BJPsych | The British Journal of Psychiatry (2018)
212, 377–385. doi: 10.1192/bjp.2018.54

Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews

Brooke Levis, Andrea Benedetti, Kira E. Riehm, Nazanin Saadat, Alexander W. Levis, Marleine Azar, Danielle B. Rice, Matthew J. Chiovitti, Tatiana A. Sanchez, Pim Cuijpers, Simon Gilbody, John P. A. Ioannidis, Lorie A. Kloda, Dean McMillan, Scott B. Patten, Ian Shrier, Russell J. Steele, Roy C. Ziegelstein, Dickens H. Akena, Bruce Arroll, Liat Ayalon, Hamid R. Baradaran, Murray Baron, Anna Beraldi, Charles H. Bombardier, Peter Butterworth, Gregory Carter, Marcos H. Chagas, Juliana C. N. Chan, Rushina Cholera, Neerja Chowdhary, Kerrie Clover, Yeates Conwell, Janneke M. de Man-van Ginkel, Jaime Delgadillo, Jesse R. Fann, Felix H. Fischer, Benjamin Fischler, Daniel Fung, Bizu Gelaye, Felicity Goodyear-Smith, Catherine G. Greeno, Brian J. Hall, John Hambridge, Patricia A. Harrison, Ulrich Hegerl, Leanne Hides, Stevan E. Hobfoll, Marie Hudson, Thomas Hyphantis, Masatoshi Inagaki, Khalida Ismail, Nathalie Jetté, Mohammad E. Khamseh, Kim M. Kiely, Femke Lamers, Shen-Ing Liu, Manote Lotrakul, Sonia R. Loureiro, Bernd Löwe, Laura Marsh, Anthony McGuire, Sherina Mohd Sidik, Tiago N. Munhoz, Kumiko Muramatsu, Flávia L. Osório, Vikram Patel, Brian W. Pence, Philippe Persoons, Angelo Picardi, Alasdair G. Rooney, Iná S. Santos, Juwita Shaaban, Abbey Sidebottom, Adam Simning, Lesley Stafford, Sharon Sung, Pei Lin Lynnette Tan, Alyna Turner, Christina M. van der Feltz-Cornelis, Henk C. van Weert, Paul A. Vöhringer, Jennifer White, Mary A. Whooley, Kirsty Winkley, Mitsuhiko Yamada, Yuying Zhang and Brett D. Thombs

Background

Different diagnostic interviews are used as reference standards for major depression classification in research. Semi-structured interviews involve clinical judgement, whereas fully structured interviews are completely scripted. The Mini International Neuropsychiatric Interview (MINI), a brief fully structured interview, is also sometimes used. It is not known whether interview method is associated with probability of major depression classification.

Aims

To evaluate the association between interview method and odds of major depression classification, controlling for depressive symptom scores and participant characteristics.

Method

Data collected for an individual participant data meta-analysis of Patient Health Questionnaire-9 (PHQ-9) diagnostic accuracy were analysed and binomial generalised linear mixed models were fit.

Results

A total of 17 158 participants (2287 with major depression) from 57 primary studies were analysed. Among fully structured interviews, odds of major depression were higher for the MINI compared with the Composite International Diagnostic Interview (CIDI) (odds ratio (OR) = 2.10; 95% CI = 1.15–3.87). Compared with semi-structured interviews, fully structured interviews (MINI excluded) were non-significantly more likely to classify participants with low-level depressive symptoms (PHQ-9 scores ≤6) as having major depression (OR = 3.13; 95% CI = 0.98–10.00), similarly likely for moderate-level symptoms (PHQ-9 scores 7–15) (OR = 0.96; 95% CI = 0.56–1.66) and significantly less likely for high-level symptoms (PHQ-9 scores ≥16) (OR = 0.50; 95% CI = 0.26–0.97).

Conclusions

The MINI may identify more people as depressed than the CIDI, and semi-structured and fully structured interviews may not be interchangeable methods, but these results should be replicated.

Declaration of interest

Drs Jetté and Patten declare that they received a grant, outside the submitted work, from the Hotchkiss Brain Institute, which was jointly funded by the Institute and Pfizer. Pfizer was the original sponsor of the development of the PHQ-9, which is now in the public domain. Dr Chan is a steering committee member 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 Hegerl declares that within the past 3 years, he was an advisory board member for Lundbeck, Servier and Otsuka Pharma; a consultant for Bayer Pharma; and a speaker for Medice Arzneimittel, Novartis, and Roche Pharma, all outside the submitted work. Dr Inagaki declares that he has received grants from Novartis Pharma, lecture fees from Pfizer, Mochida, Shionogi, Sumitomo Dainippon Pharma, Daiichi-Sankyo, Meiji Seika and Takeda, and royalties from Nippon Hyoron Sha, Nanzando, Seiwa Shoten, Igaku-shoin and Technomics, all outside of the submitted work. Dr Yamada reports personal fees from Meiji Seika Pharma Co., Ltd., MSD K.K., Asahi Kasei Pharma Corporation, Seishin Shobo, Seiwa Shoten Co., Ltd., Igaku-shoin Ltd., Chugai Igakusha and Sentan Igakusha, all outside the submitted work. All other authors declare no competing interests. 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.

https://doi.org/10.1192/bjp.2018.54 Published online by Cambridge University Press

Historically, major depression classification in research was done by clinical judgement or unstructured interviews. Lack of agreement between interviewers led to the development of standardised diagnostic interviews, including semi-structured interviews designed to be administered by clinicians, and fully structured interviews, which can be administered by lay interviewers.[1,2] Semi-structured interviews are akin to a guided diagnostic conversation. Standardised questions are asked, but interviewers may insert additional queries and use clinical judgement to decide whether symptoms are present.[2,3] Examples include the Structured Clinical Interview for DSM (SCID) and Schedules for Clinical Assessment in Neuropsychiatry (SCAN).[4,5] In contrast, fully structured interviews typically involve fully scripted, standardised questions that are read verbatim, without additional probes.[2,3] They are designed to be less subjective and provide greater standardisation, but with less flexibility and without incorporating clinical judgement.[2,3,6] Examples include the Composite International Diagnostic Interview (CIDI) and the Diagnostic Interview Schedule (DIS).[7,8] The Mini International Neuropsychiatric Interview (MINI) is also a fully structured interview, but it differs from the CIDI and DIS in that it was designed by its authors so as to be able to be administered in a fraction of the time at the cost of being over-inclusive and generating a higher rate of false-positive diagnoses.[9,10]

Although fully structured interviews are sometimes referred to as imperfect reference standards compared with semi-structured interviews,[11] both are considered appropriate reference standards for major depression classification in research.[2] Consistent with this, existing meta-analyses on depression screening tool accuracy have treated both interview types as equivalent reference standards.[12] For different interviews to be treated as equivalent diagnostic standards, the probability of being classified as meeting diagnostic criteria should not depend on the interview administered. Different interview formats, however, may lead to different diagnostic patterns. For instance, it is possible that the greater standardisation and reliability across interviews gained in fully structured interviews compared with clinician-administered semi-structured interviews could increase misclassification.

Comparing interview types

Five studies have administered validated semi-structured and fully structured interviews to the same set of participants in non-psychiatric settings within a 2-week period to assess current major depression (Supplementary Table 1 available at https://doi.org/10.1192/bjp.2018.54).[11,13–16] Most included small numbers of participants and those with major depression. Nonetheless, in the three studies with ≥100 participants, prevalence of major depression was more than twice as high when assessed with fully structured interviews compared with semi-structured interviews. To our knowledge, no studies have randomised participants to receive either a fully or semi-structured interview and compared major depression prevalence.

The high cost and burden of administering multiple diagnostic interviews to large numbers of participants or, alternatively, randomising large numbers of participants to receive semi-structured or fully structured interviews, presents a substantial barrier to testing for differences between interview types. An alternative would be to compare the probability of major depression classification using different interview types, controlling for depression symptom severity and other factors potentially related to classification. Individual participant data (IPD) meta-analysis, in which participant-level data from many studies are synthesised, offers a way to examine the association between diagnostic method and probability of major depression classification across a large number of participants, controlling for factors potentially associated with classification, including depressive symptom severity.

Study objective

The objective of this study was to examine the association between diagnostic interview method and major depression classification. First, we compared the odds of major depression classification by different diagnostic interviews: first among semi-structured interviews, and then separately among fully structured interviews, in each case controlling for depressive symptom severity and study- and participant-level characteristics. Second, we compared the odds of major depression classification between the semi-structured and fully structured interviews, including a focus on the interviews with the largest numbers of patients, the SCID and the CIDI, and controlling for depressive symptom severity and study- and participant-level characteristics. Third, we tested whether differences in the odds of classification across interview types were associated with depressive symptom severity.

Method

This study used data accrued for an IPD meta-analysis on the diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9) depression screening tool to detect major depression. Detailed methods were registered in PROSPERO (identification number CRD42014010673), and a protocol was published.[17] As an initial step, we assessed the comparability of diagnostic classifications generated by different diagnostic interviews.

Search strategy

A medical librarian searched Medline, Medline In-Process & Other Non-Indexed Citations, PsycINFO, and Web of Science from January 2000 to December 2014 on 7 February 2015, using a search strategy (Supplementary Methods 1), which was peer-reviewed with PRESS.[18] We limited our search to these databases based on research showing that adding other databases when the Medline search is highly sensitive does not identify additional eligible studies.[19] The search was limited to the year 2000 onwards because the PHQ-9 was published in 2001.[20] We 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 de-duplication, unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada), which was used to store and track search results and track the review process.

Identification of eligible studies

Data-sets from articles in any language were eligible for inclusion if they included diagnostic classification for current major depressive disorder (MDD) or major depressive episode (MDE) based on a validated semi-structured or fully structured interview conducted within 2 weeks of PHQ-9 administration, because diagnostic criteria are for symptoms in the past 2 weeks. Data-sets where not all participants were administered the PHQ-9 within 2 weeks of the diagnostic interview were included if the primary data allowed us to select participants administered the diagnostic interview and PHQ-9 within 2 weeks. Data from studies where the PHQ-9 was administered exclusively to patients known to have psychiatric diagnoses or symptoms were excluded, since screening is not done with patients already managed in psychiatric settings.[21] For defining major depression, we considered MDD or MDE based on any version of the DSM, or MDE based on any version of the ICD. (the final versions included were: ICD-10, DSM-III-R, DSM-IV and DSM-IV-TR).[22–25] If more than one was reported, we prioritised DSM over ICD, and DSM MDE over DSM MDD. We prioritised MDE over MDD because screening tests are intended to identify

https://doi.org/10.1192/bjp.2018.54 Published online by Cambridge University Press

symptoms of depression and not rule out because of bipolar disorder. We prioritised DSM over ICD because the DSM is more commonly used in existing studies. However, across all studies, there were only 23 discordant major depression classifications that depended on classification prioritisation (0.1% of participants).

Two investigators independently reviewed titles and abstracts for eligibility. If either reviewer deemed a study potentially eligible, a full-text article review was completed, also by two investigators independently. Seven members of the research team participated in the review process; however, each title and abstract and each full text was reviewed independently by only two of the seven investigators. Disagreement between reviewers after full-text review was resolved by consensus, including a third investigator (either B.L. or B.D.T.) when necessary. Titles, abstracts and full-text articles in languages other than English were translated by members of the research team or by advanced research trainees who were native speakers of the language and familiar with the topic. They were not paid for their translation services.

Data contribution and synthesis

Authors of eligible data-sets were invited to contribute de-identified primary data. Primary study country, clinical setting, language and diagnostic interview administered were extracted from published reports by two investigators independently, with disagreements resolved by consensus. Countries were categorized as 'very high', 'high' or 'low-medium' development level based on the United Nation's Human Development Index.[22] Recruitment settings were categorized as 'non-medical', 'primary care', 'in-patient specialty care' or 'out-patient specialty care'. Participant-level data included age, gender, major depression status and PHQ-9 scores. In three primary studies, multiple settings were included, thus setting was coded at the participant level.

IPD were converted to a standard format and entered into a single data-set that also included study-level data. We compared published participant characteristics and diagnostic accuracy results with results obtained using the raw data-sets. When primary data and original publications were discrepant, we identified and corrected errors when possible and resolved outstanding discrepancies in consultation with the original investigators. Two investigators assessed risk of bias of included studies independently, using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool.[23] See Supplementary Methods 2 for QUADAS-2 coding rules. Discrepancies in data extraction and risk of bias assessment were resolved by consensus.

Statistical analyses

To isolate the association between diagnostic assessment method and major depression classification, we estimated binomial generalised linear mixed models (GLMMs) with a logit link function. In all analyses, the outcome was major depression classification. The predictor of interest was either the specific diagnostic interview or interview category, depending on the analysis. Covariates were depressive symptom severity (PHQ-9 score), age, gender, country Human Development Index and clinical setting. The PHQ-9 has been shown in many studies, across diverse populations in both medical and non-medical settings, to be a valid measure of depressive symptom severity with good convergent validity and a one-dimensional factor structure.[20,24–31] Other covariates were chosen because of their potential influence on major depression classification and their availability across primary studies. To account for correlation between participants within the same primary study, a random intercept was fit for each primary study. Fixed slopes were estimated for PHQ-9 score, assessment method, age, gender, Human Development Index and clinical setting.

First, we estimated a GLMM among studies that used semi-structured interviews (SCID, SCAN and Depression Interview and Structured Hamilton (DISH)). Then, we estimated a GLMM among studies that used fully structured interviews (CIDI, Clinical Interview Schedule-Revised (CIS-R), DIS and MINI). For each model, we used the interview with the greatest number of participants as the reference category.

Second, because the MINI was intentionally designed to be a brief but overly inclusive tool,[9,10] and based on results from the first analyses that were consistent with this, we compared fully structured diagnostic interviews without the MINI, with semi-structured interviews. To do this, we estimated a GLMM to compare odds of major depression classification between the remaining semi-structured and fully structured interviews (reference = semi-structured). As a sensitivity analysis, we further restricted our analysis to studies using either the CIDI or SCID (reference = SCID), as these interviews were used substantially more often than other included interviews.

Third, we investigated a possible interaction between interview assessment method and depressive symptom severity based on categorical PHQ-9 score classifications. To do this, we separated PHQ-9 scores into three categories: low (scores 0–6; reference group), medium (scores 7–15) and high (scores 16–27). Score ranges were chosen because recent meta-analyses of the PHQ-9 have evaluated cut-off scores from 7 to 15, suggesting a mid-level range.[32] To compare models with and without the interaction term, a likelihood ratio test was used. We then replicated the model comparing semi-structured and fully structured interviews in each PHQ-9 category separately to obtain stratum-specific classification odds ratios for fully versus semi-structured interviews. Additionally, we conducted a separate interaction analysis between continuous PHQ-9 score and diagnostic interview method. As a sensitivity analysis, we further restricted our interaction analyses to studies using the CIDI or SCID.

In another set of sensitivity analyses, we re-ran all of our models adding domain scores for QUADAS-2. All analyses were run in R, using the lme4 package.

Ethics

As this study involved secondary analysis of anonymised previously collected data, the Research Ethics Committee of the Jewish General Hospital declared that this project did not require research ethics approval. However, for each included data-set, we confirmed that the original study received ethics approval and that all patients provided informed consent.

Results

Search results and inclusion of primary data

Of 5248 unique titles and abstracts identified from the database search, 5039 were excluded after title and abstract review and 113 after full-text review, leaving 96 eligible articles with data from 69 unique participant samples (Supplementary Fig. 1). Of the 69 unique samples, 55 contributed data (80%). In addition, authors of included studies contributed data from three unpublished studies, for a total of 58 data-sets. However, one primary data-set did not include data for key covariates included in analyses and was excluded, leaving 57 primary data-sets. In total, 17 158 participants (2287 with major depression) were included. Included study characteristics are shown in Supplementary Table 2a. Characteristics of eligible studies that did not provide data for the present study are shown in Supplementary Table 2b. Of the 21 171 participants in 69 eligible published data-sets,

https://doi.org/10.1192/bjp.2018.54 Published online by Cambridge University Press

Table 1 Participant data by diagnostic interview

Diagnostic interview	Studies, *N*	Participants, *N*	Major depression	
			N	%
Semi-structured				
SCID	26	4732	785	17
SCAN	2	1891	130	7
DISH	1	100	9	9
Fully structured				
CIDI	11	6271	554	9
DIS	1	1006	221	22
MINI	14	2756	524	19
CIS-R	2	402	64	16
Total	57	17 158	2287	13

CIDI: Composite International Diagnostic Interview; CIS-R: Clinical Interview Schedule-Revised; DIS: Diagnostic Interview Schedule; DISH: Depression Interview and Structured Hamilton; MINI: Mini International Neuropsychiatric Interview; SCAN: Schedules for Clinical Assessment in Neuropsychiatry; SCID: Structured Clinical Interview for DSM Disorders.

16 757 were in the 54 published studies with data included in the present study (79%).

Of the 57 total included studies, 29 used semi-structured interviews and 28 used fully structured interviews (Table 1). The SCID was the most commonly used semi-structured interview (26 studies, 4732 participants), and the CIDI (11 studies, 6271 participants) and MINI (14 studies, 2756 participants) were the most commonly used fully structured interviews.

Association of diagnostic interview and major depression classification

Semi-structured interviews

Among semi-structured interviews, compared with the SCID, odds of major depression were not significantly different for the SCAN (adjusted odds ratio (aOR) = 0.56; 95% CI = 0.18–1.78) or DISH (aOR = 1.13; 95% CI = 0.19–6.80). However, only two studies used the SCAN, and only one used the DISH.

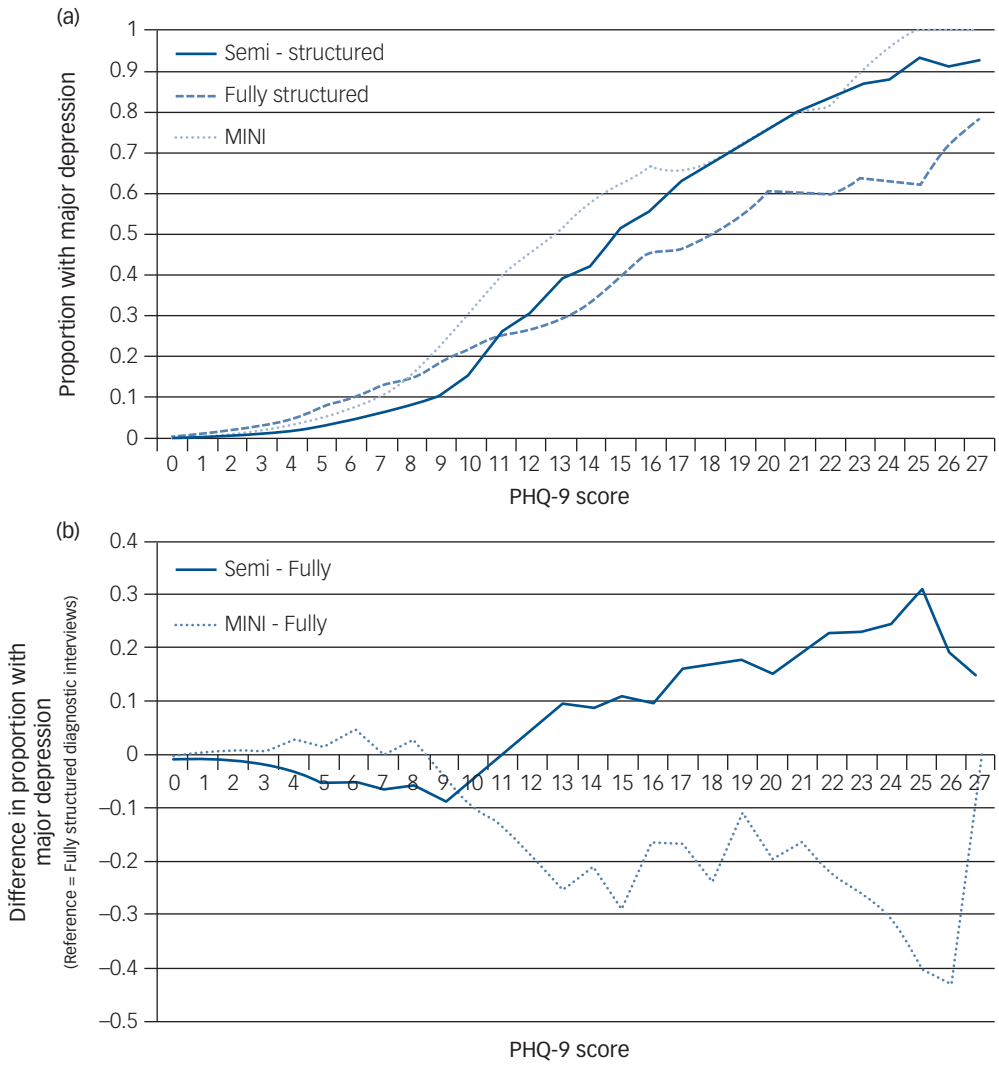

Fig. 1 (a) Probability of major depression classification by PHQ-9 score for semi-structured interviews, fully structured interviews (excluding the MINI) and the MINI. Proportions are plotted as three-point moving averages (e.g. the proportions at the PHQ-9 score of 10 are averages of the proportions at PHQ-9 scores of 9, 10, and 11). (b) Difference in probability of major depression classification by PHQ-9 score for semi-structured interviews and the MINI compared with fully structured interviews (excluding the MINI). Differences in proportions are plotted as three-point moving averages (e.g. the differences in proportions at the PHQ-9 score of 10 are averages of the differences in proportions at PHQ-9 scores of 9, 10, and 11). MINI: Mini International Neuropsychiatric Interview; PHQ-9: Patient Health Questionnaire-9.

https://doi.org/10.1192/bjp.2018.54 Published online by Cambridge University Press

Table 2 Model summary of fixed effects generalised linear mixed model considering a potential interaction between PHQ-9 score category and assessment method[a, b]

Variable	Odds ratio	95% CI
Fully structured assessment method	1.49	0.82–2.72
PHQ-9 total score	1.37	1.35–1.40
Age (years)	1.00	0.99–1.00
Male	0.89	0.77–1.03
Clinical setting	–	–
Non-medical (reference)	–	–
Primary care	0.67	0.27–1.64
Specialty care: in-patient	0.33	0.13–0.85
Specialty care: out-patient	0.64	0.26–1.54
Human Development Index	–	–
Very high (reference)	–	–
High	2.27	1.11–4.61
Low to medium	0.78	0.27–2.24
PHQ-9 score category * fully structured assessment method[c]	–	–
Low PHQ-9 (0–6) (reference)	–	–
Medium PHQ-9 (7–15)	0.73	0.57–0.92
High PHQ-9 (16–27)	0.26	0.18–0.37

PHQ-9: Patient Health Questionnaire-9.
a. Excluding the Mini International Neuropsychiatric Interview.
b. Estimate of random intercept variance = 0.58.
c. $P < 0.001$ in likelihood ratio test comparing models with and without interaction term.

Fully structured interviews

Among fully structured interviews, compared with the CIDI, odds of major depression were higher, but not significantly different for the DIS (aOR = 4.32; 95% CI = 0.95–19.62) or CIS-R (aOR = 1.53; 95% CI = 0.48–4.91), although these estimates were based on one and two studies, respectively. Participants interviewed with the MINI were substantially and statistically significantly more likely to be classified as having major depression (aOR = 2.10; 95% CI = 1.15–3.87).

Semi-structured versus fully structured interviews

Excluding the MINI, odds of major depression were similar with fully versus semi-structured interviews (aOR = 0.90; 95% CI = 0.51–1.57). In a sensitivity analysis restricted to studies that used the SCID or CIDI, odds of major depression were lower for the CIDI compared with the SCID, but this was not statistically significantly different (aOR = 0.57; 95% CI = 0.32–1.02).

Interaction between PHQ-9 scores and diagnostic interview method

The proportion of participants classified as having major depression at each PHQ-9 score for semi-structured interviews, fully structured interviews (MINI excluded) and the MINI are shown in Fig. 1a, with differences in proportions across interview types shown in Fig. 1b. As shown in Fig. 1 and Supplementary Table 3, compared with semi-structured interviews, fully structured interviews resulted in a somewhat higher probability of major depression classification for PHQ-9 scores from 0 to 10, but lower probability for PHQ-9 scores of 11–27. Consistent with this, there was a significant interaction between assessment method and PHQ-9 score category (Table 2), and the likelihood ratio test comparing models with and without the interaction term was statistically significant ($P < 0.001$). The interaction was also statistically significant when tested with the PHQ-9 as a continuous variable. The aOR for the interaction between PHQ-9 score and fully structured interview was 0.90 (95% CI = 0.88–0.92), which suggested a 10% dilution in the slope of the odds of a major depression classification across PHQ-9 scores for fully structured interviews compared with semi-structured interviews.

When stratified based on PHQ-9 score category, participants with low PHQ-9 scores (0–6) were more likely to receive a major depression classification with a fully structured interview compared with a semi-structured interview (aOR = 3.13; 95% CI = 0.98–10.00), although this was not statistically significant. Semi-structured and fully structured interviews performed similarly among participants in the medium PHQ-9 group (scores 7–15: aOR = 0.96; 95% CI = 0.56–1.66). Among participants with high PHQ-9 scores (16–27), participants were significantly less likely to be classified with major depression when using fully structured interviews (aOR = 0.50; 95% CI = 0.26–0.97; Table 3). These odds ratios corresponded to a crude prevalence of 3.2% among those administered a fully structured interview versus 1.2% among those administered a semi-structured interview in the low-range PHQ-9 group, 21.4 *v.* 20.8% in the medium-range group, and 54.2 *v.* 72.5% in the high-range group, not adjusting for PHQ-9 scores or participant characteristics.

In sensitivity analyses restricted to studies that used the SCID or CIDI, results for interaction models were similar.

Risk of bias sensitivity analyses

See Supplementary Table 4 for QUADAS-2 ratings for each included primary study. In sensitivity analyses with models that included QUADAS-2 domains, no domains were significantly associated with major depression, and the inclusion of the QUADAS-2 domains did not substantially change coefficient estimates for any variables.

Discussion

There were two main findings. First, among fully structured interviews, the adjusted odds of being classified as having major

Table 3 Generalised linear mixed model summaries for each PHQ-9 score category

PHQ-9 score category	Low PHQ-9 scores (0–6) *N* = 9339		Medium PHQ-9 scores (7–15) *N* = 3970		High PHQ-9 scores (16–27) *N* = 1093	
	OR	95% CI	OR	95% CI	OR	95% CI
OR[a] (95% CI) for fully structured assessment method	3.13	0.98–10.00	0.96	0.56–1.66	0.50	0.26–0.97
No. receiving fully structured interview	5228		1999		452	
	N	%	*N*	%	*N*	%
No. with major depression receiving fully structured interview	167	3.2	427	21.4	245	54.2
No. receiving semi-structured interview	4111		1971		641	
	N	%	*N*	%	*N*	%
No. with major depression receiving semi-structured interview	50	1.2	409	20.8	465	72.5

OR: odds ratio; PHQ-9: Patient Health Questionnaire-9.
a. Excluding the Mini International Neuropsychiatric Interview and adjusted for PHQ-9 score, age, gender, clinical setting and Human Development Index.

https://doi.org/10.1192/bjp.2018.54 Published online by Cambridge University Press

depression were approximately twice as high with the MINI compared with the CIDI. Second, excluding the MINI, there was a statistically significant interaction between fully structured versus semi-structured interview and depressive symptom severity based on the PHQ-9. Compared with semi-structured interviews, the likelihood of major depression classification increased significantly less for fully structured interviews as symptom severity increased. Fully structured interviews tended to classify more participants with low-level symptoms as having major depression, although this was not statistically significant; they performed similar to semi-structured interviews for participants with moderate symptoms, and they classified fewer participants with high-level symptoms as having major depression compared with semi-structured interviews.

The finding that odds of major depression classification were twice as high for the MINI compared with the CIDI is consistent with the interviews' designs. Whereas the CIDI and other fully structured interviews are in-depth interviews,[7,8] the MINI was developed to be able to be administered in a fraction of the time as other interviews and was described by its developers as designed to be over-inclusive.[9,10] Our findings suggest that, consistent with the developers' intent, the MINI may identify substantially higher rates of major depression if used to determine major depression status than other fully structured interviews. The probability of being classified with major depression was also high based on the DIS and CIS-R, but evidence was too limited to draw conclusions. The formats of these interviews, however, are more similar to the CIDI than the MINI.

By standardising all questions and probes and removing clinical judgement, fully structured interviews are designed to be as reliable as possible, but this may reduce advantages of semi-structured interviews related to the inclusion of a framework for incorporating clinical judgement. Consistent with this, our findings suggest that compared with semi-structured interviews, the association between symptom levels and probability of being classified as having major depression is lower for fully structured interviews (MINI excluded). Compared with semi-structured interviews, participants with low-level depressive symptoms assessed with fully structured interviews appeared more likely to be classified as having major depression, whereas participants with high-level symptoms appeared less likely. Participants with moderate symptoms were similarly likely to be classified as having major depression when semi-structured and fully structured interviews were used. This suggests that, in practice, the effect of the diagnostic interview that is selected on the prevalence that is generated likely depends on the underlying distribution of symptom levels in the population.

Existing data from other studies is roughly consistent with this. In general population samples, where depressive symptom levels are generally low, major depression prevalence has been found to be substantially higher when fully structured interviews are used versus semi-structured interviews (Supplementary Table 1).[11,13] On the other hand, in a study of patients from an alcoholic treatment unit, where depressive symptoms would be expected to be much higher, major depression prevalence was similar based on semi-structured and fully structured interviews.[15]

In research settings, semi-structured and fully structured interviews are typically used interchangeably as appropriate reference standards in depression screening tool diagnostic accuracy studies, for inclusion and exclusion in treatment trials and for determining major depression prevalence. Based on the findings of the present study, caution is warranted when deciding which interview to use. Prevalence estimates may be influenced, potentially substantially, by this choice. It is not clear to what degree estimates of screening tool accuracy may be influenced by a fully versus semi-structured interview, and this should be determined by future studies, including a replication of this study with data from IPD meta-analyses of other depression screening tools.[33,34]

This is the first study to compare fully and semi-structured interviews for major depression with an IPD meta-analysis approach. Strengths of this study include the large overall sample size and the ability to consider both study- and participant-level factors in analyses, including participant-specific depressive symptom severity. There are also limitations to consider. First, we were unable to include primary data from 15 out of 69 eligible data-sets (20% of eligible data-sets, 21% of eligible participants), and we restricted our analyses to those with complete data for all variables in our models (98% of available data). Nonetheless, this was a very large sample, many times the size of existing studies that have attempted to compare fully and semi-structured interviews for major depression. None of those studies included more than 61 participants with major depression based on a fully structured interview or 22 participants with major depression based on a semi-structured interview. Second, despite the large overall sample size, there was substantial heterogeneity across studies. We were not able to conduct subgroup analyses based on medical comorbidity or cultural aspects such as country or language because comorbidity data were not available for over half of participants, and many countries and languages were represented in few primary studies. However, studies of differential item functioning with the PHQ-9 have shown that it performs equivalently across multiple languages and between people with and without medical disorders.[35–39] Third, it is possible that residual confounding may exist, given that we were only able to consider variables collected in the original investigations, and the included study-level variables may not apply uniformly to all participants in a study. Fourth, although we coded for the qualifications of the interviewer for all semi-structured interviews as part of our QUADAS-2 rating, two studies used interviewers who did not meet typical standards, and approximately half of studies were rated unclear. This may have influenced the quality of the reference standard in some studies. Fifth, particularly for semi-structured interviews, lack of interviewer blinding may have influenced classifications. Although only two studies were coded as having non-blinded interviewers, 11 were coded as unclear. We did not query authors on interviewer characteristics and blinding if information was not published because of concern that author recollection, in some cases after over a decade had passed, may not have been accurate.

In summary, we found that the MINI diagnostic interview was associated with a substantially higher probability of major depression classification than the CIDI, controlling for depression symptom scores on the PHQ-9 and other patient characteristics. We also found that compared with semi-structured interviews, fully structured interviews tend to classify more people with low-level symptoms as depressed, but fewer people with high-level symptoms. This suggests that the choice to use either a fully structured diagnostic interview or a semi-structured interview to classify major depression may influence research findings. This is the first study that has used a large participant sample and IPD meta-analysis to compare diagnostic interview methods, and future research should replicate this study to verify results.

Brooke Levis, MSc, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada and Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada; **Andrea Benedetti**, PhD, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada; Department of Medicine, McGill University, Montréal, Québec, Canada and Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada; **Kira E. Riehm**, MSc, **Nazanin Saadat**, BSc, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; **Alexander W. Levis**, MSc, **Marleine Azar**, BSc, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec,

https://doi.org/10.1192/bjp.2018.54 Published online by Cambridge University Press

Canada and Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada; **Danielle B. Rice**, MSc, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada and Department of Psychology, McGill University, Montréal, Québec, Canada; **Matthew J. Chiovitti**, MISt, **Tatiana A. Sanchez**, BSc, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; **Pim Cuijpers**, PhD, Department of Clinical, Neuro and Developmental Psychology, EMGO Institute, VU University, Amsterdam, the Netherlands; **Simon Gilbody**, PhD, Hull York Medical School and the Department of Health Sciences, University of York, York, UK; **John P. A. Ioannidis**, MD, Department of Medicine, Department of Health Research and Policy, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, California, USA; **Lorie A. Kloda**, PhD, Library, Concordia University, Montréal, Québec, Canada; **Dean McMillan**, PhD, Hull York Medical School and the Department of Health Sciences, University of York, York, UK; **Scott B. Patten**, MD, Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada and Hotchkiss Brain Institute and O'Brien Institute for Public Health, University of Calgary, Calgary, Alberta, Canada; **Ian Shrier**, MD, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada and Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada; **Russell J. Steele**, PhD, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada and Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada; **Roy C. Ziegelstein**, MD, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; **Dickens H. Akena**, PhD, Department of Psychiatry, Makerere University College of Health Sciences, Kampala, Uganda; **Bruce Arroll**, MBChB, Department of General Practice and Primary Health Care, University of Auckland, New Zealand; **Liat Ayalon**, PhD, Louis and Gabi Weisfeld School of Social Work, Bar Ilan University, Ramat Gan, Israel; **Hamid R. Baradaran**, MD, Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; **Murray Baron**, MD, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada and Department of Medicine, McGill University, Montréal, Québec, Canada; **Anna Beraldi**, PhD, Kbo-Lech-Mangfall-Klinik Garmisch-Partenkirchen, Klinik für Psychiatrie, Psychotherapie & Psychosomatik, Lehrkrankenhaus der Technischen Universität München, Munich, Germany; **Charles H. Bombardier**, PhD, Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA; **Peter Butterworth**, PhD, Centre for Research on Ageing, Health and Wellbeing, Research School of Population Health, The Australian National University, Canberra, Australia; Centre for Mental Health, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia and Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Melbourne, Australia; **Gregory Carter**, FRANZCP, Centre for Translational Neuroscience and Mental Health, University of Newcastle, New South Wales, Australia; **Marcos H. Chagas**, MD, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; **Juliana C. N. Chan**, MD, Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China; Asia Diabetes Foundation, Prince of Wales Hospital, Hong Kong Special Administrative Region, China and Hong Kong Institute of Diabetes and Obesity, Hong Kong Special Administrative Region, China; **Rushina Cholera**, MD, Department of Pediatrics, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA; **Neerja Chowdhary**, MD, Clinical Psychiatrist, Mumbai, India; **Kerrie Clover**, PhD, Centre for Translational Neuroscience and Mental Health, University of Newcastle, New South Wales, Australia and Psycho-Oncology Service, Calvary Mater Newcastle, New South Wales, Australia; **Yeates Conwell**, MD, Department of Psychiatry, University of Rochester Medical Center, New York, USA; **Janneke M. de Man-van Ginkel**, PhD, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands; **Jaime Delgadillo**, PhD, Clinical Psychology Unit, Department of Psychology, University of Sheffield, Sheffield, UK; **Jesse R. Fann**, MD, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA; **Felix H. Fischer**, PhD, Institute for Social Medicine, Epidemiology, and Health Economics, Charité – Universitätsmedizin Berlin, Germany and Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité – Universitätsmedizin Berlin, Germany; **Benjamin Fischler**, MD, Private Practice, Brussels, Belgium; **Daniel Fung**, MD, Department of Child & Adolescent Psychiatry, Institute of Mental Health, Singapore; Yong Loo Lin School of Medicine, National University of Singapore, Singapore; Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore and Office of Clinical Sciences, Duke-NUS Medical School, Singapore; **Bizu Gelaye**, PhD, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA; **Felicity Goodyear-Smith**, MD, Department of General Practice and Primary Health Care, University of Auckland, New Zealand; **Catherine G. Greeno**, PhD, School of Social Work, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; **Brian J. Hall**, PhD, Global and Community Mental Health Research Group, Department of Psychology, Faculty of Social Sciences, University of Macau, Macau Special Administrative Region, China and Department of Health, Behavior, and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA; **John Hambridge**, Dip Clin Psych, Liaison Psychiatry Department, John Hunter Hospital, Newcastle, Australia; **Patricia A. Harrison**, PhD, City of Minneapolis Health Department, Minneapolis, Minnesota, USA; **Ulrich Hegerl**, MD, Department of Psychiatry and Psychotherapy, University Hospital Leipzig, Leipzig, Germany; **Leanne Hides**, PhD(Clin), Centre for Children's Health Research, Institute of Health & Biomedical Innovation, School of Psychology, University of Queensland, Brisbane, Queensland, Australia; **Stevan E. Hobfoll**, PhD, Department of Behavioral Sciences, Rush University Medical Center, Chicago, Illinois, USA; **Marie Hudson**, MD, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada and Department of Medicine, McGill University, Montréal, Québec, Canada; **Thomas Hyphantis**, MD, Department of Psychiatry, University of Ioannina, Ioannina, Greece; **Masatoshi Inagaki**, MD, Department of Neuropsychiatry, Okayama University Hospital, Okayama, Japan; **Khalida Ismail**, MD, Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neurosciences, King's College London Weston Education Centre, London, UK; **Nathalie Jetté**, MD, Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada; Hotchkiss Brain Institute and O'Brien Institute for Public Health, University of Calgary, Calgary, Alberta, Canada and Department of Clinical Neurosciences, University of Calgary, Calgary, Alberta, Canada; **Mohammad E. Khamseh**, MD, Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; **Kim M. Kiely**, PhD, Centre for Research on Ageing, Health and Wellbeing, Research School of Population Health, The Australian National University, Canberra, Australia; **Femke Lamers**, PhD, Department of Psychiatry, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, the Netherlands; **Shen-Ing Liu**, MD, Office of Clinical Sciences, Duke-NUS Medical School, Singapore; Department of Psychiatry, Mackay Memorial Hospital, Taipei, Taiwan; Department of Medical Research, Mackay Memorial Hospital, Taipei, Taiwan and Department of Medicine, Mackay Medical College, Taipei, Taiwan; **Manote Lotrakul**, MD, Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand; **Sonia R. Loureiro**, PhD, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; **Bernd Löwe**, MD, Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany and Schön Klinik Hamburg Eilbek, Hamburg, Germany; **Laura Marsh**, MD, Baylor College of Medicine, Houston and Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas, USA; **Anthony McGuire**, PhD, Department of Nursing, St. Joseph's College, Standish, Maine, USA; **Sherina Mohd Sidik**, PhD, Cancer Resource & Education Centre, and Department of Psychiatry, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia; **Tiago N. Munhoz**, PhD, Post-graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil; **Kumiko Muramatsu**, MD, Department of Clinical Psychology, Graduate School of Niigata Seiryo University, Niigata, Japan; **Flávia L. Osório**, PhD, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil and National Institute of Science and Technology, Translational Medicine, Ribeirão Preto, Brazil; **Vikram Patel**, MD, Department of Global Health and Social Medicine, Harvard Medical School, Boston, Massachusetts, USA; London School of Hygiene & Tropical Medicine, London, UK and Centre for Chronic Conditions and Injuries, Public Health Foundation of India, New Delhi, India; **Brian W. Pence**, PhD, Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; **Philippe Persoons**, MD, Department of Adult Psychiatry, University Hospitals Leuven, Leuven, Belgium and Department of Neurosciences, Katholieke Universiteit Leuven, Leuven, Belgium; **Angelo Picardi**, MD, Centre for Behavioural Sciences and Mental Health, Italian National Institute of Health, Rome, Italy; **Alasdair G. Rooney**, MD, Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburg, Edinburgh, Scotland, UK; **Iná S. Santos**, MD, Post-graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil; **Juwita Shaaban**, MMed (Fam. Med), Department of Family Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia; **Abbey Sidebottom**, PhD, Allina Health, Minneapolis, Minnesota, USA; **Adam Simning**, MD, Department of Psychiatry, University of Rochester Medical Center, New York, USA; **Lesley Stafford**, PhD, Centre for Women's Mental Health, Royal Women's Hospital, Parkville, Australia and Melbourne School of Psychological Sciences, University of Melbourne, Australia; **Sharon Sung**, PhD, Department of Child & Adolescent Psychiatry, Institute of Mental Health, Singapore and Office of Clinical Sciences, Duke-NUS Medical School, Singapore; **Pei Lin Lynnette Tan**, MMed (Psychiatry), Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore; **Alyna Turner**, PhD, School of Medicine and Public Health, University of Newcastle, New South Wales, Newcastle, Australia and IMPACT Strategic Research Centre, School of Medicine, Deakin University, Geelong, Victoria, Australia; **Christina M. van der Feltz-Cornelis**, MD, Clinical Center of Excellence for Body, Mind and Health, GGz Breburg, Tilburg, the Netherlands and Tilburg University, Faculty of Social Sciences, Tranzo Department, Tilburg, the Netherlands; **Henk C. van Weert**, MD, Department of General Practice, Academic Medical Centre Amsterdam, University of Amsterdam, Amsterdam, the Netherlands; **Paul A. Vöhringer**, MD, Department of Psychiatry and Mental Health, Clinical Hospital, Universidad de Chile, Santiago, Chile; Millennium Institute for Depression and Personality Research (MIDAP), Ministry of Economy, Macul, Santiago, Chile and Mood Disorders Program, Tufts Medical Center, Tufts University, Boston, USA; **Jennifer White**, PhD, Monash University, Melbourne, Australia; **Mary A. Whooley**, MD, Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA; Department of Medicine, Veterans Affairs Medical Center, San Francisco, California, USA and Department of Medicine, University of California San Francisco, San Francisco, California, USA; **Kirsty Winkley**, PhD, Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neurosciences, King's College London Weston Education Centre, London, UK; **Mitsuhiko Yamada**, MD, Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, Ogawa-Higashi, Kodaira, Tokyo, Japan; **Yuying Zhang**, PhD, Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China and Asia Diabetes Foundation, Prince of Wales Hospital, Hong Kong Special Administrative Region, China; **Brett D. Thombs**, PhD, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada and Department of Epidemiology, Biostatistics and Occupational Health, Department of Medicine, Department of Psychology, Department of Psychiatry, and Department of Educational and Counselling Psychology, McGill University, Montréal, Québec, Canada

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

First received 3 Oct 2017, final revision 17 Feb 2018, accepted 22 Feb 2018

Supplementary material

Supplementary material and author contributions are available online at https://doi.org/10.1192/bjp.2018.54.

Funding

This study was funded by the Canadian Institutes of Health Research (CIHR, KRS-134297). Ms Levis was supported by a CIHR Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Award. Dr Benedetti was supported by a Fonds de recherche du

https://doi.org/10.1192/bjp.2018.54 Published online by Cambridge University Press

Québec – Santé (FRQS) researcher salary award. Ms Riehm and Ms Saadat were supported by CIHR Frederick Banting and Charles Best Canada Graduate Scholarship Master's Awards. Mr Levis and Ms Azar were supported by FRQS Masters Training Awards. Ms Rice was supported by a Vanier Canada Graduate Scholarship. Collection of data for the study by Arroll *et al.* was supported by a project grant from the Health Research Council of New Zealand. 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 study by Bombardier *et al.* was supported by the Department of Education, National Institute on Disability and Rehabilitation Research, Spinal Cord Injury Model Systems: University of Washington (grant no. H133N060033), Baylor College of Medicine (grant no. H133N060003) and University of Michigan (grant no. H133N060032). Dr Butterworth was supported by Australian Research Council Future Fellowship FT130101444. Dr Cholera was supported by a United States National Institute of Mental Health (NIMH) grant (5F30MH096664), and the United States National Institutes of Health (NIH) Office of the Director, Fogarty International Center, Office of AIDS Research, National Cancer Center, National Heart, Blood, and Lung Institute and the NIH Office of Research for Women's Health through the Fogarty Global Health Fellows Program Consortium (1R25TW00934001) and the American Recovery and Reinvestment Act. Dr Conwell received support from NIMH (R24MH071604) and the Centers for Disease Control and Prevention (R49 CE002093). Collection of data for the primary study by Delgadillo *et al.* was supported by a grant from St. Anne's Community Services, Leeds, UK. Collection of data for the primary study by Fann *et al.* was supported by grant RO1 HD39415 from the National Center for Medical Rehabilitation Research (USA). The primary studies by Amoozegar and by Fiest *et al.* were funded by the Alberta Health Services, the University of Calgary Faculty of Medicine and the Hotchkiss Brain Institute. The primary study by Fischer *et al.* was funded by the German Federal Ministry of Education and Research (01GY1150). Dr Fischler was supported by a grant from the Belgian Ministry of Public Health and Social Affairs and a restricted grant from Pfizer Belgium. Data for the primary study by Gelaye *et al.* was supported by a grant from the NIH (T37 MD001449). Collection of data for the primary study by Gjerdingen *et al.* was supported by grants from the NIMH (R34 MH072925, K02 MH65919 and P30 DK50456). The primary study by Eack *et al.* was funded by the NIMH (R24 MH56858). 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 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. The primary study by Henkel *et al.* was funded by the German Ministry of Research and Education. 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, FRQS, the Canadian Arthritis Network and the Lady Davis Institute of Medical Research of the Jewish General Hospital, Montreal, Canada. Dr Hudson was supported by a FRQS Senior Investigator Award. Collection of data for the primary study by Hyphantis *et al.* was supported by a 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 Inagaki *et al.* was supported by the Ministry of Health, Labour and Welfare, Japan. Dr Jetté was supported by a Canada Research Chair in Neurological Health Services Research. Collection of data for the primary study by Kiely *et al.* was supported by National Health and Medical Research Council (grant 1002160) and Safe Work Australia. Dr Kiely was supported by funding from an Australian National Health and Medical Research Council fellowship (grant 1088313). The primary study by Lamers *et al.* was funded by the Netherlands Organisation for Health Research and development (grant 945-03-047). Dr Lamers received funding from the European Union Seventh Framework Programme (FP7/2007-2013, PCIG12-GA-2012-334065). The primary study by Liu *et al.* 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 49086). Dr 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.* 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 of Excellence (PRONEX/FAPERGS/CNPq, Brazil). The primary study by Muramatsu *et al.* was supported by an educational grant from Pfizer US Pharmaceutical Inc. Collection of primary data for the study by Dr Pence was provided by NIMH (R34MH084673). The primary studies by Osório *et al.* were funded by Reitoria de Pesquisa da Universidade de São Paulo (grant 09.1.01689.17.7) and Banco Santander (grant 10.1.01232.17.9). The primary study by Picardi *et al.* was supported by funds for current research from the Italian Ministry of Health. Dr Persoons was supported by a grant from the Belgian Ministry of Public Health and Social Affairs and a restricted grant from Pfizer Belgium. Dr Shaaban was supported by funding from Universiti Sains Malaysia. The primary study by Rooney *et al.* was funded by the National Health Service Lothian Neuro-Oncology Endowment Fund (UK). The primary study by Sidebottom *et al.* was funded by a grant from the United States Department of Health and Human Services, Health Resources and Services Administration (grant R40MC07840). Simning *et al.*'s research was supported in part by grants from the NIH (T32 GM07356), Agency for Healthcare Research and Quality (R36 HS018246), NIMH (R24 MH071604) and the National Center for Research Resources (TL1 RR024135). Dr Stafford received PhD scholarship funding from the University of Melbourne. The study by van Steenbergen-Weijenburg *et al.* was funded by Innovatiefonds Zorgverzekeraars. Collection of data for the studies by Turner *et al.* were funded by a bequest from Jennie Thomas through the Hunter Medical Research Institute. Dr Vöhringer was supported by the Fund for Innovation and Competitiveness of the Chilean Ministry of Economy, Development and Tourism, through the Millennium Scientific Initiative (grant IS130005). Collection of data for the primary study by Williams *et al.* was supported by an NIMH grant to Dr Marsh (RO1-MH069666). 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 Twist *et al.* was funded by the National Institute for Health Research (UK) under its Programme Grants for Applied Research Programme (grant RP-PG-0606-1142). The primary study by Thombs *et al.* was done with data from the Heart and Soul Study (Primary Investigator M.W). The Heart and Soul Study was funded by the Department of Veterans Epidemiology Merit Review Program, the Department of Veterans Affairs Health Services Research and Development service, the National Heart Lung and Blood Institute (R01 HL079235), the American Federation for Aging Research, the Robert Wood Johnson Foundation and the Ischemia Research and Education Foundation. Dr Thombs was supported by an Investigator Award from the Arthritis Society. No other authors reported funding for primary studies or for their work on the present study. The study sponsors had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report or in the decision to submit the paper for publication. B.D.T. had full access to all data in the study and had final responsibility for the decision to submit for publication.

Author contributions

BLevis, ABenedetti, P.C., S.G., J.P.A.I., L.A.K., D.M., S.B.P., I.S., R.J.S., R.C.Z. and B.D.T. were responsible for the study conception and design. D.H.A., B.A., L.A., H.R.B., M.B., ABeraldi, C.H.B., P.B., G. C., M.H.C., J.C.N.C., R.C., N.C., K.C., Y.C., J.M.G., J.D., J.R.F., F.H.F., B.F., D.F., B.G., S.G., F.G.S., C.G.G., B.J.H., J.H., P.A.H., U.H., L.H., S.E.H., M.H., T.H., M.I., K.I., N.J., M.E.K., K.M.K., F.L., S.L., M.L., S.R.L., BLöwe, L.M., A.M., S.M.S., T.N.M., K.M., F.L.O., V.P., B.W.P., P.P., A.P., A.G.R., I.S.S., J.S., ASidebottom, ASimning, L.S., S.S., P.L.L.T., A.T., C.M.vdF.C., H.C.vW., P.A.V., J.W., M.A.H., K.W., M.Y., Y.Z., and B.D.T. were responsible for collection of primary data included in this study. BLevis, K.E.R., N.S., M.A., D.B.R., M.J.C., T.A.S., and B.D.T. contributed to data extraction and coding for the meta-analysis. BLevis, ABenedetti, A.W.L., and B.D.T. contributed to the data analysis and interpretation. BLevis, ABenedetti, and B.D.T. contributed to drafting the manuscript. All authors provided a critical review and approved the final manuscript. B.D.T. is the guarantor.

References

1 Jones KD. The unstructured clinical interview. *J Couns Dev* 2010; **88**: 220–6.

2 Brugha TS, Bebbington PE, Jenkins R. A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. *Psychol Med* 1999; **29**: 1013–20.

3 Nosen E, Woody SR. Chapter 8: Diagnostic assessment in research. In *Handbook of Research Methods in Abnormal and Clinical Psychology* (ed D McKay). Sage, 2008, pp.109–124.

4 First MB. *Structured Clinical Interview for the DSM (SCID)*. John Wiley & Sons, Inc., 1995.

5 World Health Organization. *Schedules for Clinical Assessment in Neuropsychiatry: Manual*. Amer Psychiatric Pub Inc., 1994.

6 Kurdyak PA, Gnam WH. Small signal, big noise: performance of the CIDI depression module. *Can J Psychiatry* 2005; **50**: 851–6.

7 Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, et al. The Composite International Diagnostic Interview: an epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry* 1988; **45**: 1069–77.

8 Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics, and validity. *Arch Gen Psychiatry* 1981; **38**: 381–9.

9 Lecrubier Y, Sheehan DV, Weiller E, Amorim P, Bonora I, Harnett-Sheehan K, et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *Eur Psychiatry* 1997; **12**: 224–31.

10 Sheehan DV, Lecrubier Y, Harnett-Sheehan K, Janavs J, Weiller E, Keskiner A, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *Eur Psychiatry* 1997; **12**: 232–41.

11 Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychol Med* 2001; **31**: 1001–13.

12 Rice DB, Kloda LA, Shrier I, Thombs BD. Reporting completeness and transparency of meta-analyses of depression screening tool accuracy: a comparison of meta-analyses published before and after the PRISMA statement. *J Psychosom Res* 2016; **87**: 57–69.

13 Anthony JC, Folstein M, Romanoski AJ, Von Korff MR, Nestadt GR, Chahal R, et al. Comparison of the lay Diagnostic Interview Schedule and a standardized psychiatric diagnosis: experience in eastern Baltimore. *Arch Gen Psychiatry* 1985; **42**: 667–75.

14 Booth BM, Kirchner JA, Hamiltonc G, Harrell R, Smith GR. Diagnosing depression in the medically ill: validity of a lay-administered structured diagnostic interview. *J Psychiatr Res* 1998; **32**: 353–60.

15 Hesselbrock V, Stabenau J, Hesselbrock M, Mirkin P, Meyer R. A comparison of two interview schedules: the Schedule for Affective Disorders and Schizophrenia-Lifetime and the National Institute for Mental Health Diagnostic Interview Schedule. *Arch Gen Psychiatry* 1982; **39**: 674–7.

16 Jordanova V, Wickramesinghe C, Gerada C, Prince M. Validation of two survey diagnostic interviews among primary care attendees: a comparison of CIS-R and CIDI with SCAN ICD-10 diagnostic categories. *Psychol Med* 2004; **34**: 1013–24.

17 Thombs BD, Benedetti A, Kloda LA, Levis B, Nicolau I, Cuijpers P, 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

https://doi.org/10.1192/bjp.2018.54 Published online by Cambridge University Press

detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Syst Rev* 2014; **27**(3): 124.

18 McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol* 2016; **75**:40–6.

19 Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R, et al. Should meta-analysts search Embase in addition to Medline? *J Clin Epidemiol* 2003; **56**: 943–55.

20 Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16**: 606–13.

21 Thombs BD, Arthurs E, El-Baalbaki G, Meijer A, Ziegelstein RC, Steele RJ. Risk of bias from inclusion of patients who already have diagnosis of or are undergoing treatment for depression in diagnostic accuracy studies of screening tools for depression: systematic review. *BMJ* 2011; **343**: d4825.

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

23 Diagnostic and statistical manual of mental disorders: DSM-IV 4th ed. Washington, DC: American Psychiatric Association 1994.

24 Diagnostic and statistical manual of mental disorders: DSM-IV 4th ed, text revised. Washington, DC: American Psychiatric Association 2000.

25 The ICD-10 Classifications of Mental and Behavioural Disorder: Clinical Descriptions and Diagnostic Guidelines Geneva: World Health Organization 1992.

26 United Nations. *International Human Development Indicators*. UN, 2016 (http://hdr.undp.org/en/countries).

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

28 Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *J Gen Intern Med* 2006; **21**: 547–52.

29 Martin A, Rief W, Klaiberg A, Braehler E. Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. *Gen Hosp Psychiatry* 2006; **28**: 717.

30 Adewuya AO, Ola BA, Afolabi OO. Validity of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *J Affect Disord* 2006; **96**: 89–93.

31 Milette K, Hudson M, Baron M, Thombs BD. Comparison of the PHQ-9 and CES-D depression scales in systemic sclerosis: internal consistency reliability, convergent validity and clinical correlates. *Rheumatology* 2010; **49**: 789–96.

32 Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Gen Hosp Psychiatry* 2015; **37**: 567–76.

33 Thombs BD, Benedetti A, Kloda LA, Levis B, Riehm KE, Azar M, 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.

34 Thombs BD, Benedetti A, Kloda LA, Levis B, Azar M, Riehm KE, et al. Diagnostic accuracy of the Depression subscale of the Hospital Anxiety and Depression Scale (HADS-D) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open* 2016; **6**: e011913.

35 Arthurs E, Steele RJ, Hudson M, Baron M, Thombs BD, Canadian Scleroderma Research Group. Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning. *PLoS One* 2012; **7**: e52028.

36 Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *J Gen Intern Med* 2006; **21**: 547–52.

37 Chung H, Kim J, Askew RL, Jones SMW, Cook KF, Amtmann D. Assessing measurement invariance of three depression scales between neurologic samples and community samples. *Qual Life Res* 2015; **24**: 1829–34.

38 Cook KF, Kallen MA, Bombardier C, Bamer AM, Choi SW, Kim J, et al. Do measures of depressive symptoms function differently in people with spinal cord injury versus primary care patients: the CES-D, PHQ-9, and PROMIS-D. *Qual Lif Res* 2017; **26**: 139–48.

39 Leavens A, Patten SB, Hudson M, Baron M, Thombs BD, Canadian Scleroderma Research Group. Influence of somatic symptoms on Patient Health Questionnaire-9 depression scores among patients with systemic sclerosis compared to a healthy general population sample. *Arthritis Care Res* 2012; **64**: 1195–201.

EXTRA CONTENT ONLINE

https://doi.org/10.1192/bjp.2018.54 Published online by Cambridge University Press