BMJ Open

Diagnostic accuracy of the Whooley questions for the identification of depression: a diagnostic meta-analysis

Katharine Bosanquet,[1] Della Bailey,[1] Simon Gilbody,[1,2] Melissa Harden,[3] Laura Manea,[1,2] Sarah Nutbrown,[1] Dean McMillan[1,2]

To cite: Bosanquet K, Bailey D, Gilbody S, *et al*. Diagnostic accuracy of the Whooley questions for the identification of depression: a diagnostic meta-analysis. *BMJ Open* 2015;**5**:e008913. doi:10.1136/bmjopen-2015-008913

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-008913).

Received 27 May 2015
Accepted 9 October 2015

[1]Department of Health Sciences, University of York, York, UK
[2]Hull York Medical School, University of York, York, UK
[3]Centre for Reviews and Dissemination, University of York, York, UK

Correspondence to
Katharine Bosanquet; kate.bosanquet@york.ac.uk

ABSTRACT

Objectives: To determine the diagnostic accuracy of the Whooley questions in the identification of depression; and, to examine the effect of an additional 'help' question.

Design: Systematic review with random effects bivariate diagnostic meta-analysis. Search strategies included electronic databases, examination of reference lists, and forward citation searches.

Inclusion criteria: Studies were included that provided sufficient data to calculate the diagnostic accuracy of the Whooley questions against a gold standard diagnosis of major depression.

Data extraction: Descriptive information, methodological quality criteria, and 2×2 contingency tables were extracted.

Results: Ten studies met inclusion criteria. Pooled sensitivity was 0.95 (95% CI 0.88 to 0.97) and pooled specificity was 0.65 (95% CI 0.56 to 0.74). Heterogeneity was low (I^2=24.1%). Primary care subgroup analysis gave broadly similar results. Four of the ten studies provided information on the effect of an additional help question. The addition of this question did not consistently improve specificity while retaining high sensitivity as reported in the original validation study.

Conclusions: The two-item Whooley questions have high sensitivity and modest specificity in the detection of depression. The current evidence for the use of an additional help question is not consistent and there is, as yet, insufficient data to recommend its use for screening or case finding.

Trial registration number: CRD42014009695.

Strengths and limitations of this study

- An original study–the first diagnostic accuracy meta-analysis of the Whooley questions as a screening test for depression.
- Using rigorous methodology–strict inclusion/exclusion and quality assessment criteria–identified 10 studies of sufficient quality for inclusion.
- Substantial variability observed in methodological quality of included studies.
- Inconsistency in how Whooley questions are referred to means further relevant studies may have been missed.

INTRODUCTION

Depression is a highly prevalent condition that affects a substantial proportion of the population, varying from around 1 in 4 women to 1 in 10 men.[1 2] It leads to impairments in functioning that are as significant as those seen in chronic physical health conditions.[3] Although depression is a common condition, it is often hard to detect in primary care and other non-psychiatric settings. Despite the significance of the problem, there is remarkable uncertainty about the value of screening or case finding for depression. The guidance from different Western countries is contradictory,[4 5] and from a UK health perspective, recommendations offered by different UK bodies are also inconsistent.[6–10] The UK National Screening Committee[11] concluded that there is insufficient evidence to recommend the adoption of screening for depression and also identified a lack of robust evidence for case finding among populations at elevated risk. In contrast, the National Institute of Health and Care Excellence (NICE) guidance recommends that, in the UK, general practitioners (GPs) consider asking two brief questions to identify potential depression in certain patient groups[7–9] such as people with long-term conditions and women during the perinatal period; if someone responds positively to either question a more comprehensive assessment is carried out, to determine whether or not an individual is depressed.

NICE guidance recommends considering using the Whooley questions,[12] derived from the original Prime-MD,[13] to identify potential depression. The Whooley questions consist of two questions asking about low mood and loss of interest or pleasure. In the original

validation study, the questions had a sensitivity of 0.95 (0.89 to 0.98) and specificity of 0.56 (0.52 to 0.61). A subsequent validation study added a third question, which asks whether the person wants help with the difficulties identified.[14] Although NICE endorses the use of the Whooley questions, the guidance recognises that this is based on limited evidence of the diagnostic accuracy of the measure. Perhaps as a consequence of this, practitioners also have doubts about the ability of the questions to detect depression.[15] There is further uncertainty about whether the two or three-item version of the questions should be used, with some NICE guidance recommending the use of the third question,[9]—though recent policy changes have seen this removed[10]—while other guidance specifically chose not to adopt this additional question because of a lack of evidence on its effectiveness.[8]

The Whooley questions are at the centre of the UK's approach to the identification of depression, yet at the time the UK guidance was published there was limited evidence on the diagnostic performance of the test. It remains unclear whether a review of the current evidence base would lead to a revision of UK guidance. We conducted a systematic review, therefore, to identify all studies that had examined the diagnostic accuracy of the Whooley questions against a gold standard method of establishing a diagnosis of major depression according to internationally recognised criteria. A further component of the review was to assess the effect of the 'help' question in those studies that included it in the screen.

METHOD

A protocol for the systematic review was developed and published on PROSPERO (registration number: CRD42014009695 http://www.crd.york.ac.uk/PROSPERO/). We adhered to Centre for Reviews and Dissemination guidance in the conduct of the review and PRISMA guidelines in the reporting of the review.[16]

Data sources and searches

The following databases were searched to identify studies assessing the diagnostic test accuracy of the Whooley questions: MEDLINE, MEDLINE In-Process, PsycINFO, EMBASE, Cumulative Index to Nursing & Allied Health (CINAHL Plus), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Reviews of Effects (DARE), and the Health Technology Assessment (HTA) database. A number of additional sources were searched to identify studies in progress, unpublished research or grey literature: Conference Proceedings Citation Index—Science and Social Science, OAIster, ClinicalTrials.gov, Health Services Research Projects in Progress (HSRProj) and the Trip database.

Searches were conducted from 1994—the year the PRIME-MD was published from where the Whooley questions were derived—to September 2013. No language restrictions or study design filters were applied to the search strategy. In addition, a forward citation search of the Whooley 1997 paper was carried out in the Web of Science database to identify any further papers on the Whooley questions. We examined the reference lists and conducted a reverse-citation search of all included studies.

A search strategy, consisting of relevant free-text terms and subject headings, was developed in MEDLINE (OvidSP) and then adapted for use in the other databases searched. Online supplementary appendix 1 gives the full search strategy for MEDLINE. Furthermore, we contacted key experts in the field to obtain information about potential unpublished data and for clarification on aspects of their work, which consisted of six authors including Whooley *et al*,[12] Arroll and colleagues.[14 17]

An update of the searches was conducted in April 2015. No further diagnostic accuracy studies using the Whooley questions were found. However, we did observe changes to policy. NICE had amended guidance on perinatal depression (CG192).[10] It now recommends considering asking the Whooley questions alone rather than with the addition of a help question.

Study selection

Studies were selected using a prepiloted form based on the PICO inclusion criteria in the review protocol. Three reviewers assessed titles and abstracts to identify potentially eligible studies. Any queries were discussed with a second reviewer. Full text was obtained for all articles included after this initial screen. Each of these was assessed using the prepiloted form by two reviewers. At each stage any disagreements were resolved by consensus and where necessary arbitration by further reviewers.

Studies that met the following inclusion criteria were included: *Participants/population*; No restrictions were made in terms of the participants or population. *Instrument*: Studies that used either the two-item or three-item Whooley questions were included. The two-item questions had to use the standard Whooley wording, as outlined in the original article.[12]

1. "During the past month, have you often been bothered by feeling down, depressed, or hopeless?" (yes/no)
2. "During the past month, have you often been bothered by little interest or pleasure in doing things?" (yes/no)[12]

For translated versions, the wording had to be derived from the original. The questions also had to be scored as a dichotomous 'yes'/'no'. For the two-item Whooley questions, only studies that defined a positive screen as 'yes' to one or both of the questions were included. Given inconsistencies in the literature about the precise phrasing of the 'help question', all variations in phrasing were accepted. No restrictions were made in terms of mode of administration (eg, telephone or face-to-face) or the person administering the measure (eg, clinician,

researcher or self-administered). *Comparator (reference standard)*: Studies that use a gold standard diagnostic interview to establish a diagnosis of major depression according to international criteria (Diagnostic and Statistical Manual (DSM) or International Classification of Disease (ICD)) were eligible for inclusion. Studies were excluded if the target diagnosis was not solely major depression (eg, any depressive disorder). No restrictions were made in terms of who administered the gold standard or its mode of administration. *Outcome:* For a study to meet inclusion criteria, it had to report sufficient data to extract 2×2 contingency tables for either the two-item Whooley questions or the two-item questions plus an additional help question. *Study design:* No restrictions were made in the type of study design.

Data extraction and quality assessment

Two reviewers independently extracted the following data to a prepiloted standardised form: (1) descriptive characteristics of the sample and setting (country, setting, age of sample, gender of sample, sample size, proportion depressed); (2) descriptive characteristics of the Whooley (mode of administration, who administered, language); (3) descriptive characteristics of the gold standard (type of gold standard, whether DSM or ICD diagnoses); (4) quality assessment criteria (see below); and (5) the 2×2 contingency tables for the two-item Whooleys and/or two-item Whooleys plus help question against gold standard diagnosis of major depression. Any disagreements were resolved through consensus or, where necessary, arbitration by a third reviewer. Study authors were contacted to provide additional data or clarification as necessary.

Quality assessment was conducted at the study level and used criteria based on the QUADAS-II.[18] The QUADAS-II guidelines require that it is adapted for each specific review; this can involve adding or omitting questions and providing clarification about how specific questions are to be rated. We developed specific guidance on the coding of the questions in the form of a brief field guide.

We retained all of the risk of bias signalling questions and applicability questions, with the exception of one item (prespecified threshold on the index test). This item was removed because the standard method of scoring the Whooley provides a dichotomous cut-off; there is no ordinal or continuous scale that requires the prespecification of a threshold. For the signalling question 'Is the reference standard likely to correctly classify the target condition?' we operationalised this as whether the researchers who conducted the gold standard interview had received appropriate training. For the signalling question 'Was there an appropriate interval between the index test and reference standard?' we defined an appropriate interval as less than 2 weeks in keeping with how this item has been applied in previous diagnostic test accuracy studies of depression.[19]

We added two additional questions that were applied to studies using translated versions of the Whooley and reference test. For translations of the reference test, we asked whether appropriate forward and back translation methods were used and whether psychometric properties of the translated version were reported. Similarly, we asked whether appropriate translation methods were used and also applied to any translated version of the Whooley. We also added an additional question to establish whether the studies had used strategies to exclude people already known to a service to have depression. This reflects Thombs *et al*'s[20] concern that studies which include people already known to be depressed may provide an artificially inflated indication of a test's performance, because the typical aim of a screening or case finding tool is to identify depression in those not already known to be depressed. Studies met this criterion if they used strategies to exclude people already known to be depressed, such as excluding people already known to be using psychotropic medication.

Data synthesis and analysis

We constructed 2×2 contingency tables with true positive, true negative, false positive and false negative results. We performed a bivariate diagnostic meta-analysis to obtain pooled estimates of specificity, sensitivity, likelihood ratios, diagnostic ORs and their associated 95% CIs. The bivariate model is a 2-level model which takes into account the precision by which differences in sensitivity and specificity have been calculated while incorporating and estimating the amount of between-study variability in sensitivity and specificity.[21] A priori subgroup analyses were conducted on descriptive variables and quality assessment criteria.

Heterogeneity

We measured the between study heterogeneity using the I^2 statistic of the pooled diagnostic OR.[22] I^2 describes the percentage of total variation across studies, which is caused by heterogeneity rather than chance. The I^2 has a greater statistical power to detect clinical heterogeneity when fewer studies are available compared to other measures of heterogeneity. I^2 values of 25% may be considered low, 50% moderate and 75% high. We explored the causes of heterogeneity where there was significant between-study heterogeneity by visually inspecting the summary receiver operation characteristic curves and identifying the studies that were outside the 95% confidence ellipse. We also undertook a meta-regression analysis of logit diagnostic OR using a priori potential sources of heterogeneity entered as covariates in the meta-regression model.[23]

We investigated the heterogeneity resulting from sample or study design characteristics by exploring the effects of potential predictive variables.[24] For the sample we examined the effect of language (translated vs not translated), baseline prevalence of major depressive disorder in the screened population, as a proxy measure of

the spectrum of severity of disorder within the screened population, and study settings (primary care vs general hospital). For study quality, we considered blinding (of the assessor to the results of the Whooley questions as well as the gold standard) and whether the studies avoided a case–control design or an artificially inflated base rate of major depression. If these items were important sources of heterogeneity, then they would be predictive in a meta-regression analysis, and would reduce the level of between-study heterogeneity in the meta-regression model.

Analyses were conducted using STATA V.12, with the metandi, metabias, metareg and metafunnel user-written commands.

RESULTS

The initial search identified 6846 unique citations (10 589 citations before de-duplication). Twenty-two of these citations met initial inclusion criteria and were selected for further screening of the full article (figure 1). Ten of the 22 met final stage inclusion criteria. The reasons for exclusion of the 12 studies are as follows: three used the PHQ-2 not the Whooley,[25–27] for one study we were unable to establish whether the two-item questionnaire used was the Whooley,[28] four did not use a gold standard reference test,[13 29–31] two did not report data on a diagnosis of major depression alone (eg, outcome was any depression diagnosis)[32 33] and for two it was not possible to extract information to calculate a 2×2 contingency table.[34 35]

Overview of included studies

Table 1 summarises the characteristics of the included studies. The studies took place in a variety of countries and settings. The samples included adults and older adults and ranged from predominantly male[12] to entirely female samples.[36 37] Sample sizes ranged from 89[38] to over 1000[14 39] and the proportion depressed according to the gold standard ranged from 3.3%[38] to 34%.[40] Clinicians administered the Whooley questions in the majority of studies. The language of administration was English in six of the studies; translated versions were used in the remainder. A variety of gold standard measures were used, though the CIDI was used in 4 of the 10 studies.

Quality assessment

Table 2 summarises the results of the quality assessment using QUADAS-II. None of the studies was rated as at low risk of bias across all domains. A rating of an unclear risk of bias was the most common rating across the domains. All studies avoided the use of a case–control design. Only three clearly made attempts to exclude people with a known history of depression. Six of the 10 studies provided evidence of blinding in both directions (ie, Whooley interpreted blind to reference, reference interpreted blind to Whooley). In terms of the QUADAS-2 applicability criteria, all studies were rated as applicable on all three domains.

Diagnostic properties of the Whooley questions (no help question)

Ten studies reported the diagnostic properties of the Whooley questions. One study[41] reported a significantly lower sensitivity and higher specificity than other studies. In the remaining nine studies, the sensitivity ranged between and 0.90[39] and 1.00.[36–38 42] Specificity values ranged between 0.44[37 42] and 0.78.[14] Table 3 presents the individual performance of the 10 studies including sensitivity, specificity, likelihood ratios and diagnostic ORs and their corresponding 95% CIs.

The pooled sensitivity was 0.95 (CI 0.88 to 0.97), pooled specificity 0.65 (CI 0.56 to 0.74), pooled positive likelihood ratio 2.78 (CI 2.16 to3.57), pooled negative likelihood ratio 0.07 (CI 0.03 to 0.16) and diagnostic OR 36.91 (17.52 to 77.76). The level of between-study heterogeneity was low (I^2=24.1%). Figure 2 shows the Whooley questions summary receiver operating characteristic plot of major depression diagnosis. Figure 3 shows the posterior probabilities given positive and negative test results. The figure shows that, at the prevalence rate expected in the general population (less than 20%), the probability of a depressed person with a negative test result is very low; whereas the probability of a depressed person with a positive test result is around 40%.

We conducted a meta-regression to explore possible sources of heterogeneity. Descriptive variables and quality assessment criteria (setting, baseline prevalence of major depression, language, whether the study avoided a case–control design and blinding) were examined as predictors. Out of these variables, only the prevalence of major depression was significant (p=0.026).

Subgroup analyses

One of the possible reasons for heterogeneity is the various clinical settings in which the Whooley questions have been validated. On a priori grounds we conducted subgroup analyses to examine the diagnostic performance of the Whooley questions in similar clinical settings.

Five studies were conducted in primary care settings,[14 17 37 40 42] three studies recruited in hospital or out-patient-based medical settings[12 36 39] and two in community settings.[38 41] In primary care settings the Whooley questions had a pooled sensitivity of 0.96 (CI 0.91 to 0.98), pooled specificity 0.61 (CI 0.48 to 0.73), pooled positive likelihood ratio 2.53 (CI 1.80 to 3.56), pooled negative likelihood ratio 0.04 (CI 0.01 to 0.13) and diagnostic OR 52.07 (15.65 to 173.18). Heterogeneity in primary care studies was moderate I^2=49.9%.

We did not identify a sufficient number of studies (minimum of four studies for a diagnostic meta-analysis) using a comparable clinical setting to conduct further

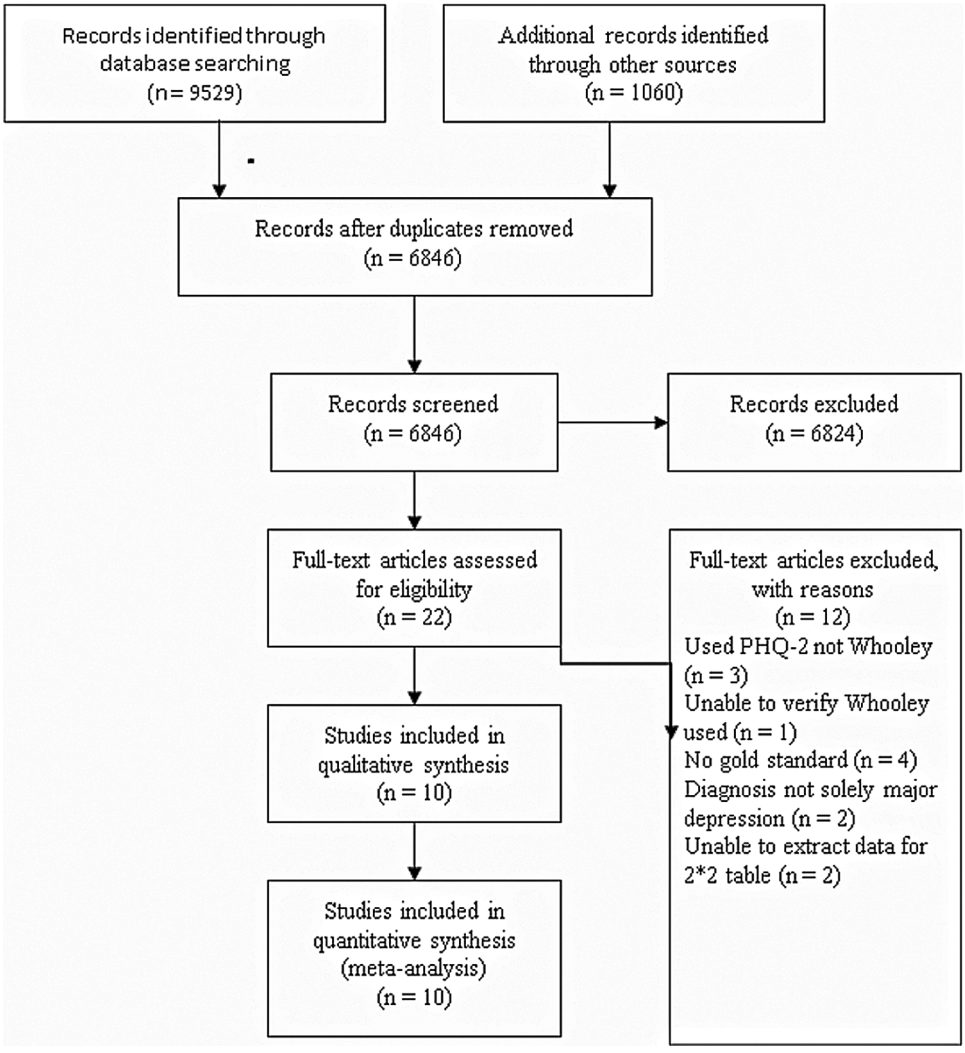

Figure 1 Overview of selection of studies (PRISMA).

subgroup analyses for other settings. There were not enough studies to pool the results separately for different age groups.

Six studies validated the original (English) version of the Whooley questions.[12 14 17 36 37 39] Pooled sensitivity for these studies was 0.95 (0.89 to 0.98), pooled specificity was 0.64 (0.54 to 0.72), positive likelihood ratio 2.67 (2.11 to 3.38), negative likelihood ratio 0.06 (0.02 to 0.15) and pooled diagnostic OR 40.64 (17.00 to 97.14). Heterogeneity in the English studies was low (7.3%).

Whooley questions and help question

Lack of consistency in the phrasing of the questions and how the data were combined meant that we were unable to combine results for a meta-analysis of the help question. Instead we described the results of the studies individually. Two studies[14 41] considered a positive screen as a positive response to either or both Whooley questions and yes to the help question (yes today; or yes, but not today). The psychometric properties of this method of scoring the Whooley questions were, as reported by Arroll *et al*[14]: sensitivity 0.95 (95% CI 0.85 to 0.99), specificity 0.89 (95% CI 0.87 to 0.91), positive likelihood ratio 9.06 (95% CI 7.41 to 11.10) negative likelihood ratio 0.04 (95% CI 0.01 to 0.18) and OR 190.00 95% (50.00—* value unable to be estimated). The psychometric properties reported by Suija *et al* showed a lower sensitivity of 0.68 (95% CI 0.46 to 0.85) but comparable specificity of 0.85 (0.82 to 0.88). Positive likelihood ratio was 4.77 (95% CI 3.36 to 6.78), negative likelihood ratio 0.37 (95% CI 0.21 to 0.66) and OR 12.80 (95% CI 5.40 to 30.20). Arroll *et al*[14] made the distinction between 'help, yes but not today' or 'yes, help today' though we were unable to extract 2×2 tables for these different responses to the help questions from the data presented in the paper.

The remaining two studies[36 42] reported the psychometric properties of the help question only in those who scored positive on either Whooley questions. Mann *et al* used the help question 'is this something you feel you need or want help with?' rather than the one proposed by Arroll *et al*[14]. Psychometric properties of a positive answer to either Whooley question and a positive answer to this question were as follows: sensitivity 0.66 (95% CI 0.38 to 0.88), specificity 0.91 (95% CI 0.78 to 0.98), positive likelihood ratio 8.22 (95% CI 2.62 to 25.80),

Table 1 Descriptive characteristics of the included studies

Study	Sample characteristics (Country, setting, age, sex)	Sample size and % depressed	Whooley characteristics	Diagnostic standard
Adachi *et al*[38]	Country: Japan Setting: community Age (years): M=38.4 (SD=6.6) Female: 9%	N=89 Depressed: 3.3	Administration: psychiatrists and clinical psychologists Language: Japanese	MINI
Arroll *et al*[17]	Country: New Zealand Setting: primary care Age (years): M=46 (range=16–90) Female: 70%	N=421 Depressed: 6	Administration: general practitioner Language: English	CIDI
Arroll *et al*[14]	Country: New Zealand Setting: primary care Age (years): not stated Female: % not stated	N=1025 Depressed: 5	Administration: not stated Language: English	CIDI
Gjerdingen *et al*[37]	Country: USA Setting: primary care Age (years): M=28.9 Female: 100%	N=506 Depressed: 4.6	Administration: doctoral-level psychology students Language: English	SCID
Mann *et al*[36]	Country: UK Setting: secondary care Age (years): M=27.4 (SD=5.8) Female: 100%	N=94 Depressed: 19	Administration: Researcher Language: English	SCID
McManus *et al*[39]	Country: USA Setting: secondary care Age (years): M=67 (SD=11) Female: 18%	N=1024 Depressed: 22	Administration: not stated Language: English	DIS
Mohd-Sidik *et al*[42]	Country: Malaysia Setting: primary care Age (years): not stated Female: 100%	N=146 Depressed: 21.2	Administration: family medicine specialist Language: Malay	CIDI
Robison *et al*[40]	Country: USA Setting: primary care Age (years): M=61 (range 50–68) Female: 71%	N=303 Depressed: 34	Administration: interviewer Language: Spanish	CIDI
Suija *et al*[41]	Country: Finland Setting: community Age (years): 72–73 Female: 58.4%	N=474 Depressed: 5.3	Administration: psychiatrist Language: not stated	MINI
Whooley *et al*[12]	Country: USA Setting: urgent care clinic Age (years): M=53 (SD=14) Female: 3%	N=536 Depressed: 18.1	Administration: self-report Language: English	DIS

MINI, Mini International Neuropsychiatric Interview; CIDI, Composite International Diagnostic Interview; DIS, Diagnostic Interview Schedule; SCID, Structured Clinical Interview for DSM Disorders; PICO, Population, Intervention, Comparator and Outcome; DOR, Diagnostic Odds Ratio; LR, Likelihood Ratio.

Table 2 Quality assessment of included studies

Study	Patient selection: Consecutive or random sample	Patient selection: avoid case–control/avoid artificially inflated base rate	Patient selection: avoided inappropriate exclusions	Patient selection: appropriately excludes those known to be depressed	Patient selection: overall risk of bias	Index test: Whooley interpreted blind to reference test	Index test: if translated, appropriate translation	Index test: overall risk of bias
Adachi *et al*[38]	✓	✓	?	?	Unclear	?	✓	Unclear
Arroll *et al*[17]	?	✓	✓	✓	Unclear	✓	NA	Low
Arroll *et al*[14]	✓	✓	✓	✓	Low	✓	NA	Low
Gjerdingen *et al*[37]	✓	✓	×	?	High	✓	NA	Low
Mann *et al*[36]	✓	✓	?	?	Unclear	✓	NA	Low
McManus *et al*[39]	✓	✓	×	?	High	?	NA	Unclear
Mohd Sidik *et al* (2011)	✓	✓	✓	✓	Low	✓	✓	Unclear
Robison *et al*[40]	?	✓	✓	?	Unclear	×	✓	High
Suija *et al*[41]	✓	✓	✓	×	High	✓	?	Unclear
Whooley *et al*[12]	?	✓	✓	?	Unclear	✓	NA	Low

Study	Reference test: Reference test correctly classifies target condition	Reference test: Reference test interpreted blind to Whooley	Reference test: If translated, appropriate translation	Reference test: If translated, psychometric properties reported	Reference test: Overall risk of bias	Flow/timing: Interval of two weeks or less	Flow/timing: All participants receive same reference test	Flow/timing: All participants included in analysis?	Flow/timing: Overall risk of bias
Adachi *et al*[38]	✓	?	✓	?	Unclear	?	✓	✓	Unclear
Arroll *et al*[17]	✓	✓	NA	NA	Low	?	✓	✓	Unclear
Arroll *et al*[14]	?	✓	NA	NA	Unclear	?	✓	?	Unclear
Gjerdingen *et al*[37]	✓	?	NA	NA	Unclear	✓	✓	×	High
Mann *et al*[36]	✓	✓	NA	NA	Low	✓	✓	×	High
McManus *et al*[39]	?	?	NA	NA	Unclear	?	✓	✓	Unclear
Mohd Sidik *et al* (2011)	✓	✓	✓	?	Unclear	✓	✓	✓	Low
Robison *et al*[40]	×	×	✓	?	High	?	✓	×	High
Suija *et al*[41]	✓	✓	?	?	Unclear	✓	✓	✓	Low
Whooley *et al*[12]	✓	?	NA	NA	Unclear	✓	✓	✓	Low

✓, criterion met; ×, criterion not met; ?, insufficient information to code whether criterion met; NA, not applicable.

Table 3 Performance of individual studies (no help question)

Study	Sensitivity (95% CI)	Specificity (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)	DOR (95% CI)
Adachi *et al*[38]	1.00 (0.29 to 1.00)	0.59 (0.48 to 0.69)	2.46 (1.90 to 3.17)	*	*
Arroll *et al*[17]	0.96 (0.82 to 0.99)	0.67 (0.62 to 0.71)	2.93 (2.51 to 3.43)	0.05 (0.01 to 0.35)	57.10 (9.71 to *)
Arroll *et al*[14]	0.95 (0.85 to 0.99)	0.78 (0.75 to 0.81)	4.43 (2.86 to 5.09)	0.05 (0.01 to 0.21)	81.70 (21.6 to *)
Gjerdingen *et al*[37]	1.00 (0.92 to 1.00)	0.44 (0.39 to 0.48)	1.79 (1.65 to 1.94)	*	*
Mann *et al*[36]	1.00 (0.78 to 1.00)	0.66 (0.57 to 0.75)	3.00 (2.31 to 3.90)	*	*
McManus *et al*[39]	0.90 (0.85 to 0.93)	0.69 (0.65 to 0.72)	2.91 (2.60 to 3.25)	0.14 (0.09 to 0.21)	20.40 (12.90 to 32.40)
Mohd-Sidik *et al*	1.00 (0.88 to 1.00)	0.70 (0.61 to 0.78)	3.83 (2.55 to 4.48)	*	*
Robison *et al*[40]	0.91 (0.78 to 0.98)	0.44 (0.37 to 0.50)	1.64 (1.42 to 1.89)	0.18 (0.13 to 0.25)	8.90 (2.83 to 27.90)
Suija *et al*[41]	0.64 (0.42 to 0.82)	0.88 (0.85 to 0.91)	5.75 (3.88 to 8.52)	0.40 (0.24 to 0.68)	14.20 (6.06 to 33.20)
Whooley *et al*[12]	0.95 (0.89 to 0.98)	0.56 (0.52 to 0.61)	2.23 (1.98 to 2.50)	0.07 (0.02 to 0.19)	30.80 (11.50 to 81.90)

*Value could not be estimated.

negative likelihood ratio 0.36 (95% CI 0.17 to 0.74) and OR 22.70 (95% CI 4.83 to 105.00).

Mohd-Sidik *et al* used the help question proposed by Arroll *et al*[14], and made the distinction between 'help, yes but not today' or 'yes, help today'. For this study we were able to ascertain how distinguishing between these two options can affect the ability of the help question to detect depression, in people who responded yes to either of the Whooley questions. If a positive answer to the help question was considered 'yes today', sensitivity was 0.61 (95% CI 0.42 to 0.78), specificity was 0.94 (95% CI 0.80 to 0.99), positive likelihood ratio was 10.4 (95% CI 2.64 to 41.1), negative likelihood ratio 0 0.41 (95% CI 0.262 to 0 0.64) and OR 25.3 (95% CI 5.55—* value unable to be estimated). If a positive answer to help question was considered a positive answer to 'yes today, or yes, but not today', sensitivity was higher at 0.87% (95% CI 0.70% to 0.96%), but specificity lower at 0.82% (95% CI 0.65% to 0.93%); positive likelihood ratio was 4.94 (95% CI 2.36 to 10.30), negative likelihood ratio was 0 0.15 (95% CI 0.06 to 0.39) and OR 31.5 (95% CI 8.22 to 120.00). In this study, therefore, answering 'yes, help today' increases the specificity of the Whooley questions when used in conjunction with the help question.

DISCUSSION

NICE guidance recommends that, in the UK, GPs consider using the Whooley questions to identify potential depression in certain patient groups[7–9] such as people with long-term conditions and women during the perinatal period. The guidance suggests that the Whooley questions are used as a case-finding tool for depression, so if an individual responds positively to one or both of the questions a more comprehensive assessment is carried out to determine whether or not that individual is depressed. The guidance acknowledges, though, that this recommendation is based on limited evidence. Furthermore, there is inconsistency between NICE guidance about whether the Whooley questions should be combined with an additional help question.

This review sought to establish the current evidence for the diagnostic performance of both the original two-item Whooley questions and their combination with an additional help question. The original validation study reported that the two-item version of the questions had high sensitivity (0.95, 95% CI 0.89 to 0.98) and modest specificity (0.56, 95% CI 0.52 to 0.61). The current review found comparable results. Pooled sensitivity was 0.95 (95% CI 0.88 to 0.97) and pooled specificity was 0.65 (95% CI 0.55 to 0.74). Similar figures were also reported in the subgroup analysis examining primary

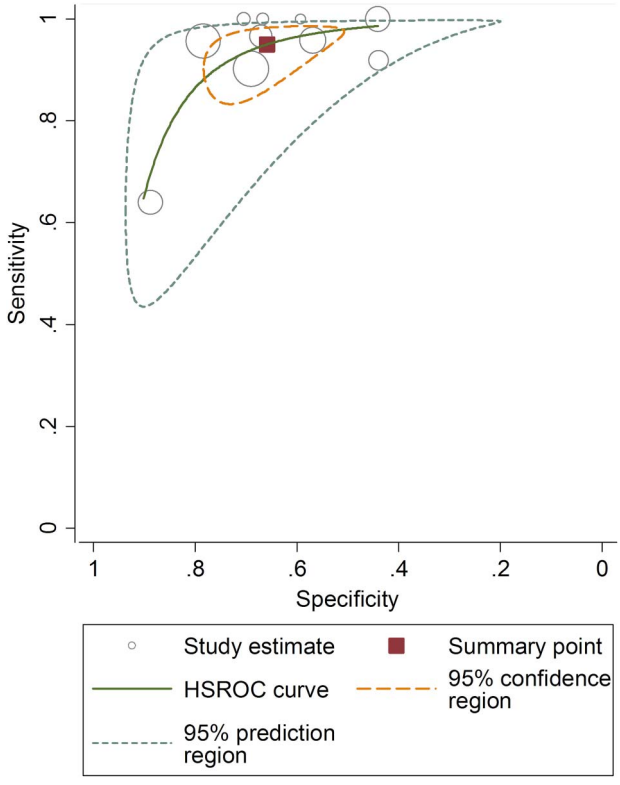

Figure 2 Whooley questions summary receiver operating characteristic plot of diagnosis of major depressive disorder. Pooled sensitivity and specificity using a bivariate meta-analysis.

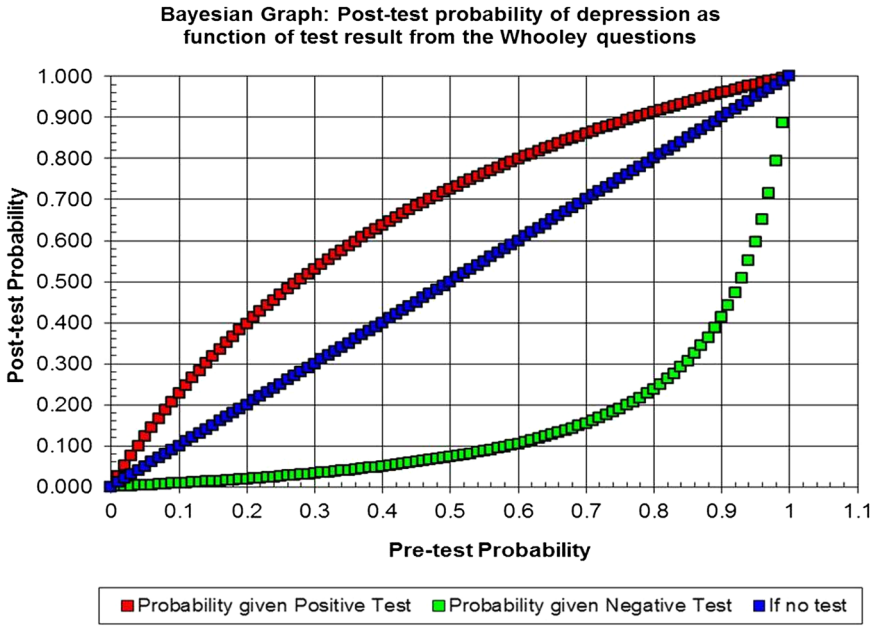

Figure 3 Bayesian graph for major depressive disorder for Whooley questions.

care studies (sensitivity: 0.96, 95% CI 0.91 to 0.98; specificity: 0.61, 95% CI 0.48 to 0.73).

Our search identified four studies that used the help questions. The authors of the original validation study[14] developed the help question in order to encourage the patient to take an active role in making decisions about their own treatment. They also suggested that the help question may improve specificity. Two categories of help were proposed in this study (help 'but not today', and help 'yes today').[14 42] However, of the four studies identified in our review, only two studies, one of which was the original validation study, distinguished between these two help categories: one study combined the two responses[41] and the fourth study[36] used a different response. Given the small number of studies and the variability in how the help question was used, we were unable to combine these studies in a meaningful way in order to ascertain the diagnostic performance of the help question when used with the original Whooley questions.

Limitations

The results of the systematic review need to be considered in light of the limitations of the primary studies used in the review and the review itself. As the QUADAS-2 ratings indicate, there are a number of limitations of the primary studies and often details about key methodological criteria were not reported. Only a small number made attempts to exclude people already known to have depression. The aim of depression screening is typically to identify depression in those not known to have that problem. It is possible that excluding those known to be depressed may alter the diagnostic performance of a test. Blinding in both directions was established in some but not all studies. Lack of blinding may artificially inflate the diagnostic performance of a test. It is possible then that the results may overestimate the performance of the Whooley.

Four of the 10 studies used the CIDI as the reference test, an instrument that has been described as an imperfect gold standard for mental health diagnosis.[43] However, the results of these studies for the two-item Whooley questions appeared broadly comparable with studies using a different gold standard. For the studies using the additional help question, the two studies that used the CIDI were the same two studies that reported increased specificity without an impact on sensitivity,[14 42] findings that were not replicated in the two studies that used other gold standards.[36 41] It is unclear to what extent these differences are linked to the use of different gold standards.

There are also a number of limitations of the review itself. First, we did not include the 'help' question in the search terms, which may have meant we missed articles focused solely on its effect. Second, although efforts were made to identify grey literature, it remains possible that unpublished studies were missed, so we cannot rule out the possibility of publication bias. Third, there is inconsistency in the published studies in how the Whooley questions are referred to, and while the inclusion of various alternative terms for the Whooley questions in the search strategy attempted to address this, it is possible that further relevant studies may have been missed.

Recommendations

The limitations suggest a number of research recommendations. Future diagnostic validation studies should report sufficient detail on the method to permit an assessment of key methodological criteria, such as those given in the QUADAS-2. Subsequent reviews of the Whooley would benefit from a more consistent method

of referring to the Whooley in primary studies. We would recommend the use of the term 'Whooley questions' and avoidance of the term 'PHQ-2'. Although the PHQ-2 shares similarities with the Whooley questions, the PHQ-2[44] asks about a different time frame and uses a different scoring system (see online supplementary appendix 2). We recommend that future studies should refer to Whooley in the title or abstract to facilitate future reviews of the measure.

CONCLUSION

This review on the diagnostic accuracy of the Whooley questions provides evidence of consistent high sensitivity and moderate specificity for the two questions across a range of settings among different populations. The Whooley questions demonstrate discriminatory power at ruling out depression: few people who answer no to both questions are depressed according to gold standard diagnostic interview. Given that depression is a common condition, this finding should be valuable to clinicians in general practice for use with patients they have concerns about. Despite its modest specificity, which means that many people who score positively will not meet diagnostic criteria for depression, the test retains value in its ability to eliminate the target condition. Although this review identified some evidence that the addition of a help question appeared to improve specificity—when used as second tier test—the inconsistency, both in how the question was phrased and how data were combined, means evidence of its performance remains limited.

Twitter Follow Simon Gilbody at @SimonGilbody

Contributors KB led on all stages of the review from development of the protocol, through screening studies, to data extraction and assessing the quality of the included studies, to production of the final report. DB involved in all stages of the review from development of the protocol, through screening studies and data extraction to synthesis and production of the final report. SG provided expert advice on methodology and approaches to assessment of the evidence base. MH devised the search strategy, carried out the literature searches and wrote the search methodology section of the report. LM reviewed the included studies and assessed their quality, performed the statistical analysis and wrote the results section of the final report. SN involved in the development of the protocol, screening studies for inclusion and data extraction. DM supervised the quality assessment, methodology and approaches to evidence synthesis and provided senior advice and support throughout the review and is guarantor. He contributed to the production of the final report. All parties were involved in drafting and/or commenting on the report.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Mental Health Foundation. Mental Health Statistics [cited 2015 07/04/15]. http://www.mentalhealth.org.uk/help-information/mental-health-statistics/
2. National Institute for Health and Clinical Excellence. *Clinical knowledge summaries: depression prevalence*. NICE, 2015. [updated Last revised in March 2015; cited 2015 07/04/15]. http://cks.nice.org.uk/depression#!backgroundsub:1
3. Moussavi S, Chatterji S, Verdes E, *et al*. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet* 2007;370:851–8.
4. Joffres M, Jaramillo A, Dickinson J, *et al*, Canadian Task Force on Preventive Health Care. Recommendations on screening for depression in adults. *CMAJ* 2013;185:775–82.
5. US Preventive Services Task Force. *Guide to clinical preventive services*. Alexandra, VA: Williams & Wilkinsy, 1996.
6. Allaby M. *Screening for depression: a report for the National Screening Committee*. Oxford: NHS PHRU, 2010.
7. National Institute for Health and Clinical Excellence. *CG90 depression: the Nice Guideline on the treatment and management of depression in adults*. London, 2010. http://www.nice.org.uk/guidance/cg90/evidence/cg90-depression-in-adults-full-guidance2
8. National Institute for Health and Clinical Excellence. *CG91 Depression in adults with a chronic physical health problem*. London, 2010. http://www.nice.org.uk/guidance/cg91/evidence/cg91-depression-with-a-chronic-physical-health-problem-full-guideline2
9. National Institute for Health and Clinical Excellence. *Clinical guideline 45: antenatal and postnatal mental health*. London: NICE, 2007.
10. National Institute for Health and Clinical Excellence. *NICE guidelines [CG192]: antenatal and postnatal mental health: clinical management and service guidance*. NICE, 2014. [updated December 2014; cited 2015 08/04/15]. http://www.nice.org.uk/guidance/cg192/chapter/1-recommendations#recognising-mental-health-problems-in-pregnancy-and-the-postnatal-period-and-referral-2
11. National Screening Committee. *The UK National Screening Committee's criteria for appraising the viability, effectiveness and appropriateness of a screening programme*. London: NSC, 2003.
12. Whooley M, Avins A, Miranda J, *et al*. Case-finding instruments for depression. Two questions are as good as many. *J Gen Intern Med* 1997;12:439–45.
13. Spitzer R, Williams J, Kroenke K, *et al*. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. *JAMA* 1994;272:1749–56.
14. 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;331:884.
15. Beauchamp H. What factors influence the use of the Whooley questions by health visitors? *J Health Visiting* 2014;2:378–87.
16. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Int Med* 2009;151:264–9.
17. Arroll B, Khin N, Kerse N. Screening for depression in primary care with two verbally asked questions: cross sectional study. *BMJ* 2003;327:1144–6.
18. Whiting P, Rutjes A, Westwood M, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Int Med* 2011;155:529–36.
19. Mann R, Hewitt C, Gilbody S. Assessing the quality of diagnostic studies using psychometric instruments: applying QUADAS. *Soc Psychiatry Psychiatr Epidemiol* 2009;44:300–7.
20. Thombs B, Arthurs E, El-Baalbaki G, *et al*. Risk of bias from inclusion of patients who already have diagnosis of or are undergoing treatment for depression in diagnostic accuracy studies of screening tools for depression: systematic review. *BMJ* 2011;343:d4825.
21. Reitsma J, Glas A, Rutjes AW, *et al*. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982–90.
22. Higgins J, Thompson S, Deeks J, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.
23. Thompson S, Higgins J. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21:1559–73.
24. Lijmer J, Bossuyt P, Heisterkamp S, *et al*. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21:1525–37.
25. Chagas M, Crippa J, Loureiro S, *et al*. Validity of the PHQ-2 for the screening of major depression in Parkinson's disease: two questions and one important answer. *Aging Ment Health* 2011;15:838–43.

26. Henkel V, Mergl R, Coyne J, *et al*. Screening for depression in primary care: will one or two items suffice? *Eur Arch Psychiatry Clin Neurosci* 2004;254:215–23.
27. Zuithoff N, Vergouwe Y, King M, *et al*. The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study. *BMC Fam Pract* 2010;11:98.
28. Chochinov HK, Wilson KG, Enns M, *et al*. "Are you depressed?" Screening for depression in the terminally ill. *Am J Psychiatry* 1997;154:674–6.
29. Burton C, Simpson C, Anderson N. Diagnosis and treatment of depression following routine screening in patients with coronary heart disease or diabetes: a database cohort study. *Psychol Med* 2013;43:529–37.
30. Lombardo P, Vaucher P, Haftgoli N, *et al*. The 'help' question doesn't help when screening for major depression: external validation of the three-question screening test for primary care patients managed for physical complaints. *BMC Med* 2011;9:114.
31. Shah M, Karuza J, Rueckmann E, *et al*. Reliability and validity of prehospital case finding for depression and cognitive impairment. *Am Geriatr Soc* 2009;57:697–702.
32. Biswas S, Gupta R, Vanjare H, *et al*. Depression in the elderly in Vellore, South India: the use of a two-question screen. *Int Psychogeriatr* 2009;21:369–71.
33. Ryan D, Gallagher P, Wright S, *et al*. Sensitivity and specificity of the Distress Thermometer and a two-item depression screen (Patient Health Questionnaire-2) with a 'help' question for psychological distress and psychiatric morbidity in patients with advanced cancer. *Psychooncology* 2012;21:1275–84.
34. Brody D, Hahn S, Spitzer R, *et al*. Identifying patients with depression in the primary care setting: a more efficient method. *Arch Intern Med* 1998;158:2469–75.
35. Suzuki T, Nobata R, Kim N, *et al*. Evaluation of Questionnaires (Two question case finding instrument & Beck Depression Inventory)as a tool for screening and intervention of depression in work place. *Seishin Igaku (Clinical Psychiatry)* 2003;45:699–708.
36. Mann R, Adamson J, Gilbody S. Diagnostic accuracy of case-finding questions to identify perinatal depression. *CMAJ* 2012;184:E424–30.
37. Gjerdingen D, Crow S, McGovern P, *et al*. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *Ann Fam Med* 2009;7:63–70.
38. Adachi Y, Aleksic B, Nobata R, *et al*. Combination use of Beck Depression Inventory and two-question case-finding instrument as a screening tool for depression in the workplace. *BMJ Open* 2012;2: e000596.
39. 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;96:1076–81.
40. Robison J, Gruman C, Gaztambide S, *et al*. Screening for depression in middle-aged and older puerto rican primary care patients. *J Gerontol A Biol Sci Med Sci* 2002;57:M308–14.
41. Suija K, Rajala U, Jokelainen J, *et al*. Validation of the Whooley questions and the Beck Depression Inventory in older adults. *Scand J Prim Health Care* 2012;30:259–64.
42. Mohd-Sidik S, Arroll B, Goodyear-Smith F, *et al*. Screening for depression with a brief questionnaire in a primary care setting: Validation of the two questions with help question (Malay version). *Int J Psychiatry Med* 2011;41:143–54.
43. Gelaye B, Tadesse M, Williams M, *et al*. Assessing validity of a depression screening instrument in the absence of a gold standard. *Ann Epidemiol* 2014;24:527–31.
44. Kroenke K, Spitzer R, Williams J. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care Res Rev* 2003;41:1284–92.