# 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.[9] Patients frequently do not recognise their own illness as depression and they may not disclose psychosocial problems to an unfamiliar practitioner.[10] Studies suggest that patients present with somatic (physical complaints) in as many as 70–80% of cases.[11–13] 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,


***AJ Mitchell**, BMedSci, MSc, MRCPsych, consultant in liaison psychiatry, Leicester General Hospital, Leicester; **JC Coyne**, PhD, co-leader cancer control and outcomes program, Abramson Cancer Center; professor of psychology, Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, PA, US.*

**Address for correspondence**
Alex Mitchell, Department of Liaison Psychology
Brandon Unit, Leicester General Hospital
Leicester, LE5 4PW. E-mail: alex.mitchell@leicspart.nhs.uk


**Submitted:** 16 December 2005; **Editor's response:** 6 April 2006; **final acceptance:** 27 June 2006.

and maintain high productivity expectations.[16,17] GPs may have a low index of suspicion for depression, particularly if patients with depression do not mention certain key psychological 'sign-post' symptoms.[3,18,19] Other predictors of non-recognition include less severe, non-recurrent depression,[3,20–22] and relatively low contact with patients.[23,24]

One possible solution, endorsed in recent UK and US national guidelines, is use of a suitable screening instrument.[25,26] 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.[27] Data from 18 studies of nine different instruments revealed an overall sensitivity of 84% and specificity of 72%.[28] 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.[29] In clinical practice even these short questionnaires are not routinely used in primary or secondary care.[30] 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.[26] 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.

## 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.

## 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.

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[31] 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.[32] 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.[33] A further methodological issue is the description of depression using a criterion (gold) standard. Depression can be defined as any DSM-IV/ICD[10] depression or only major depression. The definition will also affect the baseline prevalence, which is critical when considering real-world accuracy performance and also when attempting to compare different studies. Where several types of depression was studied, the validity in major depression was examined (see Supplementary Table 1).

A proposal for reporting standards of meta-analyses of diagnostic studies has been published.[34] 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

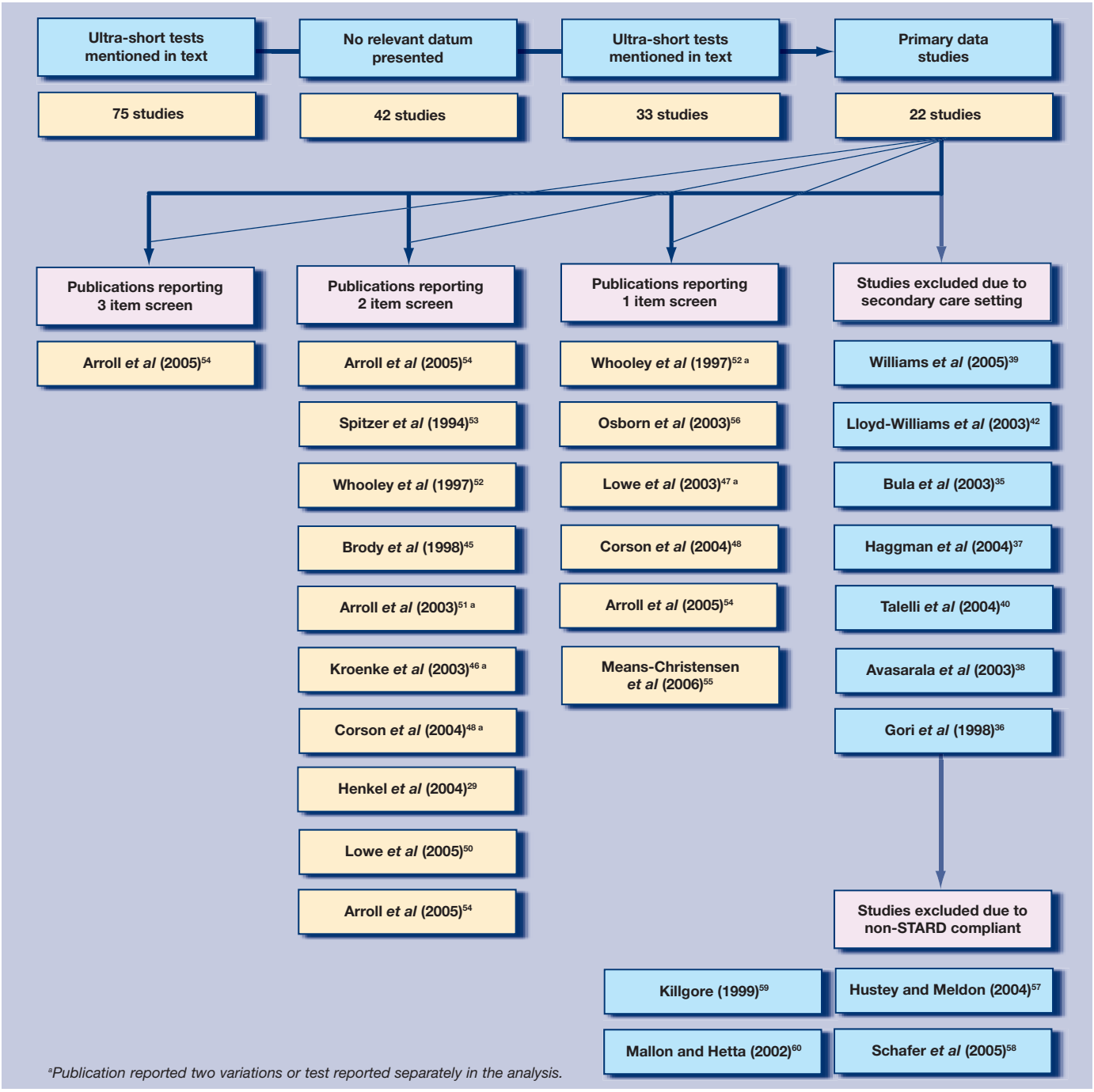


*Figure 1. Data trail of studies in systematic literature search.*

[a]*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 Youde`n 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.[35,36] These included studies in patients with back pain,[37] multiple sclerosis,[38] stroke,[39–41] cancer,[42] as well as medical inpatients. Studies of visual analogue scales were not included (although none was based in primary care).[43,44] 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.[45–56]

### Critical appraisal

Results are presented in Supplementary Table 1. Four studies were non-STARD compliant for reasons of incomplete data or inadequate sample size.[57–60] Several reports were not entirely derived from typical primary care settings. Whooley *et al* examined diagnostic accuracy in an urgent care veterans' clinic.[52] Lowe *et al* recruited a mixed sample of primary care and medical outpatients which was impossible to separate post-hoc.[47] In addition, Osborn *et al* examined a cohort aged over 75 years in primary care.[56]

### 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 17 624 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 15 743 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[52] but in a second study this was not confirmed.[47] 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, 2015 were wrongly identified as depressed, giving an overall specificity of 74.7% 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*[51] 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.[54] 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 = 21) $P<0.001$. Bias plot (Figure 3) and the Begg–Mazumdar bias statistic did not indicate conclusive publication bias (Kendall's $\tau = 0.23$ $P \leq 0.14$).

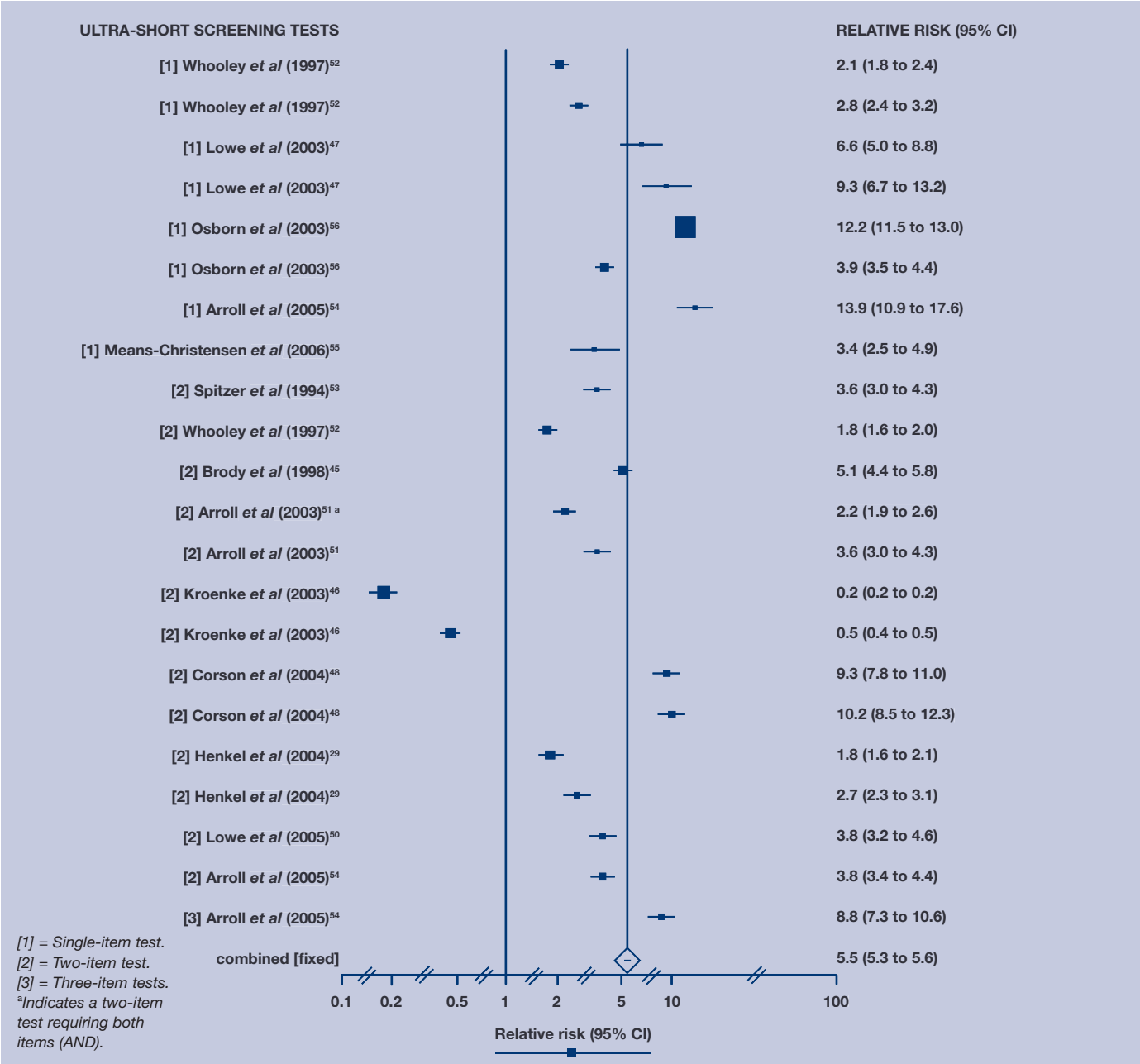


***Figure 2. Meta-analysis of ultra-short screening tests for depression in primary care.***

## 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-items tests). Only one study reported the 'AND' strategy alone with the 'AND' strategies using two-item tests having a low NPV but high PPV.[51] 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.[3] 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.[61] 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.[62,63] 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%[54] and 37%.[51] 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.[64] 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 supported screening.[25] 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.[65] 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.[66]

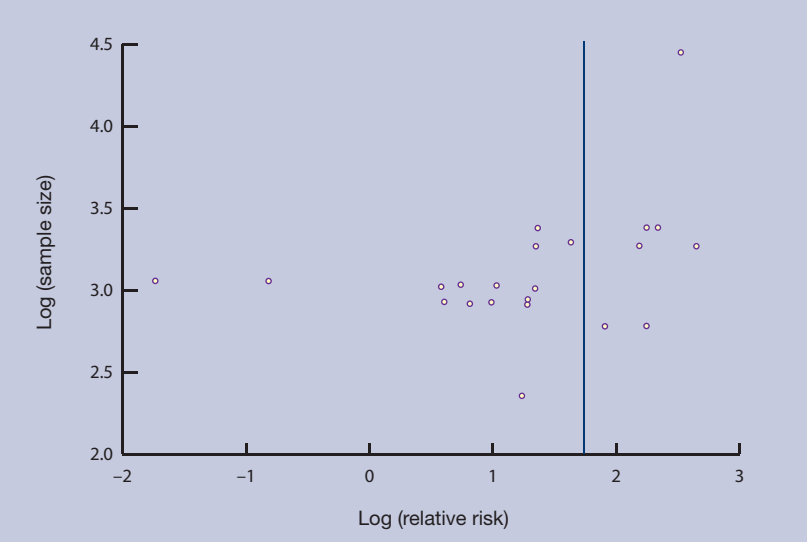


*Figure 3. Bias assessment plot of 22 ultra-short screening studies.*

### 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*[54] 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,[52] 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.[8] 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%.[67] 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.[7,68] An important unanswered question is: how do ultra-short methods compare with short and long case-finding methods when used in the same

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.[69] 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

1. Katon W, Schulberg HC. Epidemiology of depression in primary care. *Gen Hosp Psychiatry* 1992; **14(4):** 237–247.
2. Shah A. The burden of psychiatric disorder in primary care. *Int Rev Psychiatry* 1992; **4:** 243–250.
3. Wittchen H-U, Pittrow D. Prevalence, recognition and management of depression in primary care in Germany: the Depression 2000 study. *Hum Psychopharmacol Clin Exp* 2002; **17:** S1–S11.
4. Herrman H, Patrick DL, Diehr P, *et al.* Longitudinal investigation of depression outcomes in primary care in six countries: the LIDO Study. Functional status, health service use and treatment of people with depressive symptoms. *Psychol Med* 2002; **32(5):** 889–902.
5. Croudace T, Evans J, Harrison G *et al.* Impact of the ICD-10 Primary Health Care (PHC) diagnostic and management guidelines for mental disorders on detection and outcome in primary care. Cluster randomised controlled trial. *Br J Psychiatry* 2003; **182:** 20–30.
6. Thompson C, Kinmonth J, Stevens L, *et al.* Effects of a clinical-practice guideline and practice-based education on detection and outcome of depression in primary care: Hampshire Depression Project randomised controlled trial. *Lancet* 2000; **355(9199):** 185–191.
7. Christensen KS, Toft T, Frostholm, *et al.* The FIP Study: a randomised, controlled trial of screening and recognition of psychiatric disorders. *Br J Gen Pract* 2003; **53:** 758–763.
8. The Mental Health and General Practice Investigation (MaGPIe) Research Group. General practitioner recognition of mental illness in the absence of a 'gold standard.' *Aus N Z J Psychiatry* 2004; **38(10):** 789–794.
9. Tylee A. Depression in the community: physician and patient perspective. *J Clin Psychiatry* 1999; **60 (Suppl 7):** 12–6.
10. Robinson JW, Roter DL. Psychosocial problem disclosure by primary care patients. *Soc Sci Med* 1999; **48(10):** 1353–1362.
11. Yates WR, Mitchell J, Rush AJ, *et al.* Clinical features of depressed outpatients with and without co-occurring general medical conditions in STARD. *Gen Hosp Psychiatry* 2004; **26(6):** 421–429.
12. Kirmayer LJ, Robbins M, Dworkind M, *et al.* Somatization and the recognition of depression and anxiety in primary care. *Am J Psychiatry* 1993; **150(5):** 734–41.
13. Aragones E, Labad A, Pinol JL, *et al.* Somatized depression in primary care attenders. *J Psychosom Res* 2005; **58(2):** 145–151.
14. Kirmayer L, Robbins J. Patients who somatize in primary care: a longitudinal study of cognitive and social characteristics. *Psychol Med* 1996; **26(5):** 937–951.
15. Goldman LS, Nielsen NH, Champion HC. Awareness, diagnosis, and treatment of depression. *J Gen Intern Med* 1999; **14(9):** 569–580.
16. Gottschalk A, Flocke SA. Time spent in face-to-face patient care and work outside the examination room. *Ann Fam Med* 2005; **3(6):** 488–493.
17. Baik SY, Bowers BJ, Oakley LD, Susman JL. The recognition of depression: the primary care clinician's perspective. *Ann Fam Med* 2005; **3(1):** 31–37.
18. Greer J, Halgin R, Harvey E. Global versus specific symptom attributions: predicting the recognition and treatment of psychological distress in primary care. *J Psychosom Res* 2004; **57(6):** 521–527.
19. Aragones E, Pinol JL, Labad A, *et al.* Detection and management of depressive disorders in primary care in Spain. *Int J Psychiatry Med* 2004; **34(4):** 331–343.
20. Coyne JC, Schwenk TL, Fechner-Bates S. Nondetection of depression by primary care physicians reconsidered. *Gen Hosp Psychiatry* 1995; **17(1):** 3–12.
21. Tiemens BG, Ormel J, Simon GE. Occurrence, recognition and outcome of psychological disorders in primary care. *Am J Psychiatry* 1996; **153:** 636–644.
22. Aragonès E, Piñol JL, Labad A, *et al.* Detection and management of depressive disorders in primary care in Spain. *Int J Psychiatry Med* 2004; **34:** 331–343.
23. Nuyen J, Volkers AC, Verhaak PFM, *et al.* Accuracy of diagnosing depression in primary care: the impact of chronic somatic and psychiatric co-morbidity. *Psychol Med* 2005; **35:** 1185–1195.
24. Verhaak PF, Schellevis FG, Nuijen J, Volkers AC. Patients with a psychiatric disorder in general practice: determinants of general practitioners' psychological diagnosis. *Gen Hosp Psychiatry* 2006; **2:** 125–132.
25. Pignone MP, Gaynes BN, Rushton JL, *et al.* Screening for depression in adults: a summary of the evidence for the U.S. preventive services task force. *Ann Intern Med* 2002; **136(10):** 765–776.
26. National Institute for Health and Clinical Excellence. *Depression: management of depression in primary and secondary care.* Clinical Guideline 23. London: NICE, 2004.
27. Williams JW, Noel PH, Cordes JA, *et al.* Is this patient clinically depressed? *JAMA* 2002; **287:** 1160–1170.
28. Mulrow CD, Williams JWJ, Gerety MB, *et al.* Case finding instruments for depression in primary care settings. *Ann Intern Med* 1995; **122:** 913–921.
29. Henkel V, Mergl R, Kohnen R, *et al.* Use of brief depression screening tools in primary care: consideration of heterogeneity in performance in different patient groups. *Gen Hosp Psychiatry* 2004; **26:** 190–198.
30. Gilbody SM, House AO, Sheldon TA. Psychiatrists in the UK do not use outcomes measures: national survey. *Br J Psychiatry* 2002; **180(2):** 101–103.
31. Pai M, McCulloch M, Enanoria W, Colford JM. Systematic reviews of diagnostic test evaluations: what's behind the scenes? *Evid Based Med* 2004; **9:** 101–103.
32. Wells GA, Shea B, O'Connell D, *et al.* The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm (accessed 13 Dec 2006).
33. Whiting P, Rutjes AWS, Dinnes J, *et al.* Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess* 2004; **8(25):** iii, 1–234.
34. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001; **323(7305):** 157–162.
35. Bula CJ, Wietlisbach v, Yersin B, Burnand B. Does a single item question identify elderly medical inpatients who report significant depressive symptoms? *Age Aging* 2003 Mar; **32(2):** 231–233.
36. Gori C, Appollonio I, Riva GP, *et al.* Using a single question to screen for depression in the nursing home. *Arch Gerontol Geriatr* 1998; **26(Suppl 1):** 235–240.
37. Haggman S, Maher CG, Refshauge KM. Screening for symptoms of depression by physical therapists managing low back pain. *Phys Ther* 2004; **84(12):** 1157–1166.
38. Avasarala JR, Cross AH, Trinkaus K. Comparative assessment of Yale Single Question and Beck Depression Inventory Scale in screening for depression in multiple sclerosis. *Mult Scler* 2003; **9(3):** 307–310.
39. Williams LS, Brizendine EJ, Plue L, *et al.* Performance of the PHQ-9 as a

screening tool for depression after stroke. *Stroke* 2005; **36(3):** 635–638.

40. Talelli P, Lekka NP, Katsoulas G, *et al*. The verbally asked single Yale question compared with its written form as screening tool for post-stroke depression. *J Neurol* 2004; **251(Suppl 3):** III/191.
41. Watkins C, Daniels L, Jack C, *et al*. Accuracy of a single question in screening for depression in a cohort of patients after stroke: comparative study. *BMJ* 2001; **323(7322):** 1159.
42. Lloyd-Williams M, Dennis M, Taylor F, Baker I. Is asking patients in palliative care 'Are you depressed' appropriate? Prospective study. *BMJ* 2003; **327(7411):** 372–373.
43. Kertzman S, Aladjem Z, Milo R, *et al*. The utility of the visual analogue scale for the assessment of depressive mood in cognitively impaired patients. *Int J Geriatr Psychiatry* 2004; **19(8):** 789–796.
44. Di Benedetto M, Lindner H, Hare DL, *et al*. A Cardiac Depression Visual Analogue Scale for the brief and rapid assessment of depression following acute coronary syndromes. *J Psychosom Res* 2005; **59(4):** 223–229.
45. Brody DS, Hahn SR, Spitzer RL, *et al*. Identifying patients with depression in the primary care setting: a more efficient method. *Arch Intern Med* 1998; **158(22):** 2469–2475.
46. 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.
47. Lowe B, Grafe K, Kroenke K, *et al*. Predictors of psychiatric comorbidity in medical outpatients. *Psychosom Med* 2003; **65(5):** 764–770.
48. Corson K, Gerrity MS, Dobscha SK. Screening for depression and suicidality in a VA primary care setting: 2 items are better than 1 item. *Am J Manag Care* 2004; **10(11 Pt 2):** 839–845.
49. Henkel V, Mergl R, Coyne JC, *et al*. Screening for depression in primary care: Will one or two items suffice? *Eur Arch Psychiatry Clin Neurosci* 2004; **254(4):** 215–223.
50. Lowe B, Kroenke K, Kerstin Grafe. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *J Psychosom Res* 2005; **58(2):** 163–171.
51. Arroll B, Khin N, Kerse N. Screening for depression in primary care with two verbally asked questions: cross sectional study. *BMJ* 2003; **327(7424):** 1144–1146.
52. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression: two questions are as good as many. *J Gen Intern Med* 1997; **12(7):** 439–445.
53. Spitzer RL, Williams JB, Kroenke K, *et al*. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. *JAMA* 1994; **272(22):** 1749–1756.
54. Arroll B, Goodyear-Smith F, Kerse N, *et al*. 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.
55. Means-Christensen AJ, Sherbourne CD, Roy-Byrne PP, *et al*. Using five questions to screen for five common mental disorders in primary care: diagnostic accuracy of the Anxiety and Depression Detector. *Gen Hosp Psychiatry 2006*; **28(2):** 108–118.
56. Osborn DPJ, Fletcher AE, Smeeth L, *et al*. Performance of a single screening question for depression in a representative sample of 13 670 people aged 75 and over in the UK: results from the MRC trial of assessment and management of older people in the community. *Fam Pract* 2003; **20(6):** 682–684.
57. Hustey FM, Meldon SW. The use of a two-question depression screen for the detection of depression in older emergency department patients [abstract]. *Ann Emerg Med* 2004; **44(4):** 568.
58. Schafer A, Maddens M, Boyle V, Lichtenberg P. Utility of a single question depression screen in SNF-based subacute rehabilitation. *Gerontologist* 2005; **45(2):** 240-241.
59. Killgore WDS. The visual analogue mood scale: can a single-item scale accurately classify depressive mood state? *Psychological Reports* 1999; **85(3 Pt 2):** 1238–1243.
60. Mallon L, Hetta J. Detecting depression in questionnaire studies: comparison of a single question and interview data in a community sample of older adults. *Eur Psychiatry* 2002; **3**: 135–144.
61. Aragones E, Pinol JL , Labad A. The overdiagnosis of depression in non-depressed patients in primary care. *Fam Pract* 2006; **23(3):** 363–368.
62. Valenstein M, Dalack G, Blow F, *et al*. Screening for psychiatric illness with a combined screening and diagnostic instrument. *J Gen Intern Med* 1997; **12(11):** 679–685.
63. Bushnell J. Frequency of consultations and general practitioner recognition of psychological symptoms. *Br J Gen Pract* 2004; **54(508):** 838–843.
64. Brugha T, Jenkins R, Taub N, *et al*. A general population comparison of the composite diagnostic interview (CIDI) and the schedules for clinical assessment in neuropsychiatry (SCAN). *Psychol Med* 2001; **31(6):** 1001–1013.
65. Gilbody S, Whitty P, Grimshaw J, Thomas R. Improving the recognition and management of depression in primary care. *Effective Health Care Bulletin* 2002; **7(5)**. http://www.york.ac.uk/ inst/crd/ehc75.htm (accessed 2 Jan 2006).
66. Christensen KS, Toft T, Frostholm L, *et al*. Screening for common mental disorders: who will benefit? Results from a randomised clinical trial. *Fam Pract* 2005; **22(4):** 428–434.
67. The MaGPIe Research Group. The effectiveness of case-finding for mental health problems in primary care. *Br J Gen Pract* 2005; **55:** 665–669.
68. Coyne JC, Thompson R, Palmer SC, *et al*. Should we screen for depression? Caveats and potential pitfalls. *Appl Prev Psychol* 2000; **9:** 101–121.
69. Kroenke K. Depression screening is not enough. *Ann Intern Med*, 2001; **134(5):** 418–420.
70. Valenstein M, Vijan S, Zeber JE, *et al*. The cost-utility of screening for depression in primary care. *Ann Intern Med* 2001; **134(5):** 345–360.
71. Pirraglia PA, Rosen AB, Hermann RC, *et al*. Cost-utility analysis studies of depression management: a systematic review. *Am J Psychiatry* 2004; **161:** 2155–2162.
72. Mitchell AJ. Depressed patients and treatment adherence. *Lancet* 2006; **367(9528):** 2041–2043.
73. Richards JC, Ryan P, McCabe MP, *et al*. Barriers to the effective management of depression in general practice. *Aust N Z J Psychiatry* 2004; **38(10):** 795–803.
74. Wells KB, Miranda J, Bauer MS, *et al*. Overcoming barriers to reducing the burden of affective disorders. *Biol Psychiatry* 2002; **52(6):** 655–675.