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


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.

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.

Copyright and usage

© The Royal College of Psychiatrists 2018.
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. 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. Examples include the Structured Clinical Interview for DSM (SCID) and Schedules for Clinical Assessment in Neuropsychiatry (SCAN). In contrast, fully structured interviews typically involve fully scripted, standardised questions that are read verbatim, without additional probes. They are designed to be less subjective and provide greater standardisation, but with less flexibility and without incorporating clinical judgement. Examples include the Composite International Diagnostic Interview (CIDI) and the Diagnostic Interview Schedule (DIS). 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 administered in a fraction of the time at the cost of being over-inclusive and generating a higher rate of false-positive diagnoses.

Although fully structured interviews are sometimes referred to as imperfect reference standards compared with semi-structured interviews, both are considered appropriate reference standards for major depression classification in research. Consistent with this, existing meta-analyses on depression screening tool accuracy have treated both interview types as equivalent reference standards. 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). 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.

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. 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). 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

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. We limited our search to those databases based on research showing that adding other databases when the Medline search is highly sensitive does not identify additional eligible studies. The search was limited to the year 2000 onwards because the PHQ-9 was published in 2001. 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.

Method
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. 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. 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.

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. 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, 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. 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-run 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,
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, 4,732 participants), and the CIDI (11 studies, 6,271 participants) and MINI (14 studies, 2,756 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.

### Table 1  Participant data by diagnostic interview

<table>
<thead>
<tr>
<th>Diagnostic interview</th>
<th>Studies, N</th>
<th>Participants, N</th>
<th>Major depression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Semi-structured</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCID</td>
<td>26</td>
<td>4,732</td>
<td>785</td>
</tr>
<tr>
<td>SCAN</td>
<td>2</td>
<td>1,891</td>
<td>130</td>
</tr>
<tr>
<td>DISH</td>
<td>1</td>
<td>100</td>
<td>9</td>
</tr>
<tr>
<td>Fully structured</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIDI</td>
<td>11</td>
<td>6,271</td>
<td>554</td>
</tr>
<tr>
<td>DIS</td>
<td>1</td>
<td>1,006</td>
<td>221</td>
</tr>
<tr>
<td>MINI</td>
<td>14</td>
<td>2,756</td>
<td>524</td>
</tr>
<tr>
<td>CIS-R</td>
<td>2</td>
<td>402</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td>57</td>
<td>17,158</td>
<td>2,287</td>
</tr>
</tbody>
</table>

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.
The proportion of participants classified as having major depression 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, results for interaction models were similar.

There were two main findings. First, among fully structured interviews, the adjusted odds of being classified as having 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% vs. 20.8% in the medium-range group, and 54.2% vs. 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.

### Discussion

There were two main findings. First, among fully structured interviews, the adjusted odds being classified as having 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).
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.12

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; Nazanin Saadat, BSc, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Alexander W. Levis, MSc, Marlinee Azar, BSc, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Andrea Benedetti, PhD, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada; Nazanin Saadat, BSc, Lady Davis Institute for Medical Research, Jewish General Hospital, 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, Marlinee Azar, BSc, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada;
Levis et al

References


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: c4825.


