Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review

Wittkampf K A, Naeije L, Schene A H, Huyser J, van Weert H C

CRD summary
The authors concluded that the Patient Health Questionnaire (PHQ-9) is a valid method for diagnosing major depressive episodes (MDEs) in subgroups of primary care patients with a high prevalence of MDEs. The review was methodologically sound and well-reported. The authors' conclusions reflect the limited data, but should be viewed with caution given the small number of studies.

Authors' objectives
To determine the sensitivity and specificity of the 9-item mood module of the Patient Health Questionnaire (PHQ-9) for the diagnosis of depressive disorders in primary care.

Searching
EMBASE, MEDLINE and PsycINFO were searched from 1999 (date of PHQ issue) to July 2006; the search terms were reported. Additional articles were sought by checking the bibliographies of included studies. No language restrictions were applied.

Study selection
Studies assessing the diagnostic accuracy of the PHQ-9 for major depressive episodes (MDEs) were eligible for inclusion. The included studies were required to compare the PHQ-9 with a reference standard based on structured interview using the American Psychiatric Association's 'Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition' (DSM-IV) criteria. Studies administering the PHQ-9 by telephone were excluded. The included studies were conducted in a variety of populations (including both primary and secondary care). The prevalence of major depressive disorder ranged from 6.3 to 19.8% in primary care studies and from 7.9 to 33.5% in secondary care studies. Secondary care studies were conducted in a variety of patient groups, including post-stroke, dialysis, heart disease, dermatology and general medical. The components of the PHQ assessed and the diagnostic threshold used, as well as the reference standard test, varied between the studies.

Two authors independently assessed studies for inclusion, with the final decision being made by consensus with a third author.

Assessment of study quality
The methodological quality of the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool. As scoring of the index test is automated and does not involve interpretation, items relating to the blinding of operators to the reference standard results and clinical data were omitted.

Two authors independently performed the quality assessment.

Data extraction
Data to populate 2x2 contingency tables were extracted and sensitivity and specificity values calculated. The diagnostic threshold for the PHQ was defined in two ways, using a diagnostic algorithm and a summary score, and, where possible, data were extracted for both thresholds. The application of both methods was described in the article. If data could not be extracted, study authors were contacted.

The authors did not state how the data were extracted for the review, or how many reviewers performed the data extraction.
Methods of synthesis
Sensitivity and specificity values for individual studies were presented in forest plots. The studies were grouped by setting (primary or secondary care) and diagnostic threshold (algorithm or summary score). A random-effects method, weighted by the inverse variance, was used to generate summary estimates of sensitivity and specificity for the 4 studies conducted in primary care using a diagnostic algorithm threshold for the PHQ-9. These estimates were then used to generate a pre-test post-test probability modifying plot.

Results of the review
Twelve studies with a total of 6,394 participants were included in the review.

In all 12 studies the index test formed part of the reference standard, indicating a potential for incorporation bias.

Primary care.
For studies using a diagnostic algorithm threshold (4 studies, n=3,053), the sensitivity ranged from 0.71 (specificity 0.92) to 0.86 (specificity 0.94), and the specificity ranged from 0.90 (sensitivity 0.83) to 0.98 (sensitivity 0.73). The pooled estimates for sensitivity and specificity were 0.77 (95% confidence interval, CI: 0.71, 0.84) and 0.94 (95% CI: 0.90, 0.97), respectively. There was significant between-study heterogeneity in the estimates of specificity. The one study using a summary score threshold reported a sensitivity and specificity of 0.88.

Secondary care.
For studies using a diagnostic algorithm threshold (4 studies, n=1,411), the sensitivity ranged from 0.43 (specificity 0.92) to 0.84 (specificity 0.92), and the specificity ranged from 0.91 (sensitivity 0.55) to 0.94 (sensitivity 0.69). For studies using a summary score threshold (3 studies, n=1,402), the sensitivity ranged from 0.54 (specificity 0.90) to 0.94 (specificity 0.91), and the specificity ranged from 0.89 (sensitivity 0.91) to 0.91 (sensitivity 0.94).

Authors’ conclusions
The PHQ-9 is a useful diagnostic tool when applied in selected primary care populations with a high prevalence of depressive disorder. In the general primary care population (estimated prevalence of MDE = 10%), the PHQ-9 is a useful tool in avoiding overdiagnosis (high specificity), but may miss some patients (poor sensitivity).

CRD commentary
The review addressed a clearly stated research question that was defined by limited, but appropriate inclusion criteria. The review methodology was generally robust and the methods and results were reported clearly. The methodological quality of the included studies was assessed, and the results of this assessment reported and discussed. The authors did not specifically search for unpublished studies and they acknowledged that publication bias remains a possibility. The analyses were severely limited by the available data, which were few and heterogeneous, as the authors acknowledged. Within these limitations the methods used were appropriate and the results were reported clearly. The authors’ interpretation of the data presented is reasonable, but should be viewed cautiously given the paucity of data.

Implications of the review for practice and research
Practice: The authors stated that the PHQ-9 is a valid instrument to detect patients with MDEs. The PHQ-9 should not be used in populations with a low pre-test probability (<10%) because of the risk of overdiagnosis and overtreatment. In these populations, positive PHQ-9 results suggest ‘possible’ MDE, and should be followed by further investigations.

Research: The authors stated that in future meta-analyses of diagnostic data, bivariate meta-regression models should be used to more accurately estimate the sensitivity and specificity.

Funding
Not reported.

Bibliographic details

**PubMedID**
17888804

**DOI**
10.1016/j.genhosppsych.2007.06.004

**Indexing Status**
Subject indexing assigned by NLM

**MeSH**
Depression /diagnosis; Diagnostic Errors /prevention & control; Humans; Netherlands; Primary Health Care; Psychometrics; Surveys and Questionnaires

**AccessionNumber**
12007003316

**Date bibliographic record published**
07/02/2008

**Date abstract record published**
03/11/2008

**Record Status**
This is a critical abstract of a systematic review that meets the criteria for inclusion on DARE. Each critical abstract contains a brief summary of the review methods, results and conclusions followed by a detailed critical assessment on the reliability of the review and the conclusions drawn.