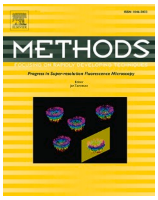
Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth



External validation of a shortened screening tool using individual participant data meta-analysis: A case study of the Patient Health Questionnaire-Dep-4

Daphna Harel[a,b,*], Brooke Levis[c], Ying Sun[d], Felix Fischer[e], John P.A. Ioannidis[f,g,h,i], Pim Cuijpers[j], Scott B. Patten[k], Roy C. Ziegelstein[l], Sarah Markham[m], Andrea Benedetti[n,o,p], Brett D. Thombs[d,n,o,q,r,s,t], the DEPRESsion Screening Data DEPRESSD PHQ Collaboration[1]

[a] *Department of Applied Statistics, Social Science, and Humanities, New York University, United States*
[b] *Center for the Promotion of Research at the Intersection of Information, Society, and Methodology, New York University, United States*
[c] *Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, United Kingdom*
[d] *Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada*
[e] *Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité –Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität Zu Berlin, and Berlin Institute of Health, Berlin, Germany*
[f] *Department of Medicine, Stanford University, Stanford, CA, USA*
[g] *Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA*
[h] *Department of Biomedical Data Science, Stanford University, Stanford, CA, USA*
[i] *Department of Statistics, Stanford University, Stanford, CA, USA*
[j] *Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, the Netherlands*
[k] *Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada*
[l] *Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[m] *Department of Biostatistics and Health Informatics, King's College London, London, UK*
[n] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada*
[o] *Department of Medicine, McGill University, Montréal, Québec, Canada*
[p] *Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada*
[q] *Department of Psychiatry, McGill University, Montréal, Québec, Canada*
[r] *Department of Psychology, McGill University, Montréal, Québec, Canada*
[s] *Department of Educational and Counselling Psychology, McGill University, Montréal, Québec, Canada*
[t] *Biomedical Ethics Unit, McGill University, Montréal, Québec, Canada*

ARTICLE INFO

Keywords:
Optimal test assembly
Sensitivity
Specificity
Equivalence testing
Self-report questionnaire

ABSTRACT

Shortened versions of self-reported questionnaires may be used to reduce respondent burden. When shortened screening tools are used, it is desirable to maintain equivalent diagnostic accuracy to full-length forms. This manuscript presents a case study that illustrates how external data and individual participant data meta-analysis can be used to assess the equivalence in diagnostic accuracy between a shortened and full-length form. This case study compares the Patient Health Questionnaire-9 (PHQ-9) and a 4-item shortened version (PHQ-Dep-4) that was previously developed using optimal test assembly methods. Using a large database of 75 primary studies (34,698 participants, 3,392 major depression cases), we evaluated whether the PHQ-Dep-4 cutoff of ≥ 4 maintained equivalent diagnostic accuracy to a PHQ-9 cutoff of ≥ 10. Using this external validation dataset, a PHQ-Dep-4 cutoff of ≥ 4 maximized the sum of sensitivity and specificity, with a sensitivity of 0.88 (95% CI 0.81, 0.93), 0.68 (95% CI 0.56, 0.78), and 0.80 (95% CI 0.73, 0.85) for the semi-structured, fully structured, and MINI reference standard categories, respectively, and a specificity of 0.79 (95% CI 0.74, 0.83), 0.85 (95% CI 0.78, 0.90), and 0.83 (95% CI 0.80, 0.86) for the semi-structured, fully structured, and MINI reference standard categories, respectively. While equivalence with a PHQ-9 cutoff of ≥ 10 was not established, we found the sensitivity of the PHQ-Dep-4 to be non-inferior to that of the PHQ-9, and the specificity of the PHQ-Dep-4 to be marginally smaller than the PHQ-9.

* Corresponding author at: 246 Greene Street, 3rd floor, New York, NY 10003, USA.
E-mail address: daphna.harel@nyu.edu (D. Harel).

[1] The members of the DEPRESsion Screening Data DEPRESSD PHQ Collaboration are listed in Appendix A at the end of the article.

https://doi.org/10.1016/j.ymeth.2021.11.005
Received 16 June 2021; Received in revised form 15 October 2021; Accepted 9 November 2021
Available online 12 November 2021

Please cite this article as: t, *Methods*, https://doi.org/10.1016/j.ymeth.2021.11.005

1. Introduction

Self-reported symptom measures are used to assess mental health symptoms and may also be used to screen for mental disorders. However, in clinical practice and research, individuals may be asked to complete several measures, each with multiple items or domains, which can be demanding on their time, and sensitive items, such as asking about suicidal ideation, may be emotionally burdensome [1–4]. Long measures can result in poor data quality and high amounts of missing data. Thus, shortened forms that do not significantly reduce diagnostic accuracy can provide meaningful data while reducing respondent burden and potentially increasing data quality.

The Patient Health Questionnaire-9 (PHQ-9) is a 9-item, self-report questionnaire that measures depressive symptoms [5–7]. Scores on each item on the PHQ-9 range reflect symptoms in the last 2 weeks and range from 0 (“not at all”) to 3 (“every day”). Scores range from 0 to 27 with higher scores indicating higher levels of depressive symptomatology.

An individual participant data meta-analysis (IPDMA) on the accuracy of the PHQ-9 to screen for major depression was conducted on 29 studies with a semi-structured diagnostic interview as the reference standard (6,725 participants, 924 major depression cases). This study found that the standard and most commonly used for the PHQ-9, cutoff threshold of ≥ 10, maximized the combination of sensitivity (0.88, 95% CI 0.83, 0.92) and specificity (0.85, 95% CI 0.82, 0.88) [8].

Using a subset of data from the IPDMA, a previous study developed a 4-item shortened form of the PHQ-9, known as the PHQ-Dep-4, through optimal test assembly (OTA) methods. As with the PHQ-9, scores on each item of the PHQ-Dep-4 reflect symptoms in the last 2 weeks and range from 0 (“not at all”) to 3 (“every day”). PHQ-Dep-4 scores range from 0 to 12 with higher scores indicating higher levels of depressive symptomatology.

The initial development study used 20 primary studies (7,850 participants, 863 major depression cases), which we refer to as the development sample, that administered the English version of the PHQ-9 and used a validated semi-structured or fully structured diagnostic interview (Mini International Neuropsychiatric Interview [MINI] excluded) to classify major depression. The PHQ-Dep-4 includes items 1, 2, 6, and 8 from the PHQ-9, representing depressed mood, loss of interest/pleasure, low self-esteem/guilt and psychomotor agitation [9]. OTA is a mixed-integer programming procedure that uses an estimated item response theory model to select the subset of items that best satisfies pre-specified constraints. In the case of the PHQ-Dep-4 development study, there were pre-specified constraints on the concurrent validity, reliability, and equivalency of diagnostic accuracy of the shortened form with the full-length form [10]. Although more commonly used in the development of high-stakes educational tests [11], recent studies have demonstrated that OTA can be used to develop shortened versions of patient-reported outcome measures [9,12–17]. This procedure was shown in a simulation study to be replicable and reproducible, and produce shortened forms of minimal length with limited loss of information [14].

A cutoff of ≥ 4 on the PHQ-Dep-4 was found to perform equivalently to the PHQ-9 cutoff ≥ 10 in the development sample. However, accuracy of the PHQ-Dep-4 has not been externally validated outside of the development sample. It is therefore necessary to investigate whether a cutoff of ≥ 4 on the PHQ-Dep-4 continues to maintain equivalent diagnostic accuracy to the PHQ-9 cutoff ≥ 10. Conducting an external validation of this cutoff allows for the assessment of whether this cutoff was specific to the development dataset or generalizable to other studies or applications in the future. In particular, the development of the PHQ-Dep-4 was based on comparing properties of the full-length form to a set of candidate shortened forms in the development sample, and thus is susceptible to issues of overfitting or a lack of generalizability. By conducting an external validation, it is possible to see whether the equivalence in accuracy of the PHQ-Dep-4 to the PHQ-9 can be confirmed in an independent dataset.

The objective of the present study was to use data from a unique set of studies that administered the PHQ-9 as well as a validated semi-structured or fully structured diagnostic interview for major depression to validate the diagnostic accuracy of the previously developed PHQ-Dep-4. Specifically, we (1) estimated accuracy for all possible PHQ-Dep-4 cutoffs (i.e., ≥ 1 to ≥ 12), and (2) tested equivalency in accuracy for each PHQ-Dep-4 cutoff to that of a PHQ-9 cutoff of ≥ 10, with the comparison of the PHQ-Dep-4 cutoff of ≥ 4 considered the primary comparison.

2. Methods

The present validation study used data synthesized from an updated IPDMA of the screening accuracy of the PHQ-9 for major depression [8,18], excluding datasets that were included in the original PHQ-Dep-4 development project [9]. The present validation study included studies conducted in any language and using any validated semi-structured or fully structured diagnostic interview (MINI included). The main IPDMA was registered in PROSPERO (CRD42014010673) and a protocol was published [19]. The present analysis was not part of the protocol for the main IPDMA, but a separate protocol was developed and posted prior to initiation at https://osf.io/xy2b8/.

3. The main IPDMA database

3.1. Study selection

In the main IPDMA, datasets from articles in any language were eligible for inclusion if (1) they included PHQ-9 scores; (2) they included diagnostic classifications for current Major Depressive Episode (MDE) or Major Depressive Disorder (MDD) based on Diagnostic and Statistical Manual of Mental Disorders (DSM) [20–23], or International Classification of Diseases (ICD) [24] criteria, using a validated semi-structured or fully structured interview; (3) the PHQ-9 and diagnostic interview were administered within two weeks of each other, since diagnostic criteria for major depression are for symptoms in the last two weeks; (4) participants were ≥ 18 years and not recruited from youth or school-based settings; and (5) participants were not recruited from psychiatric settings or because they were identified as having symptoms of depression, since screening is done to identify unrecognized cases. Datasets where not all participants were eligible were included if primary data allowed selection of eligible participants.

3.2. Database sources and search strategy

A medical librarian searched Medline, Medline In-Process & Other Non-Indexed Citations via Ovid; PsycINFO; and Web of Science from January 1, 2000 to May 9, 2018 using a peer-reviewed search strategy (eMethods1) [25]. The search was limited to the year 2000 onwards because the PHQ-9 was first published in 2001 [7]. We also reviewed reference lists of relevant reviews and queried contributing authors about non-published studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After deduplication, remaining citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada) for processing review results.

Two investigators independently reviewed titles and abstracts for eligibility. If either investigator deemed a study potentially eligible, full-text review was done by two investigators, independently, with disagreements resolved by consensus, consulting a third investigator when necessary. Translators were consulted for languages other than those for which team members were fluent.

3.3. Data contribution and synthesis

Authors of eligible datasets were invited to contribute de-identified primary data, including PHQ-9 scores and major depression status. We emailed corresponding authors of eligible primary studies at least three times, as necessary, with at least two weeks between each email. If we did not receive a response, we emailed co-authors and attempted to contact corresponding authors by phone.

Individual participant data were converted to a standard format and synthesized into a single dataset with study-level data. We compared published participant characteristics and diagnostic accuracy results with results from raw datasets and resolved any discrepancies in consultation with the original investigators.

To define major depression, we considered MDD or MDE based on the DSM or ICD. If more than one was reported, we prioritized MDE over MDD, since screening would attempt to detect depressive episodes and further interview would determine if the episode were related to MDD, bipolar disorder, or persistent depressive disorder. When both were present, we prioritized DSM over ICD, because DSM is more commonly used in existing studies.

3.4. Data used in the present analyses

To consider an independent data source for this validation, we excluded the 20 studies that were included in the original PHQ-Dep-4 development project. We note that these 20 studies were originally used in the development paper because of their availability at the time that study was conducted, rather than a deliberate splitting of the sample. In addition, to be able to calculate PHQ-Dep-4 scores, we excluded studies and participants without item-level PHQ-9 data.

4. Statistical analyses

Using the item-level PHQ-9 data, we calculated PHQ-Dep-4 scores by summing the item scores from PHQ-9 items 1 (loss of interest), 2 (depressed mood), 6 (feeling like a failure), and 8 (physical movement). We then conducted two sets of analyses.

To assess diagnostic accuracy, we estimated sensitivity and specificity. Sensitivity, the true positive rate, refers to the probability of scoring above the cutoff in question given that the participant was classified with MDE or MDD based on DSM or ICD criteria using a validated semi-structured or fully structured interview. Specificity, the true negative rate, refers to the probability of scoring below the cutoff in question given that the participant was classified with MDE or MDD based on DSM or ICD criteria using a validated semi-structured or fully structured interview.

First, we estimated sensitivity and specificity for all possible PHQ-Dep-4 cutoffs (i.e., ≥ 1 to ≥ 12), as well as the standard PHQ-9 cutoff score of ≥ 10, which maximizes sensitivity + specificity [8,18]. For each PHQ-Dep-4 cutoff, separately, and for a PHQ-9 cutoff of ≥ 10, we fit bivariate random-effects models using adaptive Gauss-Hermite quadrature with one quadrature point [26]. This is a 2-stage meta-analytic approach that synthesizes sensitivity and specificity simultaneously and accounts for the correlation between them, as well as for precision of estimates within studies. For each analysis, this model provided estimates of pooled sensitivity and specificity.

The formulation of the model can be expressed as the following. Let $y_{s,i}^{(0)}$ be the dichotomous outcome of the screening test (PHQ-9 or PHQ-Dep-4) for the *i*-th participant in the *s*-th primary study who does not have a true depression diagnosis. Therefore, $y_{s,i}^{(0)}$ is equal to one when the participant has a high score on the screening test and zero when the participant has a low score on the screening test. Similarly, let $y_{s,i}^{(1)}$ be the dichotomous outcome of the screening test for the *i*-th participant of the *s*-th primary study who does have a true depression diagnosis. The model is formulated as:

$$y_{s,i}^{(0)} \sim Bernoulli(p_{s,i}^{(0)})$$

$$logit\left(p_{s,i}^{(0)}\right) = \mu_s^{(0)} = \mu^{(0)} + u_s^{(0)}$$

$$y_{s,i}^{(1)} \sim Bernoulli(p_{s,i}^{(1)})$$

$$logit\left(p_{s,i}^{(1)}\right) = \mu_s^{(1)} = \mu^{(1)} + u_s^{(1)}$$

$$\boldsymbol{u}_s = \begin{pmatrix} u_s^{(0)} \\ \hline u_s^{(1)} \end{pmatrix} \sim N(0, \boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \tau_0^2 & \tau_0\tau_1\rho_\tau \\ \hline \tau_0\tau_1\rho_\tau & \tau_1^2 \end{pmatrix}$$

In this case, the false positive rate (FPR), which is equal to 1 – specificity, and the true positive rate (TPR), which is the sensitivity, can be estimated for the pooled logit(FPR) and logit(TPR) through $\widehat{\mu}^{(0)}$ and $\widehat{\mu}^{(1)}$, respectively. $\widehat{\tau}^{(0)}$ and $\widehat{\tau}^{(1)}$ estimates the between-study variance of the logit-transformed parameters, and $\widehat{\rho_\tau}$ is the estimated correlation.

For these analyses, we modeled sensitivity and specificity separately among studies that used each reference standard category (semi-structured, fully structured, or MINI) as well as pooled together. We present accuracy results for the PHQ-Dep-4 separately by reference standard type because previous studies have found that there are important differences in the design and performance of different types of diagnostic interviews used as reference standards [27–30], and that PHQ-9 sensitivity and specificity vary across different reference standards [8,18]. For each reference standard category, we constructed an empirical receiver operating characteristic (ROC) plot for the PHQ-Dep-4 based on pooled sensitivity and specificity estimates from each cutoff. Separately, we marked the point in ROC-space for a PHQ-9 cutoff of ≥ 10.

Second, we tested the equivalence of the PHQ-Dep-4 and PHQ-9. The comparison of the PHQ-Dep-4 cutoff of ≥ 4 to the PHQ-9 cutoff of ≥ 10 was considered as our primary analysis. For these analyses, we pooled reference standard categories together, because although PHQ-9 and PHQ-Dep-4 sensitivity and specificity may differ by reference standard category, we did not believe that *differences* in sensitivity and specificity between PHQ-Dep-4 cutoffs and a PHQ-9 cutoff of ≥ 10 would vary by reference standard category, since each primary study compared the PHQ-Dep-4 and PHQ-9 to the same reference standard. By pooling, we increase power and therefore reduce the risk of an ambiguous outcome in the analysis. In line with this, a previous comparison of the PHQ-8 and PHQ-9 found that although accuracy differed across reference standard categories, differences in accuracy across the forms were similar across reference standard categories [31]. We estimated the crude differences in sensitivity and specificity between each PHQ-Dep-4 cutoff and a PHQ-9 cutoff of ≥ 10 and constructed confidence intervals (CI) for differences via the cluster bootstrap approach [32,33], resampling at study and subject levels with replacement. For each comparison, we ran 1000 iterations of the bootstrap. These CIs allowed us to test whether the sensitivity and specificity of each PHQ-Dep-4 cutoff are equivalent to that of the PHQ-9 based on a pre-specified minimally important difference of $\delta = 0.05$ [34], as has been done in previous studies [9,13,31]. That is, for each cutoff, for differences in sensitivity and specificity separately, we would consider the null hypothesis that there are differences large enough to be important and test that against the alternative hypothesis that there are no meaningful differences. If the entire CI is included within the interval of -0.05 to $+0.05$, we would reject the null hypothesis and conclude that equivalence is present. If the entire CI is outside of the interval, we would conclude that the accuracies are not equivalent. If the CIs cross the interval of -0.05 to $+0.05$, findings would be deemed ambiguous, and the equivalence would be found to be indeterminate. Lastly, we determined which PHQ-Dep-4 cutoff showed the smallest overall sum of absolute differences in accuracy (i.e. in sensitivity and in specificity) compared to PHQ-9 ≥ 10.

All analyses were conducted in R (R version R 3.4.1 [35], RStudio version 1.0.143) using the *glmer* function within the *lme4* package [36]. All R code used to run the analysis is included in the supplementary materials, however due to data sharing agreements, the raw data is not available.

4.1. Ethics

As this study involves secondary analysis of de-identified previously collected data, the Research Ethics Committee of the Jewish General Hospital determined that it did not require research ethics approval. However, for each included dataset, we confirmed that the original study received ethics approval and that all participants provided informed consent.

5. Results

5.1. Search results and dataset inclusion

Fig. 1 illustrates the study flow diagram. Of 9,670 unique titles and abstracts identified from database searches, 9,199 were excluded at the title and abstract review stage and 297 after full-text review. After removing duplicate samples, adding unpublished studies contributed by

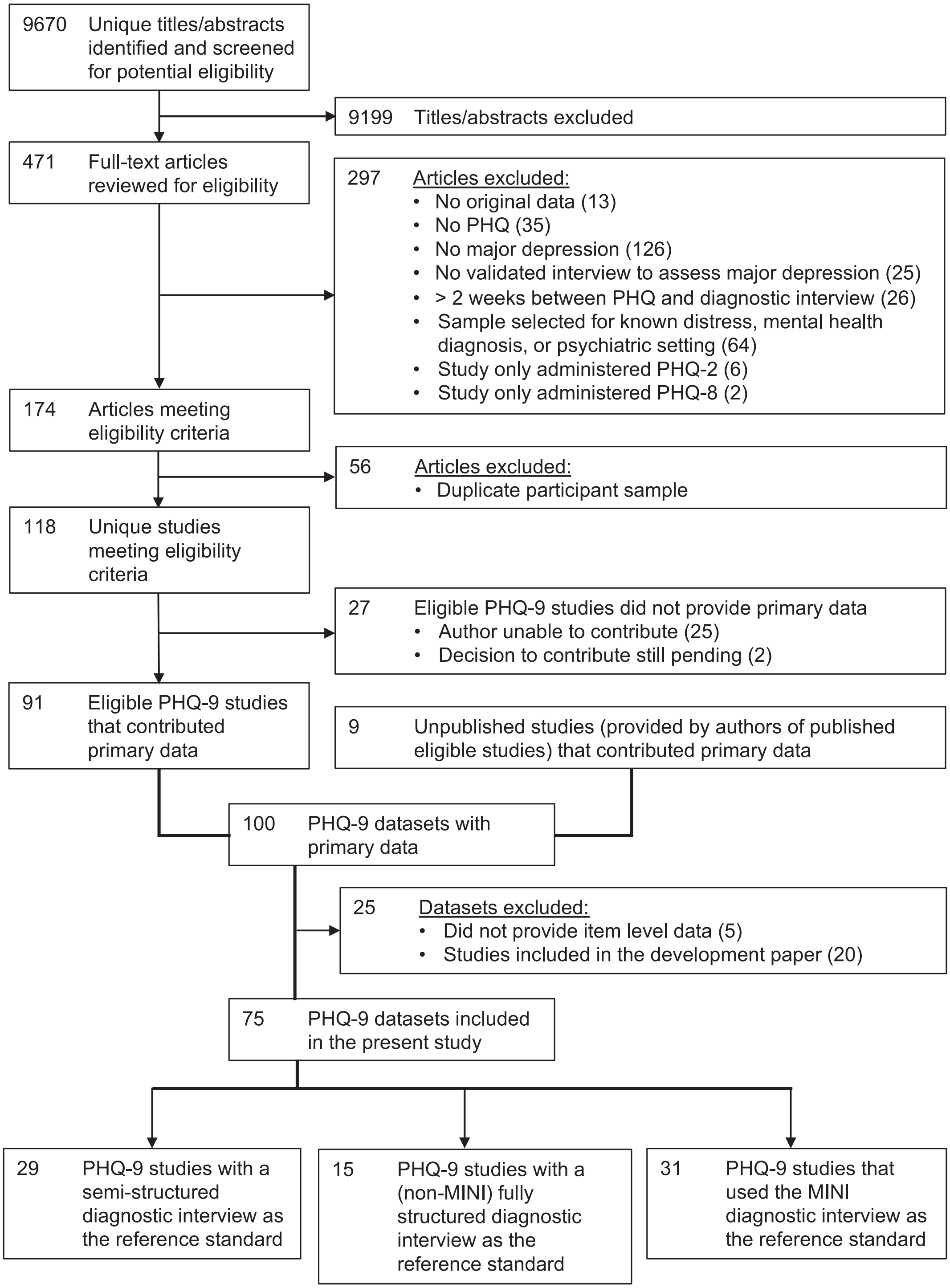

9670 Unique titles/abstracts identified and screened for potential eligibility
9199 Titles/abstracts excluded
471 Full-text articles reviewed for eligibility
297 Articles excluded:
• No original data (13)
• No PHQ (35)
• No major depression (126)
• No validated interview to assess major depression (25)
• > 2 weeks between PHQ and diagnostic interview (26)
• Sample selected for known distress, mental health diagnosis, or psychiatric setting (64)
• Study only administered PHQ-2 (6)
• Study only administered PHQ-8 (2)
174 Articles meeting eligibility criteria
56 Articles excluded:
• Duplicate participant sample
118 Unique studies meeting eligibility criteria
27 Eligible PHQ-9 studies did not provide primary data
• Author unable to contribute (25)
• Decision to contribute still pending (2)
91 Eligible PHQ-9 studies that contributed primary data
9 Unpublished studies (provided by authors of published eligible studies) that contributed primary data
100 PHQ-9 datasets with primary data
25 Datasets excluded:
• Did not provide item level data (5)
• Studies included in the development paper (20)
75 PHQ-9 datasets included in the present study
29 PHQ-9 studies with a semi-structured diagnostic interview as the reference standard
15 PHQ-9 studies with a (non-MINI) fully structured diagnostic interview as the reference standard
31 PHQ-9 studies that used the MINI diagnostic interview as the reference standard

Fig. 1. Flow diagram of study selection process.

authors, excluding studies that did not have item level data or were included in the PHQ-Dep-4 development paper, there were 75 eligible datasets (N participants = 34,698; N major depression = 3,392 [prevalence 10%]) that contributed data for our analysis.

Of the 75 included studies, 29 (7,719 participants; 923 major depression cases) used a semi-structured interview as the reference standard, 15 (12,109 participants; 873 cases) used a fully structured interview (other than the MINI), and 31 (14,870 participants; 1,596 cases) used the MINI. The Structured Clinical Interview for the DSM (SCID) was the most commonly used semi-structured interview (28 of 29 studies) and the Composite International Diagnostic Interview (CIDI) the most commonly used fully structured interview (14 of 15 studies). See Supplementary Table 1a-c for characteristics of included primary studies, eligible excluded primary studies, and the 20 studies included in the PHQ-Dep-4 development paper only. Table 1 presents participant-level descriptive statistics for the sample used in the present study.

5.2. Validation results

Fig. 2 shows receiver-operating curves for each reference standard category as well as the PHQ-9 cutoff score of ≥ 10. Table 2 shows estimated sensitivity and specificity for PHQ-Dep-4 cutoffs (≥1 to ≥ 12), as well as the standard and optimal PHQ-9 cutoff score of ≥ 10. For a PHQ-Dep-4 cutoff of ≥ 4, sensitivity was 0.88 (95% CI 0.81, 0.93), 0.68 (95% CI 0.56, 0.78), and 0.80 (95% CI 0.73, 0.85) for the semi-structured, fully structured, and MINI reference standard categories, respectively, as compared to 0.88 (0.81, 0.93), 0.64 (0.50, 0.76), and 0.73 (0.66, 0.79) for the PHQ-9 cutoff of ≥ 10, respectively. Similarly, for a PHQ-Dep-4 cutoff of ≥ 4, specificity was 0.79 (95% CI 0.74, 0.83), 0.85 (95% CI 0.78, 0.90), and 0.83 (95% CI 0.80, 0.86) for the semi-structured, fully structured, and MINI reference standard categories, respectively, as compared to 0.85 (0.80, 0.88), 0.89 (0.83, 0.93), and 0.89 (0.86, 0.91) for the PHQ-9 cutoff of ≥ 10, respectively. Fig. 2 shows the ROC plots for each reference standard category.

Table 3 shows the results of the tests of equivalence of the PHQ-Dep-4 and PHQ-9 pooled across all reference standard categories. A PHQ-Dep-4 cutoff of ≥ 4 showed the smallest overall sum of absolute differences in accuracy with PHQ-9 ≥ 10, with a difference in sensitivity of 0.03 (95% CI 0.00, 0.06) and a difference in specificity of −0.05 (95% CI −0.07, −0.04). These findings were ambiguous, as the CIs for both sensitivity and specificity crossed the interval of −0.05 to + 0.05. No other PHQ-Dep-4 cutoff indicated equivalency for both sensitivity and specificity. The next closest PHQ-Dep-4 cutoff to PHQ-9 ≥ 10 was a PHQ-Dep-4 cutoff of ≥ 5, with a difference in sensitivity of −0.07 (95% CI −0.11, −0.05) and a difference in specificity of 0.02 (95% CI 0.01, 0.03).

6. Discussion

This study used data from 75 primary studies to assess whether a previously determined PHQ-Dep-4 cutoff of ≥ 4, which was equivalent to a PHQ-9 cutoff of ≥ 10 in a development sample, would also be equivalent in a validation sample. While a PHQ-Dep-4 cutoff of ≥ 4 showed the best performance among all possible PHQ-Dep-4 cutoffs compared to the PHQ-9 cutoff of ≥ 10, the equivalence results were ambiguous, and we were unable to conclude that its specificity was equivalent to that of the PHQ-9 cutoff of ≥ 10.

We found that compared to the standard and optimal PHQ-9 cutoff of ≥ 10, a PHQ-Dep-cutoff of ≥ 4 had slightly greater sensitivity and slightly reduced specificity. The next best PHQ-Dep-cutoff of ≥ 5 had slightly greater specificity and slightly reduced sensitivity. In clinical settings, use of shortened forms such as the PHQ-Dep-4 offers the advantage of reducing respondent burden. While our study assessed the sum of sensitivity and specificity, this does not necessarily reflect local concerns such as the capacity for conducting further assessments, nor does it necessarily maximize the likelihood of patient benefits or minimize costs and harms. We note that clinicians and researchers can choose different cut-offs based on local priorities and resources using the information provided in Tables 2 and 3.

Table 1
Demographics of the study sample for patients with and without major depression.

Sociodemographic variables	Total (N = 34,698)	Participants with Major Depression (N = 3,392)	Participants without Major Depression (N = 31,306)
Age in years, *mean [median]* ± *SD (range)*[1]	47.7 [48] ± 16.3 (18, 98)	46.4 [45] ± 16.3 (18, 94)	48.9 [48] ± 16.3 (18, 98)
Women, *n (%)*[2]	20,678	2351 (11.4)	18,327 (88.6)
Men, *n (%)*[2]	13,998	1038 (7.4)	12,960 (92.6)
PHQ-9 score, *mean [median]* ± *SD (range)*	4.9 [3] ± 5.2 (0, 27)	13.1 [13] ± 6.3 (0, 27)	4.0 [3] ± 4.2 (0, 27)
Country, *n (%)*			
Netherlands	7049	494 (7.0)	6555 (93.0)
Canada	5215	190 (3.6)	5025 (96.4)
South Korea	3071	205 (6.7)	2866 (93.3)
South Africa	2300	299 (13.0)	2001 (87.0)
China	2096	136 (6.5)	1960 (93.5)
Germany	1605	147 (9.2)	1458 (90.8)
Taiwan	1532	50 (3.3)	1482 (96.7)
Latvia	1467	147 (10.0)	1320 (90.0)
USA	1247	166 (13.3)	1081 (86.7)
Greece	1036	262 (25.3)	774 (74.7)
Spain	1003	83 (8.3)	920 (91.7)
Other[3]	7077	1213 (17.1)	5864 (82.9)
Language, *n (%)*[4]			
English	8073	562 (7.0)	7511 (93.0)
Dutch	7222	522 (7.2)	6700 (92.8)
Chinese	3597	164 (4.6)	3433 (95.4)
Korean	3071	205 (6.7)	2866 (93.3)
South African languages	1838	211 (11.5)	1627 (88.5)
German	1605	147 (9.2)	1458 (90.8)
Spanish	1540	181 (11.8)	1359 (88.2)
Greek	1036	262 (25.3)	774 (74.7)
Other[5]	6611	1130 (17.1)	5481 (82.9)
General Care Setting, *n (%)*			
Outpatient care	17,624	2250 (12.8)	15,374 (87.2)
Inpatient care	2781	331 (11.9)	2450 (88.1)
Non-medical setting	14,163	806 (5.7)	13,357 (94.3)
Outpatient/inpatient mixed sample	130	5 (3.8)	125 (96.2)
Diagnostic Interview, *n (%)*			
SCID	6187	873 (14.1)	5314 (85.9)
CIDI	11,810	860 (7.3)	10,950 (92.7)
SCAN	1532	50 (3.3)	1482 (96.7)
MINI	14,870	1596 (10.7)	13,274 (89.3)
CIS-R	299	13 (4.3)	286 (95.7)
Classification system, *n (%)*			
ICD-10	909	86 (9.5)	823 (90.5)
DSM-III	1107	104 (9.4)	1003 (90.6)
DSM-IV	31,771	3089 (9.7)	28,682 (90.3)
DSM-V	911	113 (12.4)	798 (87.6)

[1] N missing = 31 participants with major depression, 216 participants without major depression.

[2] N missing = 3 participants with major depression, 19 participants without major depression.

[3] Other countries: Ethiopia, Japan, Australia, Brazil, Singapore, Malaysia, India, Israel, Mexico, Thailand, Zimbabwe, Argentina, Uganda, Iran, Kenya, Belgium, Italy, UK, Myanmar, Nepal, Hong Kong China.

[4] N missing = 8 for MDD, 97 for non-MDD.

[5] Other Languages: Amharic, Latvian, Japanese, Russian, Portuguese, Malay, Indian languages (unspecified), Malay or English, Thai, Shona, Hebrew, Farsi, Kiswahili, Italian, Burmese, Nepali, Malay, Chinese or Tamil, Filipino, Arabic, French.

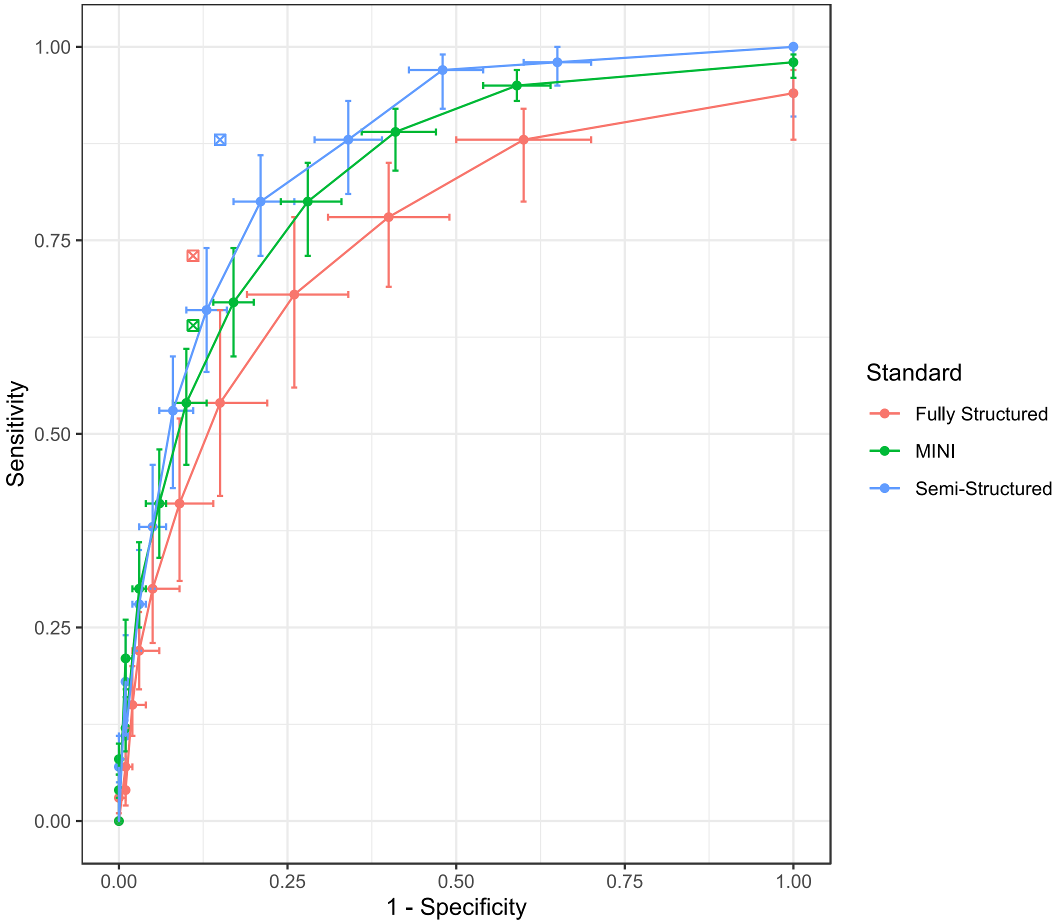


Fig. 2. Receiver-operating curve for each reference standard category. Points represent cutoffs of 0 (right) to 12 (left) for each reference standard category. X marks the PHQ-9 cutoff of ≥ 10.

While a strength of this analysis is the large number of primary studies included in the dataset, these primary studies spanned a large number of languages. This can cause concern for differential item functioning (DIF). The items for the PHQ-Dep-4 were not selected with regards to considerations of DIF. However, studies of DIF with the PHQ-9 have shown that it performs equivalently or with minimal impact of DIF across multiple languages [37–39]. We note that future research may wish to specifically investigate the impact of DIF for the PHQ-Dep-4 in comparison to the PHQ-9.

The development study tested non-inferiority rather than equivalency. The development study found a difference in sensitivity of + 0.03, and a difference in specificity of −0.03 between the two forms [9]. The present study found differences of + 0.03 and −0.05, respectively. While equivalency is therefore not established, the findings in the present study were not substantively different from the development study.

While it is not clear that the PHQ-Dep-4 performs equivalently to the PHQ-9 for specificity, clinicians screening for depression may opt to use the PHQ-Dep-4 with the understanding that depending on the cutoff used, specificity might be slightly reduced compared to the full PHQ-9 at cutoff of ≥ 10. Furthermore, clinicians should be aware that while the full PHQ-9 aligns with the nine DSM symptoms for major depression, not all PHQ-9 items may be relevant to individual presentations of a given mental disorder, and the PHQ-Dep-4 includes only a pre-specified subset of four items (1, 2, 6, and 8), thus not necessarily capturing the specific symptoms of a given patient.

There are several reasons that may explain why equivalence could not be concluded. First, although the overall sample size and number of studies used in this analysis was large, it could be that the study was underpowered, due to the design effect associated with the clustering within studies. As we do not know of methods for calculating power to establish equivalency in accuracy based on sensitivity and specificity difference for a subset of items compared to the total set, it was not possible to determine the necessary sample size needed *a priori*. Furthermore, we also did not split the data by reference standard category and conduct separate analyses. Second, we found that sensitivity in the shortened form was improved as compared to the full-length form. However, the specificity of the shortened form was lower than that of the full-length form, resulting in the inability to conclude equivalence between the two forms.

There are several other possible limitations of this study. First, for the collection of data for the full IPDMA, we were unable to obtain data from 27 eligible studies. Of the studies that provided data, five were excluded because they did not include item-level scores necessary to calculate PHQ-Dep-4, and we excluded another 20 studies from the development dataset to provide us with a set of external validation data. With the final available dataset, we were unable to investigate equivalence in specific patient populations as that would have required splitting the data even further. Second, for our first set of analyses (estimating PHQ-Dep-4 accuracy at all cutoffs), primary studies were categorized based on the diagnostic interview used, but interviewers may not have always administered the interviews as intended, which could have influenced results. This study only compared the PHQ-Dep-4 to a PHQ-9 cutoff of ≥ 10 because, although some primary studies have found other preferred cutoffs, large IPDMAs have concluded that cutoff ≥ 10 maximizes the sum of sensitivity and specificity [8,18]. Lastly, this study evaluated the items included in the PHQ-Dep-4 as previously developed and did not re-develop the shortened form. It could be that a different set of items, creating either a different form of length 4 or a potentially

Table 2
Sensitivity and specificity for each PHQ-Dep-4 cutoff and the PHQ-9 cutoff of ≥ 10.

	SEMI-STRUCTURED REFERENCE STANDARD: N studies = 29, N participants = 7719,N major depression = 923				**FULLY STRUCTURED REFERENCE STANDARD:N studies = 15, N participants = 12,109,N major depression = 873**				**MINI[2] REFERENCE STANDARD:N studies = 31, N participants = 14,870,N major depression = 1596**			
CutoffPHQ-Dep-4	**sensitivity**	**95% CI**	**specificity**	**95% CI**	**sensitivity**	**95% CI**	**specificity**	**95% CI**	**sensitivity**	**95% CI**	**specificity**	**95% CI**
>= 1	1.00	(0.91, 1.00)	0.35	(0.30, 0.40)	0.94	(0.88, 0.97)	0.40	(0.30, 0.50)	0.98	(0.96, 0.99)	0.41	(0.36, 0.46)
>= 2	0.98	(0.95, 1.00)	0.52	(0.46, 0.57)	0.88	(0.80, 0.92)	0.60	(0.51, 0.69)	0.95	(0.93, 0.97)	0.59	(0.53, 0.64)
>= 3	0.97	(0.92, 0.99)	0.66	(0.61, 0.71)	0.78	(0.69, 0.85)	0.74	(0.66, 0.81)	0.89	(0.84, 0.92)	0.72	(0.67, 0.76)
>= 4	0.88	(0.81, 0.93)	0.79	(0.74, 0.83)	0.68	(0.56, 0.78)	0.85	(0.78, 0.90)	0.80	(0.73, 0.85)	0.83	(0.80, 0.86)
>= 5	0.80	(0.73, 0.86)	0.87	(0.84, 0.90)	0.54	(0.42, 0.66)	0.91	(0.86, 0.94)	0.67	(0.60, 0.74)	0.90	(0.87, 0.92)
>= 6	0.66	(0.58, 0.74)	0.92	(0.89, 0.94)	0.41	(0.31, 0.52)	0.95	(0.91, 0.97)	0.54	(0.46, 0.61)	0.94	(0.93, 0.96)
>= 7	0.52	(0.43, 0.60)	0.95	(0.93, 0.97)	0.30	(0.23, 0.38)	0.97	(0.94, 0.98)	0.41	(0.34, 0.48)	0.97	(0.96, 0.98)
>= 8[1]	0.38	(0.30, 0.46)	0.97	(0.96, 0.98)	0.22	(0.17, 0.27)	0.98	(0.96, 0.99)	0.30	(0.25, 0.36)	0.99	(0.98, 0.99)
>= 9	0.28	(0.22, 0.35)	0.99	(0.98, 0.99)	0.15	(0.11, 0.20)	0.99	(0.98, 0.99)	0.21	(0.17, 0.26)	0.99	(0.99, 0.99)
>= 10	0.18	(0.13, 0.24)	0.99	(0.99, 1.00)	0.07	(0.04, 0.12)	0.99	(0.99, 1.00)	0.12	(0.09, 0.16)	1.00	(0.99, 1.00)
>= 11	0.11	(0.08, 0.16)	1.00	(0.99, 1.00)	0.04	(0.02, 0.07)	1.00	(0.99, 1.00)	0.08	(0.06, 0.10)	1.00	(1.00, 1.00)
>= 12	0.07	(0.05, 0.11)	1.00	(1.00, 1.00)	0.03	(0.01, 0.06)	1.00	(1.00, 1.00)	0.04	(0.03, 0.06)	1.00	(1.00, 1.00)
PHQ-9 >= 10	0.88	(0.81, 0.93)	0.85	(0.80, 0.88)	0.64	(0.50, 0.76)	0.89	(0.83, 0.93)	0.73	(0.66, 0.79)	0.89	(0.86, 0.91)

[1] BOBYQA optimizer was used to ensure model convergence for the semi-structured reference category, as the model with the default optimizer did not converge.
[2] MINI: Mini International Neuropsychiatric Interview.

Table 3
Results of the equivalence tests between the accuracy of the PHQ-Dep-4 and PHQ-9 ≥ 10.

All studies (N studies = 75, N participants = 34,698, N major depression = 3392)				
Cutoff	**Sensitivity Difference (PHQ-Dep-4 - PHQ-9 >=10)**	**95% CI**	**Specificity Difference (PHQ-Dep-4 - PHQ-9 >=10)**	**95% CI**
PHQ-Dep-4 >= 1	0.21	(0.14, 0.25)	−0.49	(-0.52, −0.46)
PHQ-Dep-4 >= 2	0.18	(0.13, 0.22)	−0.31	(-0.34, −0.28)
PHQ-Dep-4 >= 3	0.13	(0.09, 0.16)	−0.17	(-0.19, −0.15)
PHQ-Dep-4 >= 4	0.03	(0.00, 0.06)	−0.05	(-0.07, −0.04)
PHQ-Dep-4 >= 5	−0.07	(-0.11, −0.05)	0.02	(0.01, 0.03)
PHQ-Dep-4 >= 6	−0.22	(-0.27, −0.19)	0.06	(0.05, 0.08)
PHQ-Dep-4 >= 7	−0.35	(-0.41, −0.33)	0.09	(0.08, 0.11)
PHQ-Dep-4 >= 8	−0.47	(-0.53, −0.45)	0.11	(0.09, 0.13)
PHQ-Dep-4 >= 9	−0.55	(-0.62, −0.53)	0.12	(0.10, 0.14)
PHQ-Dep-4 >= 10	−0.65	(-0.72, −0.62)	0.12	(0.10, 0.15)
PHQ-Dep-4 >= 11	−0.70	(-0.77, −0.67)	0.13	(0.10, 0.15)
PHQ-Dep-4 >= 12	−0.73	(-0.80, −0.69)	0.13	(0.10, 0.15)

shorter or longer form, would result in equivalent sensitivity and specificity to the full PHQ-9.

7. Conclusion

In conclusion, this was the first study to our knowledge to externally validate the results of shortening a self-report questionnaire through the OTA method using individual participant level data. We found that the previously suggested cutoff of ≥ 4 for the PHQ-Dep-4 remained the preferred cutoff, but the specificity of the shortened form did not meet equivalency to the full PHQ-9 cutoff of ≥ 10. Clinicians may consider screening with the PHQ-Dep-4 to reduce respondent burden, but should be aware that in doing so, specificity may be slightly compromised compared to the full PHQ-9.

8. Contributions

DH, BLevis, JPAI, PC, SBP, RCZ, ABenedetti, and BDT were responsible for the study conception and design. SM contributed as a patient partner knowledge user. FF contributed an included dataset. BLevis, YS, and BDT contributed to data extraction, coding, evaluation of included studies, and data synthesis. DH, BLevis, FF, ABenedetti, and BDT contributed to data analysis and interpretation. DH, BLevis, YS, ABenedetti, and BDT drafted the manuscript.

Members of the DEPRESSD PHQ group contributed:

To data extraction, coding, and synthesis: CH, YW, AK, PMB, ZN, MImran, DBR, KER, MA, AWL. Via the design and conduct of database searches: JTB, LAK. As members of the DEPRESSD Steering Committee, including conception and oversight of collaboration: SG, DM. By contributing included datasets: DA, LA, HRB, ABeraldi, CNB, ABhana, RIB, MHC, JCNC, LFC, DC, AC, FMD, JMdMvG, CDQ, SF, JRWF, DF, ECG, BG, LG, LJG, EPG, BJH, LHantsoo, EEH, MHärter, UH, LHides, SEH, SH, MHudson, TH, MInagaki, HJJ, NJ, MEK, SK, BAK, YK, FL, MAL, HFLA, SIL, ML, SRL, BLöwe, NPL, CL, RAM, BPM, SMS, TNM, KM, JEMN, LN, FLO, PP, AP, SLP, TJQ, ER, SDR, KR, HJR, ISS, MTS, JS, EHS, LSpangenberg, LStafford, SCS, KS, PLLT, MTR, TDT, CMvdFC, TvH, HCvW, LIW, JLW, DW, KW, MY, QZZ, YZ.

All authors, including group authors, provided a critical review and approved the final manuscript. DH is the guarantor; she had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analyses. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted

Funding

This study was funded by the Canadian Institutes of Health Research (CIHR; KRS-134297, PJT-162206). Dr. Levis and Dr. Wu were supported by Fonds de recherche du Québec - Santé (FRQ-S) Postdoctoral Training Fellowships. The primary study by Fischer et al. was funded by the German Federal Ministry of Education and Research (01GY1150). The primary studies by Patten et al. and Prisnie et al. were supported by the Cumming School of Medicine, University of Calgary, and Alberta Health Services through the Calgary Health Trust, as well as the Hotchkiss Brain Institute. Dr. Patten was supported by a Senior Health Scholar award from Alberta Innovates Health Solutions. Dr. Benedetti was supported by a FRQ-S researcher salary award. Dr. Thombs was supported by a Tier 1 Canada Research Chair. Ms. Rice was supported by a Vanier Canada Graduate Scholarship. The primary study by Amtmann et al. was supported by a grant from the Department of Education (NIDRR grant number H133B080025) and by the National Multiple Sclerosis Society (MB 0008). Data collection for the study by Ayalon et al. was supported from a grant from Lundbeck International. The primary study by Khamseh et al. was supported by a grant (M-288) from Tehran University of Medical Sciences. The primary studies by Marrie et al. and Bernstein et al. were supported by CIHR (THC-135234) and Crohn's and Colitis Canada. Dr. Bernstein was supported in part by the Bingham Chair in Gastroenterology. The primary studies by Bhana et al., Kohrt et al. and Nakku et al. were output of the PRogramme for Improving Mental health carE (PRIME) and were supported by the UK Department for International Development (201446). The views expressed do not necessarily reflect the UK Government's official policies. The primary study by Buji et al. was supported by grants from the UKMMC Fundamental Research Fund (FF-2015-051) and the Fundamental Research Grant Scheme by the Malaysian Ministry of Higher Education (FRGS/2/2014/SKK09/UKM/02/1). Collection of data for the primary study by Zhang et al. was supported by the European Foundation for Study of Diabetes, the Chinese Diabetes Society, Lilly Foundation, Asia Diabetes Foundation and Liao Wun Yuk Diabetes Memorial Fund. The primary study by Chibanda et al. was supported by a grant from Grand Challenges Canada (0087-04). The primary study by Conway et al. was supported by the Institute of Health and Biomedical Innovation at Queensland University of Technology and the Sigma Theta Tau International Honour Society of Nursing (ID: 8580). The primary study by Grool et al. was supported by a programme grant from The Netherlands Heart Foundation (2007B027). The primary study by Zuithoff et al. was supported by The European Commission (PREDICTQL4-CT2002-00683) and The Netherlands Organization for Scientific Research (ZonMw 016.046.360). The primary study by Martin-Subero et al. was supported in part by a grant from the Spanish Ministry of Health's Health Research Fund (Fondo de Investigaciones Sanitarias, project 97/1184). The primary study by van Heyningen et al. was supported by the Medical Research Council of South Africa (415865), Cordaid Netherlands (Project 103/10002 G Sub 7) and the Truworths Community Foundation Trust, South Africa. The primary study by Fisher et al. was supported by grants from the National Health and Medical Research Council (APP1026550), the Australian Government Department of Social Services - Families, Housing, Community Services and Indigenous Affairs, and the Victorian Department of Education and Early Childhood Development. Dr. Fisher was supported by a Monash Professorial Fellowship and the Jean Hailes Professorial Fellowship, which is supported by a grant to the Jean Hailes Foundation from the H and L Hecht Trust managed by Perpetual Trustees. The primary study by Baron et al. was supported by the National Income Dynamics Study (NIDS), at the University of Cape Town, South Africa. The NIDS is implemented by the Southern Africa Labour and Development Research Unit, and is funded by the Department of Planning, Monitoring and Evaluation. The funding body was involved in the design of the primary study. Data for the primary study by Gelaye et al. was supported by a grant from the NIH (T37 MD001449). The primary study by Gholizadeh et al. was supported by University of Technology Sydney under UTS Research Reestablishment Grants. The primary study by Green et al. (2018) was supported by a grant from the Duke Global Health Institute (453-0751). Collection of data for the primary study by Hobfoll et al. was made possible in part from grants from NIMH (RO1 MH073687) and the Ohio Board of Regents. Dr. Hall received support from a grant awarded by the Research and Development Administration Office, University of Macau (MYRG2015-00109-FSS). The primary study by Garabiles et al. was supported by the Macao (SAR) Government, through the University of Macau RSKTO grants: MYRG-2014-111. The primary study by Hantsoo et al. was supported by K23 MH107831-02, Brain and Behavior Research Foundation NARSAD Young Investigator Award. The primary study by Haroz et al. was supported by the United States Agency for International Development Victims of Torture Fund: AID-DFD A-00-08-00308. Dr. Haroz was supported by a NIMH T32 predoctoral training grant (MH014592-38) and postdoctoral training grant (MH103210) during the conduct of primary study. Collection of data provided by Drs. Härter and Reuter was supported by the Federal Ministry of Education and Research (grants No. 01 GD 9802/4 and 01 GD 0101) and by the Federation of German Pension Insurance Institute. The primary study by Henkel et al. was funded by the German Ministry of Research and Education. The primary study by Hides et al. was funded by the Perpetual

Trustees, Flora and Frank Leith Charitable Trust, Jack Brockhoff Foundation, Grosvenor Settlement, Sunshine Foundation and Danks Trust. Data for the study by Razykov et al. was collected by the Canadian Scleroderma Research Group, which was funded by the CIHR (FRN 83518), the Scleroderma Society of Canada, the Scleroderma Society of Ontario, the Scleroderma Society of Saskatchewan, Sclérodermie Québec, the Cure Scleroderma Foundation, Inova Diagnostics Inc., Euroimmun, FRQ-S, the Canadian Arthritis Network, and the Lady Davis Institute of Medical Research of the Jewish General Hospital, Montréal, QC. Dr. Hudson was supported by a FRQ-S Senior Investigator Award. Collection of data for the primary study by Hyphantis et al. (2014) was supported by grant from the National Strategic Reference Framework, European Union, and the Greek Ministry of Education, Lifelong Learning and Religious Affairs (ARISTEIA-ABREVIATE, 1259). The primary study by Paika et al. was supported by the European Economic Area (EEA) Financial Mechanism 2009-2014 (EEA GR07/3767) and National funds as part of the program “Dissimilarity, Inequality and Social Integration” (132324/I4-25/8/2015). The primary study by Inagaki et al. was supported by the Ministry of Health, Labour and Welfare, Japan. The primary study by Kim et al. was supported by a grant from the Korean Mental Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (HM14C2567), an Institute for Information & Communications Technology Promotion grant funded by the Korea government (MSIP) (B0132- 15-1003: the development of skin adhesive patches for the monitoring and prediction of mental disorders), and by the Original Technology Research Program for Brain Science through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2016M3C7A1947307). Dr. Jetté was supported by a Canada Research Chair in Neurological Health Services Research and an AIHS Population Health Investigator Award. She is the Icahn School of Medicine at Mount Sinai Bludhorn Professor of International Medicine. The primary study by Janssen et al. was supported by the European Regional Development Fund as part of OP-ZUID; the Province of Limburg; the department of Economic Affairs of the Netherlands (31O.041); Stichting the Weijerhorst, the Pearl String Initiative Diabetes; the Cardiovascular Center Maastricht Cardiovasculair Research Institute Maastricht; School of Nutrition, Toxicology and Metabolism; Stichting Annadal; and Health Foundation Limburg. The primary study by Lamers et al. was funded by the Netherlands Organisation for Health Research and development (grant number 945-03-047). The primary study by Lara et al. was supported by the Consejo Nacional de Ciencia y Tecnología/National Council for Science and Technology (CB-2009-133923-H). The primary study by Liu et al. (2011) was funded by a grant from the National Health Research Institute, Republic of China (NHRI-EX97-9706PI). The primary study by Lotrakul et al. was supported by the Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (grant number 49086). The primary studies by Osório et al. (2012) were funded by Reitoria de Pesquisa da Universidade de São Paulo (grant number 09.1.01689.17.7) and Banco Santander (grant number 10.1.01232.17.9). Dr. Bernd Löwe received research grants from Pfizer, Germany, and from the medical faculty of the University of Heidelberg, Germany (project 121/2000) for the study by Gräfe et al. Dr. Marrie was supported by the Waugh Family Chair in Multiple Sclerosis and the Research Manitoba Chair, and CIHR grants, during the conduct of the study. Dr. Marx was supported by the Department of Defense (W81XWH-08-2- 0100/W81XWH-08-2-0102 and W81XWH-12-2-0117/W81XWH-12-2-0121). The primary study by Mohd Sidik et al. was funded under the Research University Grant Scheme from Universiti Putra Malaysia, Malaysia and the Postgraduate Research Student Support Accounts of the University of Auckland, New Zealand. The primary study by Santos et al. was funded by the National Program for Centers ofExcellence(PRONEX/FAPERGS/CNPq, Brazil). The primary study by Muramatsu et al. (2007) was supported by an educational grant from Pfizer US Pharmaceutical Inc. The primary study by Muramatsu et al. (2018) was supported by grants from Niigata Seiryo University. Dr. Osório was supported by Productivity Grants (PQ-CNPq-2 -number 301321/2016-7). The primary study by Persoons et al. was partly funded by a grant from the Belgian Ministry of Public Health and Social Affairs and supported by a limited grant from Pfizer Belgium. The primary study by Picardi et al. was supported by funds for current research from the Italian Ministry of Health. The primary study by Wagner et al. was supported by grants U10CA21661, U10CA180868, U10CA180822, and U10CA37422 from the National Cancer Institute. The study was also funded in part by a grant from the Pennsylvania Department of Health. The department specifically declaims responsibility for any analyses, interpretations, or conclusions of the primary study. The primary study by Rancans et al. was supported by The National Research Programme BIOMEDICINE 2014-2017 (5.8.1.). Dr. Shaaban was supported by funding from Universiti Sains Malaysia. The primary study by Shinn et al. was supported by grant NCI K07 CA 093,512 and the Lance Armstrong Foundation. The primary study by Spangenberg et al. was supported by a junior research grant from the medical faculty, University of Leipzig. Dr. Stafford received PhD scholarship funding from the University of Melbourne. Dr. Tran was supported by a Monash Bridging Postdoctoral Fellowship. The primary study by Volker et al. was supported by The Netherlands organization for Health Research and Development (ZonMw) and from Achmea SZ, a Dutch insurance company. The study by Wittkampf et al. was funded by The Netherlands Organization for Health Research and Development (ZonMw) Mental Health Program (nos. 100.003.005 and 100.002.021) and the Academic Medical Center/University of Amsterdam. The primary study by Liu et al. (2015) was supported by CIHR (MOP-114970). The primary study by Liu et al. (2016) was supported by Shanghai Municipal Health and Family Planning Commission Bureau-level Project - Preliminary Exploration of Depression Risk Prediction Model for Outpatients in General Hospitals (Project No. 2010105). No other authors reported funding for primary studies or for their work on this study. No funder had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Declaration of Competing Interest

All authors declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years with the following exceptions: Dr. Bernstein declares that he has consulted to Abbvie Canada, Amgen Canada, Bristol Myers Squibb Canada, Roche Canada, Janssen Canada, Pfizer Canada, Sandoz Canada, Takeda Canada, and Mylan Pharmaceuticals. He has also received unrestricted educational grants from Abbvie Canada, Janssen Canada, Pfizer Canada, and Takeda Canada; as well as been on speaker’s bureau of Abbvie Canada, Janssen Canada, Takeda Canada and Medtronic Canada, all outside the submitted work. Dr. Chan J CN is a steering committee member and/or consultant of Astra Zeneca, Bayer, Lilly, MSD, and Pfizer. She has received sponsorships and honorarium for giving lectures and providing consultancy and her affiliated institution has received research grants from these companies. Dr. Chan LF declares personal fees and non-financial support from Otsuka, Lundbeck, and Johnson and Johnson; and non-financial support from Ortho-McNeil-Janssen, and Menarini, outside the submitted work. Dr. Hegerl declares that within the last three years, he was an advisory board member for Janssen and received a research grant from Medice, all outside the submitted work. Dr. Inagaki declares that he has received personal fees from Meiji, Mochida, Takeda, Novartis, Yoshitomi, Pfizer, Eisai, Otsuka, MSD, Sumitomo Dainippon, Janssen, and Eli Lilly, all outside of the submitted work. Dr. Pugh declares that she received salary support from Pfizer-Astella and Millennium, outside the submitted work. Dr. Rancans declares that he received grants, personal fees, and non-financial support from Gedeon Richter; personal fees and non-financial support from Lundbeck, Servier, and Janssen Cilag; personal fees from Zentiva, and

Abbvie; outside the submitted work. Dr. Schram declares that the primary study by Janssen et al. was supported by unrestricted grants from Janssen, Novo Nordisk, and Sanofi. Dr. Wagner declares that she receives personal fees from Celgene, outside the submitted work. All authors declare no other relationships or activities that could appear to have influenced the submitted work. No funder had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Appendix A

The DEPRESSD PHQ Collaboration: Chen He, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Yin Wu, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Ankur Krishnan, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Parash Mani Bhandari, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Dipika Neupane, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Zelalem Negeri, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Mahrukh Imran, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Danielle B. Rice, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Kira E. Riehm, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Marleine Azar, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Alexander W. Levis, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Jill Boruff, Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montréal, Québec, Canada; Simon Gilbody, Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, UK; Lorie A. Kloda, Library, Concordia University, Montréal, Québec, Canada; Dagmar Amtmann, Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA; Liat Ayalon, Louis and Gabi Weisfeld School of Social Work, Bar Ilan University, Ramat Gan, Israel; Hamid R. Baradaran, Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; Anna Beraldi, Kbo-Lech-Mangfall-Klinik Garmisch-Partenkirchen, Klinik für Psychiatrie, Psychotherapie & Psychosomatik, Lehrkrankenhaus der Technischen Universität München, Munich, Germany; Charles N. Bernstein, University of Manitoba IBD Clinical and Research Centre, Winnipeg, Manitoba, Canada; Arvin Bhana, Centre for Rural Health, School of Nursing and Public Health, College of Health Sciences, University of KwaZulu-Natal; Ryna Imma Buji, Department of Psychiatry, Hospital Mesra Bukit Padang, Sabah, Malaysia; Marcos H. Chagas, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; Juliana C. N. Chan, Department of Medicine and Therapeutics, Hong Kong Institute of Diabetes and Obesity and Li Ka Shing Institute of Health Science, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong SAR, China; Lai Fong Chan, Department of Psychiatry, National University of Malaysia, Kuala Lumpur, Malaysia; Dixon Chibanda, Department of Community Medicine, University of Zimbabwe, Harare, Zimbabwe; Aaron Conway, Lawrence S. Bloomberg Faculty of Nursing, University of Toronto, Toronto, Canada; Federico M. Daray, Institute of Pharmacology, School of Medicine, University of Buenos Aires, Argentina; Janneke M. de Man-van Ginkel, Julius Center for Health Sciences and Primary Care, Department of Nursing Science, University Medical Center Utrecht – University Utrecht, Utrecht, the Netherlands; Crisanto Diez-Quevedo, Servei de Psiquiatria, Hospital Germans Trias i Pujol, Badalona, Spain; Sally Field, Perinatal Mental Health Project, Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town; Jane R. W. Fisher, Global and Women's Health, Public Health and Preventive Medicine, Monash University; Daniel Fung, Department of Developmental Psychiatry, Institute of Mental Health, Singapore; Emily C. Garman, Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town; Bizu Gelaye, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA; Leila Gholizadeh, Faculty of Health, University of Technology Sydney, Sydney, Australia; Lorna J. Gibson, International Statistics and Epidemiology Group, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK; Eric P. Green, Duke Global Health Institute, Duke University, Durham, North Carolina, USA; Brian J. Hall, New York University Shanghai, Shanghai, People's Republic of China; Liisa Hantsoo, Department of Psychiatry & Behavioral Sciences

, The Johns Hopkins University School of Medicine, Baltimore, Maryland; Emily E. Haroz, Center For American Indian Health, Department of International Health, Johns Hopkins Bloomberg School of Public Health; Martin Härter, Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Ulrich Hegerl, Department of Psychiatry, Psychosomatics and Psychotherapy, Goethe-Universität Frankfurt, Germany; Leanne Hides, School of Psychology, University of Queensland, Brisbane, Queensland, Australia; Stevan E. Hobfoll, STAR-Stress, Anxiety and Resilience Consultants, Chicago, Illinois, USA; Simone Honikman, Perinatal Mental Health Project, Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town; Marie Hudson, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Thomas Hyphantis, Department of Psychiatry, Faculty of Medicine, School of Health Sciences, University of Ioannina, Greece; Masatoshi Inagaki, Department of Psychiatry, Faculty of Medicine, Shimane University, Shimane, Japan; Hong Jin Jeon, Department of Psychiatry, Depression Center, Samsung Medical Center, Sungkyunkwan University School of Medicine; Nathalie Jetté, Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, New York, USA; Mohammad E. Khamseh, Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; Sebastian Köhler, Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, Netherlands; Brandon A. Kohrt, Department of Psychiatry and Behavioral Sciences, The George Washington University, Washington, DC, USA; Yunxin Kwan, Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore; Femke Lamers, Department of Psychiatry, Amsterdam Public Health Research Institute, Amsterdam UMC, Vrije Universiteit, Amsterdam, the Netherlands; Maria Asunción Lara, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz. San Lorenzo Huipulco, Tlalpan, México D. F. Mexico; Holly F. Levin-Aspenson, Department of Psychology, University of Notre Dame, Notre Dame, Indiana, USA; Shen-Ing Liu, Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore; Manote Lotrakul, Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand; Sonia R. Loureiro, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; Bernd Löwe, Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Nagendra P. Luitel, Research Department, TPO Nepal, Kathmandu, Nepal; Crick Lund, Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town; Ruth Ann Marrie, Departments of Medicine and Community Health Sciences, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada; Brian P. Marx, National Center for PTSD at VA Boston Healthcare System, Boston, MA, USA; Sherina Mohd Sidik, Department of Psychiatry, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia; Tiago N. Munhoz, Post-graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil; Kumiko Muramatsu, Department of Clinical Psychology, Graduate School of Niigata Seiryo University, Niigata, Japan; Juliet E. M. Nakku, Butabika National Referral Teaching Hospital, Kampala, Uganda; Laura Navarrete, Department of Epidemiology and Psychosocial Research, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, Ciudad de México, México; Flávia L. Osório, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; Philippe Persoons, Department of Psycho-Pedagogic Psychiatry, Healthcare Group Sint-Kamillus, Broeders van Liefde, Bierbeek, Belgium; Angelo Picardi, Centre for Behavioural Sciences and Mental Health, Italian National Institute of Health, Rome, Italy; Stephanie L. Pugh, NRG Oncology Statistics and Data Management Center, Philadelphia, PA, USA; Terence J. Quinn, Institute of Cardiovascular & Medical Sciences, University of Glasgow, Glasgow, Scotland; Elmars Rancans, Department of Psychiatry and Narcology, Riga Stradins University, Latvia; Sujit D. Rathod, Department of Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom; Katrin Reuter, Group Practice for Psychotherapy and Psychooncology, Freiburg, Germany; Heather J. Rowe, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia; Iná S. Santos, Post-graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil; Miranda T. Schram, Department of Internal Medicine, Maastricht University Medical Center, Maastricht, The Netherlands; Juwita Shaaban, Department of Family Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia; Eileen H. Shinn, Department of Behavioral Science, University of Texas M. D. Anderson Cancer Center, Houston, Texas, USA; Lena Spangenberg, Department of Medical Psychology and Medical Sociology, University of Leipzig, Germany; Lesley Stafford,Centre for Women's Mental Health, Royal Women's Hospital, Parkville, Australia; Sharon C. Sung, Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore; Keiko Suzuki, Department of General Medicine, Asahikawa University Hospital, Asahikawa, Hokkaido, Japan; Pei Lin Lynnette Tan, Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore; Martin Taylor-Rowan, Institute of Cardiovascular and Medical Science, University of Glasgow, Glasgow, Scotland; Thach D. Tran, Global and Women's Health, Public Health and Preventive Medicine, Monash University; Christina M. van der Feltz-Cornelis, Department of Health Sciences, HYMS, University of York, York, UK; Thandi van Heyningen, Division of Epidemiology & Biostatistics, School of Public Health & Family Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa; Henk C. van Weert, Department of General Practice, Institute Public Health, Amsterdam Universities Medical Centers, Amsterdam, the Netherlands; Lynne I. Wagner, Department of Social Sciences and Health Policy, Wake Forest School of Medicine, Wake Forest University, Winston-Salem, North Carolina, USA; Jian Li Wang, University of Ottawa Institute of Mental Health Research; David Watson, Dept. of Psychology, University of Notre Dame; Karen Wynter, School of Nursing and Midwifery, Deakin University, Melbourne, Australia; Mitsuhiko Yamada, Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, Ogawa-Higashi, Kodaira, Tokyo, Japan; Qing Zhi Zeng, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China; Yuying Zhang, Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymeth.2021.11.005.

References

[1] C. Goetz, et al., Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales, J. Clin. Epidemiol. 66 (7) (2013) 710–718, https://doi.org/10.1016/j.jclinepi.2012.12.015.

[2] J. Coste, F. Guillemin, J. Pouchot, J. Fermanian, Methodological approaches to shortening composite measurement scales, J. Clin. Epidemiol. 50 (3) (1997) 247–252, https://doi.org/10.1016/S0895-4356(96)00363-0.

[3] P.M. Kruyen, W.H.M. Emons, K. Sijtsma, On the Shortcomings of Shortened Tests: A Literature Review, Int. J. Test. 13 (3) (2013) 223–248, https://doi.org/10.1080/15305058.2012.703734.

[4] P.C. Stanton, J.M. Sinar, E.F. Balzer, W.K. Smith, Issues and strategies for reducing the length of self-report scales, Pers. Psychol. 55 (1) (2002) 167–194.

[5] K. Kroenke, R.L. Spitzer, The PHQ-9: a new depression diagnostic and severity measure, SLACK Incorporated Thorofare, NJ, 2002.

[6] K. Kroenke, T.W. Strine, R.L. Spitzer, J.B. Williams, J.T. Berry, A.H. Mokdad, The PHQ-8 as a measure of current depression in the general population, J. Affect. Disord. 114 (1–3) (2009) 163–173.

[7] K. Kroenke, R.L. Spitzer, J.B. Williams, The PHQ-9: validity of a brief depression severity measure, J. Gen. Intern. Med. 16 (9) (2001) 606–613.

[8] B. Levis, A. Benedetti, and B. D. Thombs, "Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis," *bmj*, vol. 365, 2019.

[9] M. Ishihara, et al., Shortening self-report mental health symptom measures through optimal test assembly methods: Development and validation of the Patient Health Questionnaire-Depression-4, Depress, Anxiety 36 (1) (2019) 82–92.

[10] W.J. Van der Linden, Linear models for optimal test design, Springer Science & Business Media, 2006.

[11] J.-T. Kuhn, T. Kiefer, Optimal test assembly in practice, Z. Für Psychol. (2015).

[12] D. Harel, et al., Shortening patient-reported outcome measures through optimal test assembly: application to the social appearance anxiety scale in the scleroderma patient-centered intervention network cohort, BMJ Open 9 (2) (2019), e024010.

[13] D. Harel, et al., Shortening the Edinburgh Postnatal Depression Scale using Optimal Test Assembly Methods: Development of the EPDS-Dep-5, Acta Psychiatr. Scand. (2020).

[14] D. Harel, M. Baron, Methods for shortening patient-reported outcome measures, Stat. Methods Med. Res. 28 (10–11) (2019) 2992–3011.

[15] A.W. Levis, et al., Using optimal test assembly methods for shortening patient-reported outcome measures: Development and Validation of the Cochin Hand Function Scale-6: A scleroderma patient-centered intervention network cohort study, Arthritis Care Res. 68 (11) (2016) 1704–1713.

[16] S. Li, et al., Nonrestorative sleep scale: a reliable and valid short form of the traditional Chinese version, Qual. Life Res. 29 (9) (2020) 2585–2592.

[17] S. Li, et al., A Short Form of the Chinese Version of the Weinstein Noise Sensitivity Scale through Optimal Test Assembly, Int. J. Environ. Res. Public Health 18 (3) (2021) 879.

[18] Z. Negeri *et al.*, "Accuracy of the Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: an updated systematic review and individual participant data meta-analysis," Under Review.

[19] B.D. Thombs, et al., The diagnostic accuracy of the Patient Health Questionnaire-2 (PHQ-2), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-9 (PHQ-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses, Syst. Rev. 3 (1) (2014) 1–16.

[20] A. P. Association, Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub, 2013.

[21] A. P. American Psychiatric Association, Diagnostic and statistical manual of mental disorders (DSM-IV) vol. 886 1994 American psychiatric association Washington Washington, DC.

[22] Diagnostic and statistical manual of mental disorders, *DSM-III*, 3rd, revised ed., American Psychiatric Association, Washington, DC, 1987.

[23] Diagnostic and statistical manual of mental disorders, *DSM-III*, 4th, revised ed., American Psychiatric Association, Washington, DC, 2000.

[24] W.H. Organization, The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines, World Health Organization, 1992.

[25] J. McGowan, M. Sampson, D.M. Salzwedel, E. Cogo, V. Foerster, C. Lefebvre, PRESS peer review of electronic search strategies: 2015 guideline statement, J. Clin. Epidemiol. 75 (2016) 40–46.

[26] R.D. Riley, S.R. Dodd, J.V. Craig, J.R. Thompson, P.R. Williamson, Meta-analysis of diagnostic test studies using individual patient data and aggregate data, Stat. Med. 27 (29) (2008) 6111–6136.

[27] B. Levis, et al., Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews, Br. J. Psychiatry 212 (6) (2018) 377–385.

[28] B. Levis, et al., Comparison of major depression diagnostic classification probability using the SCID, CIDI, and MINI diagnostic interviews among women in pregnancy or postpartum: An individual participant data meta-analysis, Int. J. Methods Psychiatr. Res. 28 (4) (2019), e1803.

[29] Y. Wu, et al., Probability of major depression diagnostic classification based on the SCID, CIDI and MINI diagnostic interviews controlling for Hospital Anxiety and Depression Scale-Depression subscale scores: an individual participant data meta-analysis of 73 primary studies, J. Psychosom. Res. 129 (2020), 109892.

[30] Y. Wu, B. Levis, J.P. Ioannidis, A. Benedetti, B.D. Thombs, Probability of Major Depression Classification Based on the SCID, CIDI, and MINI Diagnostic Interviews: A Synthesis of Three Individual Participant Data Meta-Analyses, Psychother. Psychosom. 90 (1) (2021) 28–40.

[31] Y. Wu, et al., Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis, Psychol. Med. 50 (8) (2020) 1368–1380.

[32] R. Van der Leeden, F. Busing, E. Meijer, "Bootstrap Methods for Two-Level Models: Technical Report PRM 97–04", *Leiden Univ*, Dep. Psychol. Leiden Neth. (1997).

[33] R. Van der Leeden, E. Meijer, F.M. Busing, "Resampling multilevel models", in *Handbook of multilevel analysis*, Springer (2008) 401–433.

[34] E. Walker, A.S. Nowacki, Understanding equivalence and noninferiority testing, J. Gen. Intern. Med. 26 (2) (2011) 192–196.

[35] R. C. Team R: A language and environment for statistical computing, 2013.

[36] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *ArXiv Prepr. ArXiv14065823*, 2014.

[37] E. Arthurs, R.J. Steele, M. Hudson, M. Baron, B.D. Thombs, (CSRG), Canadian Scleroderma Research Group, "Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning", PLoS ONE 7 (12) (2012), e52028.

[38] A. Teymoori, et al., Measurement invariance of assessments of depression (PHQ-9) and anxiety (GAD-7) across sex, strata and linguistic backgrounds in a European-wide sample of patients after Traumatic Brain Injury, J. Affect. Disord. 262 (2020) 278–285.

[39] H. Reich, W. Rief, and E. Brahler, "Cross-cultural validation of the German and Turkish versions of the PHQ-9: an IRT approach. BMC Psychol. 2018; 6 (26)," Epub 2018/06/07. PubMed PMID: 29871664.