The European Journal of Psychiatry 35 (2021) 1–18

THE EUROPEAN JOURNAL OF PSYCHIATRY

www.elsevier.es/ejpsy

REVIEW ARTICLE

The performance and accuracy of depression screening tools capable of self-administration in primary care: A systematic review and meta-analysis

Peter Miller[a,1], **David Newby**[a,*,2], **Emily Walkom**[b,3], **Jenny Schneider**[c,4], **Shu Chuen Li**[a,5], **Tiffany-Jane Evans**[d,6]

[a] *Pharmacy and Experimental Pharmacology, School of Biomedical Sciences and Pharmacy, Faculty of Health and Medicine, University of Newcastle, Level 1, Medical Sciences Building, University Drive Callaghan, NSW 2308, Australia*
[b] *School of Medicine and Public Health, Faculty of Health and Medicine, University of Newcastle, Level 5, Clinical Sciences Building, Calvary Mater Newcastle Hospital, Edith Street, Waratah NSW 2298, Australia*
[c] *School of Medicine and Public Health, Faculty of Health and Medicine, University of Newcastle, The Hunter Medical Research Institute (HMRI), West Wing 3, Kookaburra Circuit, New Lambton Heights, NSW 2305, Australia*
[d] *Clinical Research Design, Information Technology and Statistical Support (CReDITSS), The Hunter Medical Research Institute (HMRI), Lot 1, Kookaburra Circuit, New Lambton Heights, NSW 2305, Australia*

Received 6 February 2020; accepted 19 October 2020
Available online 23 November 2020

KEYWORDS
Depression;
Screening;
Psychometric;
Primary healthcare;
Adult

Abstract
Background and Objectives: The US Preventative Services Taskforce recommends screening adults for depression in primary care where adequate systems are established to ensure accurate diagnosis, effective treatment and follow-up. However, there is currently no consensus on which screening tool is most suitable for use in primary healthcare. We aim to systematically review the literature for operating characteristics of depression screening tools capable of self-administration in primary healthcare and meta-analyse the psychometric characteristics of these tools to determine their performance and accuracy.

* Corresponding author.
E-mail address: david.newby@newcastle.edu.au (D. Newby).
[1] PM contributed to the acquisition of data, the analysis and interpretation of data, interpretation of risk of bias, and drafting and revision of the manuscript.
[2] DN contributed to the conception of the work, the analysis and interpretation of data, interpretation of risk of bias, and drafting and revision of the manuscript.
[3] EW contributed to the acquisition of data, interpretation of risk of bias, and revision of the manuscript.
[4] JS contributed to the drafting and revision of the manuscript.
[5] SCL contributed to the conception of the work and the drafting and revision of the manuscript.
[6] TE performed the statistical analysis of the psychometric properties for screening tools included in meta-analysis

https://doi.org/10.1016/j.ejpsy.2020.10.002

Methods: An electronic literature search of EMBASE, Medline and CINAHL Complete was conducted from January 1982 to September 15, 2019 using the keywords: depression, screening, primary healthcare and adult. General and psychometric characteristics were extracted for screening tools studied in primary healthcare only when assessed against a 'reference-standard'.
Results: Eighty-one studies from 22 countries were included in the review. Forty unique depression screening tools suitable for self-administration were identified in studies yielding 138 psychometric data sets. Based on ease of administration, 18 screening tools were suitable for use in primary healthcare. Of the tools meta-analysed, only the PHQ-9 and WHO-5 displayed superior accuracy and were easily administered.
Conclusion: Although numerous depression screening tools are suitable for use in primary care based on ease of administration, the PHQ-9 was the most widely assessed tool and displayed superior DOR, a-ROC, specificity and LR+. Our review supports the use of the PHQ-9 as a brief, easily administered depression screening tool with superior discriminatory performance and robust psychometric characteristics in primary care settings.

Introduction

Depression is the most common mental health disorder in the world,[1] with a reported prevalence in primary care patients between 5 and 10%.[2] The prevalence of depression in the USA increased from 3.3% in 1991 to 8.2% in 2008,[3,4] and in Australia from 6.8% to 10.3% in the ten years to 2008.[5] If the current trend continues, depression will become the second most prevalent condition causing morbidity globally by the end of 2020.[1] Research found 20% of Americans adults and 15% of Australians aged 16–85 years will experience depressive symptoms at least once in their lifetime.[6,7]

Depression exerts a significant financial burden on health care systems and society.[4,8] The human costs are also substantial, with depression affecting relationships as well as physical and emotional health and mortality. The mean life span of depressed patients has been reported to be 5–25 years shorter than the general population.[2,9] Individuals with depression also carry an increased risk of suicide,[7,10] and alcohol and drug abuse.[11] Despite this, depression continues to be under-detected and under-treated in primary care.[12–14] Patients with depression are more likely to initially visit a primary healthcare practitioner than be treated by a mental health professional.[15] However, primary healthcare practitioners find it challenging to identify patients with depression as it often presents with other physical symptoms.[16] The use of a screening tool may assist primary healthcare practitioners in improving the detection of depression in this setting.

{Coyne, 2001 #1562}In an effort to improve detection rates and reduce the disease burden, the US Preventative Services Taskforce [USPSTF] recommends screening to identify adults with depression in primary care settings where operational strategies exist to ensure appropriate diagnosis, referral and management of patients screening positive for depression.[17] Currently however, there is no consensus on which screening tool is most suitable for use in primary healthcare. Indeed, the USPSTF does not make any specific recommendation, but suggests any commonly used depression screening tool may be used.[17] Hence, a degree of uncertainty exists around choosing an appropriately validated depression screening tool for use in primary healthcare settings.

Many self-administered screening tools for depression have been developed and validated in various settings in the past four decades, with numerous studies comparing the accuracy and performance of one or several depression screening tools in primary care.[18–20] The most recent reviews on the characteristics of depression screening tools in primary care were undertaken in 2002 and 2018[21,22] The 2002 review evaluated the general characteristics of the tools identified, including administrative characteristics such as time to complete the tool, however focused primarily on the sensitivity and specificity of the screening tools and did not evaluate other useful psychometric parameters describing performance such as likelihood ratios and predictive values.[23–25] While the review in 2018 provided a more comprehensive evaluation of the psychometric properties of a wide range of instruments, it did not examine general characteristics relating to the ease of administration such as time taken, or level of literacy required, to complete the tool. Ideally, a tool that screens for depression in primary care should be acceptable to both the individual being screened[26] and the practitioner.[27,28] To minimise the administration burden, it should be brief, easy to understand and administer. However, the assumption that a brief, simple depression screening tool is more acceptable to both patient and practitioner has not been well researched.[20] Nevertheless, the ease of administration, reflected by characteristics such as time taken to complete the tool and levels of literacy required, are important considerations when choosing a depression screening tool in primary care.[29,30] Consequently, longer tools (≥15 questions),[31] tools with complex scoring methods, or tools considered more difficult for patients to understand, may limit their utility in primary healthcare settings by adding to the burden of administration.

To date, no review has systematically evaluated both measurable psychometric properties and the administrative operating characteristics of depression screening tools capable of self-administration in primary care. For the purposes of this manuscript we have defined administrative operating characteristics as those related to the administration of the tool including literacy level required, complexity of scoring, and time to complete the tool. The aim of this systematic review and meta-analysis was to comprehensively evaluate all characteristics of depression screening tools capable of self-administration to expand the evidence around these characteristics. Adding to this body of evidence will assist primary healthcare practitioners in making a more informed choice on the most suitable depression screening tool.

Methods

Search strategy and selection criteria

An electronic literature search was conducted using EMBASE, Medline and CINAHL Complete from January 1, 1982 to 15 September 2019. The starting date was chosen because depression screening in primary care became more frequently evaluated around that time.[32] The search used the terms 'depression' or 'major depression' or 'depressive symptoms' or 'depressive disorder' and 'screening' or 'screening tool' or 'screening instrument' or' screening test' or 'screening questionnaire' or 'risk assessment' and 'adult' or 'adults' and 'primary healthcare' or 'primary care' or 'general practice'. An example of the search strategy is provided as a supplementary file. References from included articles were also searched manually for additional articles satisfying the search criteria that were not captured by the electronic search. Inclusion and exclusion criteria are presented in Box 1. Studies evaluating tools requiring clinician expertise or standardized structured psychiatric interviews were excluded as they require considerable training and skill to administer,[16] and are generally more time consuming and not suitable for primary care screening. Specific depression-related conditions such as pre-menstrual dysphoric disorder, post-partum or manic depression, isolated dysthymia, or depression with other psychiatric disorders were excluded as they were not the focus of this review. While geriatric-specific tools were also excluded, studies with geriatric patients were included.

Box 1: Inclusion Exclusion criteria.

Inclusion criteria:	Exclusion criteria:
Written in the English language	Editorials, letters, surveys, conference abstracts, case reports
Adults aged 18 years or over	Reference (gold) standard not used
Screening for depression/major depression/ depressive disorder/depressive symptoms	Screening tools requiring clinician expertise
Screening tool compared to reference (gold) standard	Screening in hospital settings
Screening tool capable of self-administration	Population-specific screening tools (e.g geriatric, cardiology patients)
Sensitivity and specificity were reported	Pre-menstrual dysphoric disorder, post-partum depression, manic depression, depression with other psychiatric disorders
Screening conducted in primary health care settings	

Data extraction and quality assessment

Two reviewers (EW and PM) independently screened and extracted data from articles obtained from databases. After duplicates were removed using EndNote X8, titles and abstracts were reviewed against the inclusion criteria. Full texts of remaining articles were independently reviewed and included in the review after any differences were resolved by consensus. Data extraction from included articles was independently undertaken on the screening tool(s) administered, the reference standard, sample size, prevalence of depression, administrative operating characteristics, cut-off values and psychometric characteristics.

Quality assessment of studies was independently assessed by two reviewers (PM and EW or DN) using the revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool.[33] All QUADAS-2 signalling questions were used to assess the risk of bias (RoB) for patient selection, index test, reference standard and flow and timing. In keeping with reference standards used in studies, the appropriate interval was defined as 'in the past month' for the signalling question 'Was there an appropriate interval between the index test and reference standard'? Any discrepancies in risk of bias assessment were resolved through consensus.

Screening tool evaluation

Administrative operating characteristics, including the number of questions, administration time, ease of scoring and required level of literacy, were extracted to assess the ease of administration and suitability for use in primary care settings. Tools containing $\geq$15 questions were considered long,[31] and tools requiring logistic regression or linear transformation were considered difficult to score. Tools requiring higher levels of literacy were considered more difficult for respondents to answer. The level of literacy required for each tool as reported in the manuscript was extracted from relevant studies or from descriptive reviews,[34] as low respondent literacy levels may influence screening tool accuracy.[31]

Psychometric characteristics were extracted for each screening tool at the recommended cut-off point determined by developers of the tool wherever possible. If unavailable, the optimum cut-off point reported by the relevant study was extracted. The optimum cut-off point, in the opinion of the author of each study, represents the most suitable trade-off between sensitivity and specificity using the receiver operating characteristic curve to maximise the correct classification.[35] Sensitivity and specificity

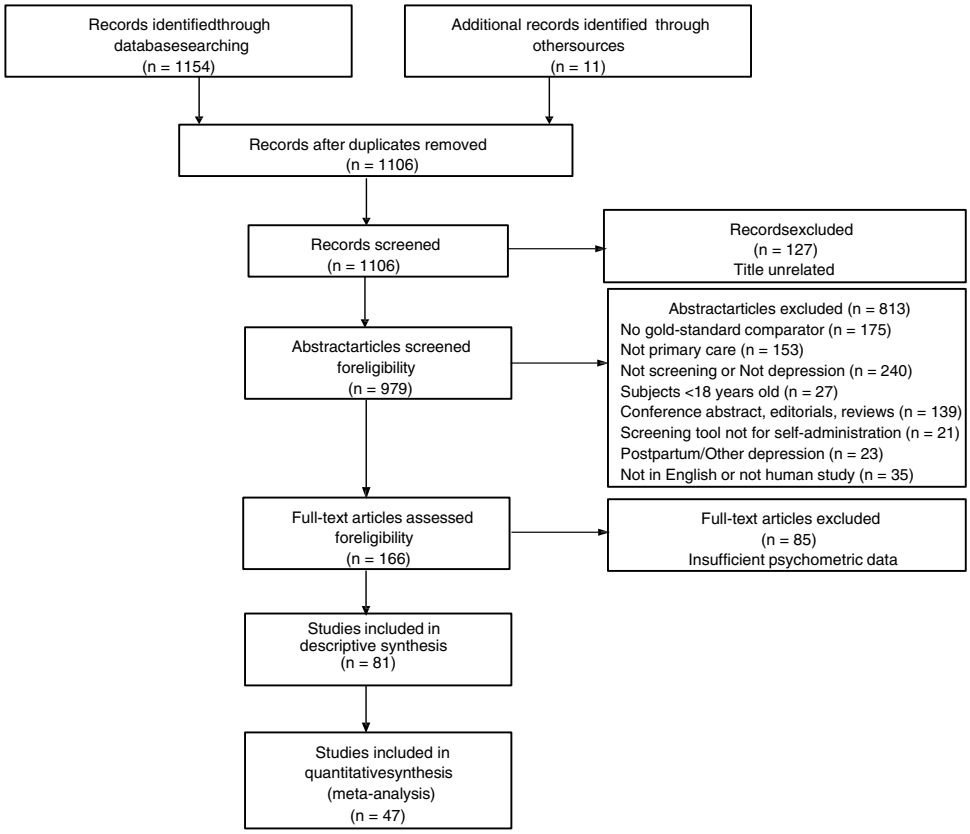

Fig. 1 PRISMA flow diagram of study selection process.

(Se%, Sp%) were extracted to identify tools with values >80%, as they are considered more effective at identifying people screening positive with depression and people screening negative without depression, respectively.[18,36] Positive and negative predictive values (PPV%, NPV%) were extracted to estimate the probability of obtaining correct screening results.[29] Area under the receiver operating characteristic curve (a-ROC) values were extracted to assess overall screening tool accuracy. Tools with a-ROC values $\geq$0.90 are considered to be highly accurate, a-ROC values of 0.7—0.9 indicate 'moderate to good' accuracy, and a-ROC values of 0.5—0.7 indicate low accuracy.[24] Positive and negative likelihood ratios (LR+, LR-) were extracted to estimate how much more (or less) likely patients with depression are to have the observed result than patients without depression.[23] A tool with a higher LR + suggests it is more likely someone with depression will have a positive screen than someone without depression and a tool with a lower LR- suggests it is more likely someone without depression will have a negative screen than someone with depression.[23] Missing predictive values and likelihood ratios were calculated using the sensitivity, specificity and prevalence of depression reported by the relevant study.

Data analysis

Meta-analysis was performed on screening tools only when evaluated by five or more studies where true positive and negative (TP, TN), and false positive and negative (FP, FN) counts were available. Only tools evaluated by $\geq$5 studies were meta-analysed because sufficient observations were required to estimate the underlying random effects with sufficient precision.[37] Diagnostic odds ratios (DOR's) were calculated by constructing 2 $\times$ 2 tables to determine TP, TN, FP and FN counts and using the formula described by Glas et al.[38] A DOR is a ratio of the odds of obtaining a positive screen if the person has depression relative to the odds of obtaining a positive screen if the person does not have depression.[38] Higher DOR values indicate the screening tool has better discriminatory performance than those with a lower DOR.[38] Meta-analyses of test accuracy used the Rutter and Gatsonis hierarchical summary ROC (HSROC) model to estimate the sensitivity, specificity, diagnostic odds ratio (DOR), positive and negative likelihood ratios, and summary ROC curves.[39] Heterogeneity was modelled by including covariates in the HSROC model and assessing whether the likelihood ratio test p-value indicated improved model fit. Statistical models were programmed using METADAS macro in SAS software v9.4,[40] in accordance with the Cochrane Handbook recommendations for meta-analyses of diagnostic test accuracy.[41] SAS METADAS macro was used to investigate heterogeneity. Heterogeneity was investigated only when $\geq$8 studies were available for that tool because sufficient studies are required to provide sufficient precision. HSROC curves and coupled forest plots were generated using RevMan5.[42]

Table 1 Studies of depression screening tools in primary healthcare.

Author. Year	Country	Tool	Cut-off (c/o)	Comparator & Criteria	Sample size	Outcome measured	Prevalence (%)
Ayalon, L. 2009	Israel	MDI PHQ-9 PHQ-1 MDI-1	≥ 21 ≥ 10 1 ≥1	SCID-I DSM-IV	153	MDD	3.9
Baer, L. 2000	USA	HANDS BDI-II Zung-SDS	≥ 9 ≥11 ≥50	SCID DSM-IV	45	Depression	64.0
Beekman, A.T. 1997	Holland	CESD-20	≥16	DIS DSM-III	487	MDD	2.0
Blank, K. 2004	USA	CESD-20 CESD-10 PHQ-2 yes/no Yale SQ	≥16 ≥ 4 ≥ 1 ≥ 1	DIS DSM-IV	125	MDD	11.0
Cameron, I.M. 2008	UK	BDI-II HADS-D PHQ-9 QID-SR_{16}	≥23 ≥ 9 ≥12 ≥13	HRSD 17 item (c/o ≥14)	282	Depression severity	47.5
Dutton, G.R. 2004	USA	BDI-II	≥14	DSM-IV interview	220	MDD	29.5
Evans, S. 1993	**UK**	GHQ-30	≥ 4	GMS-AGECAT	145	Depression	35.8
Fechner-Bates, S. 1994	USA	CESD-20	≥16	SCID DSM-IIIR	425	MDD	12.5
Gaynes, B.N. 2010	USA	M-3	≥ 5	MINI DSM-IV	647	MDD	16.0
Hanlon, C. 2015	UK	PHQ-2 PHQ-9 Kessler K10 Kessler K6	≥ 3 ≥ 6 ≥18 ≥ 9	MINI DSM-IV	306	MDD	5.9
Jirapramukpitak, T. 2009	Thailand	EURO-D	≥ 4	MINI DSM-IV	150	MDD	34.0
Klinkman, M.S. 1997	USA	CESD-20	≥ 4	DSM-IIIR	425	MDD	13.4
Lam, C.L. 1995	Hong Kong	HADS HADS-D	≥ 9 ≥ 6	CIS DSM-III	100	Depression	9.0
Lamoureux, B.E. 2010	USA	QID-SR_{16}	≥13	SCID DSM-IV	155	MDD	21.9
Lowe, B. 2004	Germany	HADS-D PHQ-9 WHO-5	≥ 8 ≥10 ≤ 8	SCID DSM-IV	501	MDD	13.2
Lowe, B. 2004	Germany	HADS-D PHQ-9 WHO-5	≥ 8 ≥10 ≤ 9	IDCL ICD-10	528	MDD	15.8
Lustman, P.J. 1997	USA	BDI-I	≥10	DIS-R DSM-III	172	MDD	36.6
Lyness, J.M. 1997	USA	CESD-20	≥21	SCID DSM-IIIR	130	MDD	9.2
McManus, D. 2005	USA	CESD-10 PHQ-9 PHQ-2 yes/no PHQ-2	≥10 ≥10 ≥ 1 ≥ 3	DIS DSM-IV	1024	MDD	21.9
Okimoto, J.T. 1982	USA	ID (Popoff) Zung-SDS	≥11 ≥60	DSM-III	55	MDD	30.9
Olsson, I. 2005	Norway	HADS-D	≥ 8	DSQ DSM-IV	1385	MDD	9.0
Rait, G. 1999	UK	BASDEC	≥ 7	GMS-AGECAT	130	Depression	10.0
Roberge, P. 2013	Canada	HADS HADS-D	≥16 ≥ 8	CIDI DSM-IV	1010* 660*	MDD	57.9*

Table 1 (*Continued*)

Author. Year	Country	Tool	Cut-off (c/o)	Comparator & Criteria	Sample size	Outcome measured	Prevalence (%)
Robison, J. 2002	USA	CESD-20 CESD-10 PHQ-2 yes/no Yale SQ	≥21 ≥ 4 ≥ 1 ≥ 1	CIDI DSM-IV	303	MDD	12.0
Schulberg, H.C. 1985	USA	CESD-20	≥16	DIS DSM-III	294	Depression	9.2
Sung, S.C. 2013	**Singapore**	QID-SR_{16} PHQ-9	≥ 9 ≥ 6	MINI DSM-IV	400	MDD	3.0
Thomas, J.L. 2001	USA	CESD-20	≥16	DIS DSM-IV	179	MDD	11.0
Upadhyaya, A.K. 1997	UK	HADS-D SelfCARE(D)	≥ 8 ≥ 5	GMS-AGECAT	72	Depression	27.8
Whooley, M.A. 1997	USA	BDI-I BDI-SF CESD-20 CESD-10 MOS-D SDDS-PC PHQ-2 yes/no	≥10 ≥ 5 ≥16 ≥10 ≥0.06 2 ≥ 1	DIS DSM-III	536	MDD	18.1
Wilhelm, K. 2004	Australia	BDI-FS	≥ 4	DSM-IV interview	212	MDD	11.8
Wilkinson, M.J. 1988	UK	HADS	≥ 8	CIS DSM-III	100	MDD	14.0
Williams, J.W. Jnr 1999	USA	CESD-20	≥16	DIS-R DSM-III	296	MDD	7.4
Williams, J.W. Jnr 1995	USA	SDS	0.043	SCID-IP DSM-III	221	MDD	4.0
Yang, Y. 2014	China	HADS-D HADS	≥ 9 ≥16	MINI(v5)DSM-IV	100	MDD	38.0
Yeung, A. 2002	USA	BDI-II	≥16	SCID-IP DSM-IIIR	180	MDD	29.4
Zich, J.M. 1990	USA	BDI-I CESD-20	≥10 ≥16	DIS DSM-III	31 34	MDD	9.7 5.9
Aragones-B, E. 2001	Spain	Zung-SDS	≥50	SCID DSM-IV	205	MDD	14.7
Arroll, B. 2003	New Zealand	PHQ-2 yes/no PHQ-1	≥1 1	CIDI DSM-IV	421	MDD	6.9
Arroll, B. 2010	New Zealand	PHQ-2 PHQ-9	≥ 3 ≥ 10	CIDI DSM-IV	2642	MDD	6.2
Awata, S 2007	Japan	WHO-5	≤13	SCID-I DSM-IV	129	MDD	10.7
N-Azah, M.N. 2005	Malaysia	PHQ-9	≥10	CIDI ICD-10	180	Depression	53.9
Azevedo-Marques, J. 2009	Brazil	WHO-5 SRQ-20	≤11 ≥8	SCID DSM-IV	120	Depression	20.8
Banerjee, S. 1998	UK	SelfCARE(D)	≥7/8	GMS-AGECAT	218	Depression	37.8
Broadhead,W.E. 1995	USA	SDDS-PC	any 2	SCID-P DSM-IIIR	388	MDD	15.7
Burnam, M.A. 1988	USA	MOS-D	≥ 0.06	DIS DSM-III	1416(ECA) 501 (PCP)	MDD	3.0 3.0
Campo-Arias, A. 2006	Colombia	Zung-SDS	≥49	SCID DSM-IV	266	MDD	16.5

Table 1 (*Continued*)

Author. Year	Country	Tool	Cut-off (c/o)	Comparator & Criteria	Sample size	Outcome measured	Prevalence (%)
Chen, S. 2010	China	PHQ-2 PHQ-9	≥ 3 ≥ 10	SCID DSM-IV	77	MDD	N/A
Chen, S, 2013	China	PHQ-9	≥10	SCID DSM-IV	280	MDD	N/A
Cheng, C.M. 2007	Hong Kong	PHQ-9 PHQ-2 yes/no	≥ 9 ≥ 1	Chinese HRSD 17 item(c/o≥16)	357	Depression	8.4
Corapcioglu, A. 2004	Turkey	PRIME-MD PHQ-9	Algorithm ≥10	DSM-IV interview	1387	MDD	6.6
Esler, D. 2008	Australia	PHQ-9 PHQ-2 yes/no PHQ-2	≥ 9 ≥ 1 ≥ 3	SCID DSM-IV	34 34 34	MDD Minor or MDD Minor or MDD	15.4 28.2 28.2
Gilbody, S. 2007	UK	PHQ-9	≥10	SCID DSM-IIIR	96	MDD	37.5
Goldberg, D.P. 1997	UK	GHQ-12	≥2	CIDI ICD-10	5438	Depression	24.0
Henkel, V. 2004	Germany	GHQ-12 WHO-5 PHQ-9	≥ 2 ≤13 ≥10	CIDI DSM-IV	448	MDD	10.2
Howe, A. 2000	UK	MHI-1	≥2	GMS-AGECAT	100	Depression	30.0
Inagaki, M. 2013	Japan	PHQ-9 PHQ-2	≥10 ≥ 3	MINI DSM-IIIR	104	MDD	7.4
Kroenke, K. 2003	USA	PHQ-2	≥ 3	SCID DSM-III	580	MDD	7.1
Kroenke, K. 2014	USA	PROMIS	≥ 8	PHQ DSM-IV	244	MDD	24.1
Kroenke, K. 2001	USA	PHQ-9	≥10	SCID DSM-III	580	MDD	7.1
Lamers, F. 2008	Holland	PHQ-9	≥ 7	MINI DSM-IV	620	MDD	N/A
Leung, K.K. 1998	Taiwan	Zung-SDS	≥55	DSM-IV interview	50	Depression	18.0
Lino, V.T. 2014	Brazil	PHQ-2	≥ 3	SCID-I DSM-IV	142	MDD	26.1
Liu, S.I. 2011	Taiwan	PHQ-9 PHQ-2	≥10 ≥ 3	SCAN DSM-IV	1954	MDD	3.3
Loerch, B. 2000	Germany	PRIME-MD	Yes	CIDI DSM-IV	704	MDD	18.6
Lotrakul, M. 2008	Thailand	PHQ-9	≥10	MINI DSM-IV	279	MDD	6.8
McQuaid, J.R. 2000	USA	CESD-20	≥16	CIDI DSM-IV	213	MDD	23.9
Means-Christensen, A.J. 2005	USA	MHI-5 MHI-1	≤23 ≤ 4	PHQ	246	MDD	6.9
Means-Christensen, A.J. 2006	USA	MHI-1	1	CIDI DSM-IV	801	MDD	39.8
Mergl, R. 2007	Germany	WHO-5 GHQ-12	≥13 ≥ 2	CIDI DSM-IV	394	Depression	22.8
Muhwezi, W.W. 2006	Uganda	SWB	≥10	MINI DSM-IV	234	MDD	31.6
Nagel, R. 1998	USA	MOS-D	n/a	DIS DSM-IIIR	147	MDD	2.8
Phelan, E. 2010	USA	PHQ-2 PHQ-9	≥ 3 ≥10	SCID DSM-IV	69	MDD	12.0
Picardi, A. 2013	Italy	PC-SAD	positive	SCID-I DSM-IV	212	MDD	N/A
Saipanish, R. 2009	Thailand	WHO-5	≤13	SCID DSM-IV	274	MDD	6.9
Schmitz, N. 1999	Germany	GHQ-12	≥ 2	SCID DSM-IIIR	408	Depression	5.2
Spitzer, R.L. 1994	USA	PRIME-MD PQ	Yes to 1 of 2 items	SCID DSM-III	431	MDD	14.0

Table 1 (*Continued*)

Author. Year	Country	Tool	Cut-off (c/o)	Comparator & Criteria	Sample size	Outcome measured	Prevalence (%)
Spitzer, R.L. 1999	USA	PHQ-9	algorithm	SCID DSM-III	585	MDD	N/A
Thapar, A 2014	UK	BDI-I	≥10	SCAN DSM-IV	272	MDD	22.2
		HADS-D	≥ 8				
		PHQ-2	≥ 3				
		PHQ-9	≥10				
		PHQ-4	≥3				
Yeung, A. 2008	USA	PHQ-9	≥15	SCID DSM-IV	184	MDD	22.8
Zhang, Y. 2013	Hong Kong	PHQ-9	≥10	MINI DSM-IV	99	MDD	23.2
Zuithoff. N 2010	Holland	PHQ-9	≥10	CIDI DSM-IV	1338	MDD	13.0
		PHQ-2	≥ 3				

Legend: MDD: Major depressive disorder, **N/A**: not available, *****: French speaking patients only.
CIDI : Composite International Diagnostic Interview, **DIS:** Diagnostic Interview Schedule, **DSM-III/IV** Diagnostic and Statistical Manual of Mental Disorders 3rd/4th edition, **GMS-AGECAT:** Geriatric Mental State-Automated Geriatric Examination for Computer Assisted Taxonomy, **HRSD :** Hamilton Rating Scale for Depression, **ICD-10 :** International Statistical Classification of Diseases, **IDCL :** International Diagnostic Checklist, **MINI:** Mini-International Neuropsychiatric Interview for DSM-IV, **SCAN:** Schedules for Clinical Assessment in Neuropsychiatry, **SCID :** The Structured Clinical Interview for DSM-IV Disorders.
BASDEC: Brief Assessment Schedule Depression Cards, **BDI-II:** Beck Depression Inventory, version 2 **BDI-FS:** Beck Depression Inventory-Fast Screen, **BDI-SF:** Beck Depression Inventory-Short Form, **CES-D:** Centre for Epidemiological Studies-Depression Scale (10 & 20 item), **Euro-D:** European Depression Scale, **GHQ:** General Health Questionnaire, **HADS:** Hospital Anxiety and Depression Scale, **HADS-D:** HADS depression subscale **HANDS:** Harvard Department of Psychiatry/National Depression Screening Day Scale **ID:** (Popoff) Index of Depression, **MDI:** Major Depression Inventory, **MHI-5:** Mental Health Inventory, **MOS-D:** Medical Outcomes Study Depression Scale, **PC-SAD:** Primary Care Screener for Affective Disorders, **PHQ:** Patient Health Questionnaire (1,2 yes/no, 2 & 9 item) **PRIME-MD:** Primary Care Evaluation of Mental Disorders, **PROMIS:** Patient Reported Outcomes Measurement Information System **QIDS-SR:** Quick Inventory of Depressive Symptomatology–Self-Report **SDDS-PC:** Symptom Driven Diagnostic System-Primary Care, **SDS:** Short Depression Screen **SelfCARE(D) SQ:** Single Question **SWB:** Subjective wellbeing subscale **WHO-5:** World Health Organization Wellbeing Index, **Zung-SDS:** Zung's Self-Rating Depression Scale.

Results

The electronic database search identified 1154 articles. Eleven additional articles were identified after examining references from full-text articles. Of these 1165 articles, 59 were duplicates and excluded, 940 were excluded after title and abstract review and eighty-five articles were excluded as psychometric data were unreported. Eighty-one studies satisfied the inclusion criteria and were included in the review. Forty-seven studies evaluated tools that were included in the meta-analysis. Fig. 1 shows the study selection process.

Twelve of the eighty-one studies evaluated three or more screening tools. Studies were conducted in 22 countries, including the USA (n = 32), UK (n = 11) and Germany (n = 6). Sample size ranged from 31 to 5438 people (median 265). Major depressive disorder (MDD) was the outcome measured in 79% of articles (n = 64), 'depression' in 20% (n = 16) and minor or major depression in 1% (n = 1) of studies. The prevalence of depression across studies in primary care varied from 2% to 64% (median = 12.8%). Table 1 summarises studies evaluating depression screening tools capable of self-administration in primary care.

The review identified forty unique depression screening tools and provided 138 psychometric data sets. The 9-item Patient Health Questionnaire (PHQ-9) was the most frequently evaluated tool (n = 27). Eleven tools assessed the presence of depressive symptoms over a 2-week period in accordance with Diagnostic and Statistical Manual of mental disorders (DSM) guidelines.[43] They included the four PHQ tools, Beck Depression Inventory version-2 (BDI-II), Beck Depression Inventory-fast screen (BDI-FS), Harvard Department of Psychiatry National Depression Screening Day Scale (HANDS), My Mood Monitor (M-3), Major Depression Inventory (MDI), Medical Outcomes Study Depression Scale (MOS-D) and World Health Organization Wellbeing Index (WHO-5). The most frequently used reference-standard was the Structured Clinical Interview for DSM Disorders (SCID) (n = 26).

Quality assessment

Twenty-three studies (28%) were considered good quality with a low RoB across all domains. Patient selection, reference standard and index test domains rated low RoB in the majority of studies. However, the flow and timing domain rated unclear RoB in 42 (53%) studies, due largely to limited reporting of the interval between administering the screening tool and receiving the reference-standard and because groups with and without depression were included in analysis in different proportions.Fig. 2 summarises the overall QUADAS-2 quality assessment results.

Administrative operating characteristics

Ten screening tools were classified as ultra-short (1–4 questions), nineteen were classified as short (5–14 questions), and eleven were classified as longer tools (≥ 15 questions).

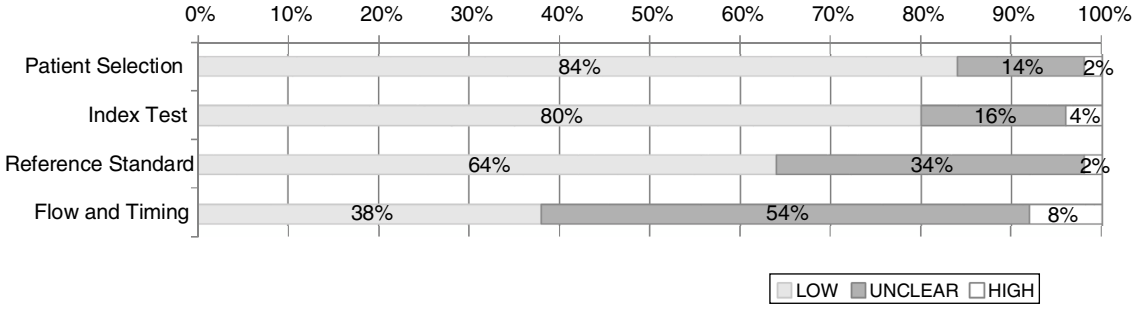

Fig. 2 QUADAS-2 risk of bias assessment summary.

Ultra-short tools took one to three minutes to complete and were easily scored except the Subjective wellbeing subscale (SWB), which did not report the ease of scoring. The level of literacy was considered 'easy' for the PHQ-1 and Yale single question tools and 'average' for the PHQ-2, PHQ-2(yes/no), PHQ-4, PROMIS, MDI-1, MHI-1, MHI-1(yes/no) and SWB. Of the ultra-short tools, only the PHQ tools met the DSM criteria. Short tools, on average, took five minutes to complete (range 2—10 min) and scoring was described as 'complex' for the MHI-5, MOS-D, SDS and SDDS-PC, 'average' for the WHO-5, and the remaining fourteen tools were simple to score. The level of literacy was classified as 'easy' for nine short tools (BDI-SF, BDI-FS, CESD-10, GHQ-12, HANDS, Kessler-K6 and K10, SDDS and SelfCARE(D)), eight were rated 'average' (EURO-D, M-3, MDI, MHI-5, MOS-D, PHQ-9, SDS and WHO-5) and two were rated 'difficult' (HADS, HADS-D). Six of the short tools (BDI-FS, HANDS, MDI, MOS-D, PHQ-9 and WHO-5) met the DSM criteria. Longer screening tools took 5—15 min to complete, were simple to score with the exception of the 37 item PC-SAD (mathematical algorithm required),[27] and required levels of literacy described as 'easy or average'. Of the longer screening tools, only the BDI-II met the DSM criteria. Overall, only the BDI-FS, HANDS, M-3, MDI, MDI-1, four PHQ tools and WHO-5 took ≤5 min to administer, had an 'easy or average' level of literacy, scoring was 'easy or average', and met the DSM criteria. Table 2 summarises the administrative operating characteristics of screening tools identified.

Psychometric characteristics

Psychometric characteristics varied between studies for the same tool and varied considerably from tool to tool. The recommended cut-off point was not applied in thirty-two screening tool evaluations. a-ROC values were available for 89 of 138 screening tool evaluations with values ranging from 0.68 to 0.97 (median = 0.88). The Zung SDS had the highest median a-ROC (0.92) across 2 studies screening 471 patients. The PHQ-9 had the highest a-ROC (0.97) in an individual study and the second highest median a-ROC (0.91) across 21 studies screening almost 6350 patients. Sensitivity values ranged from 30% to 100%, with 70 of 138 evaluations having a sensitivity >80%. Specificity values varied more widely, ranging from 12% to 98%, with 61 of 138 evaluations having a specificity >80%. PPV values (rule-in accuracy) ranged from 9% to 92%, with 5 of 138 evaluations having a PPV > 80%. NPV values (rule-out accuracy) ranged from 12% to 100%, with 121 of 138 evaluations having an NPV > 80%. Positive likelihood ratios (LR+) ranged from 1.14 to 36.50, with 41 of 138 evaluations having an LR+ ≥5, while negative likelihood ratios (LR-) ranged from <0.01 to 0.66, with 58 of 138 evaluations having an LR- ≤0.2. Table 3 summarises the psychometric characteristics of all screening tools included in the systematic review.

Authors of studies with low RoB across all domains for the tool evaluated in Table 3 are shown in bold font. Of the tools included in the meta-analysis, only the PHQ-9 and WHO-5 had a low RoB across all domains in ≥5 studies.

Meta-analysis

Six screening tools (CESD-20, HADS-D, PHQ-9, PHQ-2(scored), WHO-5 and Zung-SDS) were evaluated by five or more studies and underwent meta-analysis. Authors of studies included in the meta-analysis are shown in italic font in Table 3. Psychometric characteristics of screening tools analysed are summarised in Table 4. The PHQ-9 had the highest diagnostic odds ratio (DOR) (25.69), highest LR+ (6.79) and second-highest specificity (0.89). The WHO-5 had the second-highest DOR (19.70) and highest sensitivity (0.90), while the PHQ-2(scored) had the lowest DOR (12.78) but the highest specificity (0.91). A visual representation of the results of the meta-analysis as radar plots are provided as a supplemental file.

Considering the psychometric and administrative operating characteristics of tools meta-analysed together, only the PHQ-9 and WHO-5 could be administered in ≤5 min, had an 'easy or average' level of literacy, were 'easy or average' to score, assessed presence of symptoms over a 2 week period in accordance with DSM criteria and displayed better diagnostic accuracy than other tools meta-analysed.

Sampling method was not a source of heterogeneity in estimating pooled psychometric characteristics for the PHQ-9 but sampling method was a source of heterogeneity for CESD-20 analysis. Evaluation of heterogeneity lacked sufficient precision as estimates of random effects were zero for the PHQ-2(scored), WHO-5 and Zung-SDS and evaluation was underpowered for the HADS-D hence estimates for these four tools are of little use.

Discussion

This is the first systematic review of both the administrative operating characteristics and psychometric characteristics

Table 2 Administrative operating characteristics.

Screening Tool (Number of Studies)	Number of Items (class)	Symptom review period	Administration time (minutes)	Literacy level	Ease of scoring	Score range	Recommended cut-off ϕ
BASDEC (1)	19 (long)	Undisclosed	5—10	Average	Simple	0—19	>7
BDI-II (4)	21 (long) 21	Past 2 weeks	5—10	Easy	Simple	0–63 ¥	14+
BDI (4)	(long)	Past 7 days	5—10	Easy	Simple		10+
BDI-SF (1)	13 (short)	Past 7 days	5 – 7	Easy	Simple	0–39 ¥	≥5
BDI-FS (PC) (1)	7 (short)	Past 2 weeks	≤ 5	Easy	Simple	0—21 ¥	≥4
CESD-20 item (12)	20 (long)	Past 7 days	≤ 10	Easy	Simple	0 – 60	≥16
CESD-10 item (4)	10 (short)	Past 7 days	2 – 5	Easy	Simple	0–30	≥10
EURO-D (1)	12 (short)	Past 7 days	5	Average	Simple	0–12	≥4
GHQ-12 (4)	12 (short)	Past few weeks	5—10	Easy	Simple	0–12 ≠	≥2
GHQ-30 (1)	20 (long)	Past few weeks	10	Easy	Simple	0–30 ≠	≥4
HADS (4)	14 (short)	Past 7 days	5	Difficult	Simple	0—42	≥16
HADS-D (9)	7 (short)	Past 7 days	2—5	Difficult	Simple	0—21	≥8
HANDS (1)	10 (short)	Past 2 weeks	5—10	Easy	Simple	0–30	≥9
ID (1)	15 (long)	Recently	2 – 5	Easy	Simple	0—15	≥10
Kessler K6 (1)	6 (short)	Past 4 weeks	2—3	Easy	Simple	1 - 30	12 -14 (mild)
Kessler K10 (1)	10 (short)	Past 4 weeks	2—3	Easy	Simple	1 - 50	20—24 (mild)
M-3 (1)	7 (short)	Past 2 weeks	≤ 5	Average	Simple	0—28	≥5
MDI (1)	10 (short)	Past 2 weeks	≤ 5	Average	Simple	0—50	≥21
MHI-5 (1)	5 (short)	Past month	2—3	Average	Complex	0–30 (raw score)	≤23 Δ
MOS-D (3)	8 (short)	Past 2 weeks	2—3	Average	Complex	0 – 1€	≥0.06
PC-SAD (1)	37 (long) (includes 3 pre-screen items)	Past month	10 1–2 (pre-screen)	Easy	Complex	+ or -	Positive score (from algorithm)
PHQ-2 (12)	2 (ultrashort)	Past 2 weeks	1—2	Average	Simple	0—6	≥3
PHQ-2 Yes/No (7)	2 (ultrashort)	Past 2 weeks	1—2	Average	Simple	0—2	≥1
PHQ-4 (1)	4 (ultrashort)	Past 2 weeks	2—3	Average	Simple	0–12	≥3
PHQ-9 (27)	9 (short)	Past 2 weeks	≤ 5	Average	Simple	0 – 27 OR 1 of first 2 and ≥5 of 9 items	≥10 OR algorithm
PRIME-MD (3)	26 (long)	Past month	15 (part 1)	Average	Simple	positive responses	algorithm
PROMIS (1)	4 (ultrashort)	Past 7 days	2 – 3	Average	Simple (raw score)	4—20	≥8
QIDS-SR_{16} (3)	16 (long)	Past 7 days	5—7	Average	Simple	0—27	≥ 7
SDDS-PC (2)	5 (short)	Past month	1—2	Easy	Complex	0—5	any 2
SelfCARE(D) (2)	12 (short)	Past month	2 – 3	Easy	Simple	0–12	5—6
Single question	(ultrashort)						

Table 2 (*Continued*)

Screening Tool (Number of Studies)	Number of Items (class)	Symptom review period	Administration time (minutes)	Literacy level	Ease of scoring	Score range	Recommended cut-off ϕ
- MDI-1 (1)	1	Past 2 weeks	< 1	Average	Simple	0—5	≥2
- MHI-I (2)	1	Past year	< 1	Average	Simple	1 – 6	≥2
- MHI-1 (yes/no) (1)	1	Past year	< 1	Average	Simple	0—1	1
- Yale (2)	1	Past month	< 1	Easy	Simple	0—1	1
- PHQ-1 (2)	1	Past 2 weeks	< 1	Easy	Simple	0—1	
SDS (1)	8 (short)	Past week	unclear	Average	Complex	0—1€	0.06
SRQ-20 (1)	20 (long)	Past 30 days	unclear	Average	Simple	0—20	≥8
SWB (1)	4 (ultrashort)	Past week	unclear	Average	unclear	unclear	10
WHO-5 (7)	5 (short)	Past 2 weeks	2 – 5	Average	Average	0–25 raw score	≤13
Zung-SDS (5)	20 (long)	Past few days	5—10	Easy	Simple	25—100	≥50

Legend: BASDEC: Brief Assessment Schedule Depression Cards, **BDI-II:** Beck Depression Inventory- version 2, **BDI-FS:** Beck Depression Inventory-Fast Screen, **BDI-SF:** Beck Depression Inventory-Short Form.
CES-D: Centre for Epidemiological Studies-Depression Scale (10 & 20 item), **Euro-D:** European Depression Scale, **GHQ:** General Health Questionnaire, **HADS:** Hospital Anxiety and Depression Scale.
HADS-D: HADS depression subscale **HANDS:** Harvard Department of Psychiatry National Depression Screening Day Scale **ID:** (Popoff) Index of Depression, **MDI:** Major Depression Inventory.
MHI-5: Mental Health Inventory, **MOS-D:** Medical Outcomes Study Depression Scale, **PC-SAD:** Primary Care Screener for Affective Disorders, **PHQ:** Patient Health Questionnaire (1,2 yes/no, 2 & 9 item).
PRIME-MD: Primary Care Evaluation of Mental Disorders, **PROMIS:** Patient Reported Outcomes Measurement Information System **QIDS-SR:** Quick Inventory of Depressive Symptomatology–Self-Report.
SDDS-PC: Symptom Driven Diagnostic System-Primary Care, **SDS:** Short Depression Screen **SelfCARE(D) SQ:** Single Question **SWB:** Subjective Wellbeing scale **WHO-5:** World Health Organization Wellbeing Index, **Zung- SDS:** Zung's Self-Rating Depression Scale.
τ: Time-frame over which participants must reflect on when answering each question; **λ**, Literacy level using Fog Index; **ϕ**, Cut-off point for depression recommended by instrument developers ≠ Bimodal scoring method: Dichotomous or Likert scoring € calculated by logistic regression; **Δ**, 1–5 Likert scale raw score converted by linear transformation to 0–100; γ, BDI-II, BDI-SF & BDI-FS cut-offs for mild –moderate depression.

of depression screening tools capable of self-administration in primary care settings for fifteen years. It is also the first review to meta-analyse several psychometric characteristics of these tools where possible. Unlike previous reviews, quality assessment of all included studies was performed. We found the majority of studies had low RoB across most domains of quality and applicability, with the exception of flow and timing where a significant number were unclear primarily because of limited reporting. This provides greater confidence in the interpretation of our results.

This review identified forty unique depression screening tools capable of self-administration from eighty-one studies. In contrast, the 2002 review identified 16 screening tools from 38 studies,[21] while the 2017 review identified 55 screening tools from 60 studies.[22] We used less restrictive inclusion criteria regarding the aims of included studies therefore more studies were identified. However, we only included tools capable of self-administration where a reference standard was used to interpret screening results and only where studies reported psychometric data for depression screening. Hence, the number of studies included and the number of tools identified differed from the 2002 and 2017 reviews.

When assessing psychometric characteristics, previous reviews focused on sensitivity and specificity. Despite being useful measures of accuracy, these alone are not ideal as measurements vary according to the cut-off point chosen.[23] Unlike the previous reviews, our review systematically analysed DOR's and likelihood ratios (LR's) and evaluated a-ROC values when reported, in addition to sensitivity and specificity to provide a more thorough assessment of screening tool performance in primary care settings. Further, this is the first systematic review to meta analyse the psychometric properties including the sensitivity, specificity, likelihood ratios and diagnostic odds ratios, providing more precise estimates of these parameters. Diagnostic odds ratios, likelihood ratios and a-ROC values are considered more reliable methods of describing screening tool performance than sensitivity and specificity alone as they enable comparisons without selecting a particular cut-off point.[23,24,35,38]

Self-administered screening tools used in primary care should be brief, easily administered and easily scored.[28] They should also be easily understood by respondents, as screening tool accuracy may be influenced by lower levels of literacy.[26,31] Seventeen tools (42%) failed to satisfy one or more of these criteria, rendering them less suitable for use in primary care settings. Based solely on ease of administration, twenty-three tools were considered suitable for use in primary care. However, a meta-analysis of single-question screening tools found that, when used alone, they

Table 3 Psychometrics of included studies.

Instrument	Author Year	Cut-Off	Sensitivity (%) (95%CI)	Specificity (%) (95%CI)	PPV (%) (95%CI)	NPV (%) (95%CI)	LR +	LR -	AUC (95%CI)
BASDEC	Rait, G 1999	≥7	71.0	88.0	39.7	96.5	5.92	0.32	N/A
BDI-II	Baer, L 2000	≥11	96.0	31.0	71.0	81.0	1.39	0.13	N/A
	Cameron, IM 2008	≥20	84.0	68.0	72.0	82.0	2.65	0.24	0.85
	Dutton, GR 2004	≥14	87.7	83.9	69.5	94.2	5.44	0.15	0.91
	Yeung, A 2002	≥16	79.0	91.0	79.0	91.0	8.77	0.23	0.94
BDI	Thapar, A 2014 Lustman, PJ 1997	≥10	98.3	39.7	31.7	98.8	1.63	0.04	0.89
	Whooley, MA 1997	≥13	85.0	88.0	80.0	91.0	7.08	0.17	0.94
	Zich, JM 1990	≥10	89.0	64.0	63.9	96.4	2.50	0.17	0.87
		≥10	100.0	75.0	29.8	100.0	4.00	<0.01	N/A
BDI-SF	Whooley, MA 1997	≥ 5	92.0	61.0	34.1	97.2	2.40	0.13	0.86
BDI-FS(PC)	**Wilhelm, K 2004**	≥ 4	91.0	62.0	24.2	98.1	2.39	0.15	0.85
CES-D 20 item	*Beekman, AT 1998*	≥16	100	87.6	14.1	100.0	2.23	<0.01	N/A
	Blank, K 2004	≥16	79.0	75.0	28.1	96.6	3.10	0.29	0.86
	Fechner-Bates, S 1994	≥16	79.5	71.1	28.0	96.1	2.72	0.29	N/A
	Klinkman, MS 1997	≥16	80.7	71.7	30.7	96.0	2.85	0.27	N/A
	Lyness, JM 1997	≥21	92.0	87.0	41.8	99.1	7.08	0.09	0.94
	McQuaid, JR 2000	≥16	78.7	76.5	57.8	89.2	3.35	0.28	N/A
	Robison, J 2002	≥20	73.0	72.0	26.2	95.1	2.60	0.38	0.77
	Schulberg, HC 1985	≥16	96.3	38.6	13.7	98.9	1.57	0.10	N/A
	Thomas, JL 2001	≥16	95.0	70.0	28.4	99.1	3.17	0.07	0.88
	Whooley, MA 1997	≥16	93.0	69.0	39.7	97.8	3.00	0.10	0.89
	Williams, JW Jnr 1999	≥16	88.0	75.0	32.6	97.8	3.52	0.16	N/A
	Zich, JM 1990	≥16	100.0	53.0	12.0	100.0	2.12	<0.01	N/A
CES-D 10 item	**Blank, K 2004**	≥ 4	79.0	81.0	33.9	96.9	4.10	0.26	0.89
	McManus, D 2005	≥10	76.0	79.0	50.5	92.1	3.60	0.30	0.87
	Robison, J 2002	≥ 4	76.0	70.0	25.7	95.5	2.50	0.35	0.77
	Whooley, MA 1997	≥10	90.0	72.0	41.5	97.0	3.20	0.14	0.87
EURO-D	**Jirapramukpitak, T 2009**	≥ 5	84.3	58.6	55.9	80.2	2.00	0.27	0.78
GHQ-12	Goldberg, DP 1997	≥ 2	76.3	83.4	52.9	91.8	4.60	0.28	0.88
	Henkel, V 2004	≥ 2	85.0	63.0	34.0	95.0	2.30	0.23	0.87
	Mergl, R 2007	≥ 2	87.0	63.0	14.2	98.6	2.35	0.21	0.84
	Schmitz, N 1999	≥ 2/3	60.0	74.0	57.0	76.0	2.30	0.54	0.73
GHQ-30	Evans, S 1993	≥4	77.0	67.0	59.1	82.5	3.33	0.34	N/A
HADS	Lam, CL 1995	≥ 9	80.0	90.0	67.0	95.0	8.00	0.22	N/A
	Roberge, P 2013	≥16	62.0	77.0	55.0	82.0	2.70	0.49	0.76
	Wilkinson, MJ 1988	≥ 8	90.0	86.0	81.0	12.0	6.42	0.12	0.95
	Yang, Y 2014	≥16	66.7	76.5	32.3	92.8	2.84	0.44	0.86

Table 3 (*Continued*)

Instrument	Author Year	Cut-Off	Sensitivity (%) (95%CI)	Specificity (%) (95%CI)	PPV (%) (95%CI)	NPV (%) (95%CI)	LR +	LR -	AUC (95%CI)
HADS-D	*Cameron, IM 2008*	≥ 9	73.0	76.0	72.0	76.0	3.04	0.36	0.83
	Thapar, A 2014	≥ 8	85.2	68.2	43.3	94.2	2.68	0.22	0.86
	Lam, CL 1995	≥ 6	78.0	91.0	47.0	98.0	8.66	0.24	N/A
	Löwe, B 2004a (SCID)	≥ 8	88.0	69.0	30.1	97.4	2.84	0.17	0.89
	Löwe, B 2004b (IDCL)	≥ 8	87.0	70.0	35.2	96.6	2.90	0.18	0.88
	Olsson, L 2005	≥ 8	80.0	88.0	39.7	97.8	6.67	0.23	0.93
	Roberge, P 2013	≥ 8	56.0	80.0	41.0	88.0	2.80	0.55	0.75
	Upadhyaya, AK 1997	≥ 8/9	70.0	87.0	67.7	88.2	5.38	0.34	N/A
	Yang, Y 2014	≥ 8	80.0	90.6	41.9	97.1	8.51	0.22	0.94
HANDS	Baer, L 2000	≥ 9	96.0	60.0	81.0	89.0	2.40	0.07	N/A
ID (Popoff)	Okimoto, JT 1982	≥11	88.0	61.0	50.2	91.9	2.25	0.20	N/A
Kessler K10	Hanlon, C 2015	≥18	77.8	76.7	17.3	98.2	3.34	0.29	0.83
Kessler K6	Hanlon, C 2015	≥ 9	77.8	73.3	15.4	98.1	2.91	0.30	0.84
M-3	**Gaynes, BN 2010**	≥ 5	84.0	80.0	54.0	95.0	4.19	0.20	N/A
MDI	Ayalon, L 2009	≥21	83.3	97.2	55.0	99.0	29.75	0.17	N/A
MHI-5	Means-Christensen, A 2005	≤23	90.9	57.6	17.4	98.5	2.14	0.16	0.91
MOS-D	Burnam, MA 1988 (*PSP)	≥0.06	86.0	90.0	20.0	99.5	8.60	0.16	N/A
	Nagel, R 1998	N/A	100.0	77.0	11.0	100.0	4.34	<0.01	N/A
	Whooley, MA 1997	≥0.06	93.0	72.0	42.0	93.3	3.30	0.10	0.89
PC-SAD	Picardi, A 2013	positive	89.8	82.6	51.2	97.6	5.20	0.12	N/A
PHQ-4	Thapar, A 2014	≥3	93.4	67.9	45.4	97.3	2.91	0.10	0.91
PHQ-2 (scored)	***Arroll, B 2010***	≥ 3	61.0	92.0	33.5	97.3	7.70	0.42	N/A
	Chen, S 2010	≥ 3	84.0	90.0	N/A	N/A	8.40	0.18	0.92
	Esler, D 2008	≥ 3	30.0	70.8	30.0	70.8	1.02	0.98	N/A
	Thapar, A 2014 Hanlon, C 2015	≥ 3	72.1	82.1	53.5	91.2	4.02	0.34	0.87
	Inagaki, M 2013	≥ 3	33.3	93.1	23.1	95.7	4.82	0.72	0.78
	Kroenke, K 2003	≥ 3	61.0	98.0	66.0	97.0	24.75	0.40	0.95
	Lino, VT 2014	≥ 3	82.9	90.0	38.4	98.6	2.90	0.19	0.93
	Liu, SI 2011	≥ 3	52.0	87.0	55.0	85.0	4.00	0.13	0.77
	McManus, D 2005	≥ 3	64.0	93.6	25.4	98.7	10.00	0.38	0.90
	Phelan, R 2010	≥ 3	39.0	92.0	57.7	84.3	4.90	0.66	0.84
	Zuithoff, N 2010	≥ 3	63.0	85.0	36.4	94.4	4.20	0.44	0.81
		≥ 3	42.0	94.0	53.0	91.0	7.00	0.62	0.83
PHQ-2 yes/no	**Arroll, B 2003**	≥ 1	97.0	67.0	17.9	95.7	2.90	0.05	N/A
	Blank, K 2004	≥ 1	79.0	58.0	18.9	99.7	1.90	0.37	0.72
	Cheng, CM 2007	≥ 1	96.7	73.4	**25.0**	99.6	3.63	0.04	N/A
	Esler, D 2008	≥ 1	100.0	12.5	32.3	100.0	1.14	<0.01	N/A
	McManus, D 2005	≥ 1	90.0	69.0	44.9	95.7	2.90	0.14	0.84
	Robison, J 2002	≥ 1	92.0	44.0	18.3	97.6	1.60	0.18	0.68
	Whooley, MA 1997	≥ 1	96.0	57.0	33.0	98.5	2.20	0.07	0.82

Table 3 (*Continued*)

Instrument	Author Year	Cut-Off	Sensitivity (%) (95%CI)	Specificity (%) (95%CI)	PPV (%) (95%CI)	NPV (%) (95%CI)	LR +	LR -	AUC (95%CI)
PHQ-9	***Arroll, B 2010***	≥10	74.0	91.0	35.2	98.1	8.40	0.28	N/A
	Ayalon, L 2009	≥10	66.6	98.6	67.0	99.0	7.40	0.34	N/A
	N-Azah, N 2005 Cameron, IM 2008	≥10	60.9	80.7	38.6	78.1	3.15	0.48	0.74
	Chen, S 2010	≥10	87.0	69.0	70.0	86.0	2.80	0.19	0.88
	Chen, S 2013	≥10	75.0	89.0	N/A	N/A	6.81	0.28	0.92
	Cheng, CM 2007	≥10	87.0	81.0	N/A	N/A	5.04	0.31	0.91
	Corapcioglu, A 2004	≥ 9	80.0	92.0	N/A	N/A	10.00	0.22	N/A
	Esler, D 2008	≥10	71.4	91.9	38.2	97.9	8.81	0.31	N/A
	Gilbody, S 2007	≥ 9	80.0	71.4	33.3	95.2	2.80	0.28	N/A
	Thapar, A 2014	≥10	91.7	78.3	71.7	94.0	4.23	0.11	0.94
	Hanlon, C 2015	≥10	85.2	76.9	51.2	94.8	3.69	0.19	0.90
	Henkel, V 2004	≥ 6	77.8	80.6	20.0	98.3	4.01	0.28	0.85
	Inagaki, M 2013	algorithm	79.0	86.0	55.0	95.0	5.47	0.24	0.91
	Kroenke, K 2001	≥10	45.0	99.0	72.0	96.0	32.84	0.55	0.93
	Lamers, F 2008	≥10	88.0	88.0	35.9	99.0	7.10	0.14	0.95
	Liu, SI 2011	≥ 7	92.2	78.1	41.6	98.3	4.21	0.10	0.92
	Lotrakul, M 2008	≥10	86.0	93.9	32.5	99.5	14.1	0.15	0.96
	Löwe, B 2004a (SCID) Löwe, B 2004b (ICDL)	≥10	74.0	85.0	27.0	98.0	5.04	0.31	0.89
		≥11	98.0	80.0	42.7	99.6	4.90	0.03	0.95
	McManus, D 2005	≥10	90.0	77.0	42.3	97.6	3.91	0.13	0.92
	Phelan, E 2010	≥10	54.0	90.0	60.2	87.4	5.40	0.51	0.86
	Spitzer, RL 1999	≥10	63.0	82.0	32.3	94.2	3.50	0.46	0.87
	Sung, SC 2013	algorithm	73.0	98.0	80.2	97.0	36.5	0.27	N/A
	Yeung, A 2008	≥ 6	91.7	72.2	9.2	99.6	3.30	0.11	0.82
	Zhang, Y 2013	≥15	81.0	98.0	92.0	95.0	40.5	0.19	0.97
	Zuithoff, N 2010	≥10	56.5	84.2	42.1	90.5	3.58	0.52	0.85
		≥10	49.0	95.0	59.0	93.0	9.80	0.54	0.89
PRIME-MD (part 1)	Corapcioglu, A 2004	positive	52.2	89.7	57.1	87.6	5.07	0.53	N/A
	Loerch, B 2000	positive	68.0	84.0	50.0	92.0	4.25	0.38	N/A
	Spitzer, RL 1994	positive	57.0	98.0	80.1	99.5	28.5	0.44	N/A
PROMIS	Kroenke, K 2014	≥ 8	83.0	84.0	62.5	93.9	5.29	0.20	0.89
QIDS-SR_{16}	Cameron, IM 2008	≥11	86.0	68.0	70.0	85.0	2.68	0.21	0.89
	Lamoureux, BE 2010	≥13	76.5	81.8	54.2	92.5	4.20	0.29	0.82
	Sung, SC 2013	≥ 9	83.3	84.7	14.5	99.4	5.40	0.20	0.84
SDDS-PC	Broadhead, WE 1995	2	90.4	77.2	39.7	98.0	3.96	0.12	N/A
	Whooley, MA 1997	2	96.0	51.0	30.0	98.3	2.0	0.08	0.86

Table 3 (*Continued*)

Instrument	Author Year	Cut-Off	Sensitivity (%) (95%CI)	Specificity (%) (95%CI)	PPV (%) (95%CI)	NPV (%) (95%CI)	LR +	LR -	AUC (95%CI)
SelfCARE-D	Banerjee, S 1997	≥ 8/9	73.0	70.0	58.0	84.0	2.43	0.39	0.77
	Upadhyaya, AK 1997	≥ 5/6	95.0	86.0	48.0	97.8	2.38	0.06	N/A
Single - MDI-1	Ayalon, L 2009	≥ 1	66.6	91.6	23.0	99.0	7.92	0.36	N/A
Single - MHI-1	Howe, A 2000	≤2/3	67.0	60.0	41.7	81.0	1.68	0.55	N/A
- MHI-1	Means-Christensen, A 2005	≤ 4	88.0	62.0	14.6	98.9	2.31	0.19	0.82
- MHI-1(y/n)	Means-Christensen, A 2006	1	85.0	73.0	63.6	91.1	3.15	0.21	N/A
Single - PHQ-1	**Arroll, B 2003**	≥ 1	86.0	72.0	18.6	98.6	3.00	0.19	N/A
- PHQ-1	Ayalon, L 2009	≥ 1	83.3	93.8	45.0	99.0	13.44	0.18	N/A
Single - Yale SQ	**Blank, K 2004**	≥ 1	64.0	64.0	11.6	93.5	1.80	0.56	N/A
- Yale SQ	Robison, J 2002	≥ 1	86.0	42.0	30.0	51.0	1.48	0.34	N/A
SDS	Williams, JW Jnr 1995	0.043	100.0	72.0	13.0	100.00	3.57	<0.01	N/A
SRQ-20	**Azevedo-Marques, J 2009**	≥ 8	81.0	86.0	81.0	87.0	5.79	0.22	0.86
SWB	Muhwezi, W 2007	10	75.7	86.3	76.7	85.6	5.52	0.28	N/A
WHO-5	Awata, S 2007	≤13	100.0	74.1	31.8	100.0	3.86	<0.01	0.90
	Azevedo-Marques, J 2009 ***Henkel, V 2004***	≤11	77.0	89.0	81.0	87.0	7.00	0.26	0.83
		≤13	94.0	65.0	37.0	98.0	2.69	0.09	0.90
	Löwe, B 2004a (SCID)	≤ 8	95.0	73.0	**34.9**	98.9	3.52	0.07	0.91
	Löwe, B 2004b (ICDL)	≤ 9	87.0	70.0	35.2	96.6	2.90	0.19	0.89
	Mergl, R 2007	≤13	90.0	63.0	14.7	98.1	2.43	0.16	0.86
	Saipanish, R 2009	≤13	89.0	65.0	16.0	99.0	2.56	0.16	0.86
Zung-SDS	*Aragones, E 2001*	≥50	94.0	70.0	35.0	99.0	3.09	0.09	0.93
	Baer, L 2000	≥50	89.0	53.0	77.0	73.0	1.89	0.20	N/A
	Campo-Arias, A 2006	≥49	50.0	94.6	64.7	90.5	9.26	0.53	0.90
	Leung, KK 1998	≥55	66.7	90.2	30.2	97.7	6.80	0.37	N/A
	Okimoto, JT 1982	≥60	76.0	82.0	65.5	88.4	4.22	0.29	N/A

PPV: Positive Predictive Value, **NPV:** Negative Predictive Value, **LR+**: Positive Likelihood Ratio, **LR-**: Negative Likelihood Ratio, **AUC:** Area under the ROC curve, **N/A**: not available.
BASDEC: Brief Assessment Schedule Depression Cards, **BDI-II:** Beck Depression Inventory, version 2 **BDI-FS:** Beck Depression Inventory-Fast Screen, **BDI-SF:** Beck Depression Inventory-Short Form, **CES-D**: Centre for Epidemiological Studies-Depression Scale (10 & 20 item), **Euro-D:** European Depression Scale, **GHQ:** General Health Questionnaire, **HADS:** Hospital Anxiety and Depression Scale, **HADS-D:** HADS depression subscale **HANDS:** Harvard Department of Psychiatry/National Depression Screening Day Scale **ID:** (Popoff) Index of Depression, **MDI:** Major Depression Inventory, **MHI-5:** Mental Health Inventory, **MOS-D:** Medical Outcomes Study Depression Scale, **PC-SAD:** Primary Care Screener for Affective Disorders, **PHQ:** Patient Health Questionnaire (1,2 yes/no, 2 & 9 item) **PRIME-MD:** Primary Care Evaluation of Mental Disorders, **PROMIS:** Patient Reported Outcomes Measurement Information System **QIDS-SR:** Quick Inventory of Depressive Symptomatology–Self-Report **SDDS-PC**: Symptom Driven Diagnostic System-Primary Care, **SDS:** Short Depression Screen **SelfCARE(D) SQ:** Single Question **WHO-5:** World Health Organization Wellbeing Index, **Zung- SDS:** Zung's Self-Rating Depression Scale.
***PSP**: Primary care screening sample.
Author in bold print: QUADAS-2 RoB rated low in all domains.
Authors in *italic* print: Study included in meta-analysis.

Table 4 Summary psychometrics for screening tool meta-analysis.

Tool	Parameter			
	DOR	**Sensitivity**	**Specificity**	**LR+**	**LR-**
CESD-20	14.06	0.86	0.69	2.77	0.20
HADS-D	13.90	0.77	0.80	3.94	0.28
PHQ-9	25.69	0.77	0.89	6.79	0.26
PHQ-2(scored)	12.78	0.55	0.91	6.26	0.49
WHO-5	19.70	0.90	0.68	2.78	0.14
Zung-SDS	15.70	0.77	0.82	4.33	0.28

DOR: diagnostic odds ratio, LR+: positive likelihood ratio LR-: negative likelihood ratio.

may only identify 3 in every 10 patients with depression.[44] We also found the five single-question tools generally had higher LR- values than other tools, suggesting a negative result was more likely to occur in people with depression. Therefore, only eighteen tools were considered suitable for use in primary-care settings.

Analysis showed the PHQ-9 had the highest DOR and therefore better performance in primary care settings than other tools meta-analysed. The PHQ-9 was also one of the most accurate screening tools based on a-ROC.[24] Although the Zung-SDS had slightly higher accuracy (median a-ROC = 0.92), it was considered less suitable in primary care settings as it contained more than twice as many questions, and took around twice as long to administer.

Likelihood ratios are becoming a more popular method than sensitivity and specificity alone, in describing the clinical usefulness of a screening tool as they do not vary with prevalence of depression, unlike predictive values.[23,35] Of the tools meta-analysed, not only did the PHQ-9 have the highest DOR, it also had the highest LR + suggesting it was nearly seven times more likely to produce a positive screening result in someone with depression than in someone without depression. In contrast, although the WHO-5 had the second highest DOR, having a lower LR+, it was only three times more likely to produce a positive screen in someone with depression than in someone without depression. Hence, after a positive result, the PHQ-9 greatly increases the chance of 'ruling-in' depression and after a negative result, the WHO-5 with the lowest LR-, greatly increases the chance of 'ruling-out' depression.[24]

For screening in primary care, a tool with a higher sensitivity (fewer false-negatives) would be more useful than one with a higher specificity (fewer false positives).[45] We found the WHO-5 had the highest sensitivity (90.0%) and would be suitable for use in primary care based on ease of administration (taking $\leq$5 min to administer). A meta-analysis and several validation studies suggested another way to minimise false negatives and also reduce the burden of administration, is to use the PHQ-2(yes/no), a highly sensitive pre-screening tool, followed by the PHQ-9 after a positive PHQ-2 result.[46–48] A tool with a higher NPV is also useful for screening in primary care as it indicates a negative screen is more likely to be a true negative result, reducing the chance of a false negative result.[36,45] Of the tools meta-analysed, we found the WHO-5 had the highest median NPV (98.0%) followed by the PHQ-9 (96.5%). Despite the useful information provided by sensitivity, specificity and predictive values,[25,49] this psychometric data should be supported by more meaningful measures of screening tool performance such as DOR's, LR's and a-ROC values.

Given the current lack of guidelines to assist the choice of an appropriate depression screening tool in primary healthcare, the choice of over 40 screening tools and the considerable variability in operating characteristics, this review assists in selecting an easily administered and psychometrically sound depression screening tool. Our review showed eighteen tools were ultimately suitable for use in primary care based on brevity, ease of scoring and level of literacy. Analysis showed the PHQ-9 had the highest performance and accuracy in primary care settings, suggesting it discriminates between patients with and without depression more effectively than any other screening tool analysed. Furthermore, the PHQ-9 was the most effective tool at 'ruling-in' depression after a positive screening result based on a superior LR + . The PHQ-9 also achieved the best balance between sensitivity and specificity of the tools meta-analysed. The PHQ-9 was the most extensively evaluated tool in primary care and validated in more than nine countries, it assesses depressive symptoms in accordance with DSM guidelines and can also assist practitioners in monitoring the severity of depression and an individuals' response to treatment.[19,21] Although other tools were found to be suitable for use in primary care based on their ease of administration, further studies with these tools are required to obtain better estimates of their psychometric properties in this setting.

While ease of administration and acceptable performance are important for selecting a self-administered screening tool, these alone do not ensure the effectiveness of a screening tool. Effectiveness depends on several factors including whether the tool is acceptable to the population and whether the screening provides benefits that outweigh the harms.[50] There is evidence that screening for depression has features that may impact on the acceptability to the population, specifically the stigma associated with mental illness.[51] Further work is also required to determine what the best setting for screening with self-administered tools is. A study in 2003 found that clinic-based screening identified the largest proportion of patients with depression.[52] However, the study also found that screening at home identified an older patient population with chronic illnesses. This is important given the relationship between chronic illness and depression.[53]

There were several limitations to our systematic review. Only articles written in English were included therefore

language bias is possible. We could only meta-analyse psychometric characteristics for tools evaluated by ≥5 studies to avoid a loss of precision. Nonetheless, the most frequently evaluated tools were meta-analysed. Even though all studies were conducted in primary care settings, investigating heterogeneity for the majority of tools studied was not possible because they were evaluated by ≤8 studies. However, random effects modelling was used to adjust for possible heterogeneity between studies for accuracy and threshold of the tools meta-analysed.

Conclusion

Although we found numerous depression screening tools suitable for use in primary care based on ease of administration, the PHQ-9 was the most widely assessed tool and displayed superior DOR, LR+, a-ROC, and specificity. The PHQ-2(yes/no) may be useful as a pre-screening tool as it is an ultra-brief tool, showed good sensitivity and had a low LR- and was frequently evaluated. Our review supports the use of the PHQ-9 as a brief, easily administered depression screening tool with the most robust psychometric characteristics in primary care settings.

Ethical considerations

No ethics approval has been sought for this work as the manuscript is a systematic review of published studies conducted by other researchers.

Conflict of interest

The authors declare that they have no conflicts of interest to disclose.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements

No other acknowledgements.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ejpsy.2020.10.002.

References

1. World Health Organisation. Depression and Other Common Mental Disorders: Global Health Estimates Web page. World Health Organisation; 2017. Contract No.: WHO/MSD/MER/2017.2. Available: https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf (Accessed 31 August 2020).
2. Katon WJ. Epidemiology and treatment of depression in patients with chronic medical illness. Dialogues Clin Neurosci. 2011;13(1):7–23.
3. Compton WM, Conway KP, Stinson FS, Grant BF. Changes in the prevalence of major depression and comorbid substance use disorders in the United States between 1991-1992 and 2001-2002. Am J Psychiatry. 2006;163(12):2141–7.
4. Centers for Disease Control and Prevention. Mental Illness Surveillance Among Adults in the United States, MMWR. Report. Altlanta: U.S. Department of Health and Human Services; 2011. Available: https://www.cdc.gov/mmwr/preview/mmwrhtml/su6003a1.htm (Accessed 27 August 2020).
5. Goldney RD, Eckert KA, Hawthorne G, Taylor AW. Changes in the prevalence of major depression in an Australian community sample between 1998 and 2008. Aust N Z J Psychiatry. 2010;44(10):901–10.
6. Hasin DS, Sarvet AL, Meyers JL, Saha TD, Ruan WJ, Stohl M, et al. Epidemiology of Adult DSM-5 Major Depressive Disorder and Its Specifiers in the United States. JAMA Psychiatry. 2018;75(4):336–46.
7. Australian Bureau of Statistics. Australian Social Trends, 4102.0 2009. Mental Health. Canberra: Australian Bureau of Statistics; 2009. Contract No.: 4102.0. Available: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4102.0Main+Features30March%202009 (Accessed 31 August 2020).
8. Beyond Blue. Beyond Blue Statistics; 2018. Available: https://www.beyondblue.org.au/media/statistics (Accessed 31 August 2020).
9. Voinov B, Richie WD, Bailey RK. Depression and chronic diseases: It is time for a synergistic mental health and primary care approach. Prim Care Companion J Clin Psychiatry. 2013;15(2).
10. Jani BD, Purves D, Barry S, Cavanagh J, McLean G, Mair FS. Challenges and implications of routine depression screening for depression in chronic disease and multimorbidity: a cross sectional study. PLoS ONE. 2013;8(9):e74610.
11. Davis L, Uezato A, Newell JM, Frazier E. Major depression and comorbid substance use disorders. Curr Opin Psychiatry. 2008;21(1):14–8.
12. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. Lancet. 2009;374(9690):609–19.
13. Wittchen HU, Pittrow D. Prevalence, recognition and management of depression in primary care in Germany: The Depression 2000 study. Hum Psychopharmacol. 2002;17 SUPPL. 1:S1–11.
14. Nuyen J, Volkers AC, Verhaak PFM, Schellevis FG, Groenewegen PP, Van den Bos GAM. Accuracy of diagnosing depression in primary care: The impact of chronic somatic and psychiatric co-morbidity. Psychol Med. 2005;35(8):1185–95.
15. Coyne JC. Depression in primary care: depressing news, exciting research opportunities. APS Obs [Internet]. 2001;14.
16. Mitchell AJ. Overview Of Depression Scale and Tools. Screening For Depression In Clinical Practice. 2 ed; 2009. p. 29–56.
17. Siu AL, Bibbins-Domingo K, Grossman DC, Baumann LC, Davidson KW, Ebell M, et al. Screening for depression in adults: US preventive services task force recommendation statement. JAMA. 2016;315(4):380–7.
18. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression: Two questions are as good as many. J Gen Intern Med. 1997;12(7):439–45.
19. Lowe B, Spitzer RL, Grafe K, Kroenke K, Quenter A, Zipfel S, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. J Affect Disord. 2004;78(2):131–40.

20. Baer L, Jacobs DG, Meszler-Reizes J, Biais M, Fava M, Kessler R, et al. Development of a brief screening instrument: The HANDS. Psychother Psychosom. 2000;69(1):35–41.
21. Williams Jr JW, Pignone M, Ramirez G, Perez Stellato C. Identifying depression in primary care: A literature synthesis of case-finding instruments. Gen Hosp Psychiatry. 2002;24(4):225–37.
22. El-Den S, Chen TF, Gan Y-L, Wong E, O'Reilly CL. The psychometric properties of depression screening tools in primary healthcare settings: A systematic review. J Affect Disord. 2018;225 Supplement C:503–22.
23. Akobeng AK. Understanding diagnostic tests 2: Likelihood ratios, pre- and post-test probabilities and their use in clinical practice. Acta Paediatr. 2007;96(4):487–91.
24. Akobeng AK. Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatr. 2007;96(5):644–7.
25. Akobeng AK. Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. Acta Paediatr. 2007;96(3):338–41.
26. Allaby M, Pittam G. Appraisal of Screening for Depression: A report for the UK National Screening Committee. Oxford: Solutions for Public Health (SPH); 2014. Contract No.: 2 December 2018. Available: https://legacyscreening.phe.org.uk/policydb_download.php?doc=521 (Accessed 21 August 2020).
27. Picardi A, Adler DA, Rogers WH, Lega I, Zerella MP, Matteucci G, et al. Diagnostic accuracy of the primary care screener for affective disorder (pc-sad) in primary care. Clin Pract Epidemiol Ment Health. 2013;9:164–70.
28. Nease DE Jr, Malouin JM. Depression screening: A practical strategy. J Fam Pract. 2003;52(2):118–26.
29. Loeb DF, Corral J, Sieja A, Zehnder N, McCord M, Nease DE. The importance of training in the implementation of an outpatient depression screening and treatment protocol. J Gen Intern Med. 2014;29:S229.
30. UK National Screening Committee. Evidence and recommendations: NHS population screening. United Kingdom; 2015. Available: https://www.gov.uk/guidance/evidence-and-recommendations-nhs-population-screening#evidence-review-process (Accessed 31 August 2020).
31. Akena D, Joska J, Obuku EA, Amos T, Musisi S, Stein DJ. Comparing the accuracy of brief versus long depression screening instruments which have been validated in low and middle income countries: A systematic review. BMC Psychiatry. 2012;12(1).
32. Anderson J, Michalak E, Lam R. Depression in primary care: Tools for screening, diagnosis, and measuring response to treatment. B C Med J. 2002;44(8):415–9.
33. Whiting PF, Rutjes AWS, 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(8):529–36.
34. Agency for Healthcare Research and Quality. Table 1. Characteristics of Case-Finding Instruments Used to Detect Depression in Adults in Primary Care Settings. U.S. Department of Health & Human Services, April; 2013. Available: http://www.ahrq.gov/professionals/clinicians-providers/resources/depression/depsumtab1.html (Accessed 31 August 2020).
35. Florkowski CM. Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests. Clin Biochem Rev. 2008;29 Suppl 1. S83-S7.
36. Spitalnic S. Test Properties 1: Sensitivity, Specificity and Predictive Values. Hosp Physician. 2004:27–31.
37. Borenstein M, Hedges L, Higgins J, Rothstein H. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Method. 2010;1:97–111.
38. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56(11):1129–35.
39. Rutter CM, Gatsonis CA. A hierachical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20(19):2865–84.
40. METADAS: An SAS Macro for Meta-Analysis of Diagnostic Accuracy Studies [Computer program].
41. Macaskill P, Gatsonis C, J.J. D, Harbord RM, Takwoingi Y. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration; 2010. Available: http://srdta.cochrane.org/ (Accessed 31 August 2020).
42. The Cochrane Collaboration. Review Manager 5 (RevMan 5) [Computer program].
43. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 4th edition Washington, DC: American Psychiatric Association; 2000.
44. 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(535):144–51.
45. Ren Y, Yang H, Browning C, Thomas S, Liu M. Performance of Screening Tools in Detecting Major Depressive Disorder among Patients with Coronary Heart Disease: A Systematic Review. Med Sci Monit. 2015;21:646–53.
46. 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(8):1076–81.
47. Mitchell Aj, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych open. 2016;2(2):127–38.
48. Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. Ann Fam Med. 2010;8(4):348–53.
49. McAlpine DD, Wilson AR. Screening for depression in primary care: What do we still need to know? Depress Anxiety. 2004;19(3):137–45.
50. Andermann A, Blancquaert I, Beauchamp S, Déry V. Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years. Bull World Health Organ. 2008;86:317–9.
51. Colligan EM, Cross-Barnet C, Lloyd JT, McNeely J. Barriers and facilitators to depression screening in older adults: a qualitative study. Aging Ment Health. 2020;24(2):341–8.
52. Kanter JW, Epler AJ, Chaney EF, Liu C-F, Heagerty P, Lin P, et al. Comparison of 3 depression screening methods and provider referral in a Veterans Affairs primary care clinic. Prim Care Companion J Clin Psychiatry. 2003;5(6):245.
53. Gunn JM, Ayton DR, Densley K, Pallant JF, Chondros P, Herrman HE, et al. The association between chronic illness, multimorbidity and depressive symptoms in an Australian primary care cohort. Soc Psychiatry Psychiatr Epidemiol. 2012;47(2):175–84.