Published in final edited form as:

Gen Hosp Psychiatry. 2010 ; 32(5): 544–548. doi:10.1016/j.genhosppsy.2010.04.011.

Do the PHQ-8 and the PHQ-2 Accurately Screen for Depressive Disorders in a Sample of Pregnant Women?

Megan V. Smith, Dr.P.H. M.P.H.^{1,2}, Nathan Gotman, M.S.¹, Haiqun Lin, M.D., Ph.D.³, and Kimberly A. Yonkers, M.D.^{1,4}

¹Yale University School of Medicine, Department of Psychiatry

²Yale University School of Medicine, Department of Child Study

³Yale University School of Medicine, Department of Epidemiology and Public Health

⁴Yale University School of Medicine, Department of Obstetrics, Gynecology, and Reproductive Sciences

Abstract

Objective—The aim of this study was to assess the psychometric properties of the Patient Health Questionnaire-8 (PHQ-8), and the PHQ-2, a two-item version of the PHQ, respectively, in pregnancy. These screeners were compared to a structured diagnostic interview in a cohort of pregnant women attending prenatal care. Based upon studies documenting high sensitivity and specificity on the PHQ-8 and PHQ-2 in the general adult population, we hypothesized that both instruments would be effective in this population.

Methods—218 women, 13 of them depressed, were given the Composite International Diagnostic Interview and the PHQ-8 before 17 weeks of pregnancy. Receiver Operating Characteristic curves determined optimal thresholds and sensitivity and specificity were calculated using both dimensional and categorical approaches. Agreement between the PHQ-2 and PHQ-8 was measured using Cohen's kappa, κ .

Results—Optimal cutoffs for the PHQ-8 and PHQ-2 were 11 and 4, respectively. Using these cutoffs, the PHQ-8 had a sensitivity of 77% and a specificity of 62% while the PHQ-2 had a sensitivity of 62% and a specificity of 79%. The categorical method of scoring the PHQ-8 yielded a sensitivity of 54% and a specificity of 84%.

Conclusions—In our sample, the PHQ-8 and PHQ-2 performed almost equally in detecting probable major depressive disorder in a sample of pregnant women. The categorical scoring method for the PHQ-8 had lower sensitivity but slightly higher specificity than the dimensional version. We found the PHQ-8 and PHQ-2 to have lower sensitivity and specificity in our pregnant population as compared to findings in nonpregnant populations, however characteristics of our sample and choice of diagnostics instrument could explain these discrepant findings.

© 2010 Elsevier Inc. All rights reserved.

Corresponding Author: Megan V. Smith, DrPH, MPH., Yale University School of Medicine Departments of Psychiatry and Child Study Center, 142 Temple Street, Suite 301, New Haven, CT 06510, Phone: (203) 764-6621, Fax: (203) 764-6766, megan.smith@yale.edu.

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.

INTRODUCTION

Depression in pregnancy is common; one large meta-analysis has reported the range of antenatal depression to be between 6.5 and 12.9%.¹ However, depression in pregnancy is underdetected and undertreated.²⁻³ A growing body of research demonstrates that some of the consequences of untreated depression in pregnancy may include poor prenatal care and self care, medical and obstetrical complications, substance abuse, and an increased risk of relapse of depression in the postpartum period.^{4-6,7}

The United States Preventative Services Task Force has recommended screening for depression only if there is a comprehensive system of diagnosis and treatment available to track positive screening results.⁸ Two screening tools are used to detect depression in pregnant women, the most common is the Edinburgh Postnatal Depression Scale (EPDS).⁹ The EPDS has been validated for use in pregnant populations as well as thoroughly researched in postpartum populations.⁹ Only more recently has another screening tool, the Patient Health Questionnaire 9 (PHQ-9) been used to screen for depression in pregnancy. Unlike the EPDS, the PHQ-9 includes questions on alterations in sleep, appetite/weight, energy, and concentration, all symptoms that overlap with normative experiences of pregnancy and thus have the potential to impede an accurate assessment of a woman's mood. Although recent research has examined the PHQ-9 and the related two question PHQ-2 among postpartum women,¹⁰⁻¹¹ there is limited evidence about their performance in pregnant populations.

The aim of this study was to assess the psychometric properties of the Patient Health Questionnaire-8 (PHQ-8), an 8-item version of the PHQ, and the PHQ-2, a two-item version, to identify depression among a pregnant women attending prenatal care. This was done by comparing the PHQ-8 to a structured diagnostic interview for depressive disorders in a sample of 218 pregnant women. Based upon studies documenting high sensitivity and specificity on the PHQ-8 and PHQ-2 in the general adult population, we hypothesized that both instruments would perform well in this population.

METHODS

Recruitment

Subjects in this analysis were the first 218 women screened for participation and enrolled in the Yale Pink and Blue Study, a longitudinal cohort study investigating the effects of depression and antidepressant treatment on birth outcomes.¹³⁻¹⁴ Subjects were recruited from obstetrical offices or from hospital-based clinics in Connecticut and Western Massachusetts between 2004 and 2007. A total of 36 prenatal care sites served as sources of recruitment, 32 private obstetrician's offices and 4 publicly-funded obstetrical clinics in health centers and hospitals. Brochures and posters advertising the study targeting women in their first trimester of pregnancy were placed at each obstetrical office. Women who were interested in the study completed a form allowing staff from the Yale Pink and Blue Study to contact them. The study was approved by the Yale University School of Medicine Institutional Review Board (IRB) and IRBs at participating hospitals.

Screening

Women who agreed to be contacted were screened for eligibility by trained research assistants before 17 completed weeks of pregnancy using the PHQ-8. The PHQ-8 consists of eight of the nine criteria on which the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnosis of depressive disorders is based. The ninth DSM-IV criterion, which refers to suicidal ideation, is omitted in the PHQ-8. The criterion validity of the PHQ-8 has been demonstrated in community samples of adults¹⁵⁻¹⁶ and

with telephone administration. 15 17 Research supports the construct validity of the PHQ-8 in population-based settings to recognize symptoms of major depressive disorder. 15 18 The PHQ-2 contains the first two questions from the longer PHQ-9 and was developed for use in high volume primary care settings. 19 The instrument has been found to be sensitive and specific when used in a community sample of adults. A score of three or greater is considered a positive screen. 19 20

Inclusion/Exclusion Criteria

Women were eligible to participate in the Yale Pink and Blue Study if at the time of screening, they were intending to deliver at a participating hospital, were at least 17 years of age, had not yet completed 16 weeks of pregnancy and were willing to provide informed consent. Women were ineligible if they had a known multi-fetal pregnancy, were requiring insulin for diabetes, did not have access to a telephone, did not speak English or Spanish, were planning on relocating or intended to terminate their pregnancy. From interested volunteers, we invited participation from women who endorsed depressed mood or treatment for depression within the past five years and women who had experienced a traumatic event and had symptoms of re-experiencing that event. We also randomly selected one out of every three women who were not taking antidepressants and were neither diagnosed with nor treated for a depressive disorder in the last five years.

Assessment Procedures

Women meeting criteria for the study were interviewed in their homes prior to 17 weeks completed gestation. All interviews included the depression and anxiety disorder sections of the World Mental Health Composite International Diagnostic Interview (CIDI) 21 a valid and reliable lay interview. 22 The time frames were adjusted to obtain information longitudinally across pregnancy. Additional data collected included demographic, educational, obstetrical and treatment information. Mean time from screening with the PHQ-8 to interview with the CIDI was 1.73 weeks (S.D. 1.30). Bachelors and masters level Interviewers received a minimum of four days of didactic training followed by no less than six practice interviews and at least two supervised interviews of each type before becoming eligible to conduct independent interviews. Interviews were audiotaped, reviewed by a supervisor and coded with reference to the audiotape as needed. We assessed a random sample of 10% for reliability. This included a call by the supervisor to the subject that confirmed critical information (~5%) or a complete review of the audiotape and comparison to the written interview (~5%). Each interview took on average 67 minutes to complete.

Analytic Approach

Data were analyzed using SAS statistical software version 9.1.3. Associations between sociodemographic factors and major depressive disorder were assessed using Fisher's exact test for categorical variables and the Wilcoxon Rank-Sum Test for continuous variables. The PHQ-8 and the PHQ-2 were each assessed as screening measures compared to the CIDI interview algorithm for current (1-month) major depressive disorder (MDD). The PHQ-8 was assessed using both a dimensional (cutpoint) and categorical (DSM-IV-based) scoring algorithm. For each scoring method, we calculated sensitivity and specificity for MDD. Scoring for MDD based upon the categorical method required a minimum of 5 of the DSM-IV criteria, including either depressed mood or anhedonia, for more than half the days in a two-week interval. Additionally, we used the dimensional method of scoring to examine different cut points for the PHQ-8 and PHQ-2 and generate a receiver-operator characteristic curve (ROC). Based on this curve, we chose optimal cut points and computed sensitivity and specificity of the PHQ-8 and PHQ-2.

Traditionally, sensitivity and specificity are used to evaluate the effectiveness of instruments as screening measures. However, since these operating characteristics are usually presented without estimates of uncertainty, they are difficult to compare across studies. Cohen's Kappa statistic is a similar measure, typically used to evaluate agreement between measures of equal quality. Therefore we calculated Kappa statistics for the PHQ-8 and PHQ-2 with the CIDI. When one measure is viewed as superior (i.e. the reference standard), the Kappa statistic can still be informative. For example, when a screening measure can correctly identify positive and negative cases, agreement between the screening and the reference standard will be high; when it cannot correctly identify positive or negative cases, agreement will be low. As opposed to sensitivity and specificity, Kappa is more interpretable as a single measure and its uncertainty is easy to compute. A Kappa value of 1 implies perfect agreement and values less than 1 imply less than perfect agreement.

RESULTS

Sociodemographics

Clinical and sociodemographic characteristics of the population by a diagnosis of MDD on the CIDI are presented in Table 1. Two hundred and thirteen women were included in our analysis. Of these women, 13 met criteria for MDD based upon the CIDI. Mean maternal age was 29.3 years for depressed women and 28.9 years for nondepressed women. Both groups were well educated (54% in depressed group and 41% in nondepressed group had 16 or more years of school) and primarily white (69% in depressed group and 63% in nondepressed group) and married/cohabiting (83% in depressed group and 84% in nondepressed group). Differences in comorbid disorders (Post-traumatic Stress Disorder, Generalized Anxiety Disorder, and panic disorder) between depressed and nondepressed women were not significant.

ROC curves are shown in Figure 1. The areas under the PHQ-8 and PHQ-2 curves were virtually equivalent, 0.76 for the PHQ-2 and 0.77 for the PHQ-8. We found optimal cut points of 11 for the PHQ-8 and 4 for the PHQ-2; these were slightly higher than the recommended cut points of 10 for the PHQ-8 and 3 for the PHQ-2. In Table 2, we calculated sensitivity and specificity for each instrument using the recommended cut points in the literature and the optimal cut points derived from our data. Table 2 presents agreement (Kappa) between each instrument and the CIDI for MDD and the performance of the PHQ-8 using a categorical scoring method.

PHQ-2: For the PHQ-2 the optimal cutpoint of 4 had higher specificity (79%) but lower sensitivity (62%) than the recommended cutpoint of 3 (sens=59%, and spec.= 77%). The optimal cut off of 4 yielded a higher estimated Kappa ($\kappa=0.17$) than the traditional cutpoint ($\kappa=0.09$).

PHQ-8: The optimal cutpoint of 11 we found for the PHQ-8 resulted in a higher specificity (68%) compared to the traditional cutpoint of 10 (62%) and equal sensitivity (77%). The categorical method of scoring the PHQ-8 yielded a lower sensitivity (47%) but higher specificity (85%) compared to the dimensional scoring method.

Agreement with the CIDI for MDD was remarkably similar between the PHQ-2 optimal cut point of 4 ($\kappa=0.17$), the PHQ-8 optimal cut point of 11 ($\kappa=0.14$), and the PHQ-categorical scoring method ($\kappa=0.20$). Additionally, confidence intervals overlapped such that we did not detect a significant difference in agreement between the instruments.

DISCUSSION

This is the first study to our knowledge to compare the PHQ-8, an eight question screening tool used to detect depression in a general adult population and the PHQ-2, a two-item version of the tool, to a structured diagnostic interview in a sample of pregnant women. In our data, the PHQ-2 had similar operating characteristics and comparable sensitivity, specificity and agreement to the PHQ-8. Based on our findings, the cut points of 11 and 4 would maximize the operating characteristics of the PHQ-8 and PHQ-2 respectively. The use of the recommended cut points resulted in lower specificity in both the PHQ-8 and PHQ-2. Overall, sensitivity, specificity, and agreement were lower for our cohort compared to cohorts of general adults.

The largest validation studies of the PHQ-8 suggest a sensitivity and specificity of 88% in a general adult population.²³ For comparison, in our cohort, an instrument of this sensitivity and specificity would have a Kappa of 0.48, outside the confidence interval we found for the PHQ-8 optimal cut point (0.06—0.31). Previous studies²⁰ of the PHQ-2 report sensitivity of 87% and specificity of 78% that, in our cohort, would roughly equate to a Kappa of 0.32. This value falls into the confidence interval we found for the PHQ-2 suggesting our findings are similar to those in general adult populations.

The relatively high sensitivity we found for the PHQ-8 suggests the instrument could be clinically useful in that more depressed women will have a positive test result and therefore not be missed. Yet, potential negative consequences of low specificity include more women screening as at-risk of depression and labeled as depressed when they are not. An ideal screening tool maximizes both sensitivity and specificity but also is acceptable to the population and clinically feasible. There are notable differences between our study and other published studies including sample sizes, the diagnostic instrument used to determine criterion validity, and time frames used for comparison. In our study, we used current (1 month) MDD as the diagnostic criteria to which the screening instrument was compared. In other studies the SCID MDD current MDD algorithm was used and administered on the same day as the screening instrument. Although on average our diagnostic instrument was administered less than 1.8 weeks after the screen, studies in which the diagnostic interview was administered on the same day as the screen would most likely yield larger agreement between the screener and diagnostic interview.

The main limitation of this study is its relatively small number of cases of MDD (6%, n=13). While this would not affect our estimate, it does result in a less precise measurement as evident in the large confidence intervals surrounding the Kappa statistic.

Our finding of low sensitivity using a DSM-IV based algorithm to score the PHQ-8 could be due to the fact that the elimination of one question (suicidal ideation) means that not all DSM-IV symptoms were assessed on the instrument. This is likely to have had a minor impact, though, since prior large-scale validation studies of the PHQ have demonstrated that the scoring for the PHQ-8 and PHQ-9 is the same. Specifically, the suicide item is very informative but infrequently endorsed.¹⁶ Finally, our cohort of women receiving prenatal care was largely comprised of women who were highly educated and Caucasian, limiting the generalizability of our results.

In prior research with postpartum women and general adult medical settings, test characteristics of the PHQ-8 and 9 have been documented as favourable, leading the United States Preventative Services Task Force (USPSTF) to recommend it as a gold standard for depression screening in adult general medical settings. This is the first paper to examine the psychometrics of the PHQ-8 in a pregnant population. Our finding that the PHQ-8 and PHQ-2 performed equally in pregnancy suggests a two-item screen may be adequate for

detecting probable MDD in pregnancy. However, it is important to note that the any screening instrument can only suggest possible depression and the subsequent evaluation and management of women is the critical outcome.

Acknowledgments

This project was supported by grant 5-R01-HD045735 from the National Institutes of Child Health and Human Development

REFERENCES CITED

1. Gaynes, B.; Gavin, N.; Meltzer-Brody, S., et al. Quality AfHRA. Rockville, MD: 2005. Perinatal Depression: Prevalence, screening accuracy and screening outcomes. Evidence Report/Technology Assessment No. 119. AHRQ Publication No -5-E006-2: AHRQ Publication No. 05-E006-2.
2. Smith MV, Cavaleri MA, Howell HB, Poschman K, Rosenheck RA, Yonkers KA. Screening for and detection of depression, panic disorder, and PTSD in public-sector obstetrical clinics. *Psychiatric Services*. 2004; 55(4):407–414. [PubMed: 15067153]
3. Marcus S, Flynn H, Blow F, Barry K. Depressive symptoms among pregnant women screened in obstetrics settings. *Journal of Women's Health*. 2003; 12(4):373–380.
4. Zuckerman B, Amaro H, Bauchner H, Cabral H. Depressive symptoms during pregnancy: Relationship to poor health behaviors. *American Journal of Obstetrics and Gynecology*. 1989; 160:1107–1111. [PubMed: 2729387]
5. Kelly R, Danielsen B, Golding J, Anders T, Gilbert W, Zatzick D. Adequacy of prenatal care among women with psychiatric diagnoses giving birth in California in 1994 and 1995. *Psychiatric Services*. 1999; 50:1584–1590. [PubMed: 10577877]
6. Kelly R, Russo J, Holt V, et al. Psychiatric and substance use disorders as risk factors for low birth weight and preterm delivery. *Obstetrics & Gynecology*. 2002; 100(2):297–304. [PubMed: 12151153]
7. Suri R, Altshuler L, Hellemann G, Burt VK, Aquino A, Mintz J. Effects of antenatal depression and antidepressant treatment on gestational age at birth and risk of preterm birth. *American Journal of Psychiatry*. 2007 Aug; 164(8):1206–1213. [PubMed: 17671283]
8. Force UPST. Screening for Depression. 2002 [cited 2006 Feb]. 2002; Available for the AHRQ website, www.ahrq.gov/clinic/uspstf/uspstfdepre.htm. Available from:
9. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: Development of the 10-item Edinburgh Postnatal Depression Scale. *The British Journal of Psychiatry*. 1987; 150:782–786. [PubMed: 3651732]
10. Gjerdingen D, Crow S, McGovern P, Miner M, Center B. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *Annals of Family Medicine*. 2009; 7(1):63–70. [PubMed: 19139451]
11. Bennett I, Coco A, Coyne J, et al. Efficiency of a two-item pre-screen to reduce the burden of depression screening in pregnancy and postpartum: An implicit network study. *J Am Board Fam Med*. 2008; 21:317–325. [PubMed: 18612058]
12. Yawn B, Wilson P, Wollan P, et al. Concordance of Edinburgh Postnatal Depression Scale (EPDS) and Patient Health Questionnaire (PHQ-9) to Assess Increased Risk of Depression among Postpartum Women. *Journal of the American Board of Family Medicine*. 2009; 22(5):483–491. [PubMed: 19734393]
13. Spoozak L, Gotman N, Smith M, Belanger K, Yonkers K. Evaluation of a social support measure that may indicate risk of depression in pregnancy. *Journal of Affective Disorders*. 2008 In Press.
14. Yonkers KA, Gotman N, Smith MV, Belanger K. Typical somatic symptoms of pregnancy and their impact on a diagnosis of major depressive disorder. *Gen Hosp Psychiatry*. 2009; 31(4):327–333. [PubMed: 19555792]
15. McGuire L, Strine T, Allen R, Anderson L, Mokdad A. The patient health questionnaire 8: current depressive symptoms among U.S. older adults, 2006 Behavioral Risk Factor Surveillance System. *Am J Geriatr Psychiatry*. 2009; 17(4):324–334. [PubMed: 19307861]

16. Kroenke K, Strine T, Spitzer R, Williams J, Berry J, Mokdad A. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*. 2009; 114:163–173. [PubMed: 18752852]
17. Pinto-Meza A, Serrano-Blanco A, Penarrubia M, et al. Assessing depression in primary care with the PHQ-9: can it be carried out over the telephone? *J Gen Intern Med*. 2005; 20:738–742. [PubMed: 16050884]
18. Martin A, Rief W, Klaiberg A, et al. Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. *Gen Hosp Psychiatry*. 2006; 28:71–77. [PubMed: 16377369]
19. Kroenke K, Spitzer R, Williams J. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care*. 2003; 41:1284–1292. [PubMed: 14583691]
20. Lowe B, Kroenke K, Grafe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *Journal of Psychosomatic Research*. 2005; 58:163–171. [PubMed: 15820844]
21. WHO. Composite International Diagnostic Interview (CIDI, Version 2.1). Version 2.1 ed. Geneva, Switzerland: World Health Organization; 1997.
22. Wittchen H-U. Reliability and validity studies of the WHO Composite International Diagnostic Interview (CIDI): A critical review. *Journal of Psychiatric Research*. 1994; 28:57–84. [PubMed: 8064641]
23. Kroenke K, Spitzer R, Williams J. PHQ-9: Validity of a Brief Depression Severity Measure. *J Gen Intern Med*. 2001; 16:606–613. the. [PubMed: 11556941]

Figure 1.
Receiver Operator Characteristics (ROC) Curves for PHQ-2 and PHQ-8 Compared to CIDI Major Depressive Disorder Diagnosis in a Sample of Pregnant Women.

Table 1

Sociodemographic Characteristics of the Study Population by CIDI Major Depressive Disorder (MDD) Diagnosis

Characteristic	No MDD (N=201) N (%)	MDD (N=13) N (%)	P-value
Married/Cohabiting	173 (84)	10 (83)	1.00
Education (years)			0.14
1–11	32 (16)	4 (30)	
12	42 (21)	0 (0)	
13–15	46 (23)	2 (15)	
16+	84 (41)	7 (54)	
Race			0.57
White	126 (63)	9 (69)	
Black	42 (21)	1 (8)	
Hispanic	19 (9)	2 (15)	
Other	14 (7)	1 (8)	
Enrolled from Clinic	56 (27)	3 (25)	1.00
PTSD	7 (3)	2 (15)	0.09
GAD	2 (1)	0 (0)	1.00
Panic Disorder	4 (2)	0 (0)	1.00
	Mean (SD)	Mean (SD)	
Age	28.87 (6.72)	29.31 (5.98)	0.81
Gestational Age at Screening	10.88 (2.79)	11.17 (3.30)	0.78
PHQ2	2.18 (1.62)	3.85 (1.63)	0.001
PHQ8 (dimensional)	8.44 (4.64)	13.08 (4.97)	0.001

Table 2

Sensitivity, Specificity and Agreement of PHQ-8 and PHQ-2 Screening Instruments with CIDI MDD Diagnosis in Pregnancy

Screening Instrument	Sensitivity	Specificity	Kappa	Kappa 95% CI
PHQ2 (positive = 3 or more)	77	59	0.09	(0.02, 0.17)
PHQ2 (positive = 4 or more)	62	79	0.17	(0.04, 0.30)
PHQ8 (positive = 10 or more)	77	62	0.11	(0.02, 0.19)
PHQ8 (positive = 11 or more)	77	68	0.14	(0.04, 0.24)
PHQ8 (categorical)	54	84	0.20	(0.04, 0.36)