Do ultra-short screening instruments accurately detect depression in primary care: a pooled analysis and meta-analysis of 22 studies

Mitchell A J, Coyne J C

CRD summary
This review concluded that the sensitivity of single-question tests for detecting depression is unacceptable for use in primary care, and that two- or three-question tests have better sensitivity but higher false-positive rates. The review suffered from a number of limitations, particularly the failure to consider differences between the studies. These conclusions are therefore unlikely to be reliable.

Authors' objectives
To evaluate whether ultra-short screening tests can accurately detect depression in primary care.

Searching
MEDLINE, PsycINFO, EMBASE, CINAHL and Web of Knowledge were searched from inception to June 2006; the search terms were reported. Internet searches of online journal collections were also conducted.

Study selection
Study designs of evaluations included in the review
Inclusion criteria were not defined in terms of the study design.

Specific interventions included in the review
Studies of ultra-short screening tools for depression, defined as tools with one to four items that take less than 2 minutes to complete, were eligible for inclusion. Studies of visual analogue scales were excluded. The included studies assessed one-, two- or three-item tools.

Reference standard test against which the new test was compared
Inclusion criteria were not defined in terms of the reference standard. The reference standards used in the included studies for the diagnosis of depression were: Diagnostic Interview Schedule, Structured Clinical Interview for Depression (DSMIV), Geriatric Depression Scale, Patient Health Questionnaire (PHQ9), Composite International Diagnostic Interview (CIDI), CIDI-Auto, DSMIV(4 symptoms) and Clinician IV.

Participants included in the review
It appears that the review focused on patients in primary care, although this was not explicitly stated. Studies of medical patients, or studies exclusively in secondary care or nursing home settings, were excluded. The included studies were conducted in community veterans administration clinics, primary care alone, and a mix of primary care and medical clinics.

Outcomes assessed in the review
Inclusion criteria were not defined in terms of the outcomes. The outcomes reported in the review were the sensitivity, specificity, positive and negative predictive values, and relative risks.

How were decisions on the relevance of primary studies made?
The authors did not state how the papers were selected for the review, or how many reviewers performed the selection.

Assessment of study quality
Studies were assessed for methodological quality using the STARD guidelines for reporting diagnostic accuracy studies and the Newcastle-Ottawa Scale for non-randomised studies. The authors did not state how many reviewers performed the validity assessment.
Data extraction
The authors did not state how the data were extracted for the review, or how many reviewers performed the data extraction. The sensitivity, specificity, positive and negative predictive values, and Youde's index were calculated from raw data reported in the primary studies. Relative risks were calculated but it was unclear on what data these were based.

Methods of synthesis
How were the studies combined?
The results were analysed separately according to the number of questions in the test. The pooled sensitivity and specificity were estimated. An overall pooled relative risk was estimated using Mantel-Haenszel fixed-effect models. Publication bias was investigated visually using a funnel plot and statistically using the Begg-Mazumdar test.

How were differences between studies investigated?
Heterogeneity was investigated visually using a forest plot and statistically using the Q statistic.

Results of the review
Twelve studies reporting 22 analyses were included (20,855 patients).

Single-question tests (6 studies, 8 analyses, 17,624 patients).
The sensitivity ranged from 8 to 93% and the specificity from 62 to 99.5%. The pooled sensitivity and pooled specificity were 32% and 96%, respectively. Confidence intervals were not reported.

Two- or three-question tests (9 studies, 14 analyses, 9,653 patients).
The sensitivity ranged from 18 to 97% and the specificity from 10 to 100%. The pooled sensitivity and pooled specificity were 74% and 75%, respectively. Confidence intervals were not reported.

Authors' conclusions
The sensitivity of single-question tests for the diagnosis of depression is unacceptable for use in primary care. Two- or three-question tests have better sensitivity but higher false-positive rates.

CRD commentary
This review suffered from a number of limitations. The review question was focused but the inclusion criteria were not clearly defined and had to be deduced from other information reported. The literature search was adequate, but unpublished studies were not sought and it is unclear whether any language restrictions were applied. Details of the review process were not reported, so it is not clear whether appropriate steps were taken to minimise bias and error. Studies were assessed for methodological quality against the STARD criteria, but the results of this assessment were not reported. Limited study details were presented in a supplemental table available on the British Journal of General Practice website (accessed 28/03/2008; a subscription may be required for access), but very few study details were discussed in the text. The reliability and generalisability of the results of these studies is therefore unclear.

There was considerable heterogeneity in the estimates of sensitivity and specificity, although this was not mentioned in the text; it is only apparent from the supplemental data available on the British Journal of General Practice website. Given this heterogeneity it is questionable whether pooling was appropriate. It was certainly not appropriate to simply present the pooled figures without confidence intervals and with no consideration of differences between the studies. The predictive values presented were also misleading as prevalence varied across the studies and was not considered in this analysis. The authors also presented an overall analysis based on relative risks, which is not appropriate for diagnostic data, and it is unclear how these relative risks were calculated.

Given the limitations of this review, especially in relation to the analysis, the authors' conclusions are unlikely to be
Implications of the review for practice and research
Practice: The authors stated that ultra-short tests can be used to rule out depression but they should only be used if there are sufficient resources for more in-depth assessment of those that screen positive. Given the limitations of the review, these implications should be interpreted with extreme caution.

Research: The authors stated that further research is needed to investigate how ultra-short tests compare with short and long case-finding methods and how they compare with diagnosis by a general practitioner alone.

Bibliographic details

PubMedID
17263931

Other publications of related interest
This additional published commentary may also be of interest. Wills CE. Review: ultra-short screening tests are not highly accurate for detecting depression in primary care. Evid Based Nurs 2007;10:118.

Indexing Status
Subject indexing assigned by NLM

MeSH
Depressive Disorder /diagnosis; Family Practice; Humans; Psychiatric Status Rating Scales /standards; Sensitivity and Specificity

AccessionNumber
12007001034

Date bibliographic record published
30/04/2008

Date abstract record published
16/05/2008

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.