Isobel M Cameron, Amanda Cardy, John R Crawford, Schalk W du Toit, Steven Hay, Kenneth Lawton, Kenneth Mitchell, Sumit Sharma, Shilpa Shivaprasad, Sally Winning and Ian C Reid

Measuring depression severity in general practice:

discriminatory performance of the PHQ-9, HADS-D, and BDI-II

Abstract

Background
The UK Quality and Outcomes Framework (QOF) rewards practices for measuring symptom severity in patients with depression, but the endorsed scales have not been comprehensively validated for this purpose.

Aim
To assess the discriminatory performance of the QOF depression severity measures.

Design and setting
Psychometric assessment in nine Scottish general practices.

Method
Adult primary care patients diagnosed with depression were invited to participate. The HADS-D, PHQ-9, and BDI-II were assessed against the HRSD-17 interview. Discriminatory performance was determined relative to the HRSD-17 cut-offs for symptoms of at least moderate severity, as per criteria set by the American Psychiatric Association (APA) and NICE. Receiver operating characteristic curves were plotted and area under the curve (AUC), sensitivity, specificity, and likelihood ratios (LRs) calculated.

Results
A total of 267 were recruited per protocol, mean age = 49.8 years (standard deviation [SD] = 14.1), 70% female, mean HRSD-17=12.6 (SD = 7.62, range = 0–34). For APA criteria, AUCs were: HADS-D = 0.84; PHQ-9 = 0.90; and BDI-II = 0.86. Optimal sensitivity and specificity were reached where HADS-D ≥9 (74%, 76%); PHQ-9 ≥12 (77%, 79%), and BDI-II ≥23 (74%, 75%). For NICE criteria: HADS-D AUC = 0.89; PHQ-9 AUC = 0.93; and BDI-II AUC = 0.90. Optimal sensitivity and specificity were reached where HADS-D ≥10 (82%, 75%), PHQ-9 ≥15 (89%, 83%), and BDI-II ≥28 (83%, 80%). LRs did not provide evidence of sufficient accuracy for clinical use.

Conclusion
As selecting treatment according to depression severity is informed by an evidence base derived from trials using HRSD-17, and none of the measures tested aligned adequately with that tool, they are inappropriate for use.

Keywords
depression; primary care; sensitivity; severity; specificity.

INTRODUCTION

UK GPs are funded through the Quality and Outcomes Framework (QOF) for assessing the severity of symptoms of depression[1] with one of the following:

- the Patient Health Questionnaire 9 (PHQ-9);[2]
- the Hospital Anxiety and Depression Scale (HADS), Depression Subscale (HADS-D);[3] or
- the Beck Depression Inventory, Second Edition (BDI-II).[4]

This initiative accords with guidelines in which different treatment options are advocated according to severity.[5] Further, it has been suggested that using such tools is more reliable than relying on GPs' perceptions alone.[6] Unfortunately, the widely quoted evidence base supporting differential treatment selection on the basis of severity[5,7] is founded on studies using the clinician-rated, 17-item Hamilton Rating Scale for Depression (HRSD-17),[8] not the scales recommended by the QOF. It is, therefore, important to determine the extent to which the QOF-endorsed scales agree with the measure used to generate the original evidence that supports using severity assessments to plan treatment, such as deciding whether or not to prescribe an antidepressant drug.

The PHQ-9 and HADS-D differ significantly in categorising depressive severity in UK,[9] Swedish,[10] and Australian[11] studies. It comes as no surprise, therefore, to discover that practices using the PHQ-9 record a different prevalence of moderate and severe symptoms of depression than practices that use HADS-D.[12] It is not known, however, whether either of the scales can categorise patients appropriately; it is known only that the scales are not equivalent and, as such, cannot both be valid in this regard.

The severity cut-offs of the PHQ-9 were pragmatically derived to be 'simple for clinicians to remember and apply', following which they were verified in terms of their relative associations with variables expected to increase with severity.[2] As such, the scores were never derived in reference to a standard measure of severity. The validity of the PHQ-9 has tended to be considered in terms of its diagnostic accuracy, rather than the differentiation of severity.[13–16]

The severity cut-offs of the BDI-II were empirically derived in reference to the structured clinical interview for the Diagnostic and Statistical Manual of Mental Disorders, third edition, revised (DSM-III-R) (SCID)[17] categories of mild, moderate, and severe.[18] However this study by Beck *et al*

IM Cameron, MA, PhD, lecturer; **IC Reid,** BMed Biol, PhD, MRCPsych, professor of psychiatry, Applied Health Sciences (Mental Health); **A Cardy,** BSc, MSc, research fellow; **K Lawton,** FRCGP, senior clinical lecturer, Centre of Academic Primary Care; **JR Crawford,** BSc, MSc, PhD, professor of psychology, School of Psychology, University of Aberdeen, Aberdeen. **SW du Toit,** MRCPsych, specialist registrar in psychiatry; **S Hay,** MRCPsych, staff grade psychiatrist; **K Mitchell,** MRCPsych, consultant psychiatrist; **S Sharma,** MD, MRCPsych, specialist registrar in psychiatry; **S Shivaprasad,** MRCPsych, specialist registrar in psychiatry; **S Winning,** MBChB, staff grade psychiatrist, Royal Cornhill Hospital, Aberdeen.

Address for correspondence
Isobel M Cameron, lecturer, Applied Health Sciences (Mental Health), University of Aberdeen, Royal Cornhill Hospital, Aberdeen, AB25 2ZH.
E-mail: i.m.cameron@abdn.ac.uk

Submitted: 2 March 2011; **Editor's response:** 28 March 2011; **final acceptance:** 17 May 2011.

DOI: 10.3399/bjgp11X583209.

How this fits in

The UK Quality and Outcomes Framework (QOF) rewards practices for measuring the severity of symptoms of depression. Although demonstrably robust as case-finding tools, the QOF-endorsed scales have not been comprehensively validated for severity measurement. The severity categories of the QOF-endorsed depression measures do not adequately align with the severity categories of the 17-item Hamilton Rating Scale for Depression (HRSD-17) interview. Optimal cut-offs yielded likelihood ratios that were not adequate for clinical practice. As treatment is determined according to an evidence base derived from trials using HRSD-17, these QOF-endorsed scales are invalid.

was conducted on a sample sought entirely from a primary care site at the University of Pennsylvania, so the findings may not generalise effectively to a primary care population in the UK.

The HADS-D severity cut-offs were derived from a sample recruited from general medical outpatient clinics.[3] In terms of severity assessment, the scale was assessed against an unreferenced five-point scale administered by the researchers. As with the PHQ-9, examinations of the HADS-D have not tended to validate accuracy regarding the measurement of the severity of symptoms of depression.[16,19–21]

At present, there is an absence of objective psychometric comparison between the endorsed measures that would enable GPs to choose or reject a severity assessment tool on the basis of clinical relevance or validity. The aim of this study was to assess the discriminatory performance of the QOF-endorsed measures in categorising the severity of symptoms of depression against the HRSD-17 in primary care patients with a GP-generated diagnosis of depression.

METHOD

Participants

Patients were recruited from nine general practices in Grampian, Scotland, which were selected to yield participants with a mixed socioeconomic and urban/rural demographic. In order to be included in the study, participants had to be primary care patients aged ≥16 years old with a new or existing GP diagnosis of depression. This reflects current QOF arrangements, whereby GPs use their clinical judgement to identify depression prior to assessing severity. Individuals without the necessary spoken or written language skills to complete the questionnaires and interview were excluded.

Depression severity measures

In addition to recording demographic factors, self-complete depression severity questionnaires were applied.

HADS-D. The HADS consists of 14 items, each rated 0–3 according to the severity of difficulties experienced. Subscales for depression (HADS-D) and anxiety can be totalled, with a possible range for each of 0–21. The scores can then be interpreted as follows: mild (8–10), moderate (11–14), or severe (≥15) difficulties.

PHQ-9. The PHQ-9 consists of nine questions, rated 0–3 according to the increased frequency of difficulty experienced in each area covered. Scores, with a possible range of 0–27, are summed and can then be interpreted as follows: no depression (0), minimal (1–5), mild (6–9), moderate (10–14), moderately severe (15–19), or severe (≥20) depression.

BDI-II. The BDI-II is a depression severity questionnaire consisting of 21 items, each rated 0–3 according to severity of difficulties experienced. Scores, with a possible range of 0–63, are summed; depression can then be interpreted as minimal (0–13), mild (14–19), moderate (20–28), or severe (≥29).

Reference standard. The HRSD-17 was devised for use with patients with an existing diagnosis of depression and is intended to quantify the results of an interview assessing symptom severity. Consisting of 17 items, it was used as the 'standard' for depression severity measurement due to its wide use in intervention studies that have taken depression severity into account.[5] The American Psychiatric Association (APA)[22] and the National Institute for Health and Clinical Excellence (NICE)[23] have published different severity bandings for the HRSD-17. For APA these are: none (0–7), mild (8–13), moderate (14–18), severe (19–22), and very severe (≥23); for NICE these are: none (0–7), sub-threshold (8–13), mild (14–18), moderate (19–22), severe (≥23).

NICE's *Clinical Guideline 91* offers no evidence to support the new cut-offs they propose but states that the change was necessary, with the updated guideline now including sub-threshold depression due to changes to the diagnosis based on the fourth edition of the Diagnostic and

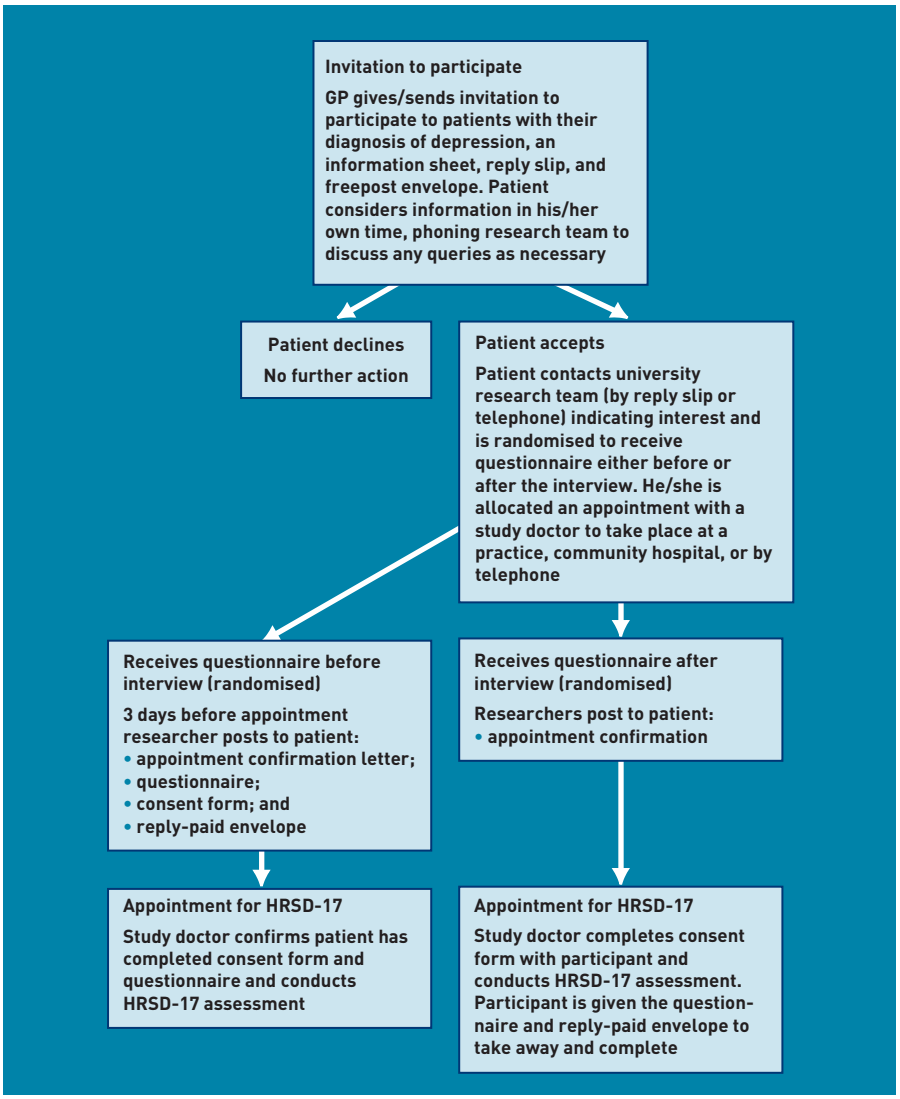


Figure 1. The recruitment process.

Table 1. Results of measurement tool assessments

Measurement tool		
Severity scores	**Mean (SD)**	**Kolmogarov-Smirnov *P*-value**
HRSD-17	12.6 (7.62)	0.22
HADS-D	8.16 (4.78)	0.32
PHQ-9	11.5 (7.29)	0.06
BDI-II	23.3 (13.0)	0.18
Parametric tests	**Minimum value**	**Maximum value**
HRSD-17	0	34
HADS-D	0	20
PHQ-9	0	27
BDI-II	0	61
Completing questionnaire before interview versus after interview	**Mean difference**	**95% CI**
HRSD-17	0.48	−1.20 to 2.52
HADS-D	0.28	−0.93 to 1.48
PHQ-9	0.59	−1.30 to 2.47
BDI-II	1.77	−2.85 to 4.11

BDI-II = Beck Depression Inventory, Second Edition. HADS-D = Hospital Anxiety and Depression Scale – Depression Subscale. HRSD-17 = 17-item Hamilton Rating Scale for Depression interview. PHQ-9 = Patient Health Questionnaire 9. SD = standard deviation.

Statistical Manual of Mental Disorders.[23]

Procedure

The HADS-D, PHQ-9, and BDI-II were assessed against both the APA and NICE severity criteria. The GRID-Hamilton Depression Rating Scale (GRID-HAMD) schedule was used as administration and scoring are standardised in this method, which helps maximise inter-rater reliability without altering the measure's original intent.

The recruitment process is outlined in Figure 1. After giving informed consent, participants completed the QOF-endorsed questionnaires and took part in a clinical interview that was conducted by one of six psychiatrists in either a doctor's surgery or a community hospital. Prospective participants were given the choice to take part in a telephone interview if preferred. In such cases the GRID-HAMD was still used, but the psychiatrists were given published instructions for the validated telephone version of HRSD-17.[24]

Six psychiatrists conducted the interviews; they were blind to the questionnaire responses. Order of administration of the interview and questionnaires was randomised, stratified by practice, to reduce any confounding by order of completion. For those randomised to receive the questionnaires first, participants were encouraged to complete them on the same day before the interview, or the day before the interview. For those randomised to receive the questionnaires after the interview, participants were encouraged to complete them on the same day after the interview, or the following day. The questionnaires were put into a booklet; their order within the booklet was systematically varied in blocks of 20.

Inter-rater reliability of HRSD-17

A random sub-sample of participants consented to have their assessment audio-recorded. Over a 12-month period commencing May 2008, five psychiatrists made a recording on five occasions and one on three occasions. After this, each psychiatrist was given five recordings, one from each of the other raters, to listen to and make their own ratings (blind to the original ratings). The psychiatrists were instructed not to attempt to rate from the recordings the two items in the HRSD-17 requiring visual observation (retardation and agitation).

Statistical analyses

Data were assessed for normality using the

Figure 2a. Percentage convergence of severity categories (HRSD-17 American Psychiatric Association Handbook of Psychiatric Measures cut-offs versus HADS-D, PHQ-9 and BDI-II).

Figure 2b. Percentage convergence of severity categories (HRSD-17 National Institute for Clinical Excellence (NICE CG91) cut-offs versus HADS-D, PHQ-9 and BDI-II).

Figure 2a.

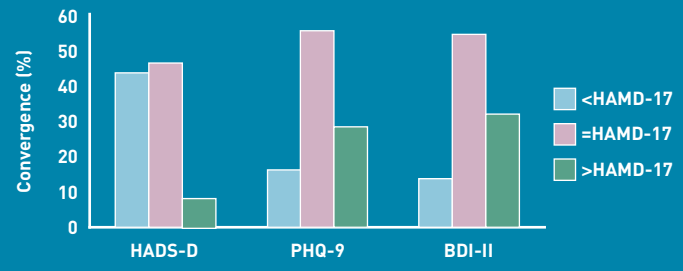
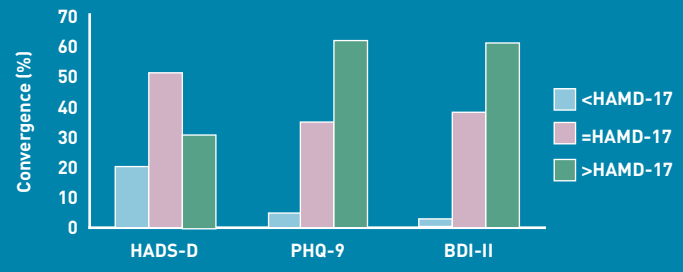


Figure 2b.



BDI-II = Beck Depression Inventory, Second Edition. HADS-D = Hospital and Anxiety Depression Scale – Depression Subscale. HRSD-17 = 17-item Hamilton Rating Scale for Depression interview. PHQ-9 = Patient Health Questionnaire 9.

one-sample Kolmogorov–Smirnov test of goodness of fit. Any confounding related to the order of administration was assessed with *t*-tests. The HADS-D, PHQ-9, and BDI-II were assessed for convergence with the HRSD-17. Data were only included where the HRSD-17 and the self-complete measures were done within 3 days of one another and where data were complete. It was considered that, with a maximum time difference of 3 days, there would be sufficient overlap in reference points.

Convergent validity was examined using Pearson correlation coefficients of each QOF-endorsed scale with the HRSD-17. Convergence of the scales' severity bandings was also investigated using the Wilcoxon signed-rank test for related samples. The established HADS-D, PHQ-9, and BDI-II severity cut-off bands for at least moderately severe symptoms of depression were assessed relative to the moderate cut-off of the APA and the NICE criteria of the HRSD-17 using receiver operating characteristic (ROC) curves.[25] ROC curve analysis was then used to determine each

Table 2. Area under the ROC curve of HADS-D, PHQ-9, and BDI-II depression severity measures, relative to HRSD-17

Questionnaire	Area under ROC curve	95% CI
APA criteria HRSD-17 ≥14 (at least moderate)		
HADS-D	0.84	0.79 to 0.89
PHQ-9	0.90	0.86 to 0.94
BDI-II	0.86	0.81 to 0.91
NICE criteria HRSD-17 ≥19 (at least moderate)		
HADS-D	0.89	0.84 to 0.93
PHQ-9	0.93	0.90 to 0.97
BDI-II	0.90	0.84 to 0.95

APA = American Psychiatric Association. BDI-II = Beck Depression Inventory, Second Edition. HADS-D = Hospital Anxiety and Depression Scale – Depression Subscale. HRSD-17 = 17-item Hamilton Rating Scale for Depression interview. NICE = National Institute for Health and Clinical Excellence. PHQ-9 = Patient Health Questionnaire 9. ROC = receiver operating characteristics.

questionnaire's optimal cut-off for at least moderately severe symptoms of depression. This cut-off was considered the most relevant focus as it is the point at which guidelines advocate the use of antidepressant medication.[5,7]

Sensitivity and specificity of the scales at detecting symptoms of at least moderate severity were calculated with accompanying confidence intervals (CIs),[26] as were positive and negative predictive values (PPVs and NPVs, respectively) and likelihood ratios (LRs).[27–28] LRs of >10 and <0.1 were considered to provide sufficient evidence for use in clinical practice.[27] Analyses were conducted using SPSS (version 17).

To assess inter-rater reliability on the HRSD-17, the 15 items rated by both a primary and secondary rater were summed. Paired *t*-tests assessed consistent differences; intraclass correlation (ICC) was calculated to express the between-pair variance as a proportion of the total variance.

RESULTS

Between October 2007 and April 2009, 1134 patients were invited to participate in the study; 286 (25%) did participate. Of those, 137 were randomised to complete the questionnaires before a HRSD-17 interview and 131 were randomised to complete the questionnaires after the interview. Eighteen participants were not randomised and a further participant was interviewed with the non-GRID version of the HRSD–17 and so these data were excluded. The mean age of responders was 49.8 years (standard deviation [SD] 14.1), 184 (70%) were female, and 244 (99%) were of white ethnicity.

Some 222 (84%) interviews were conducted face to face and 41 (16%) by telephone. Four of the participants who returned their questionnaires did not attend the interview. Assessment results are shown in Table 1. None of the scale scores differed significantly from normal distribution and no adverse events were reported.

Analysis was restricted to the 233 participants who fully completed the questionnaires and participated in the HRSD-17 interview within 3 days.

Convergent validity

The questionnaires correlated moderately with HRSD-17: HADS-D and HRSD-17, $r = 0.68$; PHQ-9 and HRSD-17, $r = 0.79$; BDI-II and HRSD-17, $r = 0.75$.

Convergence of severity banding

When compared with APA's HRSD-17 depression severity cut-offs, the HADS-D tended to categorise participants in a milder category ($P<0.001$), whereas the PHQ-9 and BDI-II tended to categorise participants in a more severe category ($P<0.01$ and $P<0.001$ respectively). Compared with NICE's HRSD-17 depression severity cut-offs, the HADS-D did not converge ($P<0.01$). The tendency for the PHQ-9 and BDI-II to categorise participants in a more severe category than the HRSD-17 was further pronounced when using the NICE cut-offs ($P<0.001$ for PHQ-9 and BDI-II). Figures 2a and 2b demonstrate the lack of alignment with the APA and NICE cut-offs respectively, (that is, only where the label indicates '= HRSD–17' was their alignment in the categorisation of the depression severity).

Figure 3a. ROC curve of HADS-D, PHQ-9, and BDI-II depression severity measures against HRSD-17 ≥14 (APA criteria of at least moderate severity).

Figure 3b. Receiver operating characteristics (ROC) curve of HADS-D, PHQ-9 and BDI-II depression severity measures against HRSD-17≥19 (NICE criteria of at least moderate severity).

Figure 3a.

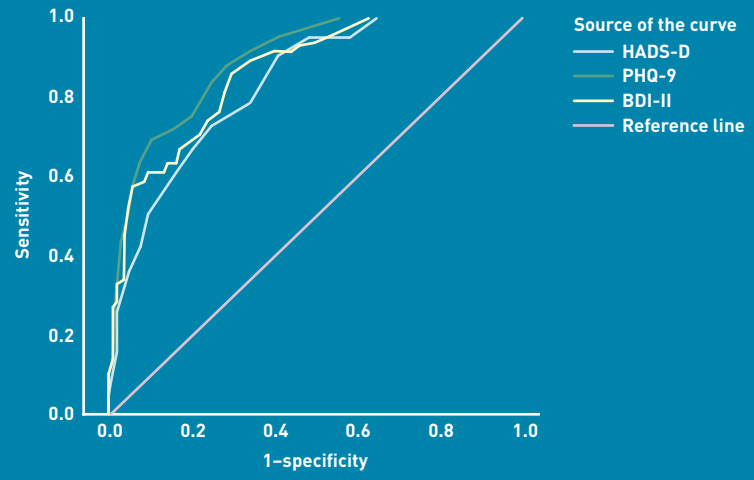
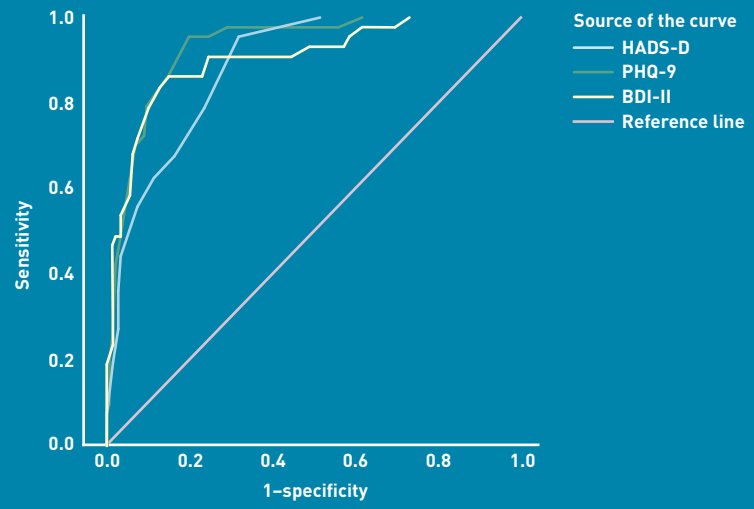


Figure 3b.



BDI-II = Beck Depression Inventory, Second Editionl. HADS-D = Hospital and Anxiety Depression Scale – Depression Subscale. HRSD-17 = 17-item Hamilton Rating Scale for Depression interview. PHQ-9 = Patient Health Questionnaire 9. ROC = receiver operating characteristic.

Empirically derived cut-offs for symptoms of depression of moderate severity

The area under the ROC curve of each self-complete depression measure for the APA and NICE cut-offs for symptoms of depression of at least moderate severity is shown in Table 2. The ROC curves are presented in Figures 3a and 3b. All three measures were shown to perform significantly better than chance at discriminating between those above and below the APA ($P<0.001$) and NICE ($P<0.001$) thresholds for moderate symptoms of depression.

For each self-complete depression severity measure, Table 3 shows the discriminatory properties at the moderate cut-off defined by the scales' developers, as well as the optimal cut-off, which was informed by the ROC analysis relative to the moderate HRSD-17 criteria of the APA and NICE. Respective PPVs and NPVs are also presented. Best sensitivity and specificity are found using the NICE criteria and cut-offs derived from the ROC curve analysis. However, the positive LR for all the measures is <10; most of the negative LRs are >0.1, with the exception of that for PHQ-9 ≥10 cut off against the NICE HRSD-17 criteria), indicating the scales are not sufficiently robust to rule in or out the presence of symptoms of depression of at least moderate severity.[27]

Inter-rater reliability

The difference between the ratings of the original assessment (primary value) and the ratings made from the audio recordings (secondary value) of the 15-item summed Hamilton scale — which excludes the two items that required visual observation — was normally distributed. The test statistic for the paired *t*-test was 0.09 (degrees of freedom = 29), $P = 0.93$, indicating that there was no evidence of any systematic difference between the primary and secondary ratings. The ICC coefficient for the summed scale was 0.95 (95% CI = 0.90 to 0.98), demonstrating acceptable agreement.

DISCUSSION

Summary

The HADS-D, PHQ-9, and BDI-II correlated moderately with the HRSD-17. All of the scales differed significantly in how they categorised depression severity relative to the HRSD-17 cut-offs, as determined by both APA and NICE criteria. Efforts to derive optimal cut-offs did not yield values with

Table 3. HAD-D, PHQ-9, and BDI-II depression severity measures: discriminatory performance of detecting at least moderate depression severity relative to HRSD-17

	% Sensitivity (95% CI)	% Specificity (95% CI)	%PPV[a] (95% CI)	%NPV[b] (95% CI)	LR+ve[c] (95% CI)	LR-ve[d] (95% CI)
Moderate severity (defined by scales' developer) by HRSD-17 ≥14 (moderate) cut-off (APA criteria)						
HADS-D ≥11	54 (44 to 63)	89 (84 to 95)	79 (69 to 89)	71 (64 to 78)	4.93 (2.91 to 8.36)	0.52 (0.42 to 0.65)
PHQ-9 ≥10	89 (83 to 95)	70 (62 to 78)	69 (61 to 77)	90 (84 to 96)	2.99 (2.26 to 3.95)	0.15 (0.09 to 0.28)
BDI-II ≥20	86 (79 to 93)	70 (61 to 79)	71 (62 to 79)	86 (78 to 93)	2.87 (2.13 to 3.86)	0.20 (0.12 to 0.34)
Optimal cut-off for moderate severity (derived from ROC curve) by HRSD-17 ≥14 cut-off (APA criteria)						
HADS-D ≥9	74 (65 to 82)	76 (69 to 83)	70 (61 to 79)	79 (72 to 86)	3.07 (2.21 to 4.26)	0.35 (0.25 to 0.49)
PHQ-9 ≥12	77 (69 to 86)	79 (72 to 86)	73 (64 to 82)	82 (76 to 89)	3.68 (2.57 to 5.27)	0.29 (0.20 to 0.43)
BDI-II ≥23	74 (65 to 83)	75 (67 to 83)	72 (63 to 81)	78 (70 to 85)	3.02 (2.13 to 4.28)	0.34 (0.24 to 0.49)
Moderate severity (defined by scales' developers) by HRSD-17 ≥19 cut- off (NICE criteria)						
HADS-D ≥11	71 (59 to 84)	82 (77 to 88)	52 (40 to 64)	91 (87 to 96)	4.00 (2.79 to 5.73)	0.35 (0.22 to 0.54)
PHQ-9 ≥10	98 (94 to 99)	57 (49 to 64)	39 (30 to 47)	99 (97 to 99)	2.27 (1.90 to 2.71)	0.04 (0.01 to 0.26)
BDI-II ≥20	91 (83 to 99)	55 (47 to 63)	37 (28 to 46)	96 (91 to 99)	2.02 (1.66 to 2.45)	0.16 (0.06 to 0.41)
Optimal cut-off for moderate severity (derived from ROC curve) by HRSD-17 ≥19 cut- off (NICE criteria)						
HADS-D ≥10	82 (71 to 92)	75 (69 to 81)	47 (36 to 58)	94 (90 to 98)	3.25 (2.44 to 4.32)	0.25 (0.14 to 0.45)
PHQ-9 ≥15	89 (81 to 98)	83 (78 to 89)	60 (49 to 71)	97 (94 to 99)	5.39 (3.79 to 7.67)	0.13 (0.06 to 0.29)
BDI-II ≥28	83 (72 to 94)	80 (73 to 86)	54 (43 to 66)	94 (90 to 98)	4.05 (2.90 to 5.67)	0.22 (0.12 to 0.41)

[a]Positive predictive value. [b]Negative predictive value. [c]Likelihood ratio for a positive result. [d]Likelihood ratio for a negative result. APA = American Psychiatric Association. BDI-II = Beck Depression Inventory, Second Edition. HAD-D = Hospital and Anxiety Depression Scale – Depression Subscale. HRSD-17 = 17-item Hamilton Rating Scale for Depression interview. NICE = National Institute for Health and Clinical Excellence. PHQ-9 = Patient Health Questionnaire 9. ROC = receiver operating characteristics.

Funding

Funding was provided by NHS Quality Improvement Scotland (QIS), which had no further role in study design; the collection, analysis and interpretation of data; the writing of the report; or in the decision to submit the paper for publication.

Ethical approval

This research was conducted with the approval of the North of Scotland Research Ethics Committee (reference number: 07/S0802/40).

Provenance

Freely submitted; externally peer reviewed.

Competing interests

Isobel M Cameron, Ian C Reid, John R Crawford and Kenneth Lawton were grantholders of funding from NHS Quality Improvement Scotland, which covered IMC's salary for work carried out on this paper.

Acknowledgements

We would like to thank the patients and staff of the nine practices in Grampian who kindly participated in this study and, in particular: Dr Martin McCrone, for facilitating the practice recruitment; Ms Kirsty Sykes and Ms Laura McQueen, for preparing research materials and assisting with data checking; Professor Richard Morriss, for drawing our attention to the GRID version of HRSD-17; Professor Ian Anderson, for advice regarding severity cut-offs of HRSD-17; the Scottish Primary Care Research Network, for support with recruitment; and NHS Quality Improvement Scotland, for funding.

Discuss this article

Contribute and read comments about this article on the Discussion Forum: **http://www.rcgp.org.uk/bjgp-discuss**

LRs that were adequate to inform clinical practice.[27]

Strengths and limitations

To the authors' knowledge, this is the first study to assess the three QOF-endorsed depression severity measurement tools in terms of their ability to measure the severity of symptoms of depression.

The HRSD-17 interview is not diagnostic and some have argued that the QOF-endorsed measures ought to be assessed against an interview such as the SCID.[12] However, the aim of using these scales in UK treatment of depression is not as case-finding tools, but as assessors of severity of depression that has already been diagnosed by a GP, in order to identify appropriate evidence-based treatment options. Although the SCID generates diagnostic categories with a crude severity dimension, the evidence base relies on the HRSD-17 in clinical trials, comparing variations in severity with outcome.[29,30] The fact that different cut-offs exist for HRSD-17 (APA and NICE) highlights ongoing uncertainty in this area of rating scales and their validity. Furthermore, NICE guidance states that its new classification should not be taken as clear cut-offs;[23] this raises the question 'What should they be taken as?'.

The HRSD-17 scores in this sample were broad in their range, but only a quarter of patients who were invited to participate did so. The priority for this sample, sought for psychometric assessment, was that, as well as covering a broad socioeconomic and urban/rural demographic, it had a distribution of patients with symptoms of depression of differing severity. The authors are satisfied that this study's sample included a wide range of primary care patients from those in remission to those with very severe symptoms. The sample was similar, in terms of sex, with patients consulting GPs for depression throughout Scotland in 2007/2008.[31]

The sample does not represent the ethnic diversity of some parts of the UK and this may impact on the generalisability of the findings.

Comparisons with existing literature

Several studies have raised concerns regarding the validity of the HADS-D and PHQ-9 in terms of their assessment of depression severity.[9–11] The current study is able to conclude that, both the HADS-D and the PHQ-9 categorise the severity of depression inaccurately when compared with HRSD-17. The HADS-D tends to place participants in a milder category of depression than the HRSD-17 and the PHQ-9 tends to place individuals in a more severe category. This latter tendency is also true of the BDI-II. These findings suggest the scales are not all measuring the same aspects of depression.

The assessment of the PHQ-9 found the measure to have better psychometric properties, in terms of severity assessment, when compared with another European study.[32] The current study had a closer time restriction between administrations of the HRSD-17 and completion of the questionnaires and should, therefore, more accurately assess concurrent measurement of mood.

The rationale for introducing depression severity measures into the QOF was partly informed by a study in which it was observed that GPs were inaccurate in their categorisation of depression severity.[6] However, the standard by which GPs were assessed in this study was the HADS-D. This study has demonstrated that the HADS-D is inaccurate in categorising the severity of symptoms of depression and it is, therefore, questionable whether the clinical judgement of GPs in assessing depression severity in patients whom they suspect to be depressed, is any better or worse. Given the present findings, GPs' intuition regarding the benefit of history-taking over applying such measures[33–35] may be well founded.

It has been emphasised by NICE that the interpretation of scores alone should not be relied upon when assessing an individual with possible depression, but that other factors — including functional impairment, history, family history, and presence of other comorbid conditions — should also be considered.[5] Nonetheless, to be of value, the scales must meet acceptable standards of accuracy.

Implications for research and practice

Although this study proposes new cut-offs for possible use in assessing depression severity, the likelihood ratios indicate they are still insufficiently precise to recommend for clinical use. As such, the time and effort expended in recording the use of the scales within the QOF mechanism seems to be something of a poor investment. More work is required to determine the rational selection of treatment strategies for depression in primary care — there is a danger that the setting of the QOF standard examined here has lent an unjustified veneer of confidence to the management of the condition, obscuring the paucity of basic research.

REFERENCES

1. NHS Employers and the General Practitioners' Committee. *Quality and Outcome Frameworks: guidance for GMS contract 2009/10.* 2009: 1–162.
2. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16(9):** 606–613.
3. Zigmond AS, Snaith P. The Hospital Anxiety and Depression Scale (HAD). *Acta Psychiatrica Scandinavica* 1983; **67(6):** 361–370.
4. Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of Beck Depression Inventories-IA and -II in psychiatric outpatients. *J Pers Assess* 1996; **67(3):** 588–597.
5. NICE. *Depression: the treatment and management in adults (update).* London: NHS National Institute for Clinical Excellence, 2009; CG90: 1–585.
6. Kendrick T, King F, Albertella L, Smith P. GP treatment decisions for patients with depression. *Br J Gen Pract* 2005; **55(513):** 280–286.
7. Anderson IM, Baldwin RC, Cowen PJ, *et al.* Evidence-based guidelines for treating depressive disorders with antidepressants: a revision of the 2000 British Association for Psychopharmacology guidelines. *J Psychopharmacol* 2008; **22(4):** 343–396.
8. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry* 1960; **23:** 56–62.
9. Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of the PHQ-9 and HADS for measuring depression severity in primary care. *Br J Gen Pract* 2008; **58(546):** 32–36.
10. Hansson M, Chotai J, Nordstom A, Bodlund O. Comparison of two self-rating scales to detect depression: HADS and PHQ-9. *Br J Gen Pract* 2009; **59(566)**: e283–288.
11. Reddy P, Dunbar J, Ford D, Philpot B. Identification of depression in diabetes: the utility of the PHQ-9 and HADS-D. *Br J Gen Pract* 2010; **60(575):** 239–245.
12. Kendrick T, Dowrick C, McBride A, *et al.* Management of depression in UK general practice in relation to scores on depression severity questionnaires: analysis of medical record data. *BMJ* 2009; **338:** b750.
13. Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007; **22(11):** 1596–1602.
14. Wittkampf KA, Naeije L, Schene AH, *et al.* Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007; **29(5):** 388–395.
15. Kroenke K, Spitzer RL, Williams JBW, Lowe B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Scales: a systematic review. *Gen Hosp Psychiatry* 2010; **32(4):** 345–359.
16. Lowe B, Spitzer RL, Kerstin G, *et al.* Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord* 2004; **78(2):** 131–140.
17. Spitzer RL, Williams JBW, Gibbon M, First MB. *User's guide for the Structured Clinical Interview for DSM-III-R.* Washington DC, US: American Psychiatric Press, 1990.
18. Beck AT, Steer RA, Brown GK. *Manual for Beck Depression Inventory-II.* San Antonio, TX: Psychological Corporation, 1996.
19. Herrmann C. International experiences with the Hospital Anxiety and Depression Scale — A review of validation data and clinical results. *J Psychosom Res* 1997; **42(1):** 17–41.
20. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale: an updated literature review. *J Psychosom Res* 2002; **52(2):** 69–77.
21. Crawford JR, Henry JD, Crombie C, Taylor EP. Normative data for the HADS from a large non-clinical sample. *Br J Clin Psychol* 2001; **40(Pt 4):** 429–434.
22. American Psychiatric Association Task Force for the Handbook of Psychiatric Measures editor. *Handbook of psychiatric measures.* Washington DC, US: American Psychiatric Association, 2000.
23. NICE. *Depression with a chronic physical health problem.* NHS National Institute for Clinical Excellence, 2009; CG91.
24. Potts MK, Daniels M, Burnam MA, Wells KB. A structured interview version of the Hamilton Depression Rating Scale: evidence of reliability and versatility of administration. *J Psychiat Res* 1990; **24(4):** 335–350.
25. Murphy JM, Berwick DM, Weinstein MC, *et al.* Performance of screening and diagnostic tests. Application of receiver operating characteristic analysis. *Arch Gen Psychiatry* 1987; 44(6): 550–555.
26. Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: a review. *BMJ* 1999; **318(7194):** 1322–1323.
27. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ* 2004; **329(7458):** 168–169.
28. Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with confidence. 2nd edn.* London: BMJ Books; 2000.
29. Kirsch I, Deacon BJ, Huedo-Medina TB, *et al.* Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008; **5(2):** e45.
30. Fournier JC, DeRubeis RJ, Hollon SD, *et al.* Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA* 2010; **303(1):** 47–53.
31. NHS National Services Scotland. General Practice — Practice Team Information (PTI.)http://www.isdscotland.org/isd/3711.html (accessed 1 Jun 2011).
32. Wittkampf K, van Ravesteijn H, Baas K, *et al.* The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. *Gen Hosp Psychiatry* 2009; **31(5):** 451–459.
33. Leydon GM, Dowrick CF, McBride AS, *et al.* Questionnaire severity measures for depression: a threat to the doctor-patient relationship? *Br J Gen Pract* 2011; **61(583):** 117–123.
34. Mitchell C, Dwyer R, Hagan T, Mathers N. Impact of the QOF and the NICE guideline in the diagnosis and management of depression: a qualitative study. *Br J Gen Pract* 2011; **61(586):** 279–289.
35. Dowrick C, Leydon GM, McBride A, *et al.* Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. *BMJ* 2009; **338:** b663.