Published in final edited form as:

Ann Epidemiol. 2014 July ; 24(7): 527–531. doi:10.1016/j.annepidem.2014.04.009.

Assessing Validity of a Depression Screening Instrument in the Absence of a Gold Standard

Bizu Gelaye, PhD^{a,b}, Mahlet G. Tadesse, ScD^c, Michelle A. Williams, ScD^a, Jesse R. Fann, MD, MPH^d, Ann Vander Stoep, PhD^b, and Xiao-Hua Andrew Zhou, PhD^e

^aDepartment of Epidemiology, Harvard School of Public Health, Boston, MA, USA

^bDepartment of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA

^cDepartment of Mathematics & Statistics, Georgetown University, Washington, DC

^dDepartments of Psychiatry and Behavioral Sciences, Rehabilitation Medicine and Epidemiology, University of Washington School of Public Health, Seattle, WA, USA

^eDepartment of Biostatistics, University of Washington School of Public Health, Seattle, WA, USA

Abstract

Background—We evaluated the extent to which use of a hypothesized imperfect gold standard, the Composite International Diagnostic Interview (CIDI), biases the estimates of diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9). We also evaluate how statistical correction can be used to address this bias.

Methods—The study was conducted among 926 adults where structured interviews were conducted to collect information about participants' current major depressive disorder (MDD) using PHQ-9 and CIDI instruments. First, we evaluated the relative psychometric properties of PHQ-9 using CIDI as a gold standard. Next, we employed a Bayesian latent-class model to correct for the bias.

Results—In comparison with CIDI, the relative sensitivity and specificity of the PHQ-9 for detecting MDD at a cut point of 10 were 53.1% (95%CI: 45.4–60.8%) and 77.5% (95%CI: 74.5–80.5%), respectively. Using a Bayesian latent-class model to correct for the bias arising from the use of an imperfect gold standard increased the sensitivity and specificity of PHQ-9 to 79.8% (95% Bayesian credible interval (BCI): 64.9–90.8%) and 79.1% (95%BCI:74.7–83.7%), respectively

© 2014 Elsevier Inc. All rights reserved.

Correspondence and requests for reprint: Bizu Gelaye, MPH, PhD, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Ave, Fifth Floor, Boston, MA 02115 USA, Telephone: 617-432-1071, Facsimile: 617-566-7805, bgelaye@hsph.harvard.edu.

Conflict of interest: None declared

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conclusion—Our results provided evidence that assessing diagnostic validity of mental health screening instrument, where application of a gold standard might not be available, can be accomplished by using appropriate statistical methods.

INTRODUCTION

Screening for mental health problems in clinical and community based settings is an important component in overall public health disease prevention and health promotion strategies. The success of screening procedures, however, is largely dependent on the accuracy of the diagnostic procedure, used as the gold standard or criterion standard, to evaluate the screening instruments [1–3]. Psychometric properties such as sensitivity and specificity are common measures used to evaluate the quality of a screening test. These psychometric properties are unbiased if the screening test results are compared with a gold standard measure [2, 3]. We recognize that there is no perfect gold standard and we use the term to refer to the best available method used to determine the presence or absence of the condition or disease of interest. However, verification of the true status using the gold standard may be impossible to obtain due to cost and human resources. In some instances the gold standard may be invasive, impractical to obtain, or unethical to conduct. For example, gold standard diagnosis of Alzheimer's disease cannot be ascertained until a patient dies and an autopsy is performed. In epidemiologic studies, where a comparison with a gold standard is not possible, validation studies often compare screening instruments with instruments that are imperfect but more precise than the screening instrument. The key assumption is that the measurement error for the imperfect reference or gold standard is unlikely to be correlated with the screening instrument [4]. If an imperfect standard is used as if it were a gold standard, the estimated accuracy of the tests would be biased due to misclassification [3, 5]. Zhou *et al* [3] call this type of bias “imperfect gold standard bias”. A number of authors have proposed model-based estimates or estimates that make use of prior information, to reduce or correct this imperfect gold standard bias without retesting subjects [1].

The Patient Health Questionnaire-9 (PHQ-9) is a very brief, easy to administer and interpret depression screening instrument [6]. Because of its brevity, the PHQ-9 is widely used as a depression screening instrument in primary care settings among racially and ethnically diverse populations. Additionally, it has been reported to be a valuable tool for the detection and management of depression [6]. Use of the PHQ-9 instrument in new clinical and research settings requires evaluating the validity of the instrument in comparison with a diagnostic ‘gold standard’. The Schedules for Clinical Assessment in Neuropsychiatry (SCAN), a semi structured clinical interview widely considered as gold standard, is used to assess and diagnose psychiatric disorders including depression among adults [7]. The instrument offers flexibility for clinicians to phrase questions about particular symptoms taking into account local context. However, it requires that clinicians make their clinical decisions following the definitions and criteria provided in the Diagnostic and Statistical Manual (DSM) [8, 9]. Although the SCAN has been reported to have excellent accuracy in depression diagnosis, it is time-demanding, expensive and requires a trained clinician, thus limiting its use in resource limited clinical settings [10].

To address these limitations, alternative diagnostic tools for the measurement of depression have been developed. One of these tools is the Composite International Diagnostic Interview (CIDI). The CIDI is a fully structured lay-administered diagnostic interview that can be used to diagnose major depressive disorder according to DSM criteria, though it is regarded as a less optimal gold standard compared to the SCAN [11, 12]. In an effort to employ a less burdensome assessment, some investigators have used the CIDI as the criterion standard to evaluate the diagnostic accuracy of the PHQ-9 [13, 14]. Use of CIDI as gold standard (an imperfect but more precise instrument than PHQ-9) is expected to bias the validity of the PHQ-9. However, no study has systematically evaluated the impact of using the CIDI as a gold standard. Therefore, we conducted this study to assess the extent to which the use of what we hypothesized to be an imperfect gold standard [3] biases the estimates of diagnostic accuracy of a screening instrument (the PHQ-9). We further sought to demonstrate how statistical methods can be used to correct the bias and improve psychometric properties of the PHQ-9.

METHODS AND MATERIALS

Participants

The study was conducted at Saint Paul General Specialized Hospital, a major referral and teaching hospital, in Addis Ababa, Ethiopia between the months of July and December, 2011. We used a two-stage study design where a total of 926 adults (≥ 18 years of age) attending outpatient departments were first interviewed by research nurses using the PHQ-9 and CIDI depression scales. Then those who screened positive for depression on the PHQ-9 questionnaire (PHQ-9 score ≥ 10), as well as a randomly selected sub-group of participants who screened negative for depression (PHQ-9 score < 10), were interviewed using the SCAN diagnostic questionnaire by licensed psychiatrists who were blind to PHQ-9 results. Following the PHQ-9 and CIDI interviews a total of 384 participants were invited to participate in the SCAN diagnostic interview (178 who screened positive on the PHQ-9 and 206 who screened negative on the PHQ-9) and 363 of them agreed to do so (94% of selected positive screens and 95% of selected negative screens). Those who refused to participate in the diagnostic interview cited reasons such as lack of time to do a follow-up interview with the psychiatrist. The questionnaire was originally written in English and then translated into Amharic. The translated version was back-translated and modified until the back-translated version was comparable with the original English version. All study participants provided informed consent, and all research protocols were approved by the institutional review boards of Addis Continental Institute of Public Health, Addis Ababa, Ethiopia and the Human Subjects Division at the University of Washington, Seattle, WA.

Study Measurements

Patient Health Questionnaire-9 (PHQ-9)—All study participants were evaluated for depressive symptoms using the PHQ-9 depression scale [15]. The PHQ-9 queries participants about the frequency with which they experience nine depressive symptoms. Each item requires participants to rate the frequency of a depressive symptom experienced in the two weeks prior to evaluation. The items include: 1) anhedonia, 2) depressed mood, 3) insomnia or hypersomnia, 4) fatigue or loss of energy, 5) appetite disturbances, 6) guilt or

worthlessness, 7) diminished ability to think or concentrate, 8) psychomotor agitation or retardation, and 9) suicidal thoughts. Scores for each item range from 0 (“not at all”) to 3 (“nearly every day”) with a total score ranging from 0 to 27 [15–17]

Composite International Diagnostic Interview (CIDI version 3.0)—Following PHQ-9 administration, participants were interviewed using the CIDI instrument. The CIDI is a fully structured diagnostic interview designed to be administered by lay interviewers. It was developed by the World Health Organization (WHO) to estimate the prevalence and identify correlates of mental disorders from large numbers of subjects across different countries. We used the depression and mania modules of the CIDI 3.0 to diagnose major depressive disorder (MDD). The mania module was used to rule out history of bipolar disorders (manic, mixed or hypomanic episodes). The CIDI interview includes three screening questions about sadness/depressed mood, feelings of discouragement, and loss of interest lasting several days or longer. Participants who endorsed any of these three questions were given the depression module. Those who failed to endorse any of the three core questions were skipped out of the depression module. In accordance with DSM-IV criteria, we defined MDD as the presence of five out of nine cardinal symptoms that persist for two weeks or longer, are present for most of the day nearly every day, and cause significant distress or impairment. These symptoms include dysphoric mood or anhedonia that persist most of the day, and clinically significant weight gain/loss or appetite disturbance, insomnia or hypersomnia, psychomotor agitation or retardation, fatigue or loss of energy, feelings of worthlessness or excessive guilt, diminished ability to concentrate or think clearly, and recurrent thoughts of death or suicide. For the purposes of this validation study, we defined current MDD as experiencing MDD in the past 12 months without plausible organic causes and without history of mania or hypomania [18, 19]. In the current study we considered the CIDI interview as an imperfect reference (or gold) standard [11, 12].

Schedules for Clinical Assessment in Neuropsychiatry (SCAN) Diagnosis

The SCAN is a semi structured clinical interview developed by the WHO for use by trained clinicians to assess and diagnose psychiatric disorders among adults [7, 20]. The use of the SCAN gives flexibility in the diagnosis of mental disorders based on ICD-10 and DSM-IV [20]. The depressed mood and ideation module of SCAN has been reported to work well in different languages and cultures including Ethiopia [21]. For the current study, we used SCAN administered in Amharic as a perfect reference (or gold) standard for MDD diagnosis in the past 12 months. Two clinicians who practiced psychiatry for many years applied the SCAN. Both had formal training on SCAN and were supervised by a senior psychiatrist who received SCAN training at WHO-designated SCAN training center. To ensure highest quality of data collection, while interviewers were in the field, they were provided strict on-site supervision and support. All SCAN interviews (and diagnoses) were conducted without knowledge of the results of initial PHQ9 and CIDI interviews.

Statistical Analysis

We first computed the psychometric properties of PHQ-9 using the CIDI as a reference or gold standard. Considering the CIDI as a reference standard subject to error, or imperfect

gold standard, we used the terms “relative sensitivity” and “relative specificity” to report test characteristics of the PHQ-9 [2]. Relative sensitivity is the percent of time that the PHQ-9 detected depression when the CIDI detected it. Similarly, relative specificity is the percent of time that the PHQ-9 did not detect depression when the CIDI did not detect it. We used a Bayesian modeling approach for diagnostic test in the absence of a gold standard developed by Branscum *et al* [22]. The Bayesian approach combines the likelihood function given the data and prior distributions for the model parameters through Bayes’ theorem to form the posterior distribution of the parameters [3, 23]. We followed this approach because prior scientific information about the psychometric properties of PHQ-9 can be incorporated in our estimation. The elicitation of prior distributions for model parameters can be done using past data, if available, or by drawing upon expert knowledge, or a combination of both [24]. As pointed out by Greenland [25], priors may vary considerably across individuals; a reasonable approach should reflect results from prior studies or reviews. The prior densities for sensitivity and specificity were assumed to be independent Beta distributions, which correspond to conjugate priors for the binomial sampling model. When evaluating two tests on a single population, as we are doing here, the model is not identifiable and informative priors need to be specified for some of the parameters [22]. We chose priors based on a recent diagnostic meta-analysis conducted by Manea *et al* [26]. We reasoned this will provide a less subjective prior than a single expert opinion. In their study, Manea and colleagues summarized the psychometric properties of the PHQ-9 across 18 validation studies ($n=7,180$) conducted in various clinical settings around the world. The authors provided psychometric properties of PHQ-9 for detecting MDD using cut-off scores between 8 and 11. For instance the prior values for sensitivity and specificity for diagnosing MDD using a cut-off score of 9 and above have a modal value of 83% (95% confidence interval (CI): 68–92%) and 86% (95% CI: 76–92%), respectively. These two values were used to generate the corresponding Beta prior distributions using the BetaBuster software (<http://www.epi.ucdavis.edu/diagnostictests/betabuster.html>). Because both the PHQ-9 and CIDI are based on DSM-IV criteria and are measuring the same psychological phenomenon (depression) and symptoms, the results are expected to be correlated [27]. Hence, a Bayesian model to assess test accuracy for two conditionally dependent tests was used [22] to estimate the sensitivity and specificity of the instruments. The model was fit using a WinBUGS code that implements this Bayesian latent class model [21]. Convergence of the Markov chain Monte Carlo (MCMC) sampler was checked by running two chains with different initial values. Mixing of the chains was assessed by the lack of autocorrelation among the sampled values. [28]. The posterior distributions were computed based on 10,000 iterations after discarding the initial 1,000 iterations as burn-in, and 95% Bayesian credible intervals (95% BCI) were evaluated.

RESULTS

The median PHQ-9 score was 5 (range 0–27). A total of 258 participants fulfilled DSM-IV criteria for MDD on the PHQ-9 (27.8%; 95% CI 24.9–30.7%) using a score of ≥ 10 , while 668 were classified as non-depressed (Table 1). In comparison to CIDI, a total of 86 subjects were classified as depressed and 592 as non-depressed on both instruments. Distributions for other PHQ-9 score cut-offs are also displayed in Table 1. As shown in Figure 1, PHQ-9

scores were significantly higher (Wilcoxon rank-sum, $p < 0.001$) among individuals classified as depressed on CIDI (median=10; 1st quartile=6; 3rd quartile=14; interquartile range =8, while for non-depressed participants the corresponding values were median=4; 1st quartile =2; 3rd quartile =9; interquartile range =7).

Relative psychometric properties of PHQ-9 using CIDI as a gold standard are shown in Table 2. In comparison with CIDI the relative sensitivity and specificity of PHQ-9 for a cut point of 9 were 61.7% (95%CI: 54.2–69.2%) and 73.8% (95%CI: 70.6–76.9%), respectively. As expected, lower sensitivities and improved specificities were noted for increased PHQ-9 score cut points. For instance the sensitivity and specificity of the PHQ-9 for MDD diagnosis at a cut point of 10 were 53.1% (95%CI: 45.4–60.8%) and 77.5% (95%CI: 74.5–80.5%), respectively. After applying a Bayesian latent class model, the sensitivity and specificity of PHQ-9 for a cut point of 9 increased to 81.0% (95% BCI: 66.6–91.2%) and 75.4% (95% BCI: 70.6–80.7%), respectively (Table 2). Using a cut point of 10 the sensitivity and specificity were estimated to be 79.8% (95% BCI: 64.9–90.8%) and 79.1% (95% BCI: 74.7–83.7%), respectively. The psychometric properties remained stable for PHQ-9 cut-point of 11. The results were robust to different prior prevalence assumptions. Based on the Youden Index, the diagnostic accuracy of the PHQ-9 was optimized at a cutoff of 10 or more (Youden index=0.59).

For comparison psychometric properties from our validation study of the PHQ-9 that used psychiatrist administered SCAN diagnosis (gold standard) are also presented in Table 2 [29]. Given the stage study design these estimates were evaluated using inverse probability weighting approaches (Begg and Greenes adjusted estimates). For instance, a score of 10 on the PHQ-9 is associated with 71 % sensitivity (61.2–83.9) and 77% specificity (74.8–79.5) in diagnosing MDD compared with psychiatrist administered SCAN diagnosis.

DISCUSSION

Overall, improvements in psychometric properties of the PHQ-9 (when compared against CIDI, an imperfect gold standard) were noted after employing Bayesian latent class modeling approaches compared to the relative estimates. Overall, the sensitivity estimates calculated using Bayesian modeling approaches were closer to the values obtained using the SCAN gold standard (regarded as the true or perfect gold standard in this study) while the relative estimates of the specificity were closer to the SCAN gold standard.

Brief, reliable and valid screening instruments help health care professionals determine the magnitude of mental health problems such as depression. As noted earlier, because of its brevity, the PHQ-9 [15, 17] has gained increased recognition as a preferred instrument for depression screening in primary care settings among racially and ethnically diverse populations [30, 31]. Use of screening and diagnostic instruments, such as the PHQ-9, in new settings requires evaluating their psychometric properties in comparison with a gold or criterion standard. In many developing countries such evaluations of the diagnostic validity of instruments remain a main challenge due to cost, time, and human resources limitation of using the gold or reference standard. One strategy to overcome this problem is the use of structured interviews that do not require a specially trained clinician to administer them.

Indeed, some investigators have used the CIDI as a gold standard to evaluate the diagnostic accuracy of the PHQ-9 [13]. Use of the CIDI as a gold standard results in errors due to the fact that it is a recognized imperfect gold standard [11, 12]. To our knowledge, this is the first study to examine the impact of using an imperfect gold standard (CIDI interview) to determine the validity of the PHQ-9 and to use statistical methods to correct for the bias. After applying Bayesian statistical modeling, we note the psychometric properties of the PHQ-9 were closer to what would be obtained if the SCAN gold standard were used.

Some caveats should be considered when interpreting the results of our study. Due to the self-report nature of psychological conditions, it is possible that the SCAN clinical diagnosis may not be the best gold standard or reference test. Furthermore, although we used prior estimates based on a previously conducted meta-analysis, one might arrive at slightly different estimates using different informative priors. As stated by Zou *et al* [3], estimates can be greatly affected by a change in prior distribution. Our sensitivity analysis, in which different prior distribution were used, showed that the posterior distributions for sensitivity estimates are influenced by the choice of the prior distribution. However, specificity and prevalence estimates remained quite stable. As shown in Figure 2, we speculate that the lower prevalence of depression in the study population results in sensitivity estimates close to the prior as there is less information. However for specificity; there is more information in the data for non-depressed individuals; so the influence of the prior is weaker. In addition, evaluating the posterior distributions using the SCAN gold standard estimates as priors yielded similar results. Finally, due to our use of one study population, the model is non-identifiable and requires the use of informative priors [22]. Some investigators [32] have noted that one can get more accurate estimates using a non-identifiable model with informative prior compared to an identifiable but less realistic model with strong assumptions. It will be useful in future studies to evaluate the impact of using two study populations with different disease prevalence estimates to correct for an imperfect gold standard. Finally, our study was conducted in an urban hospital; therefore, the sensitivity and specificity estimates may not be generalizable to populations in rural and remote areas.

In conclusion, our findings underscore the importance of evaluating the feasibility and validity of using an imperfect gold standard. In addition, using Bayesian statistical approaches in the absence of a gold standard may result in improvements of psychometric properties of mental health screening instruments. The methods presented here are useful in assessing the diagnostic validity of screening tests in the absence of a gold standard. The Bayesian modeling framework allowed incorporation of additional information from prior literature. Such additional information is valuable for assessing diagnostic accuracy of screening tests when gold standard is not available. This analytical approach may thus represent an important tool for adjusting the bias that might result from the use of an imperfect gold standard.

Acknowledgments

This research was supported, in part, by an award from the National Institutes of Health, National Institute of Minority Health and Health Disparities (T37-MD001449). The authors wish to thank the staff of Addis Continental Institute of Public Health for their expert technical assistance. The authors would also like to thank Saint Paul Hospital for granting access to conduct the study. This research was done as partial fulfillment for the requirements

of a PhD degree by one of the authors (B.G.) in the Department of Epidemiology, University of Washington, School of Public Health, Seattle, Washington, USA.

References

1. Bouma A, Stegeman JA, Engel B, de Kluijver EP, Elbers AR, De Jong MC. Evaluation of diagnostic tests for the detection of classical swine fever in the field without a gold standard. *Journal of veterinary diagnostic investigation : official publication of the American Association of Veterinary Laboratory Diagnosticians, Inc.* 2001; 13(5):383–388.
2. Pepe, M. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003.
3. Zhou, XH.; McClish, DK.; Obuchowski, NA. Statistical methods in diagnostic medicine. 2. Oxford: Wiley-Blackwell; 2011.
4. White, E.; Armstrong, BK.; Saracci, R.; Armstrong, BK. Principles of exposure measurement in epidemiology : collecting, evaluating, and improving measures of disease risk factors. Oxford; New York: Oxford University Press; 2008.
5. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health technology assessment.* 2007; 11(50):iii, ix–51. [PubMed: 18021577]
6. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine.* 2001; 16(9):606–613. [PubMed: 11556941]
7. Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, Jablenski A, Regier D, Sartorius N. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Archives of general psychiatry.* 1990; 47(6):589–593. [PubMed: 2190539]
8. Janca A, Ustun TB, Sartorius N. New versions of World Health Organization instruments for the assessment of mental disorders. *Acta Psychiatr Scand.* 1994; 90(2):73–83. [PubMed: 7976462]
9. Andrews G, Peters L, Guzman AM, Bird K. A comparison of two structured diagnostic interviews: CIDI and SCAN. *Aust N Z J Psychiatry.* 1995; 29(1):124–132. [PubMed: 7625961]
10. Hirsch, SR.; Weinberger, DR. Schizophrenia. Oxford; Cambridge, Mass., USA: Blackwell Science; 1995.
11. Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychological medicine.* 2001; 31(6):1001–1013. [PubMed: 11513368]
12. Wittchen HU. Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): a critical review. *J Psychiatr Res.* 1994; 28(1):57–84. [PubMed: 8064641]
13. Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, Falloon K, Hatcher S. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med.* 2010; 8(4):348–353. [PubMed: 20644190]
14. Sherina MS, Arroll B, Goodyear-Smith F. Criterion validity of the PHQ-9 (Malay version) in a primary care clinic in Malaysia. *The Medical journal of Malaysia.* 2012; 67(3):309–315. [PubMed: 23082424]
15. Spitzer RL, Williams JB, Kroenke K, Hornyak R, McMurray J. Validity and utility of the PRIME-MD patient health questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. *Am J Obstet Gynecol.* 2000; 183(3):759–69. [PubMed: 10992206]
16. American Psychiatric Association. Diagnostic and statistical manual of mental disorders DSM-IV. <http://www.psychiatryonline.com/DSMPDF/dsm-iv.pdf>
17. 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. [PubMed: 10568646]
18. Kessler RC, Abelson J, Demler O, Escobar JI, Gibbon M, Guyer ME, Howes MJ, Jin R, Vega WA, Walters EE, et al. Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview

- (WMHCIDI). *International Journal of Methods in Psychiatric Research*. 2004; 13(2):122–139. [PubMed: 15297907]
19. Kessler RC, Ustun TB. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research*. 2004; 13(2):93–121. [PubMed: 15297906]
 20. Aboraya A, Tien A, Stevenson J, Crosby K. Schedules for Clinical Assessment in Neuropsychiatry (SCAN): introduction to WV's mental health community. *W V Med J*. 1998; 94(6):326–328. [PubMed: 9868378]
 21. Alem A, Kebede D, Shibre T, Negash A, Deyassa N. Comparison of computer assisted scan diagnoses and clinical diagnoses of major mental disorders in Butajira, rural Ethiopia. *Ethiopian medical journal*. 2004; 42(2):137–143. [PubMed: 16895030]
 22. Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive veterinary medicine*. 2005; 68(2–4):145–163. [PubMed: 15820113]
 23. Broemeling, LD. Bayesian biostatistics and diagnostic medicine. Boca Raton: Chapman & Hall/CRC; 2007.
 24. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*. 1995; 141(3):263–272. [PubMed: 7840100]
 25. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol*. 2006; 35(3):765–775. [PubMed: 16446352]
 26. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2012; 184(3):E191–196.
 27. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001; 57(1):158–167. [PubMed: 11252592]
 28. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS \– A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*. 2000; 10(4):325–337.
 29. Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, Wondimagegn D, Lemenhe A, Fann JR, Vander Stoep A, et al. Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. *Psychiatry research*. 2013; 210(2):653–661. [PubMed: 23972787]
 30. Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *J Gen Intern Med*. 2006; 21(6):547–552. [PubMed: 16808734]
 31. Lichtman JH, Bigger JT Jr, Blumenthal JA, Frasure-Smith N, Kaufmann PG, Lesperance F, Mark DB, Sheps DS, Taylor CB, Froelicher ES. 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: endorsed by the American Psychiatric Association. *Circulation*. 2008; 118(17):1768–1775. [PubMed: 18824640]
 32. Gustafson P. The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. *Stat Med*. 2005; 24(8):1203–1217. [PubMed: 15558709]

Figure 1. Distribution of PHQ-9 scores across major depressive disorder status based on CIDI Box plots comparing PHQ-9 scores among those classified as having major depressive disorder (right side) and those without depression (left side) based on the CIDI depression module. The central box shows the data between the upper and lower quartiles, with the median represented by the middle line. The “whiskers” (lines on either side of central box) extend from the upper and lower quartiles to a distance of 1.5xIQR (interquartile range) away of the most extreme data point within that range, whichever is smaller.

Figure 2.
A Prior and Posterior Distribution Plot.

Table 1

Distribution of Patient Health Questionnaire-9(PHQ-9) and Composite International Diagnostic Interview (CIDI) results

	PHQ-9 Score cut-off		
	9	10	11
CIDI positive			
PHQ-9 positive	100	86	76
PHQ-9 negative	62	76	86
CIDI negative			
PHQ-9 positive	200	172	146
PHQ-9 negative	564	592	618
Total	926	926	926

Table 2

Relative and Bayesian model-adjusted sensitivity and specificity values along with associated 95% confidence intervals/credible intervals (CIs) for major depressive disorder diagnosis across various cut-off points of the PHQ-9

	Sensitivity	Specificity	NPV	PPV
Relative estimates				
Score 9	61.7 (54.2–69.2)	73.8 (70.6–76.9)	33.3	90.1
Score 10	53.1 (45.4–60.8)	77.5 (74.5–80.5)	33.3	88.6
Score 11	46.9 (39.2–54.6)	80.9 (78.1–83.7)	34.2	87.8
Bayesian estimates				
Score 9	81.0 (66.6–91.2)	75.4 (70.6–80.7)	26.9	97.7
Score 10	79.8 (64.9–90.8)	79.1 (74.7–83.7)	33.8	97.4
Score 11	78.9 (63.4–90.3)	82.3 (78.2–86.5)	39.4	96.5
Using SCAN gold standard				
Score 9	79.0 (68.2–89.1)	72.0 (68.5–74.9)	23.0	96.4
Score 10	71.1 (61.2–83.9)	76.6 (74.8–79.5)	23.5	94.5
Score 11	57.5 (41.6–72.0)	79.8 (78.0–81.5)	33.3	90.1