BJPsych | The British Journal of Psychiatry (2017)
210, 255–260. doi: 10.1192/bjp.bp.116.186932

Review article

Comparison of diagnostic performance of Two-Question Screen and 15 depression screening instruments for older adults: systematic review and meta-analysis

Kelvin K. F. Tsoi, Joyce Y. C. Chan, Hoyee W. Hirai and Samuel Y. S. Wong

Background
Screening for depression in older adults is recommended.

Aims
To evaluate the diagnostic accuracy of the Two-Question Screen for older adults and compare it with other screening instruments for depression.

Method
We undertook a literature search for studies assessing the diagnostic performance of depression screening instruments in older adults. Combined diagnostic accuracy including sensitivity and specificity were the primary outcomes. Potential risks of bias and the quality of studies were also assessed.

Results
A total of 46 506 participants from 132 studies were identified evaluating 16 screening instruments. The majority of studies (63/132) used various versions of the Geriatric Depression Scale (GDS) and 6 used the Two-Question Screen. The combined sensitivity and specificity for the Two-Question Screen were 91.8% (95% CI 85.2–95.6) and 67.7% (95% CI 58.1–76.0), respectively; the diagnostic performance area under the curve (AUC) was 90%. The Two-Question Screen showed comparable performance with other instruments, including clinician-rated scales. The One-Question Screen showed the lowest diagnostic performance with an AUC of 78%. In subgroup analysis, the Two-Question Screen also had good diagnostic performance in screening for major depressive disorder.

Conclusions
The Two-Question Screen is a simple and short instrument for depression screening. Its diagnostic performance is comparable with other instruments and, therefore, it would be favourable to use it for older adult screening programmes.

Declaration of interest
None.

Depression is a common disorder in older adults. The prevalence of depression in elderly people has been reported to be between 10 and 20%.[1,2] Older adults with physical illnesses or living in residential care facilities showed higher prevalence, from 14 to 44%.[3,4] Depression is associated with an increased risk of suicide, decline in functioning and quality of life.[5–7] It also increases the utilisation of healthcare services.[2,5,8] A wide range of pharmaceutical treatments and psychosocial interventions can relieve the symptoms of depression[5] and early detection and management of the disease can alter the disease prognosis. The United States Preventive Services Task Force (USPSTF) has recommended screening for depression in primary care settings.[8–11] However, detection of depression in older adults is more difficult.[12] The somatic symptoms of depression such as loss of appetite, weight loss, decreased energy and disturbed sleep are similar to the symptoms of other physical illness.[3] Moreover, older adults often complain of physical discomfort instead of low mood, and therefore the diagnosis of depression in older adults is often missed.[13] An effective screening instrument to identify older adults at risk or with clinically relevant depressive symptoms is important.

There are over 20 screening instruments used for detection of depression and studies have used a variety of screening instruments. The Geriatric Depression Scale (GDS)[14] and the Even Briefer Assessment Scale for Depression (EBAS-DEP)[15] were designed specifically for older adults and the Cornell Scale for Depression in Dementia (CSDD)[16] was designed specifically for patients with dementia. The recent report from USPSTF showed that the GDS was the most common screening instrument used in depression screening programmes for older adults.[17] Other screening instruments such as the Beck Depression Inventory (BDI)[18] were not originally designed for older adults although they are also commonly used for screening in older adults. The National Institute for Health and Care Excellence (NICE) has recommended using the Two-Question Screen for screening of depression in primary care and general hospital settings since 2004.[19] The Two-Question Screen is a self-rating screening instrument that consists of just two questions and can be completed in 1–2 min. The two questions asked for symptoms in the past month are: (a) 'Have you been troubled by feeling down, depressed or hopeless?' and (b) 'Have you experienced little interest or pleasure in doing things?' The rating method is only 'Yes' and 'No' answers. Although the Two-Question Screen is very short, some studies have demonstrated its accuracy in detecting depression.[20,21] Two meta-analyses that were conducted in patients with chronic physical health problems or cancer revealed that the instrument had a high level of acceptability.[22,23] Other studies have shown that the GDS-30,[14] GDS-15,[24] the Center for Epidemiological Depression Scale (CEDS)[25] and the SelfCARE(D)[26] had good sensitivity and specificity in depression screening for older adults.[3,13] The Two-Question Screen is relatively simple to use when compared with other instruments, in addition to being recommended by NICE. The objective of this systematic review was therefore to evaluate the diagnostic accuracy of the Two-Question Screen for older adults and to compare it with other available screening instruments used in screening for depression.

Method

This study was performed according to the standard guidelines for systematic review of diagnostic studies, including the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA)[27] and guidelines from the Cochrane Diagnostic Test Accuracy Working Group.[28,29]

Search strategy

A list of screening instruments for depression was identified from previous studies.[3–4,13,22,23] Literature searches were performed using the electronic databases Medline, EMBASE and PsycINFO from the earliest available dates stated in each database and searched until 31 October 2015. Each screening instrument was searched with the general keywords of 'depression' and 'elderly'. Diagnostic studies comparing the accuracy of screening instruments for depression were identified from the search records. The literature search was extended to Google Scholar with the names of individual screening instruments for depression. The relevancy of the citation was ranked in the search results of Google Scholar, so we scanned the first ten pages of all search records. The selection was limited to peer-reviewed articles published in the English language. A manual search was also performed on the bibliographies of review articles and any research studies cited in the eligible studies.

Inclusion and exclusion criteria

Studies were included if they met the following inclusion criteria: (a) included older adults as participants for the detection of depression in any clinical or community settings, and the mean or median age of the participants was 60 or older; (b) used standard diagnostic criteria as the gold standard for defining depression, including DSM (for example, DSM-IV-TR[30]), ICD (for example ICD-10[31]), Geriatric Mental State – The Automated Geriatric Examination for Computer Assisted Taxonomy (GMS–AGECAT),[32,33] Provisional Diagnostic Criteria for Depression in Alzheimer's Disease (PDC-dAD);[34] and (c) reported the number of participants with depression and evaluated the accuracy of the screening instruments, including sensitivity, specificity or data that could be used to derive those values. Studies were excluded if (a) they were not written in English; or (b) they included an uncommon screening instrument that was only mentioned in three or fewer eligible studies during the literature search.

Data extraction

Two investigators (J.Y.C.C. and H.W.H.) independently assessed the relevance of search results and extracted data into a data extraction form. Data collected included year of publication, study location, number of participants, mean age, percentage of men, number of participants with depression and suggested cut-off values for depression. We also recorded the sensitivity, specificity, true-positive, false-positive, true-negative and false-negative values for each instrument. When a study reported results of sensitivities and specificities across multiple cut-off values of a screening instrument, only the results of the optimal cut-off value that was suggested in that individual paper was selected. When discrepancies were found regarding study eligibility or data extraction, the third investigator (K.K.F.T.) made the definitive decision. The main outcome was the accuracy of screening instruments in the detection of depression among older adults. All levels of depression severity were included.

Risk of bias and reporting quality

Potential risks of bias in each study were evaluated by QUADAS-2 (the Quality Assessment of Diagnostic Accuracy Studies 2 instrument),[35] which assessed (a) patient selection; (b) execution of the screening instruments; (c) execution of the reference standard; and (d) clear presentation of the patient follow-up and delayed time of reference test. An eight-point scale was designed to evaluate the study quality that showed (a) a clear definition about study population; (b) details of participant recruitment, (c) sampling of participant selection, (d) data collection plan, (e) reference standard and its rationale, (f) technical specifications, (g) rationales for cut-offs, and (h) methods for calculating diagnostic accuracy with confidence intervals.

Data synthesis and statistical analysis

The overall sensitivity and specificity of each screening instrument were pooled using a bivariate random-effects model.[36] Forest plots were used to present the pooled sensitivity and specificity. When different threshold values were used to define positive and negative likelihood ratios of the screening instruments, the results had to allow trade-off between sensitivity and specificity. Therefore, a diagnostic odds ratio (OR) was used as a single indicator of test performance.[37] A hierarchical summary receiver-operating characteristic (HSROC) curve was generated to present the summary estimates of sensitivities and specificities along with their corresponding 95% confidence intervals and prediction region.[38] The area under the HSROC curve (AUC) was calculated and the values approaching 100% indicated that the diagnostic accuracy was good.[39] When the Hessian matrix of bivariate random-effects approach was unstable or asymmetric, a random-effects model following the approach of DerSimonian & Laird was applied to estimate the pooled sensitivity and specificity, and a Moses–Littenberg summary receiver-operating characteristic (SROC) curve was generated to present the summary estimates of sensitivities and specificities with AUC presented as a summary statistic.[40,41] Statistical heterogeneity among the trials was assessed by I^2, which described the percentage of total variation across studies as a result of heterogeneity rather than chance alone. Statistical analyses were mainly performed with the Metandi and Mida procedures in Stata, version 11.

Subgroup analysis

As the severity of depression is one of the factors that may affect the diagnostic accuracy of screening instruments, studies highlighting participants with major depressive disorder were selected for subgroup analysis. Furthermore, as the studies recruited participants from different settings, subgroup analyses were also performed to assess the screening instruments in nursing homes and specialist clinic settings (i.e. recruited in specialised out-patient clinics and hospitals) and in community settings (i.e. recruited in the community or primary care).

Results

Literature search and study selection

A total of 9188 abstracts were identified, with 89 of them extracted from the bibliographies. All titles and abstracts were screened and 318 articles out of 451 relevant articles were excluded for the following reasons: studies were systematic reviews ($n = 40$); studies did not fulfil the inclusion criteria ($n = 88$); studies lacked details on sensitivity and specificity ($n = 146$); studies reported results of the screening instrument without comparing it with an appropriate gold standard ($n = 44$); a study included the same

https://doi.org/10.1192/bjp.bp.116.186932 Published online by Cambridge University Press

cohort of participants ($n = 1$) (online Fig. DS1). The definitive analysis in this systematic review included 132 studies published between 1982 and 2015 for older adults with depression from the USA, UK, Australia and another 30 countries. A total of 16 depression screening instruments were identified. Thirteen of them were self-rating scales that were either self-administered or staff-interviewed (Table 1). Two screening instruments were clinician-rated scales; the Hamilton Rating Scale for Depression – 17 items (HRSD)[42] and the Montgomery–Åsberg Depression Rating Scale (MADRS).[43] One scale was rated by the clinician and informant, the CSDD.

Study characteristics

This meta-analysis included 132 studies, with 143 cohorts, reporting the diagnostic performance of depression screening instruments for older adults. A total of 46 506 participants were included with a mean age between 60 and 87 years (online Table DS1), and 6 811 participants (14.8%) were diagnosed with depression. A total of 105 studies (79.5%) had suggested an optimal cut-off value for the screening instrument, and the other 27 studies presented the cut-off value that originally was described by the screening instrument. In terms of quality, 108 out of 132 (82%) were of good reporting quality with a score between 7 and 8, and 24 studies scored 6 (18%). The risk of bias of included studies was assessed by QUADAS-2. Fifteen studies (11.4%) and 12 studies (9.1%) across 13 screening instruments were assessed as at high risk of bias on execution for the reference standard and the index test, respectively.

Diagnostic accuracy of the Two-Question Screen

Seven cohorts from six studies (4.9%) reported the diagnostic accuracy of the Two-Question Screen for depression of older adults.[44–49] All of them used one as the cut-off value. The sensitivities ranged from 79 to 100% and the specificities ranged from 44 to 84%. The data on diagnostic accuracy were summarised by meta-analysis (Table 2). The heterogeneity among studies was large, with I^2 statistics for sensitivity and specificity of 52.7 and 94.1%, respectively. The combined data in the bivariate

Table 1 Characteristics of the 16 depression screening instruments

Depression screening instrument	Items, *n*	Score range[a]	Rating scale	Standard cut-off point[b]	Administration time, min
Self-rating scale					
Two-Question Screen	2	0–2	Yes/no	⩾1	<5
Geriatric Depression Scale (GDS)-30	30	0–30	Yes/no	⩾10	10
GDS-15	15	0–15	Yes/no	⩾5	5–10
GDS-10	10	0–10	Yes/no	⩾4	5
GDS-4	4	0–4	Yes/no	⩾1	<5
Beck Depression Inventory	21	0–63	0–3	⩾10	10
Hospital Anxiety and Depression scale – Depression subscale	7	0–21	0–3	⩾8	5
Patient Health Questionnaire (PHQ)-9	9	0–27	0–3	⩾10	5
PHQ-2	2	0–6	0–3	⩾3	<5
Center for Epidemiological Depression Scale (CEDS)-20	20	0–60	0–3	⩾16	20
CEDS-10	10	0–30	0–3	⩾10	10
Even Briefer Assessment Scale for Depression	8	0–8	Yes/no	⩾7	5
One-Question Screen[c]	1	0–1	Yes/no	⩾1	<5
Clinician-rated scale					
Hamilton Rating Scale for Depression	17	0–54	0–4	⩾8	20
Montgomery–Åsberg Depression Rating Scale	10	0–60	0, 2, 4, 6	⩾7	15
Informant and clinician-rated scale					
Cornell Scale for Depression in Dementia	19	0–38	0–2	⩾6	30

a. High scores represent more severe depression.
b. This is the first cut-off point for depression if an instrument has multiple cut-off points.
c. The one question is about sad and depressed mood.

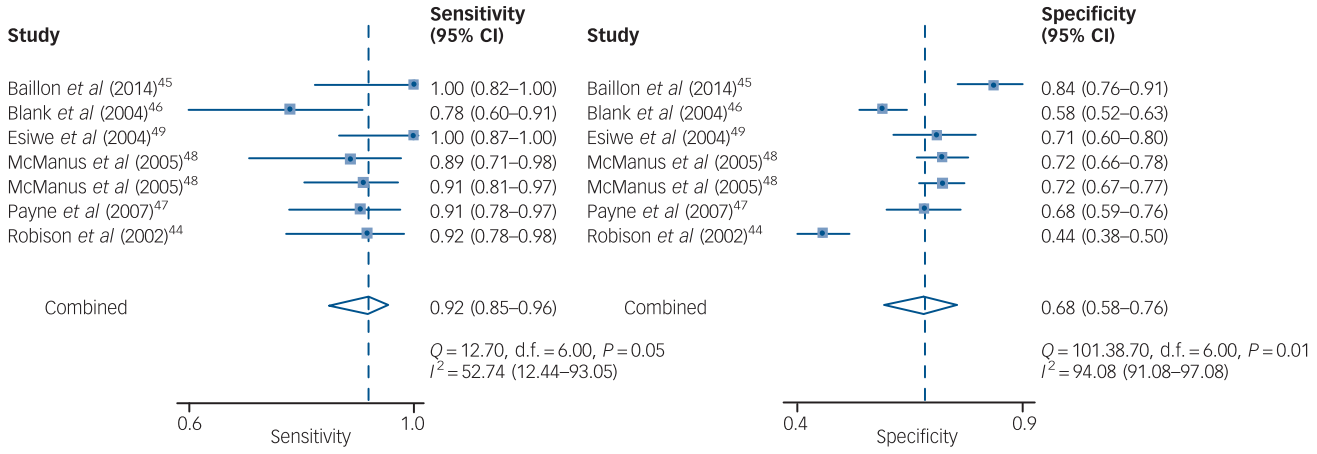

Fig. 1 Forest plot for the pooled sensitivity and specificity of the Two-Question Screen.

https://doi.org/10.1192/bjp.bp.116.186932 Published online by Cambridge University Press

Table 2 Meta-analyses for diagnostic accuracy on depression screening instruments for older adults

Screening instruments	Study cohorts, *n*	Pooled sensitivity, % (95% CI)	Pooled specificity, % (95% CI)	Pooled positive LR (95% CI)	Pooled negative LR (95% CI)	Diagnostic OR (95% CI)
Self-rating scale						
Two-Question Screen	7	91.8 (85.2–95.6)	67.7 (58.1–76.0)	2.84 (2.09–3.86)	0.12 (0.06–0.24)	23.55 (9.41–58.94)
Geriatric Depression Scale (GDS)-30	37	82.8 (80.7–87.5)	72.2 (63.1–80.0)	3.00 (2.28–3.89)	0.24 (0.19–0.30)	12.51 (8.86–17.67)
GDS-15	49	84.4 (80.5–87.4)	77.4 (72.1–82.0)	3.73 (3.00–4.65)	0.20 (0.16–0.25)	18.56 (12.72–27.1)
GDS-10	6	84.8 (58.6–93.4)	59.4 (36.8–78.6)	2.09 (1.36–3.20)	0.26 (0.16–0.42)	8.13 (5.19–12.74)
GDS-4	12	88.4 (81.1–93.2)	63.4 (51.2–74.1)	2.42 (1.79–3.26)	0.18 (0.11–0.29)	13.24 (7.21–24.30)
Beck Depression Inventory	16	85.7 (77.3–91.4)	73.5 (55.8–85.9)	3.24 (1.89–5.16)	0.19 (0.12–0.30)	16.66 (7.86–35.33)
Hospital Anxiety and Depression scale – Depression subscale	18	79.0 (70.1–85.8)	77.7 (71.5–82.9)	3.55 (2.68–4.70)	0.27 (0.18–0.40)	13.12 (7.25–23.70)
Patient Health Questionnaire (PHQ)-9	14	83.4 (77.4–88.1)	85.8 (80.3–90.0)	5.89 (4.16–8.34)	0.19 (0.14–0.29)	30.49 (17.30–53.74)
PHQ-2	11	84.6 (71.3–92.4)	79.3 (69.8–86.5)	4.09 (2.72–6.15)	0.19 (0.10–0.38)	21.15 (8.67–51.60)
Center for Epidemiological Depression Scale (CEDS)-20	16	79.7 (74.3–84.2)	76.5 (68.7–82.8)	3.39 (2.56–4.56)	0.27 (0.21–0.34)	12.79 (8.14–20.08)
CEDS-10	5	85.5 (71.0–93.4)	79.0 (68.0–87.0)	4.08 (2.73– 6.09)	0.18 (0.09–0.37)	22.24 (10.38–47.69)
Even Briefer Assessment Scale for Depression	4	82.0 (54.2–94.6)	91.2 (52.0–99.0)	9.30 (1.32–65.58)	0.20 (0.07–0.55)	47.20 (6.47–344.64)
One-Question Screen	12	66.4 (58.1–73.8)	82.1 (72.9–88.6)	3.70 (2.50–5.48)	0.41 (0.33–0.50)	9.04 (5.59–14.60)
Clinician-rated scale						
Hamilton Rating Scale for Depression	16	88.6 (82.0–93.0)	84.9 (80.6–88.3)	5.86 (4.53–7.58)	0.13 (0.08–0.21)	43.79 (24.00–79.20)
Montgomery–Åsberg Depression Rating Scale	8	81.3 (75.8–85.8)	81.5 (71.2–88.8)	4.40 (2.79–6.95)	0.23 (0.18–0.30)	19.17 (10.95–33.57)
Informant and clinician-rated scale						
Cornell Scale for Depression in Dementia	11	88.4 (79.2–93.8)	81.6 (70.0–90.7)	4.80 (2.48–9.29)	0.14 (0.07–0.27)	33.70 (10.80–105.13)

LR, likelihood ratio; OR, odds ratio.

random-effects model gave a summary point with a sensitivity of 91.8% (95% CI 85.2–95.6) and a specificity of 67.7% (95% CI 58.1–76.0) (Fig. 1). The HSROC curve was plotted with a diagnostic OR = 23.6, and the AUC was 90% (95% CI 87–92) (Fig. 2). The pooled positive likelihood ratio was 2.84 (95% CI 2.09–3.86) and the pooled negative likelihood ratio was 0.12 (95% CI 0.06–0.24).

Diagnostic accuracy of the other screening instruments

The majority of the screening instruments were self-rating scales. The GDS-30 (37 cohorts, 25.9%) and GDS-15 (49 cohorts, 34.3%) were the most frequently used screening instruments for academic studies. The pooled sensitivity and specificity were 82.8% (95% CI 80.7–87.5) and 72.2% (95% CI 63.1–80.0) for GDS-30, and 84.4% (95% CI 80.5–87.4) and 77.4% (95% CI 72.1–82.0) for GDS-15. Other short forms of the GDS, BDI, Hospital Anxiety and Depression Scale – Depression subscale (HADS-D), Patient Health Questionnaire (PHQ), CEDS and the One-Question Screen were the other common screening instruments (Table 2). For clinician-rated screening instruments, the HDRS (16 cohorts, 11.1%) and MADRS (8 cohorts, 5.6%) were found. The pooled sensitivity and specificity were 88.6% (95% CI 82.0–93.0) and 84.9% (95% CI 80.6–88.3) for the HDRS, and 81.3% (95% CI 75.8–85.8) and 81.5% (95% CI 71.2–88.8) for the MADRS, respectively. The pooled sensitivity and specificity of the CSDD (11 cohorts, 8%) were 88.4% (95% CI 79.2–93.8), 81.6% (95% CI 70.0–90.7), respectively.

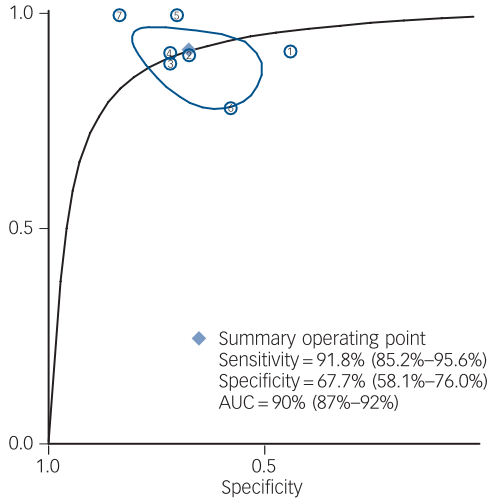

Fig. 2 Hierarchical summary receiver-operating characteristic (HSROC) curve demonstrating the summary points for sensitivity and specificity of the Two-Question Screen.

The numbers 1 to 7 represent each of the seven cohorts included in this analysis.[44–49] AUC, area under the curve.

Subgroup analyses

In total, 51 studies included participants with major depressive disorder, and 9 instruments were identified for subgroup analysis (Table 3). Three self-rating scales, including the Two-Question Screen, PHQ-2 and GDS-15, showed relative good diagnostic performance. The sensitivity and specificity were 89.8% (95% CI 84.4–93.4) and 66.2% (95% CI 56.2–74.9) for the Two-Question Screen; 96.8% (95% CI 45.2–99.9) and 76.6% (95% CI 38.4–94.5) for the PHQ-2; 89.6% (95% CI 82.8–93.9) and 75.2% (95% CI 60.6–85.6) for the GDS-15, respectively.

Most of the studies included participants recruited in nursing homes or clinic settings (online Table DS2). Seven out of 12 cohorts were screened with the GDS-4 and showed better diagnostic performance in the subgroup analysis. Compared with the overall results, the sensitivity increased from 88.0 to 89.2%; and the specificity increased from 66.8 to 77.2%. However, the changes did not reach statistical significance. Among participants recruited in community settings, only four instruments (GDS-15, GDS-30, CEDS-20 and PHQ-2) provided sufficient data for this subgroup

https://doi.org/10.1192/bjp.bp.116.186932 Published online by Cambridge University Press

Table 3 Meta-analyses for diagnostic accuracy for major depressive disorder in older adults

Screening instruments	Study cohorts, *n*	Pooled sensitivity, % (95% CI)	Pooled specificity, % (95% CI)	Pooled positive LR (95% CI)	Pooled negative LR (95% CI)	Diagnostic OR (95% CI)
Self-rating scale						
Two-Question Screen	6	89.8 (84.4–93.4)	66.2 (56.2–74.9)	2.65 (1.97–3.58)	0.15 (0.09–0.25)	17.15 (8.19–35.88)
Geriatric Depression Scale (GDS)-30	16	81.6 (67.4–90.5)	71.1 (53.0–85.6)	2.93 (1.79–4.77)	0.25 (0.16–0.40)	11.49 (7.14–18.47)
GDS-15	13	89.6 (82.8–93.9)	75.2 (60.6–85.6)	3.61 (2.20–5.93)	0.14 (0.08–0.23)	26.11 (12.21–55.83)
Beck Depression Inventory	8	85.7 (68.4–94.4)	59.8 (24.6–87.1)	2.13 (0.94–4.85)	0.24 (0.12–0.47)	8.98 (2.78–28.99)
Hospital Anxiety and Depression scale – Depression subscale	9	83.6 (77.2–88.5)	80.9 (73.9–86.4)	4.38 (3.15–6.06)	0.20 (0.14–0.29)	21.62 (12.42–37.6)
Patient Health Questionnaire-2	7	96.8 (45.2–99.9)	76.6 (38.4–94.5)	4.14 (1.27–13.54)	0.04 (0.02–1.03)	98.82 (7.52–1298.64)
Center for Epidemiological Depression Scale-20	10	87.5 (73.8–94.5)	50.5 (15.4–85.1)	1.77 (0.82–3.82)	0.25 (0.16–0.39)	7.12 (2.48–20.47)
One-Question Screen	5	66.9 (52.8–78.5)	77.1 (56.1–89.9)	2.92 (1.52–5.61)	0.43 (0.31–0.59)	6.79 (3.06–15.05)
Clinician-rated scale						
Hamilton Rating Scale for Depression	4	81.5 (74.7–86.8)	85.4 (78.3–90.4)	5.57 (3.64–8.51)	0.22 (0.15–0.30)	25.68 (13.46–49.01)

LR, likelihood ratio; OR, odds ratio.

analysis. Although the PHQ-2 showed improved sensitivity and specificity, the subgroup results were only extracted from four cohorts. The changes also did not reach statistical significance.

Discussion

Main findings

This meta-analysis included 132 studies with 143 cohorts comparing the accuracy of 16 screening instruments for detection of depression in older adults. The results demonstrated that all screening instruments, except the One-Question Screen, showed good diagnostic accuracy. Our results supported the recommendation of NICE[19] of using the Two-Question Screen for depression screening.

In this study, the GDS was found to be the most frequently used instrument for depression screening. The short form (GDS-4) and long form (GDS-15, GDS-30) showed comparable performance and thus the short form may be preferred. Both the PHQ-2[50] and the Two-Question Screen showed good diagnostic performance. Although they use the same questions, the rating method of the PHQ-2 uses four discrete possible answers to gauge severity, whereas the Two-Question Screen uses just the answers 'Yes' or 'No'. Therefore, we did not combine them as one screening instrument, and the Two-Question Screen is easier to use in clinical practice. The One-Question Screen is a shorter version but its diagnostic performance was the lowest ranked among the screening instruments. Another study has demonstrated that screening with one question had lower diagnostic performance than screening with two questions.[51]

Lower cut-off values improve diagnostic sensitivity but with a corresponding decrease in specificity. High sensitivity corresponds to high negative predictive value, which is ideal to rule out depression. We found variation in the optimal cut-off values among the studies of most of the depression screening instruments. Clinicians faced the difficult dilemma to either choose the more appropriate cut-off value to either rule in or rule out depression. In the Two-Question Screen, all of the included studies used one as the cut-off value, so the interpretation of the Two-Question Screen is simple and made it easy to compare its usefulness among various studies. It is also a self-rating instrument that does not require any input from clinicians or specialists. As a result, the Two-Question Screen is favourable in practice.

Strengths and limitations

A strength of this paper is that we carried out a comprehensive literature search and included 132 studies with 46 506 patients but there were also several limitations. First, the depression screening instruments were translated into different languages. Although it is assumed that all instruments were validated before their use for screening, there may still have been cultural differences during the interview or self-administration. Second, participants may have had different levels of depression before the screening, but the details were not well documented. We performed subgroup analyses across different recruitment settings, and hoped to reduce the heterogeneity across baseline depression levels. Third, the performances of different screening instruments were not directly compared in the same population of participants in this study and we could only find a few papers with head-to-head comparisons between different screening instruments. Since there was only a limited number of studies, we were unable to perform subgroup analysis. Finally, some unpublished studies may not have been identified through the literature searches in OVID databases and there may have been publication bias.

Implications

In conclusion, this meta-analysis shows that self-rating scales have comparable diagnostic performance with clinician-rated scales. When considering diagnostic performance and administrative convenience, the Two-Question Screen is simple and reliable when screening for depression in older adults. Therefore, it is favourable to use the Two-Question Screen in older adult screening programmes.

Kelvin K. F. Tsoi, PhD, Jockey Club School of Public Health and Primary Care and Stanley Ho Big Data Decision Analytics Research Centre, The Chinese University of Hong Kong, Shatin, Hong Kong; **Joyce Y. C. Chan**, MPH, Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, Hong Kong; **Hoyee W. Hirai**, MSc, Stanley Ho Big Data Decision Analytics Research Centre, The Chinese University of Hong Kong, Shatin, Hong Kong; **Samuel Y. S. Wong**, MPH, MD, Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, Hong Kong

Correspondence: Samuel Wong, 4/F, Jockey Club School of Public Health and Primary Care, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong. Email: yeungshanwong@cuhk.edu.hk

First received 28 Apr 2016, final revision 2 Sep 2016, accepted 13 Nov 2016

References

1 Barua A, Ghosh MK, Kar N, Basilio MA. Prevalence of depressive disorders in the elderly. *Ann Saudi Med* 2011; **31**: 620–4.

2 Lamers F, Jonkers CCM, Bosma H, Penninx BW, Knottnerus JA, van Eijk JT. Summed score of the Patient Health Questionnaire-9 was a reliable and valid

https://doi.org/10.1192/bjp.bp.116.186932 Published online by Cambridge University Press

method for depression screening in chronically ill elderly patients. *J Clin Epidemiol* 2008; **61**: 679–87.

3 Dennis M, Kadri A, Coffy J. Depression in older people in the general hospital: a systematic review of screening instruments. *Age Aging* 2012; **41**: 148–54.

4 Azulai A, Walsh CA. Screening for geriatric depression in residential care facilities: a systematic narrative review. *J Gerontol Soc Work* 2015; **58**: 20–45.

5 Phelan E, Williams B, Meeker K, Bonn K, Frederick J, Logerfo J, et al. A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam Pract* 2010; **11**: 63.

6 Pyne JM, Patterson TL, Kaplan RM, Gillin JC, Koch WL, Grant I, et al. Assessment of the quality of life of patients with major depression. *Psychiatr Serv* 1997; **48**: 224–30.

7 Creed F, Morgan R, Fiddler M, Marshall S, Guthrie E, House A. Depression and anxiety impair health-related quality of life and are associated with increased costs in general medical inpatients. *Psychosomatics* 2002; **43**: 302–9.

8 Pignone MP, Gaynes BN, Rushton JL, Burchell CM, Orleans CT, Mulrow CD, et al. Screening for depression in adults: a summary of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med* 2002; **136**: 765–76.

9 O'Connor EA, Whitlock EP, Gaynes B, Beil TL. *Screening for Depression in Adults and Older Adults in Primary Care: An Updated Systematic Review*. Evidence Synthesis No 75. AHRQ Publication no 10–05143-EF-1. Agency for Healthcare Research and Quality, 2009.

10 O'Connor EA, Whitlock EP, Beil TL, Gaynes BN. Screening for depression in adult patients in primary care settings: a systematic evidence review. *Ann Intern Med* 2009; **151**: 793–803.

11 U.S. Preventive Services Task Force. Screening for depression in adults: recommendation statement. *Am Fam Physician* 2010; **82**: 976–9.

12 Mojtabai R. Diagnosing depression in older adults in primary care. *N Engl J Med* 2014; **370**: 1180–2

13 Watson LC, Pignone MP. Screening accuracy for late-life depression in primary care: a systematic review. *J Fam Pract* 2003; **52**: 956–64.

14 Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M. Development and validation of the geriatric depression scale: a preliminary report. *J Psychiatr Res* 1983; **17**: 37–49.

15 Allen N, Ames D, Ashby D, Bennetts K, Tuckwell V, West C. A brief sensitive screening instrument for depression in late life. *Age Ageing* 1994; **23**: 213–9.

16 Alexopoulos GA, Abrams RC, Young RC, Shamoian CA. Cornell Scale for Depression in Dementia. *Biol Psychiatry* 1988; **23**: 271–84.

17 O'Connor EA, Rossom RC, Henninger M, Groom HC, Burda BU, Henderson JT, et al. *Screening for Depression in Adults: An Updated Systematic Evidence Review for the U.S. Preventive Services Task Force*. AHRQ Publication No. 14-05208-EF-1. Agency for Healthcare Research and Quality, 2016.

18 Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry* 1961; **4**: 561–71.

19 National Institute for Health and Care Excellence. *Quick Reference Guide: Depression*. NICE, 2004.

20 Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression. Two questions are as good as many. *J Gen Intern Med* 1997; **12**: 439–45.

21 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.

22 Mitchell AJ, Meader N, Davies E, Clover K, Carter GL, Loscalzo MJ, et al. Meta-analysis of screening and case finding tools for depression in cancer: evidence based recommendations for clinical practice on behalf of the Depression in Cancer Care consensus group. *J Affect Disord* 2012; **140**: 149–60.

23 Meader N, Mitchell AJ, Chew-Graham C, et al. Case identification of depression in patients with chronic physical health problems: a diagnostic accuracy meta-analysis of 113 studies. *Br J Gen Pract* 2011; **61**: 808–20.

24 Sheikh JI, Yesavage JA. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. *Clin Gerontol* 1986; **5**: 165–73.

25 Radloff LS. The CES-D Scale: a self-administrated depression scale for research in the general population. *Appl Psychol Meas* 1977; **1**: 385–401.

26 Bird A, Macdonald A, Mann AH, Philpot MP. Preliminary experience with the SelfCARE(D): a self-rating depression questionnaire for use in elderly, non-institutionalized subjects. *Int J Geriatr Psychiatry* 1987; **2**: 31–8.

27 Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009; **151**: 264–9.

28 Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PMM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008; **149**: 889–97.

29 Macaskill P, Gatsonis C, Deeks JJ, et al. Analysing and presenting results. In *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0* (eds JJ Deeks, PM Bossuyt, C Gatsonis). The Cochrane Collaboration, 2010.

30 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (4th edn, revised) (DSM-IV-TR)*. American Psychiatric Association, 2000.

31 World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. WHO, 1993.

32 Copeland JRM, Kelleher MJ, Kellett J, Gourlay AJ, Gurland BJ, Fleiss JL, et al. A semi- structured interview for the assessment of diagnosis and mental state in the elderly. *Psychol Med* 1976; **6**: 439–49.

33 Copeland JR, Dewey ME, Henderson AS, Kay DW, Neal CD, Harrison MA, et al. The Geriatric Mental State (GMS) used in the community: replication studies of the computerized diagnosis AGECAT. *Psychol Med* 1988; **18**: 219–23.

34 Olin JT, Katz IR, Meyers BS, Schneider LS, Lebowitz BD. Provisional diagnostic criteria for depression of Alzheimer disease: rationale and background. *Am J Geriatr Psychiatry* 2002; **10**: 129–41.

35 Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; **155**: 529–36.

36 Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005; **58**: 982–90.

37 Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003; **56**: 1129–35.

38 Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001; **20**: 2865–84.

39 Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; **240**: 1285–93.

40 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; **7**: 177–88.

41 Rosman AS, Korsten MA. Application of summary receiver operating characteristics (sROC) analysis to diagnostic clinical testing. *Adv Med Sci* 2007; **52**: 76–82.

42 Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry* 1960; **23**: 56–62.

43 Montgomery SA, Åsberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry* 1979; **134**: 382–9.

44 Robison J, Gruman C, Gaztambide S, Blank K. Screening for depression in middle-aged and older Puerto Rican Primary care patients. *J Gerontol A Biol Sci Med Sci* 2002; **57**: 308–14.

45 Baillon S, Dennis M, Lo N, Lindesay J. Screening for depression in Parkinson's disease: the performance of two screening questions. *Age Ageing* 2014; **43**: 200–5.

46 Blank K, Gruman C, Robison JT. Case-finding for depression in elderly people: balancing ease of administration with validity in varied treatment settings. *J Gerontol* 2004; **59A**: 378–84.

47 Payne A, Barry S, Creedon B, Stone C, Sweeney C, O' Brien T, et al. Sensitivity and specificity of a Two-Question Screening tool for depression in a specialist palliative care unit. *Palliat Med* 2007; **21**: 193–8.

48 McManus D, Pipkin SS, Whooley MA. Screening for depression for patients with coronary heart disease (data from Heart and Soul Study). *Am J Cardiol* 2005; **96**: 1076–81.

49 Esiwe C, Baillon S, Rajkonwar A, Lindesay J, Lo N, Dennis M. Screening for depression in older adults on an acute medical ward: the validity of NICE guidance in using two questions. *Age Ageing* 2015; **44**: 771–5.

50 Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care* 2003; **41**: 1284–92.

51 Mitchell AJ, Coyne JC. Do ultra-short screening instruments accurately detect depression in primary care? A pooled analysis and meta-analysis of 22 studies. *Br J Gen Pract* 2007; **57**: 144–51.

EXTRA CONTENT ONLINE

https://doi.org/10.1192/bjp.bp.116.186932 Published online by Cambridge University Press