PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Diagnostic accuracy of the Whooley questions for the identification of depression: A diagnostic meta-analysis
AUTHORS	Bosanquet, Katharine; Bailey, Della; Gilbody, Simon; Harden, Melissa; Manea, Laura; Nutbrown, Sarah; McMillan, Dean

VERSION 1 - REVIEW

REVIEWER	Bruce Arroll University of Auckland
REVIEW RETURNED	31-Dec-2014

GENERAL COMMENTS	This is an important review. However there are numerous errors of fact and interpretation. The authors use the term "Whooley" for the 2 questions. However the original paper by Whooley took the 2 questions from the Prime-MD. While I agree with the authors that they should keep away from PHQ2 perhaps it should be the Prime-MD 2 or some other name that correctly identifies the origin of the questions. They also suggest that the Nice Guideline rejects the additional help question because of a lack of sufficient evidence. I could not find any reference to that piece of information on the PDF version that is 67 pages long and referenced as October 2009. A page reference and web address needs to added or this comment removed. * Originality - does the work add enough to what is already in the published literature? If so, what does it add? If not, please cite relevant references. This study is original as I do not know of any review published on this topic. * Importance of work to general readers - does this work matter to clinicians, patients, teachers, or policymakers? Is a general journal the right place for it? This work is of importance to all of the above and a general journal is the right place. Given the pervasiveness of depression in all disciplines it is of particular importance to clinicians not involved in psychiatry. * Scientific reliability Research Question - clearly defined and appropriately answered? The research question does not completely match what was actually done. In addition to assessing the diagnostic accuracy of the Whooley questions there was also a focus on the "help" questions. However they report 4 studies that used the help questions. Only two of these distinguished help "but not today and help yes today"

	i.e. Arroll 2005 and Sidik 2011. This is a crucial point as using Bayesian analysis the likelihood ratios (LR) on table 2 (Arroll 2005) LR for yes today is 17.5 and yes but not today 7.9 and no to help 0.27 and table 3 (Sidik 2011 -not sure if this is in the final version of Sidik 2011) are the important ones: LR for yes today 10.42; LR for yes but not today 2.19 and LR for no to help 0.16). Using the Arroll 2005 figures and starting with a pre-test probability of 5% the LR of 4.43 (table 3 of this review i.e Bosanquet) gives a post-test probability of 20% and using that as the pre-test for the next LR of 17.5 for a yes today gives a post-test of 85%. A post-test probability of 85% is a very high value given that the starting point is 5%. What it means is 85 out of 100 persons testing positive for depression would truly have a major depression. It is incorrect to say that the help question evidence is inconsistent as only two studies correctly evaluated the original help questions and they found then to have LRs of greater than 10 and to quote Guyatt in Users Guide to the Medical Literature second edition page 428 “LRs greater than 10 or less than 0.1 generate large and often conclusive changes from pre-test to post-test probability.” To not take in to account the distinction of help today/not today means that valuable information is missing from the clinical encounter. The original idea of the help question was to encourage the patient to take a role in making decisions about their own treatment and this idea is supported by Prof Chris Dowrick in Beyond Depression second edition page 33. Overall design of study - adequate ? The overall design is reasonable. I cannot comment on the statistical analysis and would suggest getting a statistical opinion as that is not my strong suit. Participants studied - adequately described and their conditions defined? The authors should have contacted the authors of the original papers to obtain complete data for table 2. Some of the missing information would be readily available from those authors. None of the papers are that old so the authors are most likely still alive. Methods - adequately described? Complies with relevant reporting standard - Eg CONSORT for randomised trials ? Ethical ? The authors report following the PRISMA checklist and CRD guidance. While PRISMA does not explicitly suggest writing to original authors I feel it is essential if one is to call a review a systematic review. For a Cochrane review that is standard practice. Results - answer the research question? Credible? Well presented? There are a number of errors of fact. One is the prevalence of depression in the Arroll 2003. It is 6% not 18%. I wonder if there is confusion over the positive predictive value and the prevalence. It is also not clear where the figures for Arroll 2005 on table 4 come from. As mentioned above this table is missing crucial information and there should be a separate table for Arroll and Sidik showing their table 2 and table 3 respectively with their 3 likelihood ratios i.e help yes today, help yes but not today and help, no. Interpretation and conclusions - warranted by and sufficiently derived from/focused on the data? Message clear? The message of the validity of the 2 questions is clear and helpful. The message is wrong about the help questions and this needs to be revisited. It is not clear if the authors fully understand the use of sequential likelihood ratios in diagnostic tests. It would be helpful for clinicians to make more of the high sensitivity being good for ruling out

	depression when the answer is negative. I find clinicians frequently do not understand this point. I think figures 4,5 and 6 could be removed as they do not add much. The funnel plot could be dealt with in the text. References - up to date and relevant? Any glaring omissions? The last search was September 2013 and this needs to be updated. There are no glaring omissions. Abstract/summary/key messages/What this paper adds - reflect accurately what the paper says? There does not seem to be a section which states what this paper adds as is usual with the BMJ. From my point of view it provides greater certainty of the point estimate of the sensitivity and specificity of the two questions from the original Prime-MD. The comment on the help question is incomplete as it fails to make the distinction between the two categories of help. Correcting that omission would render a different conclusion to the paper.

REVIEWER	Felicity Goodyear-Smith University of auckland
REVIEW RETURNED	04-Jan-2015

GENERAL COMMENTS	Originality This is an original systematic review and meta-analysis of existing studies looking at the two depression screening questions with or without the additional help question. Introduction The authors mention that there is variation in advice given about screening for depression. I agree. As we identified in a 2012 paper (F. A. Goodyear-Smith, van Driel, Arroll, & Del Mar, 2012), two groups of authors, one in the UK (S. Gilbody, House, & Sheldon, 2005; S. Gilbody, Sheldon, & House, 2008; S. M. Gilbody, House, & Sheldon, 2001) and one from the US Preventative Task Force (O'Connor, Whitlock, Beil, & Gaynes, 2009; Pignone et al., 2002; U. S. Preventive Services Task Force, 2002, 2009) conducted three and two systematic reviews (+/- meta-analyses) on screening for depression respectively. All five reviews contained different combinations of RCTs. The UK reviews concluded that the evidence did not support screening whereas the US group concluded it did. Our detailed analysis of one review from each group found that the differences were largely determined by one study (Lewis, Sharp, Bartholomew, & Pelosi, 1996) pooled in the UK but not the USPTF review, and another trial (Wells et al., 2000) pooled in the USPTF but excluded from the UK review. The studies selected, and the way that data were extracted from one study in particular, influenced the recommendations in opposite directions. We concluded “Systematic reviews may be less objective than assumed. Based on this analysis of two meta-analyses we hypothesise that strongly held prior beliefs (confirmation bias) may have influenced inclusion and exclusion criteria of studies, and their interpretation. Authors should be required to declare a priori any strongly held prior beliefs within their hypotheses, before embarking on systematic reviews.” The authors should identify that they are aware of, and have considered, this issue. Importance of work

	Depression is a common condition in general practice and in hospital practice hence the BMJ is a suitable journal for this work. Research Question The stated research question was 'to identify all studies that had examined the diagnostic test 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 that the effect of an additional help question was assessed, but this was not directly stated as an objective. I note that 'help' was not one of the search terms (Appendix 10. The 'Whooley questions' ("during the past month have you often been bothered by feeling down, depressed or hopeless?" and "during the past month have you often been bothered by little interest or pleasure in doing things?") were originally from the Prime-MD (Spitzer et al., 1994) and perhaps therefore should be attributed to Spitzer rather than Whooley (Whooley, Avins, Miranda, & Browner, 1997). The Help question ("Is this something with which you would like help?" with three possible responses: "no," "yes, but not today," or "yes") is a '2nd-tier' question only asked when one or both of the initial questions has a positive response (Arroll, Goodyear-Smith, Kerse, Fishman, & Gunn, 2005). I had originally developed the help question not specifically to improve the sensitivity or specificity of the test, but as a patient-centred approach to enable patients to indicate their level of readiness to change or willingness to address any lifestyle or mood concerns and become involved in shared decision-making eg see (F. Goodyear-Smith, Warren, Bojic, & Chong, 2013). We had also found that it could improve the specificity of a general practitioner diagnosis of depression when used as a 'second tier' test (Arroll et al., 2005). Method The authors correctly follow PRISMA guidelines with respect to data sources, search strategy and study selection. Although the two questions and the help question were originally designed to be used in primary care settings with general practice / family medicine patients, the authors included all participants and populations in the selected studies. Several of the 10 included studies were conducted in secondary care settings (Gjerdingen et al; Mann et al; McManus et al). The Suija et al study was a population not primary care based one of older patients (aged 72 years and over), and two studies (Mann and Gjerdingen) were in antenatal or postnatal settings. However the authors report limited heterogeneity of findings. With respect to the sub-analysis of the help questions, LR were available for the three help question responses (help today, help later or no help) in the Arroll and Sidik studies. However in the other two studies 'yes' and 'yes but not today' were combined, and in the Mann paper patients were merely asked 'Is this something you would like help with?' and hence these four papers should not have been analysed together. Results The funnel plot Figure 6 adds little. The DOR information is already

presented in Table 3. Figures 3 and 5 appear to be the same.

Discussion
The analyses regarding the two questions appear valid and useful. The additional work on the four papers with help questions needs to be revised. The help question is only asked if one or both of the original two questions are answered positive. It is therefore a separate 'second-tier' test conducted from the position of a post-test likelihood of a positive test. The sensitivity of the two questions is already established. The addition of the help question is a second test conducted effectively after the first, therefore increases the specificity while the sensitivity remains ie addition of the help question does not generate more false negatives because the answers to the two questions are already available. In clinical terms, patients responding positively to one or both depression screening questions who also indicate that they want help (especially if they want help today) are very likely to be true cases of depression, and also are likely to be motived to engage with intervention.

References
Arroll, B., Goodyear-Smith, F., Kerse, N., Fishman, T., & Gunn, J. (2005). Effect of the addition of a "help" question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study. British Medical Journal, 331(7521), 884.
Gilbody, S., House, A. O., & Sheldon, T. A. (2005). Screening and case finding instruments for depression. Cochrane Database of Systematic Reviews, (4), CD002792.
Gilbody, S., Sheldon, T., & House, A. (2008). Screening and case-finding instruments for depression: a meta-analysis. CMAJ Canadian Medical Association Journal, 178(8), 997-1003.
Gilbody, S. M., House, A. O., & Sheldon, T. A. (2001). Routinely administered questionnaires for depression and anxiety: systematic review. British Medical Journal, 322(7283), 406-409.
Goodyear-Smith, F., Warren, J., Bojic, M., & Chong, A. (2013). eCHAT for lifestyle and mental health screening in primary care. Annals of Family Medicine, 11(5), 460-466. 10.137/afm.1512
Goodyear-Smith, F. A., van Driel, M. L., Arroll, B., & Del Mar, C. (2012). Analysis of decisions made in meta-analyses of depression screening and the risk of confirmation bias: a case study. BMC Medical Research Methodology, 12, 76. http://dx.doi.org/10.1186/1471-2288-12-76
Lewis, G., Sharp, D., Bartholomew, J., & Pelosi, A. J. (1996). Computerized assessment of common mental disorders in primary care: effect on clinical outcome. Family Practice, 13(2), 120-126.
O'Connor, E. A., Whitlock, E. P., Beil, T. L., & Gaynes, B. N. (2009). Screening for depression in adult patients in primary care settings: a systematic evidence review. Annals of Internal Medicine, 151(11), 793-803.
Pignone, M. P., Gaynes, B. N., Rushton, J. L., Burchell, C. M., Orleans, C. T., Mulrow, C. D., & Lohr, K. N. (2002). Screening for depression in adults: a summary of the evidence for the U.S. Preventive Services Task Force. Annals of Internal Medicine, 136(10), 765-776.
Spitzer, R. L., Williams, J. B., Kroenke, K., Linzer, M., deGruy, F. V., 3rd, Hahn, S. R., . . . Johnson, J. G. (1994). Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. Jama., 272(22), 1749-1756.
U. S. Preventive Services Task Force. (2002). Screening for depression: recommendations and rationale.[Summary for patients

	in Ann Intern Med. 2002 May 21;136(10):I56; PMID: 12020161]. Annals of Internal Medicine, 136(10), 760-764. U. S. Preventive Services Task Force. (2009). Screening for depression in adults: U.S. preventive services task force recommendation statement. Annals of Internal Medicine, 151(11), 784-792. Wells, K. B., Sherbourne, C., Schoenbaum, M., Duan, N., Meredith, L., Unutzer, J., . . . Rubenstein, L. V. (2000). Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial.[Erratum appears in JAMA 2000 Jun 28;283(24):3204]. JAMA, 283(2), 212-220. Whooley, M. A., Avins, A. L., Miranda, J., & Browner, W. S. (1997). Case-finding instruments for depression. Two questions are as good as many. Journal of General Internal Medicine., 12(7), 439-445.

VERSION 1 – AUTHOR RESPONSE

Reviewer #1	Response from authors
COMMENTS TO THE AUTHOR 1. This is an important review. However there are numerous errors of fact and interpretation. The authors use the term “Whooley” for the 2 questions. The original paper by Whooley took the 2 questions from the PrimeMD. 2. While I agree with the authors that they should keep away from PHQ2 perhaps it should be the PrimeMD2 or some other name that correctly identifies the origin of the questions. 3. They also suggest that the Nice Guideline rejects the additional help question because of a lack of sufficient evidence. I could not find any reference to that piece of information on the PDF version that is 67 pages long and referenced as October 2009. A page reference and web address needs to added or this comment removed.	1. Thank you for recognising this is an important review. We agree with your suggestion to refer to the PRIME-MD when first introducing the Whooley questions, to acknowledge where they derived from. This has been added. (page 4) 2. The term “Whooley” is typically how the measure is referred to in the UK and is in fact referred to as such in NICE guidance (CG91, full guideline p80) http://www.nice.org.uk/guidance/cg91/evidence. We think there are clear advantages in adopting this term to describe this measure to avoid confusion with other brief-screening measures for depression, in particular the PHQ-2. We have stated in our recommendations that future studies should refer to Whooley in the title or abstract to facilitate future reviews of the measure. See page 17. 3. Thank you for highlighting that we had mistakenly included the summary guideline rather than the full guideline in our reference list. This has been amended. The relevant text stating a lack of evidence on the “help” question can be found in the full guideline: “A single study by Arroll and colleagues (2005) added a further question to the two in the PHQ-2, asking the patient if they wanted help with their depression. This increased specificity and the GDG considered the findings of the study and the adoption of the third question, but as there was only a single study showing the effect of this approach the GDG decided not to adopt it.” (NICE CG91, p.84). See page 4. http://www.nice.org.uk/guidance/cg91/evidence/cg91-depression-with-a-chronic-physical-health-problem-full-guideline2.).
ORIGINALITY This study is original as I do not know of any review published on this topic.	Thank you for this comment.
IMPORTANCE OF WORK TO GENERAL READERS This work is of importance to all clinicians, patients, teachers, policymakers and a general journal is the right place. Given the pervasiveness of depression in all disciplines it is of particular importance to clinicians not	We agree with your comments. Given the high prevalence of depression in the general population, this work is important to a wide range of stakeholders.

involved in psychiatry.	
ABSTRACT/SUMMARY/KEY MESSAGES/WHAT THIS PAPER ADDS 1. There does not seem to be a section which states what this paper adds as is usual with the BMJ. From my point of view it provides greater certainty of the point estimate of the sensitivity and specificity of the two questions from the original PrimeMD. 2. The comment on the help question is incomplete as it fails to make the distinction between the two categories of help. Correcting that omission would render a different conclusion to the paper.	1. We agree that the meta-analysis provides greater certainty on the performance of the Whooley questions, suggesting they perform consistently across a range of settings amongst a variety of populations. 2. We have clarified the distinction between the two categories of question. See page 14.
OVERALL DESIGN OF STUDY The overall design is reasonable. I cannot comment on the statistical analysis and would suggest getting a statistical opinion as that is not my strong suit.	We have conducted a number of similar diagnostic meta-analyses and believe that the analysis has been conducted appropriately.
PARTICIPANTS STUDIED The authors should have contacted the authors of the original papers to obtain complete data for table 2. Some of the missing information would be readily available from those authors. None of the papers are that old so the authors are most likely still alive.	We sought to contact authors and did in fact contact authors of over half of the included studies to gain clarification, including Whooley.
METHODS The authors report following the PRISMA checklist and CRD guidance. While PRISMA does not explicitly suggest writing to original authors I feel it is essential if one is to call a review a systematic review. For a Cochrane review that is standard practice.	Please see above.
RESULTS 1. The research question does not completely match what was actually done. In addition to assessing the diagnostic accuracy of the Whooley questions there was also a focus on the "help" questions. However they report 4 studies that used the help questions. Only two of these distinguished help "but not today and help yes today" i.e. Arroll 2005 and Sidik 2011. This is a crucial point as using Bayesian analysis the likelihood ratios (LR) on table 2 (Arroll 2005) LR for yes today is 17.5 and yes but not today 7.9 and no to help 0.27 and table 3 (Sidik 2011 not sure if this is in the final version of Sidik 2011) are the important ones: LR for yes today 10.42; LR for yes but not today 2.19 and LR for no to help 0.16). Using the Arroll 2005 figures and starting with a pretest	1. We have distinguished between the studies which analysed the responses "help, yes but not today" or "yes, help today" separately (Arroll, 2005 and Mohd-Sidik, 2011). However, we were unable to carry out rigorous analysis as the data were not clearly presented, which made interpretation difficult. We could not perform pre-test and post-test likelihood ratios. Despite contacting the relevant author we were unable to obtain the data required to provide clarification. (pages 14/15) 2. We have corrected this mistake. See Table 1 3. This has been addressed as best we can without being able to obtain the necessary data to clarify who was asked the "help" question.

probability of 5% the LR of 4.43 (table 3 of this review i.e Bosanquet) gives a posttest probability of 20% and using that as the pretest for the next LR of 17.5 for a yes today gives a posttest of 85%. A posttest probability of 85% is a very high value given that the starting point is 5%. What it means is 85 out of 100 persons testing positive for depression would truly have a major depression. It is incorrect to say that the help question evidence is inconsistent as only two studies correctly evaluated the original help questions and they found then to have LRs of greater than 10 and to quote Guyatt in Users Guide to the Medical Literature second edition page 428 “LRs greater than 10 or less than 0.1 generate large and often conclusive changes from pretest to posttest probability.” To not take in to account the distinction of help today/not today means that valuable information is missing from the clinical encounter. The original idea of the help question was to encourage the patient to take a role in making decisions about their own treatment and this idea is supported by Prof Chris Dowrick in Beyond Depression second edition page 33. 2. There are a number of errors of fact. One is the prevalence of depression in the Arroll 2003. It is 6% not 18%. 3. I wonder if there is confusion over the positive predictive value and the prevalence. It is also not clear where the figures for Arroll 2005 on table 4 come from. As mentioned above this table is missing crucial information and there should be a separate table for Arroll and Sidik showing their table 2 and table 3 respectively with their 3 likelihood ratios i.e help yes today, help yes but not today and help, no.	
INTERPRETATION AND CONCLUSIONS 1. The message of the validity of the 2 questions is clear and helpful. 2. The message is wrong about the help questions and this needs to be revisited. It is not clear if the authors fully understand the use of sequential likelihood ratios in diagnostic tests. 3. It would be helpful for clinicians to make more of the high sensitivity being good for ruling out depression when the answer is negative. I find clinicians frequently do not understand this point. 4. I think figures 4,5 and 6 could be	1. Thank you for this acknowledgment. 2. Please see section above. 3. We agree that more should be made of the Whooley questions’ performance at ruling out depression. This is highlighted in the conclusion. (page 17) 4. Figures 4, 5 and 6 have been removed.

removed as they do not add much. The funnel plot could be dealt with in the text.	
REFERENCES The last search was September 2013 and this needs to be updated. There are no glaring omissions.	We agree that the search needed updating. In April 2015 we conducted supplementary searches. No further studies were found. However, additional policy guidance was identified and added to the introduction and references. (page 4)
Reviewer #2	
COMMENTS TO THE AUTHOR Thank you for giving me the opportunity to review this paper. As the original developer of the 'help' question I declare an interest (and potential conflict in interest) in this topic.	We appreciate you declaring an interest in the 'help' question.
ORIGINALITY This is an original systematic review and metaanalysis of existing studies looking at the two depression screening questions with or without the additional help question.	Thank you for recognising that this review and meta-analysis is original.
INTRODUCTION The authors mention that there is variation in advice given about screening for depression. I agree. As we identified in a 2012 paper (F. A. GoodyearSmith, van Driel, Arroll, & Del Mar, 2012), two groups of authors, one in the UK (S. Gilbody, House, & Sheldon, 2005; S. Gilbody, Sheldon, & House, 2008; S. M. Gilbody, House, & Sheldon, 2001) and one from the US Preventative Task Force (O'Connor, Whitlock, Beil, & Gaynes, 2009; Pignone et al., 2002; U. S. Preventive Services Task Force, 2002, 2009) conducted three and two systematic reviews (+/metaanalyses) on screening for depression respectively. All five reviews contained different combinations of RCTs. The UK reviews concluded that the evidence did not support screening whereas the US group concluded it did. Our detailed analysis of one review from each group found that the differences were largely determined by one study (Lewis, Sharp, Bartholomew, & Pelosi, 1996) pooled in the UK but not the USPTF review, and another trial (Wells et al., 2000) pooled in the USPTF but excluded from the UK review. The studies selected, and the way that data were extracted from one study in particular, influenced the recommendations in opposite directions. We concluded "Systematic reviews may be less objective than assumed. Based on this analysis of two metaanalyses we hypothesise that strongly held prior beliefs (confirmation bias) may have	Such declarations are not standard practice in systematic reviews, but we would be willing to add one should the editors feel that this is necessary.

influenced inclusion and exclusion criteria of studies, and their interpretation. Authors should be required to declare a priori any strongly held prior beliefs within their hypotheses, before embarking on systematic reviews." The authors should identify that they are aware of and have considered, this issue.	
IMPORTANCE OF WORK TO GENERAL READERS Depression is a common condition in general practice and in hospital practice hence the BMJ is a suitable journal for this work.	We agree with this comment. Given that depression is a common condition it is an important issue, particularly in primary care.
RESEARCH QUESTION 1. The stated research question was 'to identify all studies that had examined the diagnostic test 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 that the effect of an additional help question was assessed, but this was not directly stated as an objective. 2. I note that 'help' was not one of the search terms (Appendix 1). 3. The 'Whooley questions' ("during the past month have you often been bothered by feeling down, depressed or hopeless?" and "during the past month have you often been bothered by little interest or pleasure in doing things?") were originally from the PrimeMD (Spitzer et al., 1994) and perhaps therefore should be attributed to Spitzer rather than Whooley (Whooley, Avins, Miranda, & Browner, 1997). 4. The Help question ("Is this something with which you would like help?" with three possible responses: "no," "yes, but not today," or "yes") is a '2ndtier' question only asked when one or both of the initial questions has a positive response (Arroll, GoodyearSmith, Kerse, Fishman, & Gunn, 2005). I had originally developed the help question not specifically to improve the sensitivity or specificity of the test, but as a patientcentred approach to enable patients to indicate their level of readiness to change or willingness to address any lifestyle or mood concerns and become involved in shared decisionmaking eg see (F. GoodyearSmith, Warren, Bojic, & Chong, 2013). We had also found that it could improve the specificity of a	1. Thank you for highlighting this omission. It has been added to the abstract (page 1) and introduction (page 2). 2. We have recognised this as a limitation. See page 16 3. We have acknowledged that the Whooley questions derived from Spitzer's PRIME-MD 1000 study (1994). See page 4 4. Thank you for the comments informing us of your role in developing the 'help' question. We have changed the 'help' question analysis by type but were unable to look at pre-test and post-test analyses. The data were not clearly presented, which made interpretation difficult. Contacting the relevant author did not result in obtaining the relevant data. See pages 14/15

general practitioner diagnosis of depression when used as a 'second tier' test (Arroll et al., 2005).	
ABSTRACT/SUMMARY/KEY MESSAGES/WHAT THIS PAPER ADDS	N/A
OVERALL DESIGN OF STUDY	N/A
PARTICIPANTS STUDIED	N/A
METHODS 1. The authors correctly follow PRISMA guidelines with respect to data sources, search strategy and study selection. Although the two questions and the help question were originally designed to be used in primary care settings with general practice / family medicine patients, the authors included all participants and populations in the selected studies. Several of the 10 included studies were conducted in secondary care settings (Gjerdingen et al; Mann et al; McManus et al). The Suija et al study was a population not primary care based one of older patients (aged 72 years and over), and two studies (Mann and Gjerdingen) were in antenatal or postnatal settings. However the authors report limited heterogeneity of findings. 2. With respect to the subanalysis of the help questions, LR were available for the three help question responses (help today, help later or no help) in the Arroll and Sidik studies. However in the other two studies 'yes' and 'yes but not today' were combined, and in the Mann paper patients were merely asked 'Is this something you would like help with?' and hence these four papers should not have been analysed together.	1. Thank you for the statement that we followed the PRISMA guidelines correctly and for the acknowledgement that, despite considerable variation in setting and population among the Whooley questions studies, there was limited heterogeneity. 2. We agree with the comments on the 'help' question sub-analysis. We have now distinguished between the two studies which analysed the responses "help, yes but not today" or "yes, help today" separately (Arroll, 2005 and Mohd-Sidik, 2011). See pages 14/15
RESULTS The funnel plot Figure 6 adds little. The DOR information is already presented in Table 3. Figures 3 and 5 appear to be the same.	We agree the figures you highlighted are duplicated. We have removed these data.
INTERPRETATION AND CONCLUSIONS 1. The analyses regarding the two questions appear valid and useful. 2. The additional work on the four papers. With help questions needs to be revised. The help question is only asked if one or both of the original two questions are answered positive. It is therefore a separate 'secondtier' test conducted from the position of a	1. We appreciate this positive comment on the analyses of the two questions. 2. We have made these changes. See pages 14/15

posttest likelihood of a positive test. The sensitivity of the two questions is already established. The addition of the help question is a second test conducted effectively after the first, therefore increases the specificity while the sensitivity remains ie addition of the help question does not generate more false negatives because the answers to the two questions are already available. In clinical terms, patients responding positively to one or both depression screening questions who also indicate that they want help (especially if they want help today) are very likely to be true cases of depression, and also are likely to be motived to engage with intervention.	
REFERENCES	
Reviewer #3	
COMMENTS TO THE AUTHOR Greetings, Thankyou for the opportunity to review this very interesting metaanalysis on the Whooley questions. I have read the prior peer reviews and the author's response.Review for paper by Bosanquet on Diagnostic test accuracy of the Whooley questions. As a US Family Medicine physician I was initially unfamiliar with the term Whooley questions. We tend to screen with the PHQ2 and confirm with the clinical interview/PHQ9. I would like to see the manuscript perhaps include a table that shows the subtle differences b/t these three depression screeners.	We appreciate this positive comment and the information that you are based in the US not the UK.
ORIGINALITY Study was original. And seems to address the gap the authors outline in the introduction regarding the perception of a lack of evidence for the Whooley's.	Thank you for acknowledging our study is original and addresses a lack of evidence on the effectiveness of the Whooley questions.
IMPORTANCE OF WORK TO GENERAL READERS Definitely is important to primary care providers and likely to UK policy makers who make screening guidelines such as NICE. Also relevant to us in the US regarding depression screening in general still debated. Canada doesn't rec universal screening as you likely all know.	We agree with these comments that this work is important.
ABSTRACT/SUMMARY/KEY MESSAGES/WHAT THIS PAPER ADDS I have no issues w/the abstract. Agree that pending review of the data on the 3rd question, the conclusion could	Thank you for this comment.

change. However, based on the current manuscript the conclusion is appropriate.	
SCIENTIFIC RELIABILITY Yes. In my opinion, the research question was clearly defined what is the sens and spec of the Whooley's compared to the Gold Standard clinical interview. I reviewed the comments of the prior peer reviewers and agree that the large likelihood ratios frankly should be praised a bit more than the authors have. However, the heterogeneity and fact that over 6K studies were excluded cannot be ignored.	As discussed above in response to reviewer one's comments we have highlighted the large likelihood ratios in the conclusion (page 17). Heterogeneity was less than we've typically found in other diagnostic accuracy meta-analyses such as the PHQ-2.
OVERALL DESIGN OF STUDY I used the CEBM CASP worksheet for metaanalysis and systematic reviews and found the authors met all of the stated criteria. My major concern is the fact that out of 6K+ studies only 22 met the inclusion criteria and then only ten were left. I am concerned some sig studies could have been excluded. Could the authors give say the top 3 reasons some of the 6K studies were excluded?	We followed standard systematic review guidelines throughout the conduct of their review. Although there is always a chance that studies may be missed, we acknowledged this in the limitations. See page 16.
PARTICIPANTS STUDIED See above. I appreciated the details of how the process occured. Agree w/prior peer reviewer regarding contacting authors. This is mentioned in the CASP worksheet. Metaanalysis authors should contact key "experts" in the field regarding the topic looking for unpublished dataieis Whooley still alive?	As described above.
METHODS See comments above. Agree the authors followed PRISMA criteria, applied quality guidelines QUADAS, etc. my primary concern was the paucity of studies meeting criteria and was something missed that is unpublished (a known issue with all metaanalysis).	This is always a possibility. However, we carried out extensive grey literature searches under guidance from CRD information specialist (page 5).
RESULTS The prior reviewers clearly stated their concerns regarding the "help question". I thought the data regarding the Whooley's w/o the help question was clearly presented. APpreciated the tables, etc. If still concern about data validity then rec the authors include the actual 2 x 2 tables/data extraction and show specifically how the pooled sens/spec were calculated. Certainly the "similar" results to the individual studies is reassuring. I am curious how the sens dropped in the two studies that added the 3rd questiowas that a non-English	As described above.

study, unusual settings, small sample size? Explanation here would be helpful.	
INTERPRETATION AND CONCLUSIONS I think the 1st paragraph of the conclusion is the most relevant. The 2nd paragraph should be shortened and most of it placed into the discussion. The 3rd paragraph would be rendered redundant and could be removed.	We agree with these comments. We have moved the second paragraph to the discussion section (page 14) and the third paragraph has been removed (page 17).
REFERENCES Rec a updated lit search prior to publication and any adjustments made pending results.	We updated the search in April 2015. No additional studies meeting our inclusion criteria were found.
RANDOM COMMENTS OF THE TEXT 1. Page 4, line 6: describe depression prevelance., any numbers to back up the 1st sentences? 2. Page 4, line 13: describe the differnces in screening b/t UK, US and Canada. 3. Page 4, line 23: Who are the high risk groups? Would be nice to see the official screeing recs in teh UK 4. page 4, line 32: can you write out the Whooley's word for word? have atable comparing them to PrimeMD and PHQ2? 5. Page 4, line 43: care to state earlier in the manuscript the sens/spec of the Whooley's from the original article/validation study? 6. Page 4, line 55: remove "test", "recognized" sp? UK vs. US spelling? 7. Page 5, line 25: again any experts consulted/contacted as part of the lit search?? 8. Page 6, well written inclusion/exclusion criteria 9. Page 7, line 45: how did you define appropriate translation? 10. Page 8, line 2833: nice description, but prob unnecessary for BMJ. 11. Page 8, line 50: spacing error b/t 95% and confidence. 12. page 9, line 25: very concerning drop from 10K to 6K to 22 to 10. can you describe which inclusion/exclusion criteria applied most to whittle this down so much? 13. Page 11, line 6: rec you say 0.9 to 1.0 to keep same low to high for sen and spec. 14. Page 11, line 29: please list the actual percentage for the low prob for neg test result. 15. Page 12, line 717: what is the diff b/t the primary care and community settings?? are these really diff? In the US these would be the same. Please	Thank you for these specific comments most of which we have addressed: 1. Numbers added in (page 4). 2. This review was focused on screening implications for the UK. 3. High risk groups added (page 4). 4. We have added the full text of the Whooleys and the PHQ2 (Appendix 2). 5. Sens and spec stated (page 4) 6. “Test” removed. Recognised–spelling corrected. 7. See page 6. 8. Thank you for this comment. 9. We defined it as forward and back translation. See page 8 10. Thank you for this comment. 11. This has been altered. 12. See above. 13. This has been changed. 14. It was just an estimation 15. Primary care is health care provided in the community for people via a general practitioner or practice nurse. Community care is care delivered outside of a hospital and within local communities. 16. Comments added on different settings/subjects. See page 17 17. Please see page 17 and Appendix 2. 18. These paragraphs have been amended accordingly. See page 17

clarify. 16. Page 14, line 37: see prior comments. Why so diff?? Setting? Language? Number of subjects?? Please clarify and perhaps discuss w/study authors. 17. Page 15, line 1822: rec a table comparing the diff two question screeners. 18. Page 15, line 2760: see above. Like the 1st paragraph, shorten s2nd and add some of that to the discussion (as it sounds like discussion not conclusion), then delete the 3rd paragraph as reads just like the 1st.	
Reviewer #4	
COMMENTS TO THE AUTHOR This paper describes a systematic review and diagnostic metaanalysis of the Whooley 2 questions for screening for depression. This is a worthwhile investigation as there is a lack of clarity on whether these tools are of value in clinical settings. Guidance has been contradictory. Conducting this review and metaanalysis is therefore a worthwhile endeavour. It also considers whether the addition of a third question improves the diagnostic accuracy of the tool. The authors find the Whooley 2 to have high sensitivity and moderate specificity but that there is inconsistent evidence to recommend the use of the additional third question.	Thank you for acknowledging the need for a systematic review and meta-analysis of the Whooley questions and recognising its potential value in clinical settings.
ORIGINALITY	N/A
IMPORTANCE OF WORK TO GENERAL READERS I would recommend its publication in the BMJ. I have a few specific comments outlined below.	We appreciate the recommendation that this paper should be published in the BMJ.
ABSTRACT/SUMMARY/KEY MESSAGES/WHAT THIS PAPER ADDS Abstract clearly written. Suggest that under Data Extraction they specify that QUADASII was used to assess quality – as this is a further strength.	Thank you for stating these positive comments.
OVERALL DESIGN OF STUDY This study has been well designed and is clearly reported.	See above.
INTRODUCTION Easy to follow. The authors make a good case for conducting this study – as uncertainty is evident. It would be of benefit, in paragraph 1, to make a statement about why considering a screening tool at all is necessary – ie that recognising depression in primary care and other nonpsychiatric settings is	We agree that the case for this study was strong given the lack of pooled evidence on the effectiveness of the Whooley questions. Though we have not added a statement in the introductory paragraph about why considering a screening tool is necessary– as we have preferred to focus on the debate around screening tools which exists in the UK–we have stated that depression is a common condition, often under detected in primary care. (page 4)

notoriously difficult for clinicians. Metaanalyses show that whilst clinicians (without screening / diagnostic tools) do reasonably well at ruling out when depression is absent (high Sp), they fair poorly at identifying when depression is present (low Se) (see Mitchell AJ et al., Lancet 2009; Cepoiu M et al., JGIM 2008) – as such tools are needed to help – possibly tools such as the Whooley 2.	
PARTICIPANTS STUDIED	N/A
METHODS 1.This study has several methodological strengths: prespecified protocol; broad range of databases searched; other sources sought; sensitive search strategy unrestricted by language and design filters. The authors should state why search began in 1994 – must relate to creation of Whooley 2 but be good to state this. 2. The prepilot form was completed by 2 reviewers – it is unclear whether they were working independently and then seeking agreement or if they each did a proportion. This could be made clearer. Presently the QUADASII citation does not appear in the reference list. 3. The authors specify that up to a 2 week interval is acceptable for administration of test and the gold standard. They cite their previous work to justify this. I think this is a long interval for symptoms that may well naturally fluctuate and have altered within a two week period. How do the authors justify this? 4. Statistical methods are clearly described.	1. We appreciate the positive comments on the methodology. We have added in why the search began in 1994 (page 5). 2. The 2 reviewers who completed the prepilot form worked independently and sought agreement from a third reviewer where necessary. Thank you for highlighting that the QUADASII citation did not appear in the reference list. This has been rectified (page 7). 3. We acknowledge it is a limitation not to have the index test and reference standard conducted on the same day. See page 16 4. Thank you for acknowledging this.
RESULTS Clearly presented. Good use of Figures. Authors should consider removing funnel plot if they believe there are too few studies for it to be interpretable in a meaningful way.	We have removed the funnel plot.
INTERPRETATION AND CONCLUSIONS 1. Discussion well written. Authors fairly represent limitations. This could be extended to consider the two week interval for test and gold standard administrations. 2. In the conclusion, the authors state that “many who score positive on the test will not meet diagnostic criteria for depression” and they refer to this in the following paragraph as a “problem”. I this a problem? – as the test is a screening tool it is designed to ascertain	1. Added as limitation. See page 16 2. As a screening tool the Whooley questions are good at ‘ruling out' depression and we have stated that more should be made of that. See page 17

the POSSIBILITY of depression, not CERTAINTY. Whooley would therefore be acceptable because the tool has high sensitivity. The problem would only occur if clinicians were using the Whooley 2 as a diagnostic instrument – which is not proposed. Perhaps a statement should be made to explain that moderate Sp is not such a concern in a screening tool.	
REFERENCES The search ended in 2013 – would it benefit from updating?	We updated the search in April 2015. No additional studies meeting our inclusion criteria were found.

VERSION 2 – REVIEW

REVIEWER	Jens Klotsche German Rheumatism Research Center Berlin, Epidemiology unit
REVIEW RETURNED	11-Sep-2015

GENERAL COMMENTS	The reviewer completed the checklist but made no further comments.