

# Validation of PHQ-2 and PHQ-9 to Screen for Major Depression in the Primary Care Population

Bruce Arroll, MBCChB, PhD,  
FRNZCGP<sup>1</sup>

Felicity Goodyear-Smith, MBCChB,  
MGP, FRNZCGP<sup>1</sup>

Susan Crengle, MBCChB, PhD,  
FRNZCGP<sup>2</sup>

Jane Gunn, MBBS, PhD, FRACGP<sup>3</sup>

Ngaira Kerse, MBCChB, PhD,  
FRNZCGP<sup>1</sup>

Tana Fishman, FRNZCGP<sup>1</sup>

Karen Falloon, MBCChB, FRNZCGP<sup>1</sup>

Simon Hatcher, MBBs, FRANZCP<sup>4</sup>

<sup>1</sup>Department of General Practice and Primary Health Care University of Auckland, Auckland, New Zealand

<sup>2</sup>Department of Maori Health, University of Auckland, Auckland, New Zealand

<sup>3</sup>Department of General Practice, University of Melbourne, Melbourne, Australia

<sup>4</sup>Department of Psychological Medicine, University of Auckland, Auckland, New Zealand

*Conflicts of interest: none reported*

## CORRESPONDING AUTHOR

Bruce Arroll, MBChB, PhD, FRNZCGP  
Department of General Practice & Primary Health Care  
Private Bag 92019  
Auckland, New Zealand  
b.arroll@auckland.ac.nz

## ABSTRACT

**PURPOSE** Although screening for unipolar depression is controversial, it is potentially an efficient way to find undetected cases and improve diagnostic acumen. Using a reference standard, we aimed to validate the 2- and 9-question Patient Health Questionnaires (PHQ-2 and PHQ-9) in primary care settings. The PHQ-2 comprises the first 2 questions of the PHQ-9.

**METHODS** Consecutive adult patients attending Auckland family practices completed the PHQ-9, after which they completed the Composite International Diagnostic Interview (CIDI) depression reference standard. Sensitivities and specificities for PHQ-2 and PHQ-9 were analyzed.

**RESULTS** There were 2,642 patients who completed both the PHQ-9 and the CIDI. Sensitivity and specificity of the PHQ-2 for diagnosing major depression were 86% and 78%, respectively, with a score of 2 or higher and 61% and 92% with a score 3 or higher; for the PHQ-9, they were 74% and 91%, respectively, with a score of 10 or higher. For the PHQ-2 a score of 2 or higher detected more cases of depression than a score of 3 or higher. For the PHQ-9 a score of 10 or higher detected more cases of major depression than the PHQ determination of major depression originally described by Spitzer et al in 1999.

**CONCLUSIONS** We report the largest validation study of the PHQ-2 and PHQ-9, compared with a reference standard interview, undertaken in an exclusively primary care population. The PHQ-2 score of 2 or higher had good sensitivity but poor specificity in detecting major depression. Using a PHQ-2 threshold score of 2 or higher rather than 3 or higher resulted in more depressed patients being correctly identified. A PHQ-9 score of 10 or higher appears to detect more depressed patients than the originally described PHQ-9 scoring for major depression.

*Ann Fam Med* 2010;8:348-353. doi:10.1370/afm.1139.

## INTRODUCTION

Unipolar depression is second only to stroke as the leading cause of disability-adjusted life years.<sup>1</sup> The incidence of "any depressive condition in the past 12 months" is estimated at 18.1% in a New Zealand family practice study.<sup>2</sup> Given that 80% of the population visits their family physicians each year,<sup>3</sup> family physicians are in an excellent position to improve the diagnosis and management of depressive disorders.

In the absence of systematic screening, family physicians miss at least 50% of cases of major depression.<sup>4</sup> The value of screening for depression in primary care is under debate, with the United States (US) Preventive services task force making the case for screening,<sup>4</sup> and the Cochrane review coming to the opposite conclusion.<sup>5</sup> The 9-item Patient Health Questionnaire (PHQ-9)<sup>6</sup> has been recommended for depression screening in primary care.<sup>7,8</sup>

Subsequent to screening in primary care is the issue of diagnosis. The PHQ-9 is a potentially valuable tool for diagnosis and management

of depression because it can generate a diagnosis of major depression, as well a continuous score to monitor treatment. Of recent interest has been the use of fewer screening questions,<sup>9</sup> including the use of the first 2 questions of the PHQ-9.<sup>10</sup> These 2 questions, known as the PHQ-2, ask about the frequency of the symptoms of depressed mood and anhedonia, scoring each as 0 (not at all) to 3 (nearly every day). The validation study of the PHQ-2 by Kroenke et al included a sample of 580 primary care patients and a reference standard interview conducted 48 hours later.<sup>10</sup> We believe this number of participants to be modest, and there are methodological issues associated with the delay of the reference standard interview. What is not fully known are the relative benefits of initial use of the PHQ-2 or PHQ-9 for screening in primary care.

The aim of our study was to validate the PHQ-2 and PHQ-9, using the computerized Composite International Diagnostic Interview (CIDI) as the reference standard,<sup>11-13</sup> in a larger cohort of primary care patients by administering the reference standard immediately after the screening test, and to compare the validations of the PHQ-2 and the PHQ-9. Specifically we wished to investigate the yields obtained with the PHQ-2 and the PHQ-9 at a range of thresholds compared with the scoring system originally described by Spitzer et al in 1999. We use the term PHQ major depression (original) to describe the Spitzer scoring system, which requires a score of 2 or higher on at least 1 of the first 2 questions and then a minimum score of 2 or higher on 5 of the questions. Clinically this scoring system is onerous to calculate, and we wished to see how these criteria compared with simple additive scores.

## METHODS

We report data from 1 arm of a 3-armed randomized control trial of screening for depression in primary care: one group received the PHQ-9<sup>8</sup>; a second group received the Two Questions With Help Question (TQWHQ)<sup>9</sup>; and the third, a control group, received no screening. The CIDI was administered as a reference standard to all groups.<sup>11,14</sup>

The setting was those family practices in Auckland who were willing to participate and able to provide a separate room for patient interviews. The study took place from 2006 to 2009. Patients were approached in the waiting room consecutively by the research assistant and asked to participate in the study.

### Recruitment

Family physicians in Auckland who worked more than 2 days a week in practice were eligible for the study. A fee of NZ\$9 per patient was paid to each family

physician to compensate for time spent on the study and on reassessing patients found to be suicidal on the questionnaires.

All eligible patients who gave informed consent were enrolled in the study. Eligible patients included those aged 16 years older who were able to communicate in English and who were not suffering from any brain injury, dementia, terminal illness, or intoxication. Patients were recruited consecutively to obtain an adequate spectrum of disease. Although we were more interested in screening patients who were not taking psychotropic medication, it was not feasible to exclude these patients at the time of interview.

### Measures

The PHQ-9 has 9 questions with a score ranging from 0 to 3 for each question (maximum score of 27). A threshold score of 10 or higher is considered to indicate mild major depression, 15 or higher indicates moderate major depression, and 20 or higher severe major depression. A threshold score of 15 or more is used in some settings to consider initiating treatment with antidepressants.<sup>7,15</sup>

### Procedures

The study was conducted according to the STARD guidelines.<sup>16</sup> Upon consenting to be part of the study, participants were invited by the research assistants into a private room to complete 1 of 3 randomly assigned screening questionnaires (PHQ-9, TQWHQ, or the control questionnaire eliciting demographic information only), after which they were administered the reference standard CIDI interview on a computer. The screening tools were prepackaged in brown, opaque, sealed envelopes (according to a concealed randomization code), which necessitated using practices with a spare room to facilitate an on-the-spot interview and to ensure privacy. The research assistants administering the CIDI were blinded to which arm they had been assigned and to their screening results.

The computerized CIDI is a software program that takes between 1 and 10 minutes to complete (1 minute if not depressed, or up to 10 minutes if depressed) and uses the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM IV)*<sup>17</sup> and its equivalent, the *International Classification of Diseases, Ninth Revision (ICD-9)*<sup>18</sup> diagnoses. We recorded only the *DSM IV* diagnoses, using the depression module, which indicates a diagnosis of depression and dysthymia. It also captures past episodes of depression. We did not use the bipolar (mania) module, as doing so would have lengthened the interview time. The CIDI may require a higher threshold score than the PHQ-9 for diagnosing depression.

The research assistants were instructed to remain blinded to the answers on the patients' screening questionnaires. They were also trained in administering the CIDI depression module and instructed not to assist patients with the CIDI unless there was extreme difficulty. After the CIDI interview the patients saw their family physician, who could read their results from the screening questionnaire. On occasion, a patient was seen by the physician before the CIDI could be administered, and the CIDI was administered after the consultation. A subgroup analysis determined that the timing of when the physician saw the patient did not significantly influence the CIDI results. The physician then completed a form saying whether the patient was depressed and whether the physician offered any mental health treatment. Family physicians had access to the screening results and were expected to deal with any issues of suicidality. Physician data will be reported elsewhere.

The methods and procedures used in this study were approved by the Northern Y Regional Ethics Committee, Ministry of Health (ethics approval number NTY/06/09/080).

**Statistical Methods**

All statistical analyses were carried out using the Centre for Evidence Based Medicine calculator on the University of Toronto Web site (<http://www.cebm.utoronto.ca>). Sensitivity and specificity were calculated comparing the scores on the screening tests with the reference standard. The sample size for the study was based on the power calculations for the randomized trial for a 21% difference between the 3 arms of the trial, which suggested a minimum sample size of 5,500 for the 3 groups. The final sample size of 7,757 for the 3 groups was determined by the amount of funding for the study, as we were interested in getting a well-powered study through the largest sample possible.

**RESULTS**

A total of 8,260 patients were approached (with the PHQ arm being one-third of the total); 358 refused (4.3%), and 145 did not complete their interviews (either the screening questionnaire or the CIDI interview), leaving 7,757 patients for analysis. Of these, 2,642 patients completed both a PHQ-9 and a CIDI interview, for a 95% response rate (2,642 of 2,795). The demographic characteristics of the PHQ-9 sample are shown in Table 1.

The prevalence of major depression in the past month determined from the CIDI data was 6.2% of the study population. Included were patients who were

and were not prescribed psychotropic medication. The number of patients on psychotropic medication was 242 (9.12%).

Tables 2 and 3 display the sensitivity, specificity, positive and negative likelihood ratios, positive predictive value (PPV, also known as the posttest likelihood for a positive test), and posttest likelihood of a negative test (PTL-, or 1-negative predictive value) for the PHQ-2 and the PHQ-9 when compared with the CIDI reference standard.

As an example, on Table 2, a PHQ-2 score of 2 or higher has a sensitivity of 0.86, meaning that 86% of those with a major depression will be found to be positive on the PHQ-2 screening test. The specificity of 0.78 means that 1-0.78, or 22%, of those who do not have depression will have a positive score (ie, a false-positive finding). The PPV of 21% means that of all those with a positive PHQ-2 screening test will have a major depression and that 79% will have a false-positive diagnosis. The higher the PPV, the better the test. The PTL- means that for those who have a negative PHQ-2 screening test, 1.2% will have a major depression (ie, a false-negative finding). The lower this score,

**Table 1. Demographic Characteristics of the Study Sample (N = 2,642)**

Characteristics	No.	%
Age, average (range), y	49 (17-99)	-
Sex		
Female	1,611	61.0
Male	1,031	39.0
Ethnicity		
European	1,881	71.0
Maori	191	7.2
Pacific Island	125	4.7
Chinese	33	1.3
Indian	43	1.6
Other	355	13.4
Unknown	14	0.5
Marital status		
Single	623	23.6
Married	1,372	51.9
Partnership, civil union	276	10.4
Divorced	159	6.0
Widowed	205	7.8
Unknown	7	0.3
PHQ-2 score		
≥1	1,150	44.0
≥2	679	26.0
≥3	298	11.3
PHQ-9 score		
≥10	341	13.0
≥15	137	5.2
≥19	57	2.2

**Table 2. Validity of PHQ-2 Compared With the CIDI and Posttest Probabilities**

Threshold Score for PHQ-2	Sensitivity	Specificity	LR+	LR-	PPV <sup>a</sup> %	PTL <sup>-b</sup> %
≥1	0.96 (156/163)	0.60 (1,485/2,479)	2.4	0.07	14	0.5
≥2	0.86 (140/163)	0.78 (1,940/2,479)	4.0	0.18	21	1.2
≥3	0.61 (100/163)	0.92 (2,281/2,479)	7.7	0.42	34	2.7
≥4	0.40 (66/163)	0.96 (2,388/2,479)	11.0	0.62	42	3.9

CIDI = Composite International Diagnostic Interview; LR+ = positive likelihood ratio; PHQ-2 = 2-item Patient Health Questionnaire; PPV = positive predictive value; PTL- = posttest likelihood of a negative test.

<sup>a</sup> Posttest likelihood of a positive text.

<sup>b</sup> 1 - Negative predictive value.

**Table 3. Validity of PHQ-9 Compared With the CIDI and the Posttest Probabilities**

Threshold Score for PHQ-9	Sensitivity	Specificity	LR+	LR-	PTL <sup>+</sup> <sup>a</sup> %	PTL <sup>-b</sup> %
≥8	0.82 (134/163)	0.85 (2,111/2,479)	5.8	0.16	27	1.0
≥10	0.74(121/163)	0.91 (2,259/2,479)	8.4	0.28	36	1.8
≥12	0.61(100/163)	0.94 (2,336/2,479)	10.6	0.44	41	2.6
≥15	0.45 (73/163)	0.97 (2,415/2,479)	17.3	0.57	53	3.6
PHQ major depression <sup>b,c</sup>	0.45 (73/163)	0.97 (2,405/2,479)	15.0	0.57	50	3.6

LR+ = positive likelihood ratio; LR- = negative likelihood ratio; PTL+ = posttest likelihood of a positive test; PTL- = posttest likelihood of a negative test; PHQ-9 = 9-item Patient Health Questionnaire.

<sup>a</sup> Positive predictive value.

<sup>b</sup> 1 - Negative predictive value.

<sup>c</sup> The original scoring system created by Spitzer et al in 1999<sup>6</sup> requires a score of 2 or higher on at least 1 of the first 2 questions and then a minimum score of 2 or higher on 5 of the questions for a diagnosis of PHQ major depression.

**Table 4. Comparison of Strategies**

Screening Strategy	Depressed by CIDI Diagnosis	Patients Requiring PHQ-9 %
If PHQ-2 score ≥2, complete the PHQ-9	140	26.0
If PHQ-2 score ≥3, complete the PHQ-9	100	11.3
PHQ-9 score ≥10 on all patients	121 <sup>a</sup>	100.0
PHQ-9 score major depression <sup>b</sup>	73 <sup>a</sup>	100.0

CIDI = Composite International Diagnostic Interview; PHQ-2 = 2-item Patient Health Questionnaire; PHQ-9 = 9-item Patient Health Questionnaire.

<sup>a</sup>  $\chi^2$ ,  $P < .001$ .

the better. Although the ideal test has a high PPV and a low PTL-, as the threshold rises (from 1 to 4 on Table 2) both the PPV and the PTL- rise. The likelihood ratios (positive and negative) are measures of validity (a combination of sensitivity and specificity) and are useful for clinicians when faced with situations in which the prevalence of depression may differ (eg, community, where the prevalence of depression is low, vs hospitalized patients, where it is usually higher).

When the positive likelihood ratio rises, so does the PPV, and when the negative likelihood ratio falls, so does the PTL-. Further discussion can be found in Guyatt et al.<sup>19</sup>

In Tables 3 and 4 the entry row "PHQ-9 major depression" refers to the original description of major depression, wherein 1 of the first 2 PHQ-9 questions has a score of 2 or higher, and at least 5 questions have a score of 2 or higher.<sup>8</sup> According to Table 4, which compares outcomes with the CIDI determination of depression, the most cases of depression were detected when the PHQ-2 score was 2 or higher and a PHQ-9 test score was positive, whereas the PHQ-9 major depression (original) determination detected the fewest cases. The number of patients with depression detected by a PHQ-9 score of 10 or more (n = 121) was significantly greater than number detected by the PHQ-9 major depression determination (n = 73) ( $P < .001$ ). Although depression would be detected in more patients using a PHQ-9 score of 10 or more, more patients would have false-positive scores and might receive treatment when, in fact, they do not have major depression. This situation may resolve clinically during follow-up where patients' serial PHQ scores register as not depressed or no longer depressed. The proportion of patients who would need to receive a PHQ-9 test after a positive PHQ-2 test score of 2 or higher would be 26%. When

the threshold score for the PHQ-2 is 3 or more, 11% would need to receive the PHQ-9 test.

## DISCUSSION

We report the first assessment of the PHQ-2 in an exclusively primary care population. The 2-question screen was very sensitive for a diagnosis of major depression when compared with the CIDI, with sensitiv-

ities of 0.96 and 0.86 for thresholds of 1 and greater and 2 and greater, respectively. The price paid for this high sensitivity, however, was a modest specificity of 0.60 and 0.78, respectively. At the commonly used threshold score of 3 or more, the sensitivity was 0.61 and the specificity was 0.92. The PHQ-9, in comparison, had similar sensitivities but good specificities. The finding that a score on the PHQ-9 of 10 or higher was more successful in detecting cases of major depression than the original determination of the PHQ-9 for major depression (ie, with 5 questions scoring 2 or higher, including at least 1 of the first 2 questions) suggests that the original criterion may be too strict for clinical practice.

### Strengths and Limitations

The strengths of this study are that all the patients were from primary care and they all received the CIDI reference standard assessment immediately after the PHQ-2 screening test was completed. Ours is the largest primary care study of the PHQ-2 in terms of those who received a reference standard assessment. The research assistants were blinded to the screening questionnaire, and they administered the CIDI assessment without looking at the results of the screening test. The patients were invited into the study consecutively, ensuring that there was an adequate spectrum of disease. The acceptance rate to participate in the study was high. The concern that the CIDI threshold score for depression is higher than the PHQ threshold score has been answered in that the original description for PHQ major depression is similar to a threshold score of 15 or higher on an additive score, whereas more patients with depression are detected using a threshold score of 10 or more.

A limitation to the study is that it was conducted in a New Zealand population and may not be completely generalizable to other primary care settings in which the PHQ-2 is utilized.

### Interpretation of Findings in Context of Previous Studies

Another study to include a primary care sample (but not exclusively) reported a sensitivity of 0.83 and a specificity of 0.92 when the PHQ-2 (threshold score of 3 or higher) was compared with a health professional interview in 580 patients.<sup>10</sup> The patients who received the reference standard interview had to be contacted within 48 hours of the screening interview, which may have introduced a bias into the results in that the more reliable patients returned for the reference standard interview.

A study conducted in older patients using the *DSM-IV* as a reference standard reported a sensitivity of 1.0 and a specificity of 0.77 for the PHQ-2.<sup>20</sup>

In a study of maternal depression in a low-education population with the Edinburgh postnatal depression scale as the reference standard, the sensitivity and specificity for the PHQ-2 was 0.435 and 0.972, respectively.<sup>21</sup> The sensitivity was higher for women who were educated beyond high school compared with those who were not. This finding suggests that the sensitivity at least can be influenced by demographic factors.

In a cardiology clinic the PHQ-2 was compared with the Diagnostic Interview Schedule as a reference standard, and the sensitivity and specificity were 0.39 and 0.92, respectively. The PHQ-9 also did not perform well in this setting, with a sensitivity of 0.20 and specificity of 0.90, respectively, for a threshold score of 10 or higher.<sup>22</sup>

A further study conducted in an outpatient clinic in Germany found a sensitivity and specificity of 78% and 79%, respectively, for major depression determined by a PHQ-2 score of 3 or more.<sup>23</sup> The prevalence of depression was 25.4%, which is much higher than a screened primary care population, and the reference standard was in a subpopulation of the whole group.

At a threshold score of 3 or higher and using a recognized reference standard, our sensitivity results for the PHQ-2 are generally not as high as those of other studies. This outcome may be the result of a truly consecutive sample of patients in primary care, a reference standard that was administered immediately after the screening test, or simply chance.

### Implications for Practice

For clinicians who wish to screen their patients for depression, we suggest they ask patients to respond to the first 2 questions of the PHQ-9 (ie, the PHQ-2); if their score is positive (if they score 2 or more), the patients should then complete the PHQ-9. At a PHQ-2 threshold score of 2 or more, 26% of patients will continue to complete the full PHQ-9; at a threshold score of 3 or more, 11% need will continue to complete the full PHQ-9.

That 63 patients with major depression would be missed at a threshold score of 3 or higher on the PHQ-2 would probably trouble most primary care clinicians. If clinicians prefer to miss as few cases of depression as possible, then a PHQ-2 threshold score of 2 or more rather than 3 or more would be prudent. For this reason, we recommend the 2 or higher threshold score on the PHQ-2 to be more certain that all those with depression are detected. Thus the price paid for a more complete detection of depression would be to have 26% of patients complete the full PHQ-9 (ie, an additional 7 questions).

The PHQ-2 can be a useful and time-saving tool in assisting primary care physicians with screening for

depression. Patients can be asked to complete the full PHQ-9 if their score is 2 or higher. Using a threshold score of 2 or more reduces the case-finding load, with only 26% of patients needing to progress to the PHQ-9. We believe this threshold score has clinical advantages over a threshold score of 3 or higher in that more patients with depression will be detected. A reevaluation of the original PHQ-9 criteria for major depression may also be needed, as the simple additive score PHQ-9 of 10 or higher identified more patients with depression than the originally described (and more time-consuming) method for scoring the PHQ-9.

**To read or post commentaries in response to this article, see it online at <http://www.annfammed.org/cgi/content/full/8/4/348>.**

**Key words:** Depression; physicians; family; mass screening

Submitted July 9, 2009; submitted, revised, December 29, 2009; accepted January 4, 2010.

**Funding support:** This study was funded by the Health Research Council of New Zealand with a project grant.

**Acknowledgments:** The investigators wish to thank the family practitioners from the Auckland region and the research assistants who collected the data.

## References

- World Health Organization. World Health Report. 2004. <http://www.who.int/whr/en/>.
- MaGPIe Research Group. The nature and prevalence of psychological problems in New Zealand primary healthcare: a report on mental health and general practice investigation (MAGPIE). *N Z Med J*. 2003;116(1171):U379.
- Ministry of Health. *Taking the Pulse. The 1996/97 New Zealand Health Survey*. Wellington: Ministry of Health; 1999:197-202.
- US Preventive services task force. Screening for depression: recommendations and rationale. *Ann Intern Med*. 2002;136(10):760-764. <http://www.ahrq.gov/clinic/3rduspstf/depression/depresswh.pdf>.
- Gilbody S, House AO, Sheldon TA. Screening and case finding instruments for depression. *Cochrane Database Syst Rev*. 2005; CD002792.
- Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of the prime-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA*. 1999;282(18):1737-1744.
- The Macarthur Foundation. The Macarthur Initiative on Depression and Primary Care. 2003. <http://www.depression-primarycare.org>.
- Nease DE Jr, Maloin JM. Depression screening: a practical strategy. *J Fam Pract*. 2003;52(2):118-124.
- Arroll B, Goodyear-Smith F, Kerse N, Fishman T, Gunn J. Effect of the addition of a "help" question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study. *BMJ*. 2005;331(7521):884.
- Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care*. 2003; 41(11):1284-1292.
- 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.
- Wittchen HU, Lachner G, Wunderlich U, Pfister H. Test-retest reliability of the computerized DSM-IV version of the Munich-Composite International Diagnostic Interview (M-CIDI). *Soc Psychiatry Psychiatr Epidemiol*. 1998;33(11):568-578.
- Jordanova V, Wickramesinghe C, Gerada C, Prince M. Validation of two survey diagnostic interviews among primary care attendees: a comparison of CIS-R and CIDI with SCAN ICD-10 diagnostic categories. *Psychol Med*. 2004;34(6):1013-1024.
- Peters L, Andrews G. Procedural validity of the computerized version of the Composite International Diagnostic Interview (CIDI-Auto) in the anxiety disorders. *Psychol Med*. 1995;25(6):1269-1280.
- New Zealand Guidelines Group. *Identification of Common Mental Disorders and Management of Depression in Primary Care*. 2009. <http://www.nzgg.org.nz>.
- Bossuyt PM, Reitsma JB, Bruns DE, et al; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003;326(7379):41-44.
- Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*. Washington, DC: American Psychiatric Association; 1994.
- ICD-9 –The International Classification of Diseases, Ninth Revision. <http://www.cdc.gov/nchs/icd/icd9.htm>.
- Guyatt G, Rennie D, Meade MO, Cook DJ. *User's Guide to the Medical Literature*. 2nd ed. Chicago, IL: American Medical Association, McGraw-Hill; 2008.
- Li C, Friedman B, Conwell Y, Fiscella K. Validity of the Patient Health Questionnaire 2 (PHQ-2) in identifying major depression in older people. *J Am Geriatr Soc*. 2007;55(4):596-602.
- Cutler CB, Legano LA, Dreyer BP, et al. Screening for maternal depression in a low education population using a two item questionnaire. *Arch Women Ment Health*. 2007;10(6):277-283.
- McManus D, Pipkin SS, Whooley MA. Screening for depression in patients with coronary heart disease (data from the Heart and Soul Study). *Am J Cardiol*. 2005;96(8):1076-1081.
- Löwe B, Kroenke K, Gräfe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *J Psychosom Res*. 2005; 58(2):163-171.