CRITICAL REVIEW AND INVITED COMMENTARY

Depression screening tools in persons with epilepsy: A systematic review of validated tools

*†Stephanie J. Gill, *†Sara Lukmanji, †‡§Kirsten M. Fiest, †‡¶**Scott B. Patten, *†‡Samuel Wiebe, and *†‡Nathalie Jetté

Epilepsia, **(*):1–11, 2017
doi: 10.1111/epi.13651

Stephanie Gill is a first-year medical student at the University of Calgary's Cumming School of Medicine.

Summary

Objective: **Depression affects approximately 25% of epilepsy patients. However, the optimal tool to screen for depression in epilepsy has not been definitively established. The purpose of this study was to systematically review the literature on the validity of depression-screening tools in epilepsy.**

Methods: **MEDLINE, EMBASE, and PsycINFO were searched until April 4, 2016 with no restriction on dates. Abstract, full-text review and data abstraction were conducted in duplicate. We included studies that evaluated the validity of depression-screening tools and reported measures of diagnostic accuracy (e.g., sensitivity, specificity, and negative and positive predictive values) in epilepsy. Study quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies Version 2. Medians and ranges for estimates of diagnostic accuracy were calculated when appropriate.**

Results: **A total of 16,070 abstracts were screened, and 38 articles met eligibility criteria. Sixteen screening tools were validated in 13 languages. The most commonly validated screening tool was the Neurological Disorders Depression Inventory for Epilepsy (NDDI-E) (n = 26). The Mini International Neuropsychiatric Interview (MINI) (n = 19) was the most common reference standard used. At the most common cutpoint of >15 (n = 12 studies), the NDDI-E had a median sensitivity of 80.5% (range 64.0–100.0) and specificity of 86.2 (range 81.0–95.6). Meta-analyses were not possible due to variability in cutpoints assessed, reference standards used, and lack of confidence intervals reported.**

Significance: **A number of studies validated depression screening tools; however, estimates of diagnostic accuracy were inconsistently reported. The validity of scales in practice may have been overestimated, as cutpoints were often selected post hoc based on the study sample. The NDDI-E, which performed well, was the most commonly validated screening tool, is free to the public, and is validated in multiple languages and is easy to administer, although selection of the best tool may vary depending on the setting and available resources.**

KEY WORDS: **Major depressive disorder, Mental health, Comorbidity, Diagnostic accuracy, Measurement.**

Epilepsy is the second leading neurologic cause of years lived with disability according to the Global Burden of Disease Study in 2010.[1] Depression is common in epilepsy, [2] with a reported prevalence of 23.1% according to a recent meta-analysis.[3] Depression may decrease treatment adherence,[4,5] increase the risk of suicide,[6] interfere with self-

Accepted November 25, 2016.

*Department of Clinical Neurosciences, University of Calgary, Calgary, Alberta, Canada; †Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta, Canada; ‡Department of Community Health Sciences and O'Brien Institute for Public Health, University of Calgary, Calgary, Alberta, Canada; §Department of Critical Care Medicine, University of Calgary, Calgary, Alberta, Canada; ¶Mathison Centre for Mental Health Research & Education, University of Calgary, Calgary, Alberta, Canada; and **Department of Psychiatry, University of Calgary, Calgary, Alberta, Canada

Address correspondence to Nathalie Jetté, Departments of Clinical Neurosciences and Community Health Sciences, 1403 29 Street NW, Calgary, AB T2N2T9, Canada. E-mail: nathalie.jette@albertahealthservices.ca

Key Points

- Thirty-eight studies were identified that validated 16 depression-screening tools in epilepsy in 13 languages
- The NDDI-E was the most commonly validated screening tool (n = 26) and the MINI was the most commonly used reference standard (n = 19)
- A cutpoint of >13 on the NDDI-E seems to have a better balance of median sensitivity and specificity than the recommended cutpoint of >15
- Lack of reporting of diagnostic accuracy estimates prevented meta-analysis demonstrating the importance of adhering to the Standards for Reporting of Diagnostic Accuracy (STARD) statement
- Future research should develop and/or validate screening tools for depression in children, youths, and elderly patients with epilepsy.

management, and diminish quality of life[5,7] in those with epilepsy. By promptly identifying and treating depression, epilepsy patients may have improved overall health outcomes.

A number of methods are available to detect depression or depressive symptoms, such as psychiatric or psychological assessments, structured or semistructured interviews, and self-report screening tools.[8] The use of screening tools can be effective because they are often brief, standardized, and a less resource-intensive means of assessing elevated depressive symptoms.[8] Many depression-screening tools have been developed for use in the general population, [9,10] but the only epilepsy-specific tool is the Neurological Disorders Depression Inventory for Epilepsy (NDDI-E).[11]

It is important to identify valid depression-screening tools that can be used in clinical settings, since these tools might help clinicians identify individuals who have depression, an important comorbidity of epilepsy. The objective of this study was to systematically synthesize the literature assessing the validity of screening tools for depression in persons with epilepsy.

Methods

This study was conducted according to an a priori published protocol registered with PROSPERO international prospective register of systematic reviews (CRD42015027425). All findings were reported using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards (Appendix S1).[12] Our primary research objective was to establish the criterion validity of depression-screening tools in persons with epilepsy. We defined criterion validity as the ability of screening tools to correctly identify depression when calibrated against a known reference standard. Criterion validity was operationalized using reported measures of diagnostic accuracy (defined below).

Eligibility

Studies were included based on the following criteria: (1) results presented on original data (i.e., not a review paper); (2) validation studies; (3) reporting on the diagnostic accuracy of any depression screening tool relative to a comparison measure (e.g., gold standards [or tool that could reasonably be considered a gold standard], other screening tools, clinical diagnostic interviews, and so on) in persons with epilepsy. In order to allow a comprehensive summary of available criterion validation data, no restrictions were placed on the reference standard used for validation, although only studies using a gold standard were included in the statistical analyses.

Search strategy

The MEDLINE, EMBASE, and PsycINFO databases were searched from inception until April 4, 2016, with no restrictions on country or language of publication. The search strategy included subject headings and keywords related to the following terms: epilepsy, depression, and validity (Search Strategy, Appendix S2). The reference lists of previously published reviews and all studies included in this review were hand searched to ensure no papers were missed. Abstracts and conference proceedings were excluded.

Study selection

A two-step process was used to select studies for inclusion in the review. After duplicates were removed (e.g., same study from two different databases), titles and abstracts were screened to identify articles meeting the predetermined eligibility criteria. The second step consisted of a full-text article review of all abstracts identified in the first stage. Studies were excluded at the full-text stage if the validation did not include a reference standard that measured depression specifically. All steps were conducted independently by two reviewers (SG and SL) and disagreements were resolved by discussion with a third reviewer. Speakers of the respective language screened non-English articles using the same process.

Data abstraction

A standardized form was used to abstract data in duplicate and included patient demographics and study details. When available, the following data were abstracted: study region, ascertainment source (i.e., hospital or tertiary care clinic), age, sex, number of participants, screening tool(s) under validation, cutpoints assessed, reference standard used for validation, study-specific prevalence of depression based on the reference standard, and the following measures of diagnostic accuracy (when reported): sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative

predictive value (NPV), true positives (TPs), false positives (FPs), true negatives (TNs), false negatives (FNs), area under the receiver operating characteristic (ROC) (AUC), binomial regression coefficient, Cronbach's alpha, Kappa, likelihood ratios, any effect modifiers/confounders assessed, and any recommended/optimal cutpoints.

Risk of bias/quality assessment

Risk of bias and applicability were assessed using the Quality Assessment of Diagnostic Accuracy Studies, version 2 [QUADAS-2 (www.quadas.org)] (Appendix S3).[13] Overall assessment of bias was based on responses to four domains: (1) patient selection; (2) index test; (3) reference standard, and (4) flow and timing, for which there were multiple signaling questions to guide the assessment of each domain. If one or more of the four domains was considered as having a high or unclear risk of bias, the overall classification was rated as having a high risk of bias. The overall risk of bias was only considered low if all domains were rated as having a low risk of bias. The level of applicability was also assessed using a single signaling question for the first three domains previously listed to identify if the domain of interest was consistent with the review question.[13]

Data synthesis and analysis

Findings from all included studies were summarized using medians, ranges, and frequencies (%). Meta-analyses were not conducted due to study heterogeneity (e.g., different cutpoints assessed, reference standards used, and so on). Only reference standards generally considered to be a "gold standard" were considered for use in the median estimate calculations. If a study validated a screening tool against more than one reference standard, the best reference standard was used, and only this estimate was included in the summary statistic to avoid excessive emphasis. The "best" reference standard was determined based on its known clinical utility for diagnosing depression, or if it was the tool that is more widely known as a gold standard according to the literature and a psychiatric expert on our team. If two or more studies assessed the same version of a screening tool at the same cutpoint (with an appropriate reference standard, i.e., not another screening tool) the Se, Sp, PPV, NPV, and AUC estimates were summarized using medians and ranges. For example, the NDDI-E at a cutpoint of >11 was validated by five studies; therefore, these estimates were summarized using medians and ranges. All estimates regardless of the cutpoint assessed are included in Appendix S4. In addition, the medians and ranges were estimated separately for the studies that had a high or unclear risk of bias on any of the QUADAS-2 domains versus studies that had a low risk of bias on each QUADAS-2 domain when feasible (Appendix S5). An attempt was made to contact study authors when details were lacking in their published papers but not all responded.

Results

Results of the search

A total of 16,070 unique abstracts were identified, of which 91 were reviewed in full-text and 38 met all eligibility criteria. Reasons for full-text article exclusion are listed in Figure 1.

Description of studies

The 38 studies [11,14–20] identified were published between 1998 and 2016 in 20 countries (Appendix S4); 34 were published in English,[11,14,15,18–20,23,24] 3 in Portuguese,[16,21,22] and one in German.[17] All studies included male and female participants, and sample sizes ranged from 44 to 575 (median 143). Twenty-seven studies examined participants ≥18 years old,[11,15–21] seven studies >16 years old,[14,24,28,39,42,44,46] and only one study examined youth (10–17 years old).[48] In three studies it was inferred from the reported mean and median ages that these participants were adults.[26,31,33] All studies ascertained their samples from outpatient settings such as tertiary or university clinics. Finally, the median prevalence of depression based on the reference standard assessment was 21.0% (range 5.7% to 30.6%).

Screening tools and reference standards

Validity was assessed in 16 screening tools, against 7 reference standards. A description of all screening tools examined is found in Box 1. A number of studies did more than one validation for each tool (i.e., against different reference standards), or validated different versions of a screening tool. The most commonly validated screening tools were the NDDI-E (n = 26 validations[11,15,17,19,20,22–24]), Beck Depression Inventory (BDI; n = 12 validations, including the modified BDI, BDI-Fast Screen, BDI-I, BDI-II, and the Cognitive Affective Subscale for the BDI[17,18,20,21,31–33,39]), Hospital Anxiety and Depression Scale (HADS; n = 10 validations, including the HADS total score, HADS-Depression [HADS-D] Score, and one study used the HADS-Anxiety [HADS-A] score to assess depression[14,16,18,20,25,27,39,42]), Emotional Thermometers (ETs; n = 5 validations, including the ET7, ET4, and individual questions on anxiety and depression[24,42]), Patient Health Questionnaire 9 (PHQ-9; n = 4 validations[25,39,43,49]), Patient Health Questionnaire 2 (PHQ-2; n = 2 validations[25,35]), and the Hamilton Rating Scale for Depression (HAM-D or HRSD; n = 4 validations, including the HRSD 17 and 21 question versions[18,21,41]). The NDDI-E was validated in 13 languages including Arabic,[15] Chinese,[29,47] Danish,[30] English,[11,20,24,27,28,35,37,42] French,[38] German,[17] Greek,[50] Italian,[40] Japanese,[45] Korean,[34] Portuguese,[19,22] Spanish,[23,46] and Serbian.[44]

The various gold and/or reference standards used to validate the screening tools were the Mini International Neuropsychiatric Interview (MINI; multiple versions in

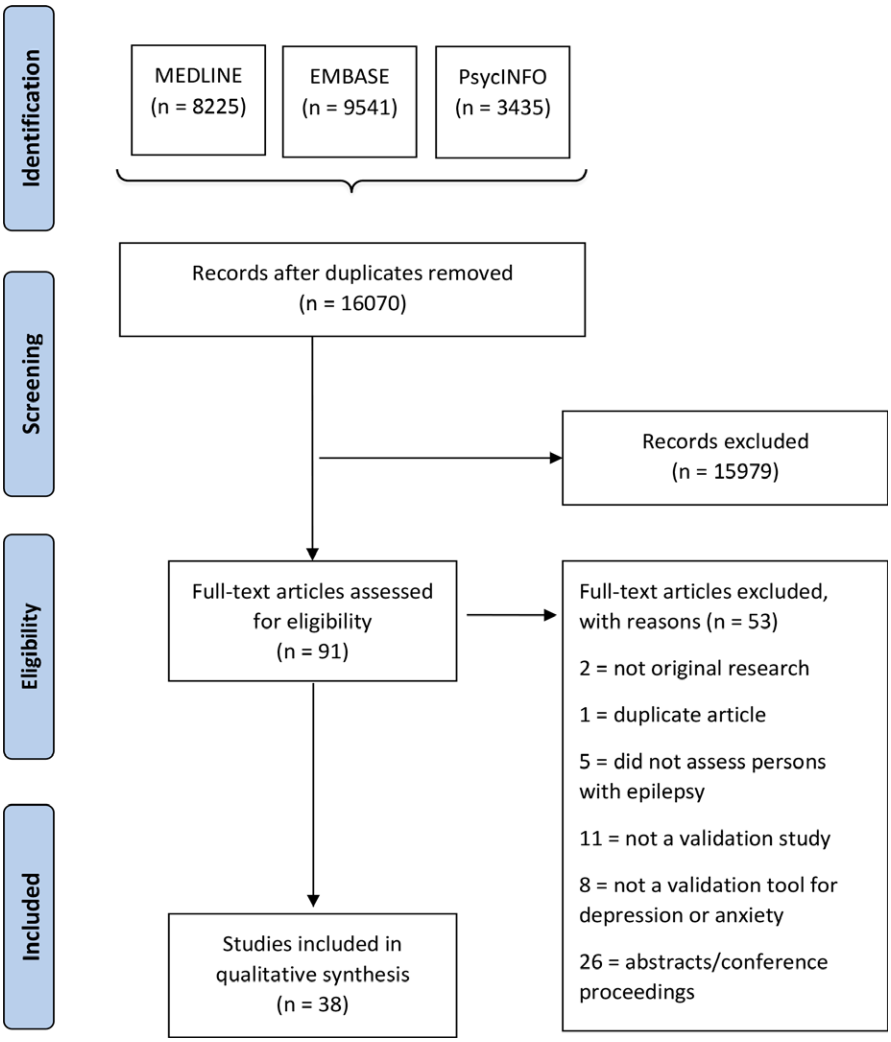

Figure 1.
PRISMA flow diagram.
Epilepsia © ILAE

different languages, n = 19 validations, including the MINI, MINI-Plus, MINI 5.0.0, and MINI 6.0.0[17,19–23,27,29]), undifferentiated structured or semi-structured psychiatric/psychology interviews including expert opinions based on psychiatric consultation (n = 10 validations [16,18,26,28,31,33,36,37]), Composite International Diagnostic Interview (CIDI; n = 1 validation[14]), Schedule for Affective Disorders and Schizophrenia for School Aged Children- Present (K-SADS-P; n = 1 validation[48]), and Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition and Fourth Edition-Text Revision (SCID DSM-IV and IV-TR; n = 3 validations[11,25,32]). The BDI (n = 2 validations[15,44]) and the Major Depression Inventory (MDI; n = 3 validations[24,39,42]) were also used as reference standards.

Performance characteristics

A number of screening tools including the BDI, BDI-II, HRSD-17, PHQ-9, and PHQ-2 had two studies that validated the same cutpoint (Table 1). Based on single estimates with the highest combination of Se and Sp, the recommended cutpoint for the BDI is >16 (Se, 94.4; Sp, 90.6),[21] for the BDI-II is >11 (Se, 96.2; Sp, 80.0),[32] for the HRSD-17 is >6 (Se, 94.4; Sp, 80.4),[41] for the PHQ-9 is ≥10 (Se, 92.0; Sp, 74.0), [43] and for the PHQ-2 is ≥3 (Se, 80.0; Sp, 100.0).[35] Other screening tools, such as the HAM-D and ET7, were validated by multiple studies; however, the medians and ranges were not calculated because studies did not validate the same cutpoints or did not assess the same version of the screening tool. The highest combination of Se and Sp based on single estimates for the HAM-D is a cutpoint of >16 (Se, 95.0; Sp, 75.5)[21] and for the ET7 is ≥29 (Se, 85.4; Sp, 79.2).[42] The fact that many studies did not validate the same cutpoints raises concerns about post hoc selection of cutpoints assessed because generally these scales have preestablished cutpoints. All calculated medians and ranges for the estimates of diagnostic accuracy are found in Table 1.

The HADS-D was validated at the greatest number of cutpoints (n = 10), ranging from 3 to 13, with the most

Table 1. Summary of diagnostic accuracy estimates

Tool	Cutpoint	Sensitivity median (range) n	Specificity median (range) n	PPV median (range) n	NPV median (range) n	AUC median (range) n
BDI	>16	0.915 (0.886–0.944) n = 2	0.909 (0.906–0.912) n = 2	0.793 (0.791–0.795) n = 2	0.966 (0.954–0.977) n = 2	–
BDI AUC Overall (any reference standard used)						0.874 (0.784–0.963) n = 2
BDI-II	>15	0.889 (0.840–0.938) n = 2	0.810 (0.744–0.875) n = 2	0.558 (0.495–0.620) n = 2	0.975 (0.967–0.983) n = 2	0.926 (0.925–0.927) n = 2
HADS-D	3	1.000 (1.000) n = 2	0.316 (0.214–0.418) n = 2	–	–	–
	4	0.964 (0.928–1.000) n = 2	0.457 (0.296–0.618) n = 2	–	–	–
	5	0.928 (0.880–1.000) n = 3	0.428 (0.406–0.846) n = 3	0.526 (0.349–0.702) n = 2	0.952 (0.903–1.000) n = 2	–
	6	0.928 (0.846–1.000) n = 3	0.779 (0.520–0.891) n = 3	0.581 (0.393–0.769) n = 2	0.984 (0.968–1.000) n = 2	–
	7	0.928 (0.808–1.000) n = 3	0.857 (0.620–0.909) n = 3	0.644 (0.488–0.800) n = 2	0.982 (0.964–1.000) n = 2	–
	8	0.771 (0.420–0.875) n = 7	0.853 (0.724–0.991) n = 7	0.581 (0.536–0.972) n = 5	0.944 (0.887–0.956) n = 5	0.869 (0.720–0.900) n = 3
	9	0.725 (0.600–0.786) n = 3	0.855 (0.806–0.991) n = 3	0.784 (0.600–0.967) n = 2	0.882 (0.855–0.908) n = 2	–
	10	0.534 (0.425–0.643) n = 2	0.919 (0.847–0.991) n = 2	–	–	–
	11	0.472 (0.300–0.643) n = 2	0.969 (0.938–1.000) n = 2	–	–	–
	13	0.325 (0.150–0.500) n = 2	0.990 (0.979–1.000) n = 2	–	–	–
HADS-D AUC Overall (any reference standard used)						0.907 (0.792–0.989) n = 3
HADS-T	15	0.819 (0.680–0.957) n = 2	0.712 (0.681–0.743) n = 2	0.440 (0.436–0.444) n = 2	0.922 (0.855–0.988) n = 2	–
HRSD-17	>5	0.893 (0.842–0.944) n = 2	0.893 (0.842–0.944) n = 2	–	–	–
HRSD-17 AUC Overall (any reference standard used)						0.835 (0.774–0.896) n = 2
NDDI-E	>11	0.930 (0.846–0.975) n = 4	0.727 (0.529–0.853) n = 4	0.510 (0.448–0.611) n = 3	0.980 (0.953–0.982) n = 3	–
	>12	0.926 (0.722–1.000) n = 10	0.709 (0.535–0.905) n = 10	0.441 (0.100–0.633) n = 9	0.976 (0.944–1.000) n = 8	–
	>13	0.837 (0.654–0.923) n = 10	0.863 (0.696–0.952) n = 10	0.658 (0.385–0.813) n = 10	0.948 (0.909–0.989) n = 10	0.929 (0.900–0.943) n = 5
	>14	0.787 (0.444–0.879) n = 10	0.885 (0.780–0.967) n = 10	0.643 (0.432–0.744) n = 8	0.932 (0.899–0.970) n = 8	0.838 (0.790–0.886) n = 2
	>15	0.805 (0.640–1.000) n = 12	0.862 (0.810–0.956) n = 12	0.593 (0.230–0.860) n = 11	0.960 (0.881–1.000) n = 11	0.900 (0.869–0.940) n = 5
	>16	0.615 (0.000–0.920) n = 3	0.901 (0.890–1.000) n = 3	0.498 (0.000–0.687) n = 3	0.939 (0.836–0.990) n = 3	–
	>17	0.513 (0.425–0.600) n = 2	0.951 (0.950–0.951) n = 2	0.714 (0.000–0.733) n = 3	0.864 (0.808–0.919) n = 2	–
	>18	0.440 (0.350–0.580) n = 6	0.970 (0.944–0.980) n = 6	0.733 (0.360–0.900) n = 7	0.881 (0.789–0.970) n = 6	–
NDDI-E AUC Overall (any reference standard used)						0.908 (0.850–0.985) n = 8
NDDI-E AUC Overall (using any version of the MINI as the reference standard)						0.908 (0.850–0.958) n = 6
PHQ-9	10	0.830 (0.739–0.920) n = 2	0.802 (0.740–0.863) n = 2	0.460 (0.459–0.460) n = 2	0.963 (0.955–0.970) n = 2	–
	11	0.710 (0.609–0.810) n = 2	0.852 (0.820–0.884) n = 2	0.486 (0.451–0.520) n = 2	0.943 (0.935–0.950) n = 2	–
	12	0.636 (0.522–0.750) n = 2	0.891 (0.820–0.911) n = 2	0.525 (0.480–0.570) n = 2	0.932 (0.924–0.940) n = 2	–
PHQ-9 AUC Overall (any reference standard used)						0.911 (0.905–0.917) n = 2
PHQ-2	>2	0.612 (0.423–0.800) n = 2	0.937 (0.873–1.000) n = 2	0.678 (0.355–1.000) n = 2	0.871 (0.840–0.901) n = 2	–

validations at a cutpoint of 8 (n = 7). At the HADS cutpoint of 8, the median Se (n = 7) was 77.1 (range 42.0–87.5), Sp (n = 7) was 85.3 (range 72.4–99.1), PPV (n = 5) was 58.1 (range 53.6–97.2), NPV (n = 5) was 94.4 (range 88.7–95.6), and the AUC estimate for the cutpoint of 8 (n = 3) was 0.869 (range 0.72–0.90). Based on single estimates with the highest combination of Se and Sp, the recommended cutpoint for the HADS-D would be 7 (Se, 91.0; Sp, 100.0).[14]

The NDDI-E had the second highest number of evaluated cutpoints (n = 8), but it had the most validations at a single cutpoint of >15 (n = 12). At the most validated cutpoint of >15, the median Se (n = 12) was 80.5 (range 64.0–100.0), Sp (n = 12) 86.2 (range 81.0–95.6), PPV (n = 11) 59.3 (range 23.0–86.0), and NPV (n = 11) 96.0 (range 88.1–100.0). The NDDI-E cutpoint of >13 was the second most highly validated cutpoint (n = 10) with a median Se of 83.7 (range 65.4–92.3), Sp of 86.3 (range 69.6–95.2), PPV of 65.8 (range 38.5–81.3), and NPV of 94.8 (range 90.9–98.9). For a better understanding of the relationship between the median Se and median Sp for the different cutpoints of the NDDI-E, the median values were plotted against the cutpoints (Fig. 2). It appears that the best balance of Se and Sp occurs at a cutpoint of >13. Based on single estimates with the highest combination of Se and Sp, the recommended cutpoint for the NDDI-E would be >15 (Se 93.3; Sp 94.4).[15] Median AUC estimates were only calculated when two or more studies validating the same tool calculated the AUC for the same cutpoints or if the study estimated an AUC for the overall tool. These estimates are included in Table 1. Median AUC estimates for the overall screening tools were calculated for the BDI (n = 2) (AUC 87.4%, range 78.4–96.3%), HADS-D (n = 3) (AUC 90.7%, range 79.2–98.9%), HRSD-17 (n = 2) (AUC 83.5%, range 77.4–89.6%), NDDI-E (n = 8) (AUC 90.8%, range 85.0–98.5%), and PHQ-9 (n = 2) (AUC 91.1%, range 90.5–91.7%). Studies that validated the NDDI-E had the highest number of overall AUC estimates, and it was possible to calculate the median AUC according to the reference standard used. Of the eight studies that calculated an overall AUC for the NDDI-E, six of these studies used a version of the MINI as the reference standard.

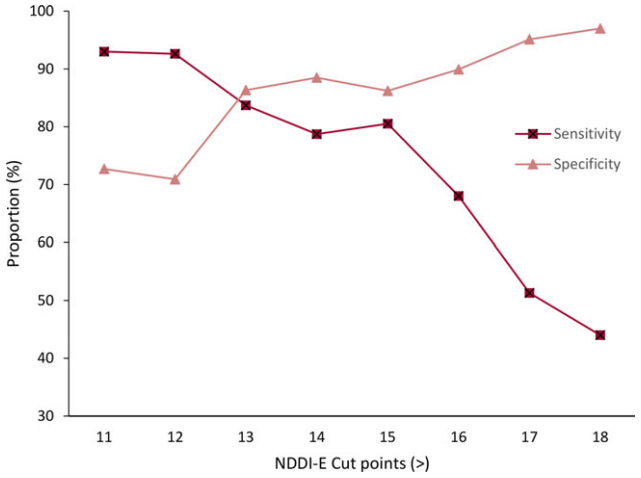

Figure 2.
Median sensitivities and specificities for different NDDI-E cutpoints.
Epilepsia © ILAE

Analyzing the data based on the study quality assessment was possible only for the NDDI-E (Appendix S5). Generally, there is not a consistent trend in median estimates of diagnostic accuracy for the low- versus high-risk studies. For example, for the NDDI-E at a cutpoint of >13, the low risk of bias studies have a median Se of 84.0 and Sp of 86.6, whereas the high risk of bias studies have a median Se of 82.5 and Sp of 73.0. In contrast, at a cutpoint of >15m the low risk of bias studies have a median Se of 76.3 and Sp of 85.8, whereas the high risk of bias studies have a median Se of 80.5 and Sp of 86.2.

Risk of bias assessment

Of the 38 studies included, 13 had unclear risk of bias in at least one of the four categories,[11,17,21,22,24,31,36,39] whereas six studies had a high risk of bias in at least one of the four QUADAS-2 rating system categories[26,28,35,38,48,49] (Appendix S3). Generally, this unclear or high risk of bias was in the "index test" category. Specifically, it was unclear if the screening tool was interpreted without knowledge of the reference standard results. Nineteen studies were in the low risk of bias in all categories.[14–16,18–20,23,25]

Discussion

This systematic review summarizes the published literature on depression-screening tool validations in persons with epilepsy. When treating persons with epilepsy, it is important to choose screening tools that accurately detect depression so that it can be treated appropriately and lead to improved outcomes. To guide the proper selection of screening tools for use in clinical practice, it is essential to assess the diagnostic accuracy of these tools and to ensure that the results are interpreted correctly.[51] Screening tools are seldom 100% accurate; however, a depression-screening tool in epilepsy will preferably provide adequate information that will maximize the correct identification of depression but minimize the potential harm that could result from treatment if a tool is too sensitive.

The most validated HADS-D cutpoint of 8, is the cutpoint that is recommended to detect depression in general populations.[52] Although fewer studies validated the HADS-D in persons with epilepsy and the test requires a paid license for administration, theoretically it still may be a beneficial tool for detecting depression in persons with epilepsy, since it does not include somatic symptoms of depression that may overlap with side effects from medication or the disease

Box 1. Description of included depression index tests

1 *__Neurological Disorders Depression Inventory for Epilepsy__ (NDDI-E n = 26). The NDDI-E is an epilepsy-specific self-rating inventory designed to rapidly determine major depressive episodes in neurology clinics.[11] It was designed to identify depressive symptoms that can be set apart from antiepileptic drug side effects.[11] It has been validated in a number of different languages.

2 __Beck Depression Inventory__ (BDI n = 4; mBDI n = 1; BDI-FS n = 1; BDI-II n = 5; BDI Cognitive-Affective subscale n = 1). The BDI is a self-reported inventory for assessing depressive symptoms within the preceding 2 weeks.[61] The first version of this inventory is based on the earlier DSM-III criteria for diagnosing depression, whereas the BDI-II is based on the DSM-IV criteria for diagnosing depression—both have 21 questions. The modified BDI (mBDI) is a modified version of the original BDI that includes positive feeling choices in addition to the negative choices.[62] The BDI-FS is the seven-item fast screen assessment for medical patients based on the DSM-5 criteria for major depression.[63] The BDI cognitive-affective subscale assesses the mental component of depression.

3 __Hospital Anxiety and Depression Scale__ (HADS-Total n = 2; HADS-Depression n = 7; HADS-Anxiety n = 1). The HADS-Total is a 14-item tool comprising a 7-item depression subscale and a 7-item anxiety subscale. The HADS-Total score and the HADS-Depression score can both be used to assess depression in a number of different populations.[64]

4 *__Emotional Thermometers__ (ET7 n = 2; ET4 n = 1; AnxT n = 1; DepT n = 1). A visual analog tool for patients to subjectively rate using a Likert scale, composed of seven pictorial thermometers for (1) distress, (2) anxiety, (3) depression, (4) anger, (5) duration, (6) burden, and (7) need help. This tool was originally developed by the National Comprehensive Cancer Network to screen for psychological distress in oncology.[65] The ET7 uses all seven domains; the ET4 is a previous version of the tool and uses distress, anxiety, depression, and anger domains. The AnxT is the individual thermometer for anxiety and DepT is depression.

5 *__Patient Health Questionnaire 9 Questions__ (PHQ-9 n = 4). The PHQ is a self-rated questionnaire composed of nine items to screen for depression based on DMS-IV diagnostic criteria.[66] The PHQ-9 assesses depressive symptoms over the previous 2 weeks.

6 *__Patient Health Questionnaire 2 Questions__ (PHQ-2 n = 2). The PHQ-2 uses the first two questions from the PHQ-9 assessing the two cardinal symptoms of depression: diminished interest or pleasure and depressed mood.[67]

7 *__Hamilton Rating Scale for Depression__ (HAM-D n = 1; HRSD-17 n = 2; HRSD-21 n = 1). This is a clinician-rated depression-screening tool that is widely used and available in a number of languages.[41,68] The HRSD-17 consists of 17 questions and HRSD-21 has the same first 17 questions, with questions 18–21 used to obtain further information about depressive symptoms. Additional versions of the HRSD can have up to 29 questions, with the HAM-D in question having 24 items.

8 *__Center for Epidemiologic Studies Depression Scale__ (CES-D n = 1). A 20 item self-report scale that was developed from multiple other validated depression scales for use in the general population.[69]

9 __Child Behavior Checklist__ (CBCL n = 1). The Achenbach System of Empirically Based Assessment (ASEBA) is a group of forms created for assessing behavioral, emotional, and adaptive issues and characteristics.[70] The CBCL is a component of ASEBA and can be used to detect DSM-oriented recent depression and is completed by a parent.

10 __Adult Self-Report__ (ASR n = 1). Another form from the ASEBA, the ASR, is a self-completed report for individuals ages 18–59 and can be used to detect DSM-oriented recent depression.[70]

11 __Mbewe et al. 10-Item Screening Tool__ (n = 1). Mbewe and colleagues developed a 10-item questionnaire administered by a primary health care worker to detect anxiety and depression in people with epilepsy (PWE) in Zambia.[36] The questionnaire was developed in two languages: Nyanja and Bemba.

12 __Parent Questionnaire for Psychiatric Disorders__ (n = 1). Determines if a child has ever had depression or attention-deficit/hyperactivity disorder (ADHD) using single questions inquiring about lifetime depression and lifetime ADHD.[31]

13 __Young Adult Questionnaire for Psychiatric Disorders__ (n = 1). Determines if the child was ever diagnosed with depression, anxiety, or ADHD using single questions. Administered by a trained interviewer.[31]

14 __World Health Organization index for psychological well-being__ (WHO-5 n = 1). A 5-item questionnaire that uses positive statements to measure an individual's mental well-being over the preceding 2 weeks.[71] It has been validated for detecting depression in a number of clinical populations.

15 **Minnesota Multiphasic Personality Inventory-2** (MMP1 Scale 2 n = 1). The MMPI-2 is the second version of the questionnaire and is a known measure of psychopathology. It has 567 items to determine nine clinical scales and a number of subscales.[72] Scale 2 is the depression subscale consisting of five subscales for the evaluation of different characteristics of depression. This questionnaire has been validated in the general population with increasing focus in neurologic conditions.
16 **Major Depression Inventory** (MDI n = 1). Depression scale consisting of items corresponding to ICD-10 symptoms of depression over the preceding 2 weeks.[71]
*Freely available in the public domain.

itself.[53] The NDDI-E has the highest number of validations at each cutpoint (i.e., the most validated tool), which is not surprising as it was developed specifically for the rapid detection of depressive symptoms in persons with epilepsy.[11] The NDDI-E was designed to be a brief, straightforward screening tool for use in outpatient neurology settings as a means to overcome barriers of depression management in persons with epilepsy.[11] In this review, the NDDI-E cutpoint of >15 was the most validated cutpoint. The NDDI-E was originally developed and validated against the SCID (a gold standard for detecting depression).[11] The authors found that a cutpoint of >15 had the highest Se, Sp, and PPV for detecting major depression, indicating it was the optimal cutpoint for detection of depression.[11]

Although the NDDI-E cutpoint of >15 has been recommended, when plotting the median Se and Sp for the NDDI-E, it appears that the cutpoint of >13 may also be optimal for detecting depression in persons with epilepsy because the Se and Sp curves converge at this point, indicating the best balance of Se and Sp. Our findings show that NDDI-E cutpoints of >15 and >13 both have higher Sp than Se [(>15: Sp, 86.2; Se, 80.5) (>13: Sp, 86.3; Se, 83.7)], but both Se and Sp are slightly higher at a cutpoint of >13. However, of the 11 studies validating the NDDI-E at a cutpoint of >13, not all used the SCID for detecting depression, and therefore the varied reference standards used may be under- or overrepresenting true positive cases, resulting in a different threshold for the NDDI-E. However, this is likely not a significant concern, as the median AUC did not differ between NIDDI-E validation studies that used any reference standard versus a gold standard. Mental health resources are often limited in many epilepsy care settings whether in low-, middle-, or high-income countries. As such, a tool or cutpoint with higher Sp is desirable to minimize false positives, whereas in the rare clinical settings where greater mental health resources are available, detection of depression in patients with epilepsy can be optimized using a tool or cutpoint with a higher Se.

Furthermore, the majority of the NDDI-E validation studies were conducted to validate the NDDI-E for use in a language other than English. The authors indicated the NDDI-E was directly translated from English into the respective language and then backtranslated to ensure reliability between the two versions,[15,17,19,23,29,34,37,38] also known as the Brislin Technique.[54] Although this step should ensure reliable data for comparison, it must be noted that these different language validations were combined when determining the medians and ranges for the estimates of diagnostic accuracy.

The most common reference standard used was the MINI. The MINI has been validated previously against the SCID in persons with epilepsy and was found to have high concordance with the SCID in the identification of current major depressive episodes ($\kappa = 0.86$).[55] The SCID is widely accepted as a gold standard for detecting depression in research[56]; however, it may not be the best tool for use in clinical settings as it is time consuming and is supposed to be administered by a mental health professional. Although usually regarded as a brief screening instrument and not a "true gold standard," it is understandable that the MINI, which can be administered by trained staff, possesses face validity, and has a high concordance with the SCID, was so widely used as a reference standard. Most importantly, when assessing the diagnostic accuracy of a screening tool it is important that a recognized reference or gold standard is used (e.g., the SCID or the MINI), as opposed to another screening tool (e.g., the BDI), as this may result in an inaccurate assessment of depression.

The lack of studies validating depression screening tools in children and youths with epilepsy was identified as a significant gap in the literature, with only one of the 38 studies in this review performed in youths with epilepsy. In that study,[48] the NDDI-E was adapted for use in youths with epilepsy ages 10–17. The prevalence of depression in youth with epilepsy has been reported to range between 10% and 30%,[57] with a recent study from the United States reporting a prevalence of 23% in youth and adolescent.[58] Considering the prevalence of depression in youth is similar to, or potentially higher than that, of the prevalence of depression in adults with epilepsy, it is imperative that depression-screening tools also be validated in these younger populations. Future research should focus on the development of new screening tools for youth or the adaptation of existing screening tools traditionally used in adults, such as the NDDI-E youth. Hopefully these tools can then be disseminated for clinical use to ensure improved detection of depression in youths with epilepsy.

In this review we identified a great deal of variability in screening tool cutpoints, language of screening tools assessed, and reference standards, as well as a lack of standardized reporting of diagnostic accuracy estimates precluding a meta-analysis. Furthermore, studies more often reported AUC statistics and corresponding precision estimates; however, these estimates provide less clinical utility, as they are more difficult to interpret and apply in clinical practice.[51] Another concern was the lack of reporting on the variability of estimates. It is notable that Se and Sp were reported in 34 studies with only 6 of these studies reporting corresponding confidence intervals. PPVs and NPVs were reported in 32 studies (plus one study reported only PPVs), with only 4 studies reporting the corresponding confidence intervals. Furthermore, in a number of studies, the PPV and NPV were reported only for the cutpoint with the best Se and Sp and not for all cutpoints assessed. However, predictive values are less generalizable, since they depend on base-rates, whereas Se and Sp are primary test characteristics, which may explain the lack of reporting for predictive values. Based on these observations, it is evident that incorporating adequate standards of reporting, such as the variability of estimates, is important as it allows for an in-depth critical appraisal by readers and researchers, and therefore provides a greater contribution to the field.[59] To increase the value of research that assesses diagnostic accuracy of screening tools, manuscripts should adhere to published guidelines such as the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement.[59] The STARD statement provides a comprehensive list of essential items to be included in a published manuscript.

Although a number of studies reported the subtypes of epilepsy patients included in their overall sample, they did not stratify the results by epilepsy subtype (e.g., incident vs. prevalent, temporal lobe vs. frontal lobe). Future studies should address the validity of depression-screening tools based on epilepsy subtypes. Another limitation identified in the literature is the selection of cutpoints post hoc based on the study sample. By using a sample-based approach to choose the cutpoints reported, there may be an overestimation in the validity of the tools due to selective reporting around cutpoints where better performance is partially due to sampling variability.[60] Therefore, these screening tools may not work as well in practice as they do for the samples used in the studies. To avoid this selective reporting issue, we suggest that studies report all cutpoints they assessed and not only the high performing cutpoints. Another interesting finding was that when PPVs and NPVs were reported, it was apparent that the NPV was always substantially higher than the PPV, regardless of the Se and Sp within or between instruments. This points to the relatively low prevalence of depression in the individual studies (as low as 5.8%). This is important because higher prevalence can inflate predictive values, and therefore the data presented must be interpreted cautiously in studies where the prevalence of depression in persons with epilepsy is lower or higher than expected in that setting. However, most studies with lower prevalence of depression in epilepsy than expected did not report predictive values, and those that did could not be combined in any analyses due to limited data reporting. Therefore, it is not possible to provide more precise information about which settings or samples may be more or less accurate.

This systematic review used an expansive search strategy in three large databases with no restrictions on publication language or date and used a well-documented methodology, including PRISMA standards for reporting. Our methodology also included a quality assessment specific for studies assessing the diagnostic accuracy of screening tools, the QUADAS-2, allowing us to critically appraise the risk of bias and applicability of included studies. We observed a high risk of bias for the index test domain. Specifically, it was unclear in a number of studies if their interpretation of the screening tool and reference standard results were blinded. In many of the cases, we believe it was a matter of unclear reporting rather than methodologic issues. Specifically, due to the nature of the self-reported scales being evaluated, it is possible that authors did not see the need to clarify this point. All articles were included in the systematic review, regardless of the QUADAS-2 assessment. In fact, the median estimates of diagnostic accuracy do not appear to reveal better or worst performance whether one looks at low versus high risk of bias studies. Therefore, we do not feel that including all studies in our analysis influenced the final conclusions of the study.

Another limitation of this systematic review was the inability to synthesize the data using meta-analysis, and consequently we were unable to make recommendations about the best cutpoints for the different screening tools identified. Finally, although we followed a well-documented and rigorous methodology when conducting our systematic reviews for these studies of diagnostic accuracy, these types of systematic reviews may still be limited to identifying only areas of weakness in the studies rather than providing pooled estimates that can be used to make strong recommendations.[51]

A number of factors must be considered when determining the optimal tool to screen for depression in persons with epilepsy. Ultimately the ideal tool will depend on the properties of the screening tool and the clinical and health care needs of the setting in which it is used. Generally, the optimal tool will depend on resource availability. Although we have insufficient evidence to make a definitive suggestion on the best tool and cutpoint, we believe the NDDI-E may be most practical for a variety of settings as it is freely available in the public domain, relatively easy to score, and it has been validated in many languages. This review provides a thorough overview of the current literature looking at validations of depression-screening tools in persons with

epilepsy, and we recommend that screening tools be chosen based on available resources and the intentions of the clinic (i.e., importance placed on highest Se, Sp, or PPV). Future research should focus on validating depression-screening tools against the best possible reference standard with adherence to the STARD statement, and consider the development and/or validation of screening tools in children, youth, and the elderly where serious gaps in knowledge exist.

Acknowledgements

N. Jette is supported by a Canada Research Chair Tier 2 in Neurological Health Services Research. This study was partly supported by funding from the Hotchkiss Brain Institute Neuro and Mental Health Research Clinic Initiative.

Disclosure

Stephanie Gill, Scott Patten, Samuel Wiebe, Kirsten Fiest, and Sara Lukmanji have no conflicts of interest to disclose. Nathalie Jette in an Associate Editor of *Epilepsia*. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

References

1. Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;380:2163–2196.
2. Tellez-Zenteno JF, Patten SB, Jette N, et al. Psychiatric comorbidity in epilepsy: a population-based analysis. *Epilepsia* 2007;48:2336–2344.
3. Fiest KM, Dykeman J, Patten SB, et al. Depression in epilepsy: a systematic review and meta-analysis. *Neurology* 2013;80:590–599.
4. DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. *Arch Intern Med* 2000;160:2101–2107.
5. Kobau R, Gilliam F, Thurman DJ. Prevalence of self-reported epilepsy or seizure disorder and its associations with self-reported depression and anxiety: results from the 2004 HealthStyles Survey. *Epilepsia* 2006;47:1915–1921.
6. Hesdorffer DC, Ishihara L, Webb DJ, et al. Occurrence and recurrence of attempted suicide among people with epilepsy. *JAMA Psychiatry* 2016;73:80–86.
7. Boylan LS, Flint LA, Labovitz DL, et al. Depression but not seizure frequency predicts quality of life in treatment-resistant epilepsy. *Neurology* 2004;62:258–261.
8. Fiest KM, Patten SB, Jette N. Screening for depression and anxiety in epilepsy. *Neurol Clin* 2016;34:351–361.
9. Pettersson A, Bostrom KB, Gustavsson P, et al. Which instruments to support diagnosis of depression have sufficient accuracy? A systematic review *Nord J Psychiatry* 2015;69:497–508.
10. Lakkis NA, Mahmassani DM. Screening instruments for depression in primary care: a concise review for clinicians. *Postgrad Med* 2015;127:99–106.
11. Gilliam FG, Barry JJ, Hermann BP, et al. Rapid detection of major depression in epilepsy: a multicentre study. *Lancet Neurol* 2006;5:399–405.
12. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
13. 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;155:529–536.
14. Al-Asmi A, Dorvlo AS, Burke DT, et al. The detection of mood and anxiety in people with epilepsy using two-phase designs: experiences from a tertiary care centre in Oman. *Epilepsy Res* 2012;98:174–181.
15. Alkhamees HA, Selai CE, Shorvon SD, et al. The use of the NDDI-E in Arabic to identify symptoms of depression of moderate or greater severity in people with epilepsy. *Epilepsy Behav* 2014;32:55–58.
16. Botega NJ, Ponde MP, Medeiros P, et al. Validation of the Hospital Anxiety and Depression Scale in ambulatory epileptic patients. *J Bras Psiquiatr* 1998;47:285–289.
17. Brandt C, Labudda K, Illies D, et al. Rapid detection of a depressive disorder in persons with epilepsy. Validation of a German version of the NDDI-E. *Nervenarzt* 2014;85:1151–1155.
18. de Lemos Zingano B, Guarnieri R, Diaz AP, et al. Validation of diagnostic tests for depressive disorder in drug-resistant mesial temporal lobe epilepsy. *Epilepsy Behav* 2015;50:61–66.
19. de Oliveira GN, Kummer A, Salgado JV, et al. Brazilian version of the Neurological Disorders Depression Inventory for Epilepsy (NDDI-E). *Epilepsy Behav* 2010;19:328–331.
20. de Oliveira GN, Lessa JM, Goncalves AP, et al. Screening for depression in people with epilepsy: comparative study among neurological disorders depression inventory for epilepsy (NDDI-E), hospital anxiety and depression scale depression subscale (HADS-D), and Beck depression inventory (BDI). *Epilepsy Behav* 2014;34:50–54.
21. de Oliveira GNM, de Araujo Filho GM, Kummer A, et al. Beck Depression Inventory (BDI) and Hamilton Rating Scale for Depression (HAM-D) in patients with epilepsy. *J Bras Psiquiatr* 2011;60:131–134.
22. De Oliveira GNM, De Araujo Filho GM, Kummer A, et al. Neurological disorders depression inventory for epilepsy (NDDI-E): Brazilian version of a screening instrument [Portuguese]. *JECN* 2011;17:49–53.
23. Di Capua D, Garcia-Garcia ME, Reig-Ferrer A, et al. Validation of the Spanish version of the Neurological Disorders Depression Inventory for Epilepsy (NDDI-E). *Epilepsy Behav* 2012;24:493–496.
24. Drinovac M, Wagner H, Agrawal N, et al. Screening for depression in epilepsy: a model of an enhanced screening tool. *Epilepsy Behav* 2015;44:67–72.
25. Fiest KM, Patten SB, Wiebe S, et al. Validating screening tools for depression in epilepsy. *Epilepsia* 2014;55:1642–1650.
26. Friedman DE, Kung DH, Laowattana S, et al. Identifying depression in epilepsy in a busy clinical setting is enhanced with systematic screening. *Seizure* 2009;18:429–433.
27. Gandy M, Sharpe L, Perry KN, et al. Assessing the efficacy of 2 screening measures for depression in people with epilepsy. *Neurology* 2012;79:371–375.
28. Gross DW, Davies M, Jirsch JD, et al. Screening for depression in a tertiary epilepsy clinic. *Can J Neurol Sci* 2014;41:525–527.
29. Guo Y, Chen ZM, Zhang YX, et al. Reliability and validity of the Chinese version of the Neurological Disorders Depression Inventory for Epilepsy (C-NDDI-E). *Epilepsy Behav* 2015;45:225–228.
30. Hansen CP, Amiri M. Combined detection of depression and anxiety in epilepsy patients using the Neurological Disorders Depression Inventory for Epilepsy and the World Health Organization well-being index. *Seizure* 2015;33:41–45.
31. Hesdorffer DC, Baldin E, Caplan R, et al. How do we measure psychiatric diagnoses? Implications of the choice of instruments in epilepsy. *Epilepsy Behav* 2014;31:351–355.
32. Jones JE, Hermann BP, Woodard JL, et al. Screening for major depression in epilepsy with common self-report depression inventories. *Epilepsia* 2005;46:731–735.
33. Karzmark P, Zeifert P, Barry J. Measurement of depression in epilepsy. *Epilepsy Behav* 2001;2:124–128.
34. Ko PW, Hwang J, Lim HW, et al. Reliability and validity of the Korean version of the Neurological Disorders Depression Inventory for Epilepsy (K-NDDI-E). *Epilepsy Behav* 2012;25:539–542.
35. Margrove K, Mensah S, Thapar A, et al. Depression screening for patients with epilepsy in a primary care setting using the Patient Health Questionnaire-2 and the Neurological Disorders Depression Inventory for Epilepsy. *Epilepsy Behav* 2011;21:387–390.
36. Mbewe EK, Uys LR, Nkwanyana NM, et al. A primary healthcare screening tool to identify depression and anxiety disorders among people with epilepsy in Zambia. *Epilepsy Behav* 2013;27:296–300.

37. Metternich B, Wagner K, Buschmann F, et al. Validation of a German version of the Neurological Disorders Depression Inventory for Epilepsy (NDDI-E). *Epilepsy Behav* 2012;25:485–488.
38. Micoulaud-Franchi JA, Barkate G, Trebuchon-Da Fonseca A, et al. One step closer to a global tool for rapid screening of major depression in epilepsy: validation of the French NDDI-E. *Epilepsy Behav* 2015;44:11–16.
39. Mitchell AJ, Ioannou N, Rampling JM, et al. Which symptoms are indicative of depression in epilepsy settings? An analysis of the diagnostic significance of somatic and non-somatic symptoms. *J Affect Disord* 2013;150:861–867.
40. Mula M, Iudice A, La Neve A, et al. Validation of the Italian version of the Neurological Disorders Depression Inventory for Epilepsy (NDDI-E). *Epilepsy Behav* 2012;24:329–331.
41. Mula M, Iudice A, La Neve A, et al. Validation of the Hamilton Rating Scale for Depression in adults with epilepsy. *Epilepsy Behav* 2014;41:122–125.
42. Rampling J, Mitchell AJ, Von Oertzen T, et al. Screening for depression in epilepsy clinics. A comparison of conventional and visual-analog methods. *Epilepsia* 2012;53:1713–1721.
43. Rathore JS, Jehi LE, Fan Y, et al. Validation of the Patient Health Questionnaire-9 (PHQ-9) for depression screening in adults with epilepsy. *Epilepsy Behav* 2014;37:215–220.
44. Ristic AJ, Pjevalica J, Trajkovic G, et al. Validation of the Neurological Disorders Depression Inventory for Epilepsy (NDDI-E) Serbian version. *Epilepsy Behav* 2016;57(Part A):1–4.
45. Tadokoro Y, Oshima T, Fukuchi T, et al. Screening for major depressive episodes in Japanese patients with epilepsy: validation and translation of the Japanese version of Neurological Disorders Depression Inventory for Epilepsy (NDDI-E). *Epilepsy Behav* 2012;25:18–22.
46. Thomson AE, Calle A, Fontela ME, et al. Screening of major depression in epilepsy: the Neurologic Depression Disorders Inventory in Epilepsy-Spanish version (Argentina). *Epilepsia* 2014;55:331–334.
47. Tong X, An D, Lan L, et al. Validation of the Chinese version of the Neurological Disorders Depression Inventory for Epilepsy (C-NDDI-E) in West China. *Epilepsy Behav* 2015;47:6–10.
48. Wagner JL, Smith G, Ferguson PL, et al. Preliminary psychometrics of the Neurological Disorders Depression Inventory for Epilepsy-Youth. *J Child Neurol* 2013;28:1392–1399.
49. Walker ER, Engelhard G Jr, Thompson NJ. Using Rasch measurement theory to assess three depression scales among adults with epilepsy. *Seizure*. 2012;21:437–443.
50. Zis P, Yfanti P, Siatouni A, et al. Validation of the Greek version of the Neurological Disorders Depression Inventory for Epilepsy (NDDI-E). *Epilepsy Behav* 2013;29:513–515.
51. Egger M, Davey-Smith G, Altman D. *Systematic reviews in health care: meta-analysis in context*. London: John Wiley & Sons; 2008.
52. Bjelland I, Dahl AA, Haug TT, et al. The validity of the hospital anxiety and depression scale. an updated literature review. *J Psychosom Res* 2002;52:69–77.
53. Snaith RP, Zigmond AS. The hospital anxiety and depression scale. *Br Med J (Clin Res Ed)* 1986;292:344.
54. Brislin RW. Back-translation for cross-cultural research. *J Cross Cult Psychol* 1970;1:185–216.
55. Jones JE, Hermann BP, Barry JJ, et al. Clinical assessment of Axis I psychiatric morbidity in chronic epilepsy: a multicenter investigation. *J Neuropsychiatry Clin Neurosci* 2005;17:172–179.
56. Sanchez-Villegas A, Schlatter J, Ortuno F, et al. Validity of a self-reported diagnosis of depression among participants in a cohort study using the Structured Clinical Interview for DSM-IV (SCID-I). *BMC Psychiatry*. 2008;8:43.
57. Jones JE, Austin JK, Caplan R, et al. Psychiatric disorders in children and adolescents who have epilepsy. *Pediatr Rev* 2008;29:e9–e14.
58. Guilfoyle SM, Monahan S, Wesolowski C, et al. Depression screening in pediatric epilepsy: evidence for the benefit of a behavioral medicine service in early detection. *Epilepsy Behav* 2015;44:5–10.
59. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clin Chem* 2015;61:1446–1452.
60. Thombs BD, Benedetti A, Kloda LA, et al. Diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for detecting major depression in pregnant and postnatal women: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open* 2015;5:e009742.
61. Beck AT, Ward CH, Mendelson M, et al. An inventory for measuring depression. *Arch Gen Psychiatry* 1961;4:561–571.
62. Dori GA, Overholser JC. Evaluating depression severity and remission with a modified Beck Depression Inventory. *Pers Individ Dif* 2000;28:1045–1061.
63. Kliem S, Mößle T, Zenger M, et al. Reliability and validity of the beck depression inventory-fast screen for medical patients in the general German population. *J Affect Disord* 2014;156:236–239.
64. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;67:361–370.
65. Mitchell AJ, Baker-Glenn EA, Granger L, et al. Can the Distress Thermometer be improved by additional mood domains? Part I. Initial validation of the Emotion Thermometers tool. *Psychooncology* 2010;19:125–133.
66. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606–613.
67. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care* 2003;41:1284–1292.
68. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry* 1960;23:56–62.
69. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Appl Psychol Meas* 1977;1:385–401.
70. Achenbach T. Child Behavior Checklist. In Kreutzer JS, DeLuca J, Caplan B (Eds) *Encyclopedia of clinical neuropsychology*. New York: Springer; 2011:546–552.
71. Bech P. *Clinical psychometrics*. Oxford: John Wiley & Sons; 2012.
72. Butcher JN. *Minnesota Multiphasic Personality Inventory-2: Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press; 2001.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. PRISMA check list.

Appendix S2. Search strategy.

Appendix S3. QUADAS-2 assessment.

Appendix S4. Summary of studies assessing the validity of depression-screening tools in persons with epilepsy.

Appendix S5. Summary of Diagnostic Accuracy estimates for low- versus high-risk studies according to the QUADAS-2 assessment.