM-IV | 10 to 14 | 521 | 71 | 14 |
| **Lamers, 2008** (42) | Netherlands | Elderly primary care patients with type 2 diabetes or chronic obstructive pulmonary disease | MINI | DSM-IV | 6 to 8 | 611 | 277 | 45 |
| **Lotrakul, 2008** (43) | Thailand | Primary care patients | MINI | DSM-IV | 6 to 15 | 279 | 19 | 7 |
| **Stafford, 2007** (40) | Australia | Coronary artery disease patients | MINI | DSM-IV | 5, 6, 10 | 193 | 35 | 18 |
| **Thombs, 2008** (41) | United States | Cardiology outpatients | C-DIS | DSM-IV | 4 to 10 | 1024 | 224 | 22 |
| **Williams, 2005** (49) | United States | Stroke patients | SCID | DSM-IV | 10 | 316 | 106 | 34 |
| **Wittkampf, 2009** (50) | Netherlands | Primary care patients | SCID | DSM-IV | 10, 15 | 435 | 77 | 18 |
| **Yeung, 2008** (31) | United States | Chinese American primary care patients | SCID | DSM-IV | 15 | 184 | 37 | 20 |

**Abbreviations**: PHQ-9: Patient Health Questionnaire-9; C-DIS: Computerized Diagnostic Interview Schedule; CIDI: Composite International Diagnostic Interview; DSM: Diagnostic and Statistical Manual of Mental Disorders; ICD: International Classification of Diseases; MINI: Mini International Neuropsychiatric Interview; SCID: Structured Clinical Interview for DSM Disorders.

**Table 2.** Comparison of Results from Meta-Analyses of Studies Published 2004-2009 with Published Data Only Versus All Data1

|  |  |  |
| --- | --- | --- |
| **Published Data (Conventional Meta-analysis)** |  | **All Data (IPD Meta-analysis)**a |
| **Cutoff** | **N Studies** | **N Patients** | **N MDD** | **% MDD** | **Sensitivity** | **95% CI** | **Specificity** | **95% CI** |  | **Sensitivity**  | **95% CI** | **Specificity**  | **95% CI** |
| 7 | 4 | 2094 | 550 | 26 | 0.85 | 0.70-0.94 | 0.73 | 0.62-0.81 |  | 0.97 | 0.91-0.99 | 0.73 | 0.67-0.78 |
| 8 | 4 | 2094 | 550 | 26 | 0.79 | 0.63-0.89 | 0.78 | 0.71-0.85 |  | 0.93 | 0.85-0.97 | 0.78 | 0.74-0.82 |
| 9 | 4 | 1579 | 309 | 20 | 0.78 | 0.56-0.90 | 0.82 | 0.75-0.88 |  | 0.89 | 0.79-0.95 | 0.83 | 0.80-0.86 |
| 10 | 11 | 3794 | 723 | 19 | 0.85 | 0.71-0.93 | 0.88 | 0.85-0.91 |  | 0.87 | 0.75-0.94 | 0.88 | 0.85-0.90 |
| 11 | 5 | 1253 | 216 | 17 | 0.92 | 0.58-0.99 | 0.90 | 0.81-0.95 |  | 0.83 | 0.68-0.92 | 0.90 | 0.88-0.92 |
| 12 | 6 | 1388 | 261 | 19 | 0.82 | 0.65-0.92 | 0.92 | 0.87-0.96 |  | 0.77 | 0.63-0.87 | 0.92 | 0.90-0.94 |
| 13 | 4 | 1073 | 186 | 17 | 0.82 | 0.75-0.87 | 0.94 | 0.84-0.98 |  | 0.67 | 0.56-0.77 | 0.94 | 0.92-0.95 |
| 14 | 3 | 977 | 150 | 15 | 0.71 | 0.57-0.83 | 0.97 | 0.87-0.99 |  | 0.59 | 0.48-0.70 | 0.96 | 0.94-0.97 |
| 15 | 4 | 1075 | 193 | 18 | 0.61 | 0.52-0.70 | 0.98 | 0.96-0.99 |  | 0.52 | 0.42-0.62 | 0.97 | 0.96-0.98 |

1 10 is the standard cutoff for identifying MDD with the PHQ-9 (8, 10-14)

a For all cutoffs, N Studies = 13; N Patients = 4589; N MDD = 1037

**Abbreviations**: CI: confidence interval; IPD: individual patient data; MDD: major depressive disorder

**Table 3.** Discrepancies in Sensitivity and Specificity Across Cutoffs from Meta-Analyses of Studies Published 2004-2009

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cutoff** | **Percentage of Patients Included in Published Results** | **Percentage of Cases Included in Published Results** | **SensitivityPublished – SensitivityAll Data****(Conventional - IPD)** | **SpecificityPublished – SpecificityAll Data****(Conventional - IPD)** |
| **Estimate** | **95% CI** | **Estimate** | **95% CI** |
| 7 | 46 | 53 | -0.12 | -0.30, -0.01 | 0.00 |  -0.08, 0.12 |
| 8 | 46 | 53 | -0.14 |  -0.33, -0.01 | 0.00 |  -0.06, 0.09 |
| 9 | 34 | 30 | -0.11 |  -0.37, 0.03 | -0.01 |  -0.04, 0.09 |
| 10 | 83 | 70 | -0.02 |  -0.09, 0.03 | 0.00 |  -0.02, 0.00 |
| 11 | 27 | 21 | 0.09 |  -0.15, 0.18 | 0.00 |  -0.05, 0.08 |
| 12 | 30 | 25 | 0.05 |  -0.09, 0.16 | 0.00 |  -0.04, 0.05 |
| 13 | 23 | 18 | 0.15 |  0.01, 0.25 | 0.00 |  -0.05, 0.07 |
| 14 | 21 | 14 | 0.12 |  -0.10, 0.25 | 0.01 |  -0.04, 0.05 |
| 15 | 23 | 19 | 0.09 |  -0.09, 0.25 | 0.01 |  -0.03, 0.01 |

**Abbreviations**: CI: confidence interval; IPD: individual patient data

**Figure legend/titles**

**Figure 1.** Receiver Operating Characteristic Curves Produced by A) Conventional Meta-Analysis Using Published Data Only and B) Individual Patient Data Meta-Analysis Using All Data

Numbers within Receiver Operating Characteristic curves indicate each of the Patient Health Questionnaire-9 cutoffs between 7 (right) and 15 (left).

**Figure 2.** Patient Health Questionnaire-9 Reporting Patterns in the Original Studies

Cells shaded in grey represent cutoffs for which diagnostic accuracy results were reported in the original reports. Cells marked with an “O” represent each study’s optimal cutoff (the cutoff maximizing Youden’s J). For the Azah et al study, the optimal cutoff determined in the original study was using unweighted data and based on “any depression” rather than MDD. This table replicates the original decision on which cutoffs to publish. The study by de Lima Osório et al published cutoffs 10 through 21. For the study by Yeung et al, the optimal cutoff was determined using unweighted data rather than weighted, as the original study did not include weighted analyses.

**Figure 3.** Forest Plots of Sensitivity Estimates Using A) the Standard Cutoff 10 for Each Study and B) the Optimal Cutoff from Each Study. The square boxes represent the point estimates of sensitivity for each study, while the bars extending from each box represent the 95% confidence intervals.

**Figure 4.** Forest Plots of Specificity Estimates Using A) the Standard Cutoff 10 for Each Study and B) the Optimal Cutoff from Each Study. The square boxes represent the point estimates of specificity for each study, while the bars extending from each box represent the 95% confidence intervals.