Do ultra-short screening instruments accurately detect depression in primary care?

A pooled analysis and meta-analysis of 22 studies

Alex J Mitchell and James C Coyne

ABSTRACT

Background
Guidance from the National Institute for Health and Clinical Excellence recommends one or two questions as a possible screening method for depression. Ultra-short (one-, two-, three- or four-item) tests have appeal due to their simple administration but their accuracy has not been established.

Aim
To determine whether ultra-short screening instruments accurately detect depression in primary care.

Design of study
Pooled analysis and meta analysis.

Method
A literature search revealed 75 possible studies and from these, 22 STARD-compliant studies (Standards for Reporting of Diagnostic Accuracy) involving ultra-short tests were entered in the analysis.

Results
Meta-analysis revealed a performance accuracy better than chance ($P<0.001$). More usefully for clinicians, pooled analysis of single-question tests revealed an overall sensitivity of 32.0% and specificity of 97.0% (positive predictive value [PPV] was 55.6% and negative predictive value [NPV] was 92.3%). For two- and three-item tests, overall sensitivity on pooled analysis was 73.7% and specificity was 74.7% with a PPV of only 38.3% but a pooled NPV of 93.0%. The Youden index for single-item and multiple item tests was 0.289 and 0.47 respectively, suggesting superiority of multiple item tests. Re-analysis examining only ‘either or’ strategies improved the ‘rule in’ ability of two- and three-question tests (sensitivity 79.4% and NPV 94.7%) but at the expense of being able to rule out a possible diagnosis if the result was negative.

Conclusion
A one-question test identifies only three out of every 10 patients with depression in primary care, thus unacceptable if relied on alone. Ultra-short two- or three-question tests perform better, identifying eight out of 10 cases. This is at the expense of a high false-positive rate (only four out of 10 cases with a positive score are actually depressed). Ultra-short tests appear to be, at best, a method for ruling out a diagnosis and should only be used when there are sufficient resources for second-stage assessment of those who screen positive.

Keywords
depression; diagnostic techniques and procedures; meta-analysis; screening; sensitivity and specificity.

INTRODUCTION

Approximately 7% of consultations in primary care are for depressive disorder. Depression is the third most common reason for consultation.$^{1,2}$ In one large survey, 90% of GPs said that patients with depression require a lot more time than patients with other disorders.$^3$

Although major depression has received most attention, milder forms of depression, including symptoms of depression insufficient to warrant a syndromal diagnosis, are at least as common and also linked with poor quality of life.$^4$ Numerous publications draw attention to the low detection rates of depression in primary care. Even motivated clinicians typically achieve a true positive case recognition rate (sensitivity of clinical detection alone) of between 36 and 56%.$^5–8$ Clinicians are better at ruling out non-depressed cases by achieving a true negative non-case specificity approaching 90%.$^7$ Barriers to correct detection are related to patients and clinicians.$^2$ Patients frequently do not recognise their own illness as depression and they may not disclose psychosocial problems to an unfamiliar practitioner.$^9$ Studies suggest that patients present with somatic (physical complaints) in as many as 70–80% of cases.$^{10–12}$ In addition, many patients prefer a medical to a psychiatric explanation.$^{14,15}$

Doctors have to consider many possible diagnoses during short appointments, averaging 8–20 minutes,
and maintain high productivity expectations. GPs may have a low index of suspicion for depression, particularly if patients with depression do not mention certain key psychological ‘sign-post’ symptoms. Other predictors of non-recognition include less severe, non-recurrent depression, and relatively low contact with patients.

One possible solution, endorsed in recent UK and US national guidelines, is use of a suitable screening instrument. This raises two important questions. Firstly, do screening tests for depression work accurately and, secondly, is the screening tool practical in primary care? A number of standardised diagnostic instruments with robust psychometric properties have been developed and validated in primary care. Data from 18 studies of nine different instruments revealed an overall sensitivity of 84% and specificity of 72%. However, these questionnaires typically take more than 5 minutes to complete.

To improve acceptability, a number of tools have been developed with less than 15 items and a completion time of less than 5 minutes. Examples include the 5-item World Health Organisation (WHO) Well-Being Index Questionnaire (WHO-5) and the 9 item Patient Health Questionnaire (PHQ). On testing, the positive predictive value of these instruments appears to be modest and the status of the instruments is uncertain. In clinical practice even these short questionnaires are not routinely used in primary or secondary care. This has led to the development of ultra-short questionnaires consisting of three-, two-, or even a single-detection question. Perhaps the most well known example is the PHQ-2.

The National Institute for Health and Clinical Excellence (NICE) has released guidelines for the management of unipolar depression in primary and secondary care. This included the recommendation of screening for at-risk groups and suggests that two simple screening questions will suffice. These are the PHQ-2 questions, namely: ‘During the last month, have you often been bothered by feeling down, depressed or hopeless?’; and, ‘During the last month, have you often been bothered by having little interest or pleasure in doing things?’. No specific evidence was cited by NICE; therefore, the study aims were to examine the diagnostic validity of these two questions and others that have been used to screen for depression.

METHOD

Definitions
See Box 1 for definitions of screening tools by length.

Search
A systematic literature search, critical appraisal of the collected studies, and a meta and pooled analysis were conducted.

How this fits in

The National Institute for Health and Clinical Excellence (NICE) recommends use of one- and two-item screening instruments for depression, but the validity of such brief methods has not been established. One-item tests miss over half (70%) of patients with depression, which is an unacceptable proportion. Two-item tests perform considerably better, but with a high false positive rate. One- and two-item tests can be used as a rule-out method but clinicians relying on ultra-short screening instruments must follow up those who initially screen positive with a more accurate case-finding method.

The following abstract databases were searched. Medline 1966–June 2006, PsycINFO 1887–June 2006, EMBASE 1980–June 2006, and CINAHL 1982–June 2006. In these databases the following keywords were searched (MeSH terms): ‘depress$ or mood’ and ‘screen or detect or diagnose or recognise’ and ‘short or brief or 1 item or single item or single question or two item or two question or three item or three question or patient health questionnaire’. A number of full text collections including Science Direct, Ingenta Select, Ovid Full text, and Wiley Interscience were searched. In these online databases the same search terms were used but as a full text search and citation search. The abstract database Web of Knowledge (version 3.0, ISI) was searched, using the above terms as a text word search, and using key papers in a reverse citation search.

Critical appraisal

Previously outlined review guidelines for diagnostic tests were followed and the primary studies were examined. In summary, data were extracted from the full text copy of the reports for review against STARD (Standards for Reporting of Diagnostic Accuracy) criteria. In addition the Newcastle-Ottawa Scale criteria for assessing the quality of non-randomised studies in meta-analyses were used. Questions for each report included the setting, the data integrity, the choice of reference criterion, the drop-out rate, the method of application of the screening questionnaire, and the type of depression examined.

Pooled and meta-analysis

In examining studies of ultra-short tests, a number of methodological issues can be anticipated. Detection

Box 1. Definitions of screening tools by length.

- Ultra-short screening tools were defined as those with 1–4 items, taking less than 2 minutes to complete.
- Short screening tools were defined as those with 5–14 items, taking between 2 and 5 minutes to complete.
- Standard screening tools were defined as those with 15 or more items, taking more than 5 minutes to complete.
strategies based on only two questions may require answers to one or both questions to be affirmative to ‘rule in’ depression. Similarly an answer to one or neither question may rule out depression. In effect, even two simple questions can be used with a categorical cut-off in three variations (Yes and No; Yes and Yes; No and No). The performance of a test will vary with the baseline prevalence of the condition. A further methodological issue is the description of depression using a criterion (gold) standard. Depression can be defined as any DSM-IV/ICD 

A proposal for reporting standards of meta-analyses of diagnostic studies has been published. The meta-analysis calculated the proportion of true cases (true positives plus true negatives) to the proportion of false cases (false positives plus false negatives) based on raw data from primary studies. Thus a ratio of 1 is.

Figure 1. Data trail of studies in systematic literature search.

*Publication reported two variations or test reported separately in the analysis.
equivalent to a chance detection. In addition to calculating overall meta-analytic effect size, tests for ‘non-combinability’ of studies (heterogeneity) and bias were performed. Statsdirect (version 2.2.6, 2006) was used for all analysis.

Where a meta-analysis reveals the relative risk of correct versus incorrect identification, a more clinically useful analysis is gained by pooled examination of the primary data. In the pooled analysis the raw numbers from each study reveal overall accuracy for each test and can be divided into sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The summary Youden’s index (sensitivity + specificity − 1) can also be calculated.

RESULTS

Systematic literature search

The search identified 33 papers of interest from over 75 possible ‘hits’ (Figure 1). Data included publications in non-peer reviewed sources, such as conference posters or abstracts. Eight studies of ultra-short screening tests in medical patients, or studies exclusively in secondary care or nursing home settings were excluded. These included studies in patients with back pain, multiple sclerosis, stroke, cancer, as well as medical inpatients. Studies of visual analogue scales were not included (although none was based in primary care). Several studies of short but not ultra-short tests were found and excluded. After excluding review articles and editorials, 22 individual analyses of ultra-short diagnostic tests reported in 12 unique publications were identified.

Critical appraisal

Results are presented in Supplementary Table 1. Four studies were non-STARD compliant for reasons of incomplete data or inadequate sample size. Several reports were not entirely derived from typical primary care settings. Whooley et al examined diagnostic accuracy in an urgent care veterans’ clinic. Lowe et al recruited a mixed sample of primary care and medical outpatients which was impossible to separate post-hoc. In addition, Osborn et al examined a cohort aged over 75 years in primary care.

Pooled analysis

Single-question tests. Eight analyses from six publications examined single-question tests for the diagnosis of depression in primary care. In total these studies examined 17624 participants of whom 1881 were depressed using the criterion standard; baseline prevalence was 10.7% (range = 5.0 to 36.0% [Supplementary Table 2]). From the pooled analysis 601 of 1881 cases of depression were correctly identified, giving an overall sensitivity of 31.9%. Of 15743 non-depressed cases, 479 were wrongly identified as depressed, giving an overall specificity of 96.0%. When accuracy was considered by proportion of positive or negative answers, then the overall PPV was 55.6% and overall NPV was 92.3%; therefore, the Youden index was 0.289. In one study, the PHQ question 1 alone appeared to have superior sensitivity and NPV but in a second study this was not confirmed. However, in both of these studies, the PHQ question 2 alone had superior sensitivity and NPV, suggesting question 2 may be worth further study.

Two- or three-question tests. Fourteen analyses from nine publications examined two- or three-question tests for the diagnosis of depression in primary care. In total, these studies examined 9653 participants of whom 1700 were depressed using the criterion standard; the baseline prevalence was 17.6%.

From the pooled analysis 1253 of 1700 cases of depression were correctly identified by two- or three-question tests, giving an overall sensitivity of 73.7% which is significantly better than single-question sensitivity. Of 7953 non-depressed cases, 2019 were wrongly identified as depressed, giving an overall specificity of 77.7% which was significantly worse than single-question sensitivity of 87.0%. Further, the overall PPV was 38.3% and overall NPV was 93.0%; therefore, the Youden index was 0.47, higher than the single-question performance. On further analysis, Arroll et al compared two compulsory questions (‘AND’ strategy) with positive responses on one of two questions (‘either or’ strategy). They found that requiring positive answers to both questions produced high PPV and specificity at the expense of NPV and sensitivity. That is, the ‘AND’ strategy works well as ‘rule in test’, but a negative answer cannot exclude a significant number of false negatives. More recently, Arroll et al examined whether the addition of a third item (‘the help question’) would enhance performance. Results suggest a modest enhancement of PPV performance.

Meta-analysis

The meta-analysis demonstrated that ultra-short strategies had a highly significant ability to identify depression/no depression in primary care compared with chance (Figure 2). The overall estimate of effect (Mantel–Haenszel, Rothman–Boice pooled relative risk) was 5.46 (95% confidence interval = 5.30 to 5.62; P<0.001). The test for ‘non-combinability’ for relative risk (Q) was 4529 (degrees of freedom = 11) P<0.001. Bias plot (Figure 3) and the Begg–Mazumdar bias statistic did not indicate conclusive publication bias (Kendall’s τ = 0.23 P=0.14).
ULTRA-SHORT SCREENING TESTS


Relative risk (95% CI)

2.1 (1.8 to 2.4)
2.8 (2.4 to 3.2)
6.6 (5.0 to 8.8)
9.3 (6.7 to 13.2)
12.2 (11.5 to 13.0)
3.9 (3.5 to 4.4)
13.9 (10.9 to 17.6)
3.4 (2.5 to 4.9)
3.6 (3.0 to 4.3)
1.8 (1.6 to 2.0)
5.1 (4.4 to 5.8)
2.2 (1.9 to 2.6)
3.6 (3.0 to 4.3)
0.2 (0.2 to 0.2)
0.5 (0.4 to 0.5)
9.3 (7.8 to 11.0)
10.2 (8.5 to 12.3)
1.8 (1.6 to 2.1)
2.7 (2.3 to 3.1)
3.8 (3.2 to 4.6)
3.8 (3.4 to 4.4)
8.8 (7.3 to 10.6)
5.5 (5.3 to 5.6)

DISCUSSION

Summary of main findings

The pooled analysis reveals a low overall sensitivity of 32.0% for a single-question strategy but 73.7% for two- and three-item tests. The specificity is 97% for single items and 74.7% for two items. PPVs were 55.6% (single-item tests) and 38.5% (two- and three-item tests; combines studies using ‘AND’ plus ‘either or’ strategies). NPVs were 92.3% (single-item tests) and 93.0% (two- and three-item tests). Only one study reported the ‘AND’ strategy alone with the ‘AND’ strategies using two-item tests having a low NPV but high PPV.¹ With this study removed from the pooled analysis, the overall sensitivity improves to 1225/1543 (79.4%) and the NPV also improved to 94.7%.

Thus, one-question tests identify only three out of every 10 patients with depression in primary care, so seven out of 10 cases would go unrecognised (these would remain lost even if a two-stage screen were applied). This performance is not acceptable. Ultra-short two- or three-question tests have better accuracy, identifying eight out of 10 depressed cases (two going unrecognised compared with a full interview).

However, this acceptable level of sensitivity is accompanied by a number of false-positive cases who could have been inappropriately referred or treated if these questionnaires were relied on alone. Moreover, even when a diagnosis of depression has been ruled out, additional time may be required for resolving the symptoms that have been uncovered.¹ Pooled PPV
results for two- and three-item tests show that four out of 10 participants who score positive are actually depressed and six out of 10 are false positives. This is significantly greater than the 1:4 false-positive rate typically generated by GPs when unassisted. Given the recent concern about over-treatment, this is also unlikely to be acceptable. Thus, to make a diagnosis a clinician would be required to use a second stage method (such as a standard diagnostic tool) in patients who screen positively on first pass.

It remains uncertain whether GPs have the time or inclination to use a multi-step algorithm approach. There is also a danger that competent physicians could abandon clinical diagnostic criteria and simply rely on screening scores in the midst of a formal implementation of screening. Where these ultra-short questionnaires appear to perform best is in ruling out a diagnosis. One-, two-, and three-item methods essentially perform well (NPV >90%) at excluding a diagnosis if the initial result is negative. By using an ultra-short method, only one in 10 patients who answer negatively will have a hidden diagnosis of depression.

**Strengths and limitations of the study**

This is the first study to examine systematically the merits of ultra-short diagnostic methods for depression in primary care. Its conclusions are based on a comprehensive literature search and meta-analysis of a very large pooled sample. Limitations to this study nonetheless need to be considered. Data have been collected from individual studies, in different settings where the prevalence varies sevenfold, between 5% and 37%. In eight out of 22 comparisons, the Composite International Diagnostic Interview (CIDI) was used as the criterion standard.

The CIDI was developed for use by non-clinically qualified interviewers in large epidemiological surveys. It has been found to have poor sensitivity when compared with clinical assessments of depression. A high proportion of patients with depression have mild disorders that do not reach the cut-off number of symptoms or the clinical significance criteria set in this meta-analysis. Further work is needed to examine pooled diagnostic accuracy in mild cases. Finally, physical illness is often present, particularly in older patients with depression. Effects of physical co-morbidity have not formally been studied here.

**Comparison with existing literature**

Two systematic reviews reached conflicting results about the value of routine screening using longer instruments. After pooling data, the US preventive task force recommended screening. However, this result was dependent on inclusion of a single large positive study in which substantial clinical resources were introduced along with screening. Using meta-analysis, Gillbody et al did not recommend routine screening but their data illustrated that feedback of high scoring patients was effective in increasing the rate of recognition of depression. In a recent large scale randomised trial incorporating screening score feedback, detection and follow-up rates improved, at least for those who had baseline low rates of recognition.

**Implications for clinical practice**

In clinical practice the use of a very simple ‘rule out’ measure will have appeal. The question regarding to what degree performance is different from routine clinical abilities, particularly of GPs who perform better than chance level, remains unanswered. Only one group has attempted to compare the result of ultra-short questionnaires with GP diagnosis alone. Arroll et al reported that GPs’ ability to eliminate depression was comparable to questionnaire methods alone; however, this study appeared to be contaminated by allowing GPs to see questionnaire data. Without help, Whooley et al found that GPs recognised only 8.8% of depression, but this exceptionally low rate may be due to the fact that the study was conducted at an urban, urgent care veterans’ clinic.

In the large MaGPIe survey (part of the Mental Health and General Practice Investigation study) from New Zealand, the overall GP detection rate was 56.4% in a sample of 775 primary care attenders. In those diagnosed as depressed by three independent instruments, GP recognition rate was 85.1% and in those patients who were CIDI positive it was 70.3%. The current authors’ suggest that future studies of screening tests should be measured against clinicians’ unassisted ability to detect depression; this would help to determine the added value of the instrument beyond usual care. An important unanswered question is: how do ultra-short methods compare with short and long case-finding methods when used in the same context?
population? Provisional results from Henkel and colleagues suggest that the PHQ-9, General Health Questionnaire-12, and WHO-5 may be only modestly superior to ultra-short tests.29

In the wider context of effective treatment of depression, screening is not enough on its own. It could be considered a first step to improving outcomes.68 Further steps include feedback of outlying scores, an agreed action plan for positive results, and a comprehensive treatment plan, including follow-up.70,71

It is important to acknowledge that a positive screen does not equate to the need for antidepressants, and that most patients prefer alternative options if available.72 In conclusion, ultra-short screening tests may have practical appeal for busy GPs but perform adequately only for ruling out a diagnosis. In settings where ultra-short questionnaires are being considered, a longer follow-up case-finding method, effective interpretation of results and effective treatment options must also be established.73,74

Supplementary information
Additional information accompanies this article at http://www.rcgp.org.uk/bjgp_suppinfo

Funding body
Not applicable

Ethics committee
Not applicable

Competing interests
The authors have stated that there are none.

REFERENCES

British Journal of General Practice, February 2007

Original Papers