Canadian Psychiatric Association
Association des psychiatres du Canada

Systematic Review

The Canadian Journal of Psychiatry / La Revue Canadienne de Psychiatrie
2016, Vol. 61(12) 746-757

DOI: 10.1177/0706743716651833
TheCJP.ca | LaRCP.ca
SAGE

Accuracy of Depression Screening Tools to Detect Major Depression in Children and Adolescents: A Systematic Review

Exactitude des instruments de dépistage de la dépression pour détecter la dépression majeure chez les enfants et les adolescents: une revue systématique

Michelle Roseman, MSc[1], Lorie A. Kloda, PhD[2], Nazanin Saadat, BSc[1], Kira E. Riehm, BA[1], Abel Ickowicz, MD[3], Franziska Baltzer, MD[4,5], Laurence Y. Katz, MD[6], Scott B. Patten, MD, PhD[7], Cécile Rousseau, MD[5], and Brett D. Thombs, PhD[1,5]

Abstract

Objective: Depression screening among children and adolescents is controversial, and no clinical trials have evaluated benefits and harms of screening programs. A requirement for effective screening is a screening tool with demonstrated high accuracy. The objective of this systematic review was to evaluate the accuracy of depression screening instruments to detect major depressive disorder (MDD) in children and adolescents.

Method: Data sources included the MEDLINE, MEDLINE In-Process, EMBASE, PsycINFO, HaPI, and LILACS databases from 2006 to September 30, 2015. Eligible studies compared a depression screening tool to a validated diagnostic interview for MDD and reported accuracy data for children and adolescents aged 6 to 18 years. Risk of bias was assessed with QUADAS-2.

Results: We identified 17 studies with data on 20 depression screening tools. Few studies examined the accuracy of the same screening tools. Cut-off scores identified as optimal were inconsistent across studies. Width of 95% confidence intervals (CIs) for sensitivity ranged from 9% to 55% (median 32%), and only 1 study had a lower bound 95% CI $\geq$80%. For specificity, 95% CI width ranged from 2% to 27% (median 9%), and 3 studies had a lower bound $\geq$90%. Methodological limitations included small sample sizes, exploratory data analyses to identify optimal cut-offs, and the failure to exclude children and adolescents already diagnosed or treated for depression.

Conclusions: There is insufficient evidence that any depression screening tool and cut-off accurately screens for MDD in children and adolescents. Screening could lead to overdiagnosis and the consumption of scarce health care resources.

[1] Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec
[2] Libraries, Concordia University, Montreal, Quebec
[3] Department of Psychiatry, Hospital for Sick Children, University of Toronto, Toronto, Ontario
[4] Montreal Children's Hospital, Montreal, Quebec
[5] McGill University, Montreal, Quebec
[6] University of Manitoba, Winnipeg, Manitoba
[7] University of Calgary, Calgary, Alberta

Corresponding Author:
Brett D. Thombs, PhD, Jewish General Hospital, 4333 Cote Ste Catherine Road, Montréal, QC H3T 1E4, Canada.
Email: brett.thombs@mcgill.ca

Abrégé

Objectif : Le dépistage de la dépression chez les enfants et les adolescents est controversé, et aucun essai clinique n'a évalué les avantages et les inconvénients des programmes de dépistage. Une condition essentielle pour un dépistage efficace est un instrument de dépistage ayant démontré une grande exactitude. L'objectif de cette revue systématique était d'évaluer l'exactitude des instruments de dépistage de la dépression pour détecter le trouble dépressif majeur (TDM) chez les enfants et les adolescents.

Méthode : Les sources de données comprenaient les bases de données MEDLINE, MEDLINE In-Process, EMBASE, PsycINFO, HaPI, et LILACS, de 2006 au 30 septembre 2015. Les études admissibles comparaient un instrument de dépistage de la dépression avec une entrevue diagnostique validée pour le TDM, et rendaient compte des données d'exactitude pour les enfants et les adolescents de 6 à 18 ans. Le risque de biais a été évalué à l'aide de QUADAS-2.

Résultats : Nous avons identifié 17 études comportant des données sur 20 instruments de dépistage de la dépression. Peu d'études ont examiné l'exactitude de ces mêmes instruments de dépistage. Les seuils d'inclusion identifiés comme étant optimaux étaient irréguliers dans toutes les études. L'étendue des intervalles de confiance (IC) à 95% pour la sensibilité allait de 9% à 55% (moyenne 32%), et une seule étude avait une limite inférieure de l'IC à 95% $\geq$ 80%. Plus précisément, l'étendue des IC à 95% allait de 2% à 27% (moyenne 9%), et 3 études avaient une limite inférieure $\geq$ 90%. Les limitations méthodologiques comprenaient de petites tailles d'échantillons, des analyses de données exploratoires pour identifier les seuils d'inclusion optimaux, et l'omission d'exclure les enfants et les adolescents déjà diagnostiqués ou traités pour la dépression.

Conclusions : Il n'y a pas suffisamment de données probantes pour affirmer que tout instrument de dépistage et seuil d'inclusion de la dépression dépistent avec exactitude le TDM chez les enfants et les adolescents. Le dépistage pourrait entraîner le surdiagnostic et la consommation des maigres ressources de santé.

Systematic Review Registration : PROSPERO; CRD42012003194

Keywords

depression, screening, children, adolescents, diagnostic accuracy

Screening children and adolescents for depression is controversial. In 2009, the United States Preventive Services Task Force (USPSTF) recommended that adolescents, but not younger children, should be routinely screened for depression in primary care settings when depression care systems are in place to ensure accurate diagnosis, treatment, and follow-up.[1] The USPSTF recently reiterated this recommendation in its 2016 guideline.[2] By contrast, depression screening among children and adolescents has not been recommended in the United Kingdom or Canada.[3,4] No clinical trials have evaluated depression screening programs among children or adolescents,[2] and there are no examples of well-conducted trials among adults that have shown that depression screening would improve mental health outcomes.[5-8]

Depression screening, if initiated in practice, would involve the use of self-report questionnaires to identify children or adolescents who may have depression but have not otherwise been identified as possibly depressed by health care professionals or via self-report.[9,10] Health care professionals would need to administer a screening tool and use a predetermined cut-off score to separate children and adolescents who may have depression from those unlikely to have depression. Screening, which would be done with all children and adolescents who are not suspected of having depression, is different from case finding, which is only done with patients who health care professionals believe are at risk.[10]

In screening, tools must be accurate enough to identify a large proportion of unrecognized depression cases and to effectively rule out noncases to avoid unnecessary mental health assessments and the possibility of overdiagnosis and overtreatment. Thus, although screening may not improve mental health outcomes, it would consume scarce resources and further burden an already financially strapped mental health care system that struggles to provide adequate care for children and adolescents with obvious mental health needs. There is increasing attention to the problem of overdiagnosis and overtreatment across areas of medicine.[11] In depression screening, overdiagnosis could result in the prescription of psychotropic medications to an increased number of children, who would be exposed to the adverse effects of these medications, even if they did not experience benefits from screening.[6]

Few systematic reviews have assessed the accuracy of screening tools for detecting major depressive disorder (MDD) in children and adolescents, including data on screening tool sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). A 2009 United States Agency for Healthcare Research and Quality (AHRQ) review,[12] upon which the 2009 USPSTF guidelines[1] were based, included 9 studies, of which 5 compared a depression screening tool to a diagnosis of MDD based on a validated diagnostic interview. An updated 2016 AHRQ review,[13] which formed the basis for the USPSTF's recent guidelines,[2] identified no new eligible diagnostic accuracy studies. The 2016 AHRQ review included only a subset of 5 studies from the 2009 review, of which 3 compared a screening tool to a validated diagnostic interview as the reference standard for MDD.

A 2015 systematic review and meta-analysis[14] included 52 articles on 4 commonly used depression screening tools among children and adolescents. Thirty-three studies reported diagnostic accuracy data, but approximately half were conducted with children or adolescents in mental health treatment or who were referred for mental health evaluation. Children already referred for treatment or receiving treatment, however, would not be screened in actual practice, since screening is done to identify depression among patients who have not otherwise been identified as possibly depressed. Screening accuracy should be evaluated among undiagnosed and untreated patients.[15] Furthermore, in the meta-analyses conducted for each included screening tool, the authors used sensitivity and specificity results for each primary study based on an "optimal" cut-off threshold that maximized accuracy in the particular primary study, rather than using the same cut-off across included studies. For example, their meta-analysis of the accuracy of the Beck Depression Inventory (BDI) combined results from studies using cut-offs ranging from $\geq$11 to $\geq$23. As a result, synthesized accuracy values did not reflect what would be achieved in practice if the BDI were used for screening, since in practice, a cut-off must be chosen prior to screening.

The objective of the present systematic review was to evaluate the accuracy of depression screening instruments to detect MDD in children and adolescents.

Method

Detailed methods were registered in the PROSPERO prospective register of systematic reviews (CRD42012003194), and a review protocol was published.[16]

Search Strategy

The MEDLINE, MEDLINE In-Process, EMBASE, PsycINFO, HaPI, and LILACS databases were searched on September 30, 2015, using a peer-reviewed search strategy (Supplementary File 1). Searches included articles published January 2006 or later because the 2009 AHRQ systematic review on depression screening in children and adolescents,[7] which included studies on the diagnostic accuracy of depression screening tools, searched through May 2006. Studies included in the 2009 and 2016 AHRQ reviews[12,13] were evaluated for possible inclusion in the present review. Search results were downloaded into the citation management database RefWorks (RefWorks-COS, Bethesda, MD, USA), and the software's duplication check was used to identify citations retrieved from multiple sources.

Identification of Eligible Studies

Eligible articles were original studies in any language with data on children and adolescents aged 6 to 18 years, conducted in general medicine clinics, schools, and community settings. Studies of college and university populations were excluded. Studies with mixed population samples were eligible if data for children or adolescents aged 6 to 18 years were reported separately or if at least 80% of the sample were aged 18 years or younger.

Eligible diagnostic accuracy studies had to report data that allowed determination of the sensitivity, specificity, PPV, and NPV of a self-report depression screening tool compared to a current *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) diagnosis of MDD or major depressive episode (MDE) or *International Classification of Diseases* (*ICD*) depressive episode, established with a validated diagnostic interview administered within 2 weeks of the screening tool. Study authors were contacted to determine eligibility if this interval was not specified. Studies that reported only parent or teacher-completed depression measures were excluded. Studies that assessed broader diagnostic categories, such as any depressive disorder, were included only if they reported screening accuracy for MDD separately or if at least 80% of cases of depression, however defined, had a *DSM* diagnosis of MDD or MDE or an *ICD* diagnosis of depressive episode.

Two investigators independently reviewed titles/abstracts for eligibility, with full-text review of articles that were identified as potentially eligible by one or both investigators. Disagreements after full-text review were resolved by consensus. All titles/abstracts and full-text articles were available in English, Spanish, German, Portuguese, or Chinese and reviewed by investigators fluent in those languages. Non-English articles were reviewed by a single investigator.

Evaluation of Eligible Studies

Two investigators independently extracted data into a standardized spreadsheet (Supplementary File 2). Risk of bias was assessed based on published information with the revised Quality Assessment for Diagnostic Accuracy Studies–2 (QUADAS-2) tool.[17] QUADAS-2 incorporates assessments of risk of bias across 4 core domains: patient selection, the index test, the reference standard, and the flow and timing of assessments (see Supplementary File 3). Any discrepancies in data extraction and risk of bias assessment were resolved by consensus.

Data Presentation and Synthesis

Data on the accuracy of screening tools were extracted with 95% confidence intervals[18] based on "optimal" cut-offs identified by primary study authors. We also determined the lower bound of confidence intervals for each study, which is important for clinical decision making. For example, if at least 80% sensitivity and 90% specificity are deemed necessary to consider screening, the lower bound of 95% confidence intervals of accuracy estimates should be at least 80% for sensitivity and at least 90% for specificity.[19] Studies were heterogeneous in terms of patient samples, screening tools

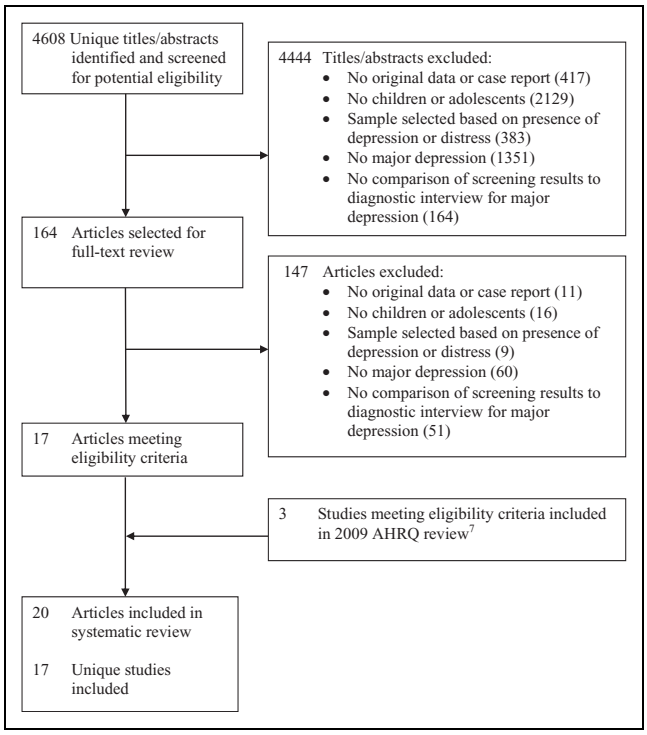

Figure 1. PRISMA flow diagram of study selection process.

and cut-offs, and criterion standards. Thus, results were not pooled quantitatively.

Results

Selection of Eligible Studies

Of 4608 unique titles/abstracts identified from the database search, 4444 were excluded after title/abstract review and 147 after full-text review, leaving 17 eligible articles (Figure 1).[20-36] Three additional eligible articles[37-39] published prior to our search were identified from the 2009 AHRQ review,[12] resulting in a total of 20 included articles reporting on 17 unique studies.

Study Characteristics and Diagnostic Accuracy Results

Of the 17 included studies, 9 were conducted in school settings,[20-22,30-32,37-39] 5 in primary care or specialty medicine settings,[23-29,36] 2 in programs for adolescent mothers,[33,34] and 1 as part of a population-based longitudinal study.[35] Ten studies[20,22,26-29,31,33-35,37,38] restricted participants to adolescents (aged 12 years and older), whereas 3 studies[21,30,39] recruited study samples exclusively from high school settings but did not explicitly report on the age range of included participants. Four studies[23-25,32,36] included both children and adolescents. Sample sizes in the 17 studies ranged from 49 to 4027 (median 290) and MDD cases from 4 to 305 (median 19). There were 2 German-language articles[25,27] (see Table 1).

The 17 included studies reported diagnostic accuracy for 20 different depression screening instruments, including subscales and alternate-length versions of standard instruments. Diagnostic accuracy was based on exploratory methods, in which the same data were used to both identify an "optimal" screening cut-off and assess accuracy in 16 studies[20-34,36-39] and not specified in 1 study[35] (Table 2). Only 2 screening tools, the BDI (4 studies) and the Patient Health Questionnaire–9 (PHQ-9; 3 studies), had diagnostic accuracy results reported in 3 or more studies. Of all included studies, only 2 studies,[37,38] a study from the United States with 5 MDD cases and a study from Spain with 10 MDD cases, identified the same optimal cut-off for a screening tool (BDI $\geq$16).

The 4 studies of the BDI[20,37-39] included 5 to 76 MDD cases per study. "Optimal" screening cut-offs identified ranged from $\geq$11 to $\geq$18. The width of 95% confidence intervals ranged from 13% to 43% (median 30%) for sensitivity and from 3% to 16% (median 5%) for specificity. For sensitivity, only 1 study from Nigeria had a lower bound for the 95% confidence interval of at least 80%, and no studies had a lower bound $\geq$90%. For specificity, 3 studies had lower bounds $\geq$80% with 2 were $\geq$90%.

Three studies of the PHQ-9,[28,29,31] which included 18 to 31 MDD cases, reported optimal cut-off scores that ranged from $\geq$5 to $\geq$15. The width of the 95% confidence intervals ranged from 24% to 39% (median 28%) for sensitivity and from 7% to 11% (median 8%) for specificity. None of the studies had a lower confidence interval bound of at least 80% for sensitivity (maximum 71%), with only 1 study over 80% for specificity.

For all other screening tools with accuracy results, number of MDD cases ranged from 4 to 305 (median 20). Estimates of sensitivity were generally imprecise, with 95% confidence intervals widths of 9% to 55% (median 33%). For specificity, the width of 95% confidence intervals ranged from 2% to 27% (median 11%).

Risk of Bias

As shown in Table 3, risk of bias was high for 16 of 17 studies that did not prespecify a screening test cut-off and unclear for the remaining study.[35] Only 1 study[33] excluded children and adolescents with already diagnosed or treated depression who would not be screened in practice. Thus, for patient selection applicability, 15 studies were rated as unclear risk of bias, and 1 study[32] was rated as high risk since 25% of study participants were already receiving psychosocial services. Risk of bias was unclear in 12 of 17 studies for methods of sample selection and unclear or high in 5 studies for the blinding of interviewers to screening test results. In addition, 6 studies were rated as unclear risk for issues related to patient flow and timing, including administration of the reference standard to only a subset of the sample, handling of missing data, and the interval between the index test and reference standard (see Supplementary File 4

Table 1. Characteristics of Studies of Diagnostic Accuracy.

First Author, Year, Country	Age Group, y	Setting	*N*	Mean Age, y	Males, %	Major Depression Criterion Standard	Major Depression, *n* (%)	Instrument(s)	Instrument Language
BDI									
Adewuya, 2007, Nigeria[20]	13-18	School	454	15[a]	58[a]	K-SADS	76 (17)	BDI	NR
Barrera,[b] 1988, United States[37]	12-18	School	49	15	45	CAS	5 (10)	BDI	English
Canals, 2001, Spain[38]	17-18	School	290	18[c]	50[c]	SCAN	10 (3)	BDI	Spanish
Roberts, 1991, United States[39]	High school grades 9-12	School	1704	17	47	K-SADS	43 (3)	BDI	English
BDI-II									
Araya, 2013, Chile[21]	High school grade 10	School	571[d]	16	46	MINI-KIDS	301 (53)	BDI-II	Spanish
Pietsch,[e] 2012, Germany[27]	13-16	Medical clinics	314	14	40	Kinder-DIPS	21 (7)	BDI-II, BDI-FS	German
CDI									
Bang, 2015, Korea[22]	12-16	School	468	13	44	K-SADS	63 (13)	CDI	Korean
Butwicka, 2012, Poland[23]	8-18	Inpatient diabetes care	163	14	57	K-SADS	4 (3)	CDI	Polish
CES-D									
Logsdon, 2010, United States[34]	13-18	Teen parent program	59	16	0	K-SADS	10 (17)	CES-D, CES-D-30	English
Roberts, 1991, United States[39]	High school grades 9-12	School	1704	17	47	K-SADS	43 (3)	CES-D	English
Pietsch,[e,f] 2013, Germany[26]	13-16	Medical clinics	327	14	41	Kinder-DIPS	22 (7)	CES-D-15	German
EPDS									
Logsdon, 2010, United States[34]	13-18	Teen parent program	59	16	0	K-SADS	10 (17)	EPDS	English
Venkatesh,[g] 2014, United States[33]	13-18	Prenatal clinic	96	NR[h]	0	KID-SCID	8 (8)	EPDS, EPDS-7, EPDS-3, EPDS-2	English
PHQ									
Ganguly, 2013, India[31]	14-18	School	233	16	54	K-SADS	31 (13)	PHQ-9	English
Richardson, 2010, United States[28,29]	13-17	Medical clinics	442	15	40	DISC	19 (4)	PHQ-9, PHQ-2	English
Tsai, 2014, Taiwan[30]	High school students	School	165	17[i]	40[i]	K-SADS	18 (11)	PHQ-9, PHQ-2, PHQ-1	Chinese
MFQ-SF									
Katon, 2008, United States[36]	11-17	Primary care	1375	14	53	DISC	83[j] (6)	MFQ-SF	English
Turner, 2014, United Kingdom[35]	17-18	Population-based longitudinal study	4027	18[k]	43	CIS-R	305 (8)	MFQ-SF[l]	English

(continued)

Table 1. (continued)

First Author, Year, Country	Age Group, y	Setting	*N*	Mean Age, y	Males, %	Major Depression Criterion Standard	Major Depression, *n* (%)	Instrument(s)	Instrument Language
Other									
Katon, 2008, United States[36]	11-17	Primary care	1375	14	53	DISC	83[j] (6)	ASI	English
Fruhe,[m] 2012, Germany[24,25]	9-12	Medical clinics	228-246	11	57	Kinder-DIPS	11[n] (4) to 12[o] (5)	ChilD-S,[p] DIKJ, DTK Dysphoria subscale	German
Ventevogel,[q] 2014, Burundi[32]	10-15	School	61	13	55	K-SADS	11[r] (18)	DSRS	Kirundi

ASI, Childhood Anxiety Sensitivity Index; BDI, Beck Depression Inventory; BDI-II, Beck Depression Inventory–Second Edition; BDI-FS, Beck Depression Inventory–Fast Screen; CAS, Child Assessment Schedule; CDI, Children's Depression Inventory; CES-D, Center for Epidemiological Studies Depression Scale; CES-D-15, 15-item version of the Center for Epidemiological Studies Depression Scale; CES-D-30, 30-item version of the Center for Epidemiological Studies Depression Scale; ChilD-S, Children's Depression Screener; CIS-R, Clinical Interview Schedule–Revised form; DIKJ, German version of the Children's Depression Inventory (Depressionsinventar für Kinder und Jugendliche); DISC, Diagnostic Interview Schedule for Children; DSRS, Depression Self-Rating Scale; DTK, Depression Test for Children (Depressionstest für Kinder); EPDS, Edinburgh Postnatal Depression Scale; EPDS-2, 2-item subscale of the Edinburgh Postnatal Depression Scale; EPDS-3, 3-item anxiety subscale of the Edinburgh Postnatal Depression Scale; EPDS-7, 7-item depressive symptoms subscale of the Edinburgh Postnatal Depression Scale; Kinder-DIPS, German structured diagnostic interview for mental disorders in children and adolescents; KID-SCID, Structured Clinical Interview for *DSM-IV* Childhood Diagnoses; K-SADS, Schedule for Affective Disorders and Schizophrenia for School-Age Children; MFQ-SF, Mood and Feelings Questionnaire–Short Form; MINI-KIDS, Mini International Neuropsychiatric Interview for Children and Adolescents; NR, not reported; PHQ, Patient Health Questionnaire; PHQ-1, 1-item version of Patient Health Questionnaire; PHQ-2, 2-item version of Patient Health Questionnaire; PHQ-9, 9-item version of Patient Health Questionnaire; SCAN, Schedule for Assessment of Neuropsychiatric Disorders.

[a]Demographic data based on overall study sample of $N = 1095$, rather than the 454 adolescents included in the analyses reported in the table.

[b]Of the 2 study samples reported in Barrera et al.[37] (i.e., psychiatric hospital inpatients and secondary school students), only the school sample was eligible for inclusion.

[c]Demographic data based on overall study sample of $N = 304$, rather than the 290 adolescents included in the analyses reported in the table.

[d]Article reported $N = 592$, but diagnostic accuracy data were reported for $N = 571$ (sum of depressed and nondepressed cases).

[e]Pietsch et al.[26,27] report on the same cohort. However, the BDI-II and BDI-SF were employed as the screening instruments in Pietsch et al.,[27] and the CES-D-15 was employed as the screening instrument in Pietsch et al.[26]

[f]Diagnostic accuracy data were provided by the authors, as published results did not include patients with dysthymia or minor depression among noncases of major depressive disorder (MDD).

[g]Diagnostic accuracy data extracted for 6 weeks postpartum visit (first administration of depression measures) for identified optimal overall cut-offs for each screening instrument, rather than for identified optimal time point cut-offs.

[h]Median age of sample = 16.

[i]Demographic data based on overall study sample of $N = 2257$, rather than the 165 adolescents included in the analyses reported in the table.

[j]Diagnostic accuracy data were reported for N "depressive disorders" = 83, of which there were 82 cases of MDD and 1 case of dysthymia.

[k]Mean age based on study sample of $N = 4503$ who attended the cohort assessment day, rather than the 4027 adolescents included in the analyses reported in the table.

[l]Short form of Mood and Feelings Questionnaire abbreviated as "SMFQ" in Turner et al.[35] but reported above as "MFQ-SF" for consistency with the labeling of the same questionnaire in Katon et al.[36]

[m]Fruhe et al.[24,25] report on the same cohort. However, the *DSM-IV* diagnostic system was employed in Fruhe et al.,[24] and the *ICD-10* diagnostic system was employed in Fruhe et al.,[25] resulting in different numbers of cases of depressive disorders.

[n]Diagnostic accuracy data were reported for N "any depressive disorder" = 11, of which there were 9 cases of MDD and 2 cases of dysthymia.

[o]Diagnostic accuracy data were reported for N "any depressive disorder" = 12, of which there were 10 cases of MDD and 2 cases of dysthymia.

[p]Data on the accuracy of the ChilD-S in this cohort were also reported in Fruhe et al.[25] Per protocol, data on the accuracy of this instrument were extracted from the larger sample.

[q]Data were provided by the authors to correct inconsistencies in the published manuscript.

[r]Diagnostic accuracy data were reported for N "depressive disorders" = 11, of which there were 9 or 10 cases of MDD and 1 or 2 cases of adjustment disorder with depressive symptoms.

Table 2. Diagnostic Accuracy Results.

First Author, Year, Country	Instrument/ Cut-off	Derivation of Cut-off	Range of Cut-offs Reported	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Positive Predictive Value, % (95% CI)	Negative Predictive Value, % (95% CI)
BDI							
Adewuya, 2007, Nigeria[20]	BDI ≥18	Exploratory	≥15-21	91 (82-95)	97 (95-98)	86 (77-92)	98 (96-99)
Barrera,[a] 1988, United States[37]	BDI ≥16	Exploratory	≥6, 11, 16, 21, 26	100 (57-100)	93 (82-98)	63 (31-86)	100 (91-100)
Canals, 2001, Spain[38]	BDI ≥16	Exploratory	≥10, 11, 14, 16	90 (60-98)	96 (93-98)	45 (26-66)	100 (98-100)
Roberts, 1991, United States[39]	BDI ≥11	Exploratory	≥11	84 (70-92)	81 (79-83)	10 (7-14)	99 (99-100)
BDI-II							
Araya, 2013, Chile[21]	BDI-II ≥17	Exploratory	≥14, 17, 20	79 (74-83)	70 (64-75)	74 (69-79)	75 (69-80)
Pietsch,[b] 2012, Germany[27]	BDI-II ≥19	Exploratory	≥16-22	86 (65-95)	93 (89-95)	46 (32-61)	99 (97-100)
	BDI-FS ≥6	Exploratory	≥4-8	81 (60-92)	90 (86-93)	37 (25-51)	99 (96-99)
CDI							
Bang, 2015, Korea[22]	CDI ≥20	Exploratory	≥15, 17, 20, 25	83 (71-90)	89 (85-92)	54 (44-63)	97 (95-98)
Butwicka, 2012, Poland[23]	CDI ≥53	Exploratory	≥53	100 (51-100)	82 (75-87)	12 (5-27)	100 (97-100)
CES-D							
Logsdon, 2010, United States[34]	CES-D ≥16	Exploratory	≥4-31	70 (40-89)	45 (32-59)	21[c] (10-37)	88[c] (70-96)
	CES-D-30 ≥16	Exploratory	≥11-46	100 (72-100)	27 (16-40)	22[c] (12-36)	100[c] (77-100)
Roberts, 1991, United States[39]	CES-D ≥24	Exploratory	≥24	84 (70-92)	75 (73-77)	8 (6-11)	99 (99-100)
Pietsch,[b,d] 2013, Germany[26]	CES-D-15 ≥14	Exploratory	≥14	95 (78-99)	80 (75-84)	26 (17-36)	100 (98-100)
EPDS							
Logsdon, 2010, United States[34]	EPDS ≥5	Exploratory	≥1-10	80 (49-94)	59 (45-72)	29[c] (15-47)	94[c] (79-98)
Venkatesh,[e] 2014, United States[33]	EPDS ≥9	Exploratory	≥8-10	75 (41-93)	86 (78-92)	33 (16-56)	97 (91-99)
	EPDS-7 ≥7	Exploratory	≥7, 8, 10	100 (68-100)	84 (75-90)	36 (20-57)	100 (95-100)
	EPDS-3 ≥10	Exploratory	≥10, 13	63 (31-86)	67 (57-76)	15 (6-30)	95 (87-98)
	EPDS-2 ≥10	Exploratory	≥10	88 (53-98)	80 (70-87)	28 (14-48)	99 (92-100)
PHQ							
Ganguly, 2013, India[31]	PHQ-9 ≥5	Exploratory	≥1-15, 17, 21	87 (71-95)	80 (74-85)	40 (29-52)	98 (94-99)
Richardson, 2010, United States[28,29]	PHQ-9 ≥11	Exploratory	≥6-13, Algorithm	89 (69-97)	78 (73-81)	15 (10-23)	99 (98-100)
	PHQ-2 ≥3	Exploratory	≥1-6	74 (51-88)	75 (71-79)	12 (7-19)	98 (96-99)
Tsai, 2014, Taiwan[30]	PHQ-9 ≥15	Exploratory	≥9-16	72 (49-88)	95 (91-98)	65 (43-82)	97 (92-99)
	PHQ-2 ≥3	Exploratory	≥2-4	94 (74-99)	82 (75-88)	40 (26-54)	99 (96-100)
	PHQ-1 ≥2	Exploratory	≥1-3	61 (39-80)	88 (81-92)	38 (23-56)	95 (90-97)
MFQ-SF							
Katon, 2008, United States[36]	MFQ-SF ≥6	Exploratory	≥6	80 (70-87)	81 (79-83)	22 (17-27)	98 (97-99)
Turner, 2014, United Kingdom[35]	MFQ-SF[f] ≥11	Unclear	≥11	71 (66-76)	83 (82-84)	26 (23-29)	97 (97-98)
Other							
Katon, 2008, United States[36]	ASI ≥13	Exploratory	≥13	73 (63-82)	66 (63-69)	12 (10-15)	97 (96-98)

(continued)

Table 2. (continued)

First Author, Year, Country	Instrument/ Cut-off	Derivation of Cut-off	Range of Cut-offs Reported	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Positive Predictive Value, % (95% CI)	Negative Predictive Value, % (95% CI)
Fruhe,[g] 2012, Germany[24,25]	ChilD-S[h] ≥11	Exploratory	≥9-12	91 (62-98)	89 (85-93)	29 (16-45)	100 (97-100)
	DIKJ ≥12	Exploratory	≥9-15	92 (65-99)	82 (76-87)	22 (13-35)	99 (97-100)
	DTK Dysphoria subscale ≥10	Exploratory	≥5-11	75 (47-91)	90 (85-93)	29 (16-47)	98 (96-99)
Ventevogel,[i] 2014, Burundi[32]	DSRS ≥19	Exploratory	≥13, 15, 17, 19, 21	64 (35-85)	88 (76-94)	54 (29-77)	92 (80-97)

ASI, Childhood Anxiety Sensitivity Index; BDI, Beck Depression Inventory; BDI-II, Beck Depression Inventory–Second Edition; BDI-FS, Beck Depression Inventory–Fast Screen; CDI, Children's Depression Inventory; CES-D, Center for Epidemiological Studies Depression Scale; CES-D-15, 15-item version of the Center for Epidemiological Studies Depression Scale; CES-D-30, 30-item version of the Center for Epidemiological Studies Depression Scale; ChilD-S, Children's Depression Screener; CI, confidence interval; DIKJ, German version of the Children's Depression Inventory (Depressionsinventar für Kinder und Jugendliche); DSRS, Depression Self-Rating Scale; DTK, Depression Test for Children (Depressionstest für Kinder); EPDS, Edinburgh Postnatal Depression Scale; EPDS-2, 2-item subscale of the Edinburgh Postnatal Depression Scale; EPDS-3, 3-item anxiety subscale of the Edinburgh Postnatal Depression Scale; EPDS-7, 7-item depressive symptoms subscale of the Edinburgh Postnatal Depression Scale; MFQ-SF, Mood and Feelings Questionnaire–Short Form; PHQ, Patient Health Questionnaire; PHQ-1, 1-item version of Patient Health Questionnaire; PHQ-2, 2-item version of Patient Health Questionnaire; PHQ-9, 9-item version of Patient Health Questionnaire.

[a]Of the 2 study samples reported in Barrera et al.[37] (i.e., psychiatric hospital inpatients and secondary school students), only the school sample was eligible for inclusion.

[b]Pietsch et al.[26,27] report on the same cohort. However, the BDI-II and BDI-SF were employed as the screening instruments in Pietsch et al.,[27] and the CES-D-15 was employed as the screening instrument in Pietsch et al.[26]

[c]The 2 × 2 tables (number of true positives, false positives, true negatives, and false negatives) for each screening instrument at the optimal screening threshold could not fully be replicated based on published sensitivity, specificity, positive predictive value, and negative predictive value. Study authors were unable to provide original diagnostic data to resolve discrepancies. For the EPDS ≥5, the published positive predictive value was 31% and negative predictive value 93%. For the CES-D ≥16, the published positive predictive value was 23% and negative predictive value 87%. For the CES-D-30 ≥16, the published positive predictive value was 24% and negative predictive value 100%. Confidence intervals were not reported in the published data, and negative predictive value was reported as (negative predictive value − 1).

[d]Diagnostic accuracy data were provided by the authors, as published results did not include patients with dysthymia or minor depression among noncases of major depressive disorder.

[e]Diagnostic accuracy data extracted for 6 weeks postpartum visit (first administration of depression measures) for identified optimal overall cut-offs for each screening instrument, rather than for identified optimal time point cut-offs.

[f]Short form of Mood and Feelings Questionnaire abbreviated as "SMFQ" in Turner et al.[35] but reported above as "MFQ-SF" for consistency with the labeling of the same questionnaire in Katon et al.[36]

[g]Fruhe et al.[24,25] report on the same cohort. However, the *DSM-IV* diagnostic system was employed in Fruhe et al.,[24] and the *ICD-10* diagnostic system was employed in Fruhe et al.,[25] resulting in different numbers of cases of depressive disorders.

[h]Data on the accuracy of the ChilD-S in this cohort were also reported in Fruhe et al.[25] Per protocol, data on the accuracy of this instrument were extracted from the larger sample.

[i]Data were provided by the authors to correct inconsistencies in the published manuscript.

for detailed QUADAS-2 coding notes for all included studies).

Excluded Studies and Comparison with Previous Systematic Reviews

Of the 9 diagnostic accuracy studies included in the 2009 AHRQ systematic review,[12] 3 were included in the present review.[37-39] Of the other 6 studies, 3 did not administer a validated diagnostic interview as the reference standard for MDD, 1 did not administer a self-report screening instrument as the index test, and 1 compared the screening instrument to the diagnosis of any depressive disorder but did not report the number of patients diagnosed with MDD. In another study, the diagnostic interview was consistently administered more than 2 weeks after the screening instrument, per author report. Of the 5 diagnostic accuracy studies included in the 2016 AHRQ systematic review,[13] 2 were included in the present review.[38,39] Of the other 3 studies, 1 did not administer a validated diagnostic interview as the reference standard for MDD, 1 did not administer a self-report screening instrument as the index test, and 1 consistently administered the diagnostic interview more than 2 weeks after the screening instrument, per author report (Supplementary File 5).

The 2015 Stockings et al.[14] review included 33 studies that reported on the diagnostic accuracy of depression screening instruments, 5 of which were included in the present review.[20,34,37-39] Of these 5 studies, 3[37-39] were also included in the 2009 AHRQ review and 2[38,39] in the 2016 AHRQ review. Sixteen studies in the Stockings et al.[14] review were excluded from the present review because samples were recruited from psychiatric settings or selected on the basis of distress or depression (e.g., referred for mental health evaluation). The remaining 12 studies were excluded for other reasons, including not using a validated

Table 3. Quality Assessment of Studies of Diagnostic Accuracy (QUADAS-2).

	QUADAS-2 Domains[a]						
	Risk of Bias				Applicability Concerns		
First Author, Year, Country	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Adewuya, 2007, Nigeria[20]	Low	High	Low	Unclear	Unclear	Low	Low
Araya, 2013, Chile[21]	Unclear	High	Low	Unclear	Unclear	Low	Low
Bang, 2015, Korea[22]	Unclear	High	Low	Unclear	Unclear	Low	Low
Barrera, 1988, United States[37]	Unclear	High	Unclear	Unclear	Unclear	Low	Low
Butwicka, 2012, Poland[23]	Low	High	Unclear	Low	Unclear	Low	Low
Canals, 2001, Spain[38]	Unclear	High	Low	Low	Unclear	Low	Low
Fruhe, 2012, Germany[24,25]	Low	High	Low	Low	Unclear	Low	Unclear
Ganguly, 2013, India[31]	Low	High	Low	Low	Unclear	Low	Low
Katon, 2008, United States[36]	Unclear	High	Unclear	Low	Unclear	Low	Low
Logsdon, 2010, United States[34]	Unclear	High	Low	Low	Unclear	Low	Low
Pietsch, 2012, 2013, Germany[26,27]	Low	High	Low	Low	Unclear	Low	Low
Richardson, 2010, United States[28,29]	Unclear	High	High	Unclear	Unclear	Low	Low
Roberts, 1991, United States[39]	Unclear	High	Unclear	Low	Unclear	Low	Low
Tsai, 2014, Taiwan[30]	Unclear	High	Low	Unclear	Unclear	Low	Low
Turner, 2014, United Kingdom[35]	Unclear	Unclear	Low	Low	Unclear	Low	Low
Venkatesh, 2014, United States[33]	Unclear	High	Unclear	Low	Low	Low	Low
Ventevogel, 2014, Burundi[32]	Unclear	High	Low	Low	High	Low	Unclear

[a]See Supplementary File 3 for QUADAS-2 risk of bias and applicability judgments. See Supplementary File 4 for detailed QUADAS-2 coding notes. Items are rated "low," "high," and "unclear" based on the QUADAS-2 guidelines and reflect the risk of bias or the degree of concern about applicability. Quality ratings were based only on published information, with the exception of information on the interval between index test and reference standard, for which information was obtained from study authors to determine study eligibility.

diagnostic interview as the reference standard, not comparing index test results to MDD diagnoses or reporting the number of patients diagnosed with MDD, or administering the screening test and diagnostic interview more than 2 weeks apart.

Discussion

The main findings of this systematic review were that there are relatively few studies on the accuracy of depression screening tools to detect MDD in children and adolescents and that existing studies have reported on a large number of different depression screening instruments in heterogeneous patient populations and settings. Only 2 screening tools, the standard versions of the BDI and PHQ-9, had diagnostic accuracy results reported in 3 or more studies.

Results on the performance of individual depression screening tools differed substantially across studies and require cautious interpretation. In all but 1 study, in which the derivation of the cut-off score was not specified,[35] exploratory data analysis methods were used to both set an "optimal" cut-off score and determine the accuracy of that cut-off score in the same patient sample. When data-driven methods are used to maximize diagnostic accuracy, studies generally overestimate screening tool performance, sometimes substantially.[40,41] Cut-off scores identified as "optimal" using these data-driven methods were inconsistent across included studies and varied too widely to provide health care professionals with an indication as to the most accurate cut-off score for any single screening tool. Only 2 studies included in our review[37,38] identified the same "optimal" cut-off for a screening tool (BDI $\geq$16). Furthermore, with only 1 exception, all included studies failed to appropriately exclude children and adolescents already diagnosed or treated for depression who would not be screened in clinical practice to identify new cases, which can also lead to inflated estimates of screening tool accuracy.[15]

Another important methodological consideration is that sample sizes in most included studies were small for the purpose of estimating diagnostic accuracy, with a median of 19 MDD cases per study. Estimates of screening tool sensitivity were imprecise, as reflected in wide 95% confidence intervals. Of the 20 results reported for sensitivity, only 1 study reported a lower confidence interval bound for sensitivity of at least 80%. While confidence interval widths were narrower for estimates of specificity, only 3 studies reported a lower confidence interval bound of at least 90%.

The 2016 systematic review,[13] which was done for the USPSTF guideline,[2] included only 5 studies on the accuracy of depression screening tools, which represent a subset of the 9 diagnostic accuracy studies included in the 2009 USPSTF review.[12] Among the factors that may explain why the AHRQ review did not identify numerous screening accuracy studies included in the present review are the use of a single,

combined search strategy for the review's 6 key questions rather than a search designed for diagnostic test accuracy studies, the exclusion of non-English language studies and studies conducted in developing countries, the exclusion of studies conducted in specialty medicine settings, and the decision to exclude otherwise eligible studies on the basis of quality ratings.[42] Quality exclusions were based on a list of possible quality indicators but not on a validated system for rating quality or risk of bias, such as QUADAS-2. Three of 5 studies included in the 2016 AHRQ systematic review did not meet eligibility criteria for the present review. Of the 2 studies that were included in the present review, both were rated, using QUADAS-2, as having unclear risk of bias related to patient selection and high risk related to the failure to prespecify an index test threshold.[38,39]

The 2016 USPSTF guidelines suggest the use of the Patient Health Questionnaire for Adolescents (PHQ-A) and the Beck Depression Inventory–Primary Care version (BDI-PC) as screening tools for adolescents in primary care settings.[2] The PHQ-A is similar to the PHQ-9 for adults with minor adaptations in wording.[43] The BDI-PC is a 7-item depression screening tool, derived from the cognitive items of the BDI-II.[44] This recommendation was based on only 1 study of the PHQ-A[43] and no evidence on the accuracy of the BDI-PC in children or adolescents.[13] The PHQ-A study[43] was excluded from the present review because it did not compare the PHQ-A to a validated diagnostic interview to determine MDD status.

The USPSTF recommends routine depression screening for adolescents in primary care settings when integrated depression care systems are in place.[2] This recommendation was made, even though no trials among children or adult patients have found that patients who are screened have better outcomes than patients who are not screened when both groups have access to similar depression treatments.[7,45] Screening is sometimes implemented even without direct evidence of effectiveness. TeenScreen, an American program based at Columbia University, urged implementation of universal depression screening for adolescents and was reportedly active at over 2800 sites in the United States and internationally before the project's unexplained closure in 2012.[46] In Canada, several provincial governments have called for widespread depression screening in school settings and medical practices.[47-49] In the absence of trials, the findings of the present review suggest important reasons why depression screening may be less effective than anticipated and could result in more harm than benefit. If the evidence base for depression screening tools overestimates their accuracy, the use of these questionnaires in screening programs would likely lead to high false-positive rates, unnecessary labeling, overtreatment in some cases, and the consumption of scarce mental health resources that could otherwise be used to provide better care for children and adolescents with undertreated mental health problems.[6]

A possible limitation of the systematic review is that we did not search for unpublished studies. Given the findings of the systematic review, it is unlikely that this would have changed the findings or conclusions. Another possible limitation is that we did not conduct a de novo search for studies prior to 2006 but rather used studies included in a previously published systematic review. It is possible that there could have been eligible early studies that were not identified, although the existence of multiple systematic reviews on this topic suggests that this is unlikely. Finally, although validated diagnostic interviews are considered the gold standard for establishing psychiatric diagnoses, there is not robust evidence establishing their degree of accuracy or replicability.

Conclusions

In summary, this systematic review found that there is insufficient evidence of the ability of depression screening instruments to accurately detect MDD in children and adolescents. Few studies have examined the accuracy of the same screening tools in comparable settings and populations, and there is inadequate evidence to recommend any single cut-off score for any of the instruments evaluated in the included studies. Significant methodological concerns, including small sample sizes, the use of data-driven exploratory methods to identify ''optimal'' cut-off scores, and the failure to exclude patients already diagnosed or treated for depression, raise concerns that existing studies may overestimate screening tool accuracy. Well-conducted studies with large sample sizes that present results across the range of possible cut-offs and follow guidance from key sources, including the Cochrane Handbook for Diagnostic Test Accuracy Meta-Analyses[50] and the STARD statement,[51] are needed. The absence of any evidence from clinical trials that depression screening would improve mental health outcomes, along with the results from this systematic review, suggests that screening children and adolescents could lead to more harm than benefit and would consume scarce mental health resources that could otherwise be used to provide treatment for underserved youth with mental disorders.

Supplemental Material

The online supplementary files are available at http://cpa.sagepub.com/supplemental.

Acknowledgments

We thank Yue Zhao, MSc, Concordia University, Montreal, Quebec, and Linda Kwakkenbos, PhD, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, for assistance with translation. They were not compensated for their contributions.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Dr. Patten reported that he received a research grant from a competition cosponsored by the Hotchkiss Brain Institute and

Pfizer Canada. All other authors declare that they have no competing interests.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the Canadian Institutes for Health Research (KA1-119795). MR was supported by a Murray R. Stalker Primary Care Research Bursary and a Mach-Gaensslen Foundation of Canada Student Grant as part of the McGill University Faculty of Medicine Research Bursary Program. BDT was supported by an Investigator Award from the Arthritis Society. No funding body had any involvement in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

References

1. US Preventive Services Task Force. Screening and treatment for major depressive disorder in children and adolescents: US Preventive Services Task Force recommendation statement. Pediatrics. 2009;123(4):1223-1228.
2. Siu AL; US Preventive Services Task Force. Screening for depression in children and adolescents: US Preventive Services Task Force recommendation statement. Ann Intern Med. 2016 Feb 9 (Epub ahead of print; DOI:10.7326/M15-2957).
3. MacMillan HL, Patterson CJ, Wathen CN, et al. Screening for depression in primary care: recommendation statement from the Canadian Task Force on Preventive Health Care. CMAJ. 2005;172(1):33-35.
4. National Collaborating Centre for Mental Health. Depression in children and young people: identification and management in primary, community, and secondary care. London (UK): National Institute for Health and Clinical Excellence; 2015.
5. Thombs BD, Ziegelstein RC. Does depression screening improve depression outcomes in primary care? BMJ. 2014; 348:g1253.
6. Thombs BD, Coyne JC, Cuijpers P, et al. Rethinking recommendations for screening for depression in primary care. CMAJ. 2012;184(4):413-418.
7. Thombs BD, Ziegelstein RC, Roseman M, et al. There are no randomized controlled trials that support the United States Preventive Services Task Force guideline on screening for depression in primary care: a systematic review. BMC Med. 2014;12:13.
8. Thombs BD, Arthurs E, Coronado-Montoya S, et al. Depression screening and patient outcomes in pregnancy or postpartum: a systematic review. J Psychosom Res. 2014;76:433-446.
9. UK National Screening Committee. Second report of the UK National Screening Committee. London (UK): Departments of Health for England, Scotland, Northern Ireland and Wales; 2000.
10. Raffle A, Gray M. Screening: evidence and practice. London (UK): Oxford University Press; 2007.
11. Moynihan R, Doust J, Henry D. Preventing overdiagnosis: How to stop harming the healthy. BMJ. 2012;344:e3502.
12. Williams SB, O'Connor EA, Eder M, et al. Screening for child and adolescent depression in primary care settings: a systematic evidence review for the US Preventive Services Task Force. Pediatrics. 2009;123(4):e716-e735.
13. Forman-Hoffman V, McClure E, McKeeman J, et al. Screening for major depressive disorder in children and adolescents: a systematic review for the U.S. Preventive Services Task Force. Ann Intern Med. 2016 Feb 9 (Epub ahead of print; DOI:10.7326/M15-2259).
14. Stockings E, Degenhardt L, Lee YY, et al. Symptom screening scales for detecting major depressive disorder in children and adolescents: a systematic review and meta-analysis of reliability, validity and diagnostic utility. J Affect Disord. 2015;174: 447-463.
15. Thombs BD, Arthurs E, El-Baalbaki G, et al. Risk of bias from inclusion of already diagnosed or treated patients in diagnostic accuracy studies of depression screening tools: a systematic review. BMJ. 2011;343:d4825.
16. Thombs BD, Roseman M, Kloda LA. Depression screening and mental health outcomes in children and adolescents: a systematic review protocol. Syst Rev. 2012;1:58.
17. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-536.
18. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat. 1998; 52(2):119-126.
19. Bachmann LM, Puhan MA, ter Riet G, et al. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ. 2006; 332:1127-1129.
20. Adewuya AO, Ola BA, Aloba OO. Prevalence of major depressive disorders and a validation of the Beck Depression Inventory among Nigerian adolescents. Eur Child Adolesc Psychiatry. 2007;16(5):287-292.
21. Araya R, Montero-Marin J, Barroilhet S, et al. Detecting depression among adolescents in Santiago, Chile: sex differences. BMC Psychiatry. 2013;13:122.
22. Bang YR, Park JH, Kim SH. Cut-off scores of the children's depression inventory for screening and rating severity in Korean adolescents. Psychiatry Investig. 2015;12(1):23-28.
23. Butwicka A, Fendler W, Zalepa A, et al. Efficacy of metabolic and psychological screening for mood disorders among children with type 1 diabetes. Diabetes Care. 2012;35(11): 2133-2139.
24. Fruhe B, Allgaier AK, Pietsch K, et al. Children's Depression Screener (ChilD-S): development and validation of a depression screening instrument for children in pediatric care. Child Psychiatry Hum Dev. 2012;43(1):137-151.
25. Fruhe B, Allgaier AK, Pietsch K, et al. Depression screening in pediatric patients—a comparison of the concurrent validity of the German version of the Children's Depression Inventory, the German Depression Test for Children, and the new Children's Depression Screener [in German]. Z Kinder Jugendpsychiatr Psychother. 2012;40(3):161-169.
26. Pietsch K, Allgaier A, Fruhe B, et al. Screening for adolescent depression in paediatric care: validity of a new brief version of the Center for Epidemiological Studies Depression Scale. Child Adolesc Ment Health. 2013;18(2):76-81.

27. Pietsch K, Hoyler A, Fruhe B, et al. Early detection of major depression in paediatric care: validity of the Beck Depression Inventory–Second Edition (BDI-II) and the Beck Depression Inventory–Fast Screen for Medical Patients (BDI-FS) [in German]. Psychother Psychosom Med Psychol. 2012;62(11): 418-424.
28. Richardson LP, McCauley E, Grossman DC, et al. Evaluation of the Patient Health Questionnaire-9 Item for detecting major depression among adolescents. Pediatrics. 2010;126(6): 1117-1123.
29. Richardson LP, Rockhill C, Russo JE, et al. Evaluation of the PHQ-2 as a brief screen for detecting major depression among adolescents. Pediatrics. 2010;125(5):e1097-e1103.
30. Tsai FJ, Huang YH, Liu HC, et al. Patient Health Questionnaire for school-based depression screening among Chinese adolescents. Pediatrics. 2014;133(2):e402-409.
31. Ganguly S, Samanta M, Roy P, et al. Patient Health Questionnaire-9 as an effective tool for screening of depression among Indian adolescents. J Adolesc Health. 2013;52(5):546-551.
32. Ventevogel P, Komproe IH, Jordans MJ, et al. Validation of the Kirundi versions of brief self-rating scales for common mental disorders among children in Burundi. BMC Psychiatry. 2014; 14:36.
33. Venkatesh KK, Zlotnick C, Triche EW, et al. Accuracy of brief screening tools for identifying postpartum depression among adolescent mothers. Pediatrics. 2014;133(1):e45-e53.
34. Logsdon MC, Myers JA. Comparative performance of two depression screening instruments in adolescent mothers. J Womens Health (Larchmt). 2010;19(6):1123-1128.
35. Turner N, Joinson C, Peters TJ, et al. Validity of the Short Mood and Feelings Questionnaire in late adolescence. Psychol Assess. 2014;26(3):752-762.
36. Katon W, Russo J, Richardson L, et al. Anxiety and depression screening for youth in a primary care population. Ambul Pediatr. 2008;8(3):182-188.
37. Barrera M Jr, Garrison-Jones CV. Properties of the Beck Depression Inventory as a screening instrument for adolescent depression. J Abnorm Child Psychol. 1988;16(3):263-273.
38. Canals J, Bladé J, Carbajo G, et al. The Beck Depression Inventory: psychometric characteristics and usefulness in non-clinical adolescents. Eur J Psychol Assess. 2001;17(1):63-68.
39. Roberts RE, Lewinsohn PM, Seeley JR. Screening for adolescent depression: a comparison of depression scales. J Am Acad Child Psychiatry. 1991;30(1):58-66.
40. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol. 2006;59(8):798-801.
41. Leeflang MM, Moons KG, Reitsma JB, et al. Bias in sensitivity and specificity caused by data-driven selection of optimal cut-off values: mechanisms, magnitude, and solutions. Clin Chem. 2008;54(4):729-737.
42. Forman-Hoffman VL, McClure E, McKeeman J, et al. Screening for major depressive disorder among children and adolescents: a systematic review for the U.S. Preventive Services Task Force. Evidence Synthesis No. 116. AHRQ Publication No. 13-05192-EF-1. Rockville (MD): Agency for Healthcare Research and Quality; 2016.
43. Johnson JG, Harris ES, Spitzer RL, et al. The Patient Health Questionnaire for Adolescents: validation of an instrument for the assessment of mental disorders among adolescent primary care patients. J Adolesc Health. 2002;30(3):196-204.
44. Steer RA, Cavalieri TA, Leonard DM, et al. Use of the Beck Depression Inventory for Primary Care to screen for major depression disorders. Gen Hosp Psychiatry. 1999;21(2): 106-111.
45. Canadian Task Force on Preventive Health Care, Joffres M, Jaramillo A, et al. Recommendations on screening for depression in adults. CMAJ. 2013;185(9):775-782.
46. Lenzer J. Controversial mental health program closes down. BMJ. 2012;345:e8100.
47. Alberta Health and Wellness Communications. Positive futures—optimizing mental health for Alberta's children & youth: a framework for action (2006–2016). Edmonton: Government of Alberta; 2006 [cited 2016 Feb 21]. Available from: http://www.health.alberta.ca/documents/mental-health-framework-child-06.pdf.
48. British Columbia Guidelines and Protocols Advisory Committee. Anxiety and depression in children and youth—diagnosis and treatment. Victoria: B.C. Government; 2010 [cited 2016 Feb 21]. Available from: http://www2.gov.bc.ca/assets/gov/health/practitioner-pro/bc-guidelines/depressyouth.pdf.
49. Manitoba Healthy Living. Reclaiming hope: Manitoba's youth suicide prevention strategy. Winnipeg: Manitoba Government; 2008 [cited 2016 Feb 21]. Available from: http://www.gov.mb.ca/healthyliving/mh/docs/hope.pdf.
50. Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy: Version 0.9. The Cochrane Collaboration; 2013 [cited 2016 Apr 16]. Available from: http://srdta.cochrane.org/.
51. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527.