Making Sense of Variations in Prevalence Estimates of Depression in Cancer: A Co-Calibration of Commonly Used Depression Scales Using Rasch Analysis

Sylvie D. Lambert, RN, PhD[a]; Kerrie Clover, PhD, MPsychClin, MAPS[b,c]; Julie F. Pallant, PhD[d]; Benjamin Britton, DPsych[b]; Madeleine T. King, PhD[e]; Alex J. Mitchell, MBBS, MSc, MD, MRCPsych[f]; and Gregory Carter, PhD, MBBS, FRANZCP[b,c]

Abstract

Background: The use of different depression self-report scales warrants co-calibration studies to establish relationships between scores from 2 or more scales. The goal of this study was to examine variations in measurement across 5 commonly used scales to measure depression among patients with cancer: Hospital Anxiety and Depression Scale-Depression subscale (HADS-D), Centre for Epidemiologic Studies Depression Scale (CES-D), Patient Health Questionnaire-9 (PHQ-9), Beck Depression Inventory-II (BDI-II), and Depression Anxiety and Stress Scale-Depression subscale (DASS-D). **Methods:** The depression scales were completed by 162 patients with cancer. Participants were also assessed by the major depressive episode module of the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, 4th Edition. Rasch analysis and receiver operating characteristic curves were performed. **Results:** Rasch analysis of the 5 scales indicated that these all measured depression. The HADS and BDI-II had the widest measurement range, whereas the DASS-D had the narrowest range. Co-calibration revealed that the cutoff scores across the scales were not equivalent. The mild cutoff score on the PHQ-9 was easier to meet than the mild cutoff score on the CES-D, BDI-II, and DASS-D. The HADS-D possible cutoff score was equivalent to cutoff scores for major to severe depression on the other scales. Optimal cutoff scores for clinical assessment of depression were in the mild to moderate depression range for most scales. **Conclusions:** The labels of depression associated with the different scales are not equivalent. Most markedly, the HADS-D possible case cutoff score represents a much higher level of depression than equivalent scores on other scales. Therefore, use of different scales will lead to different estimates of prevalence of depression when used in the same sample. (J Natl Compr Canc Netw 2015;13:1203–1211)

Background

Approximately 15% of patients with cancer report clinically significant depression, with prevalence rates ranging from 0% to 58%[1–6]—this makes depression one of the most common psychological symptoms for this patient population.[4] Variations in depression prevalence estimates have been associated with medical (eg, cancer type), personal (eg, demographics), and social (eg, social support) factors.[2–4,7] Generally, younger age, low social support, and advanced disease have been associated with higher incidences of depression.[3] An additional risk factor is the use of maladaptive coping strategies.[8] Some studies have found that women are more at risk of reporting depression than men,[2] although results are not consistent.[3]

Beyond these population factors, variation in prevalence estimates may be attributed to the wide range of self-report scales used to assess depression.[3] Variations in the item content and scoring algorithms mean that different scales are not directly comparable. This variability makes comparisons across studies difficult and creates barriers in confidently interpreting to what degree different scales identify patients who are over thresholds for depression.

Commonly used scales of depression among patients with cancer include the Hospital Anxiety and Depression Scale (HADS),[9] the Centre for Epidemiologic Stud-

From [a]Ingram School of Nursing, McGill University, Montreal, Quebec, Canada; [b]Psycho-Oncology Service, Calvary Mater Newcastle, New South Wales, Australia; [c]Centre for Translational Neuroscience and Mental Health, University of Newcastle, New South Wales, Australia; [d]Rural Health Academic Centre, University of Melbourne, Victoria, Australia; [e]School of Psychology and Sydney Medical School, University of Sydney, New South Wales, Australia; and [f]University of Leicester, Cancer Studies & Molecular Medicine, Leicester, United Kingdom.

Submitted May 6, 2015; accepted for publication July 30, 2015.

The authors have disclosed that they have no financial interests, arrangements, affiliations, or commercial interests with the manufacturers of any products discussed in this article or their competitors.

Correspondence: Sylvie Lambert, RN, PhD, Ingram School of Nursing, McGill University, Wilson Hall, 3506 Sherbrooke Street, Montreal, Quebec, Canada H3A 2A7. E-mail: sylvie.lambert@mcgill.ca

ies Depression Scale (CES-D),[10] the Beck Depression Inventory-II (BDI-II),[11] the Depression Anxiety Stress Scales-21 (DASS-21),[12] and the Patient Health Questionnaire-9 (PHQ-9).[13] A recent review by Luckett et al[14] evaluated scales of anxiety, depression, and distress commonly used in psychosocial oncology trials. The evaluation criteria included evidence of reliability and validity; documented responsiveness; criterion validity against a diagnostic interview; availability of comparison data; length; number of constructs captured; and ease of administration. Although this review found that the HADS performed best overall, conceptual and psychometric concerns about the HADS-Depression subscale (HADS-D) (eg, emphasis on anhedonia) led to the recommendation that the CES-D is a better measure of depression among patients with cancer. However, as a range of scales continue to be used, co-calibration studies (or test-equating studies) are needed to establish relationships between scores from 2 or more scales.

The goal of this study was to examine variations in measurement among 5 commonly used scales to measure depression: HADS-D, CES-D, PHQ-9, DASS-Depression subscale (DASS-D), and BDI-II. Specifically, the objectives were to (1) test whether selected scales measure the same construct (ie, depression), (2) compare the scales' measurement range, (3) co-calibrate the scales' cutoff scores, and (4) further examine the accuracy statistics of the scales' cutoff scores using comparison with the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (SCID) as the gold standard.

Methods

Patients and Setting

This cross-sectional study was conducted at a cancer center in New South Wales, Australia, with a convenience sample of adult outpatients from medical oncology, radiation oncology, hematology, or psycho-oncology clinics. Adult patients with sufficient English language skills who were well enough to participate were eligible. Patients attending their first clinic visit were excluded. The Hunter Area Health Service Ethics Committee approved this study.

Procedures

The study was introduced to potential participants in the waiting room by a research assistant. The research assistant then provided interested patients with a detailed study information sheet and later contacted them by telephone to invite consenting patients to make a time to come into the hospital to complete a computer-administered survey[15–17] and the SCID. Data collection coincided, as much as possible, with a scheduled follow-up clinical appointment. Written consent was obtained from all participants.

Data Collection

The computerized survey included self-administered scales of anxiety and depression; however, only the following depression scales were included in this analysis: HADS-D, CES-D, PHQ-9, DASS-D, and BDI-II. Table 1 details each scale's cutoff scores.

HADS-D: The HADS-D is a 7-item self-administered questionnaire in which each item is rated on a 4-point response scale.[9] The maximum score on this subscale is 21. The HADS was originally developed to detect anxiety and depression in the context of medical outpatient clinics and has since been widely used across a number of illness contexts, including cancer. Somatic symptoms (eg, dizziness, headache) were purposefully excluded from the HADS to avoid confounding psychological symptoms with disease or treatment. The psychometric properties of this scale have been well established.[14,18,19]

CES-D: The CES-D is a 20-item scale developed to measure depressive symptomatology in the general population. The CES-D captures depressive symptoms in the past week, with the total score on this scale ranging from 0 to 60.[20] The CES-D has been established as a valid and reliable measure of depressive symptomatology among individuals with cancer.[21,22]

PHQ-9: The PHQ-9 is a 9-item scale developed from its predecessor, the PHQ, to assess depression severity in primary care. The PHQ-9 yields a maximum total score of 27.[13] The reliability of the PHQ-9 among patients with cancer has also been supported.[22]

DASS-D: The DASS-21 is a 21-item measure developed to provide maximum discrimination between these constructs of anxiety and depression. Although the DASS was originally developed to measure anxiety and depression, in psychometric testing a third factor was identified and was labeled *stress*. The DASS was initially tested among nonclinical samples; however, it has since been used across a wide range of clinical populations, and has established psychometric properties.[12] Similar to the HADS,

Table 1 Prevalence of Depression According to Various Scales

	Mean	SD	Scale's Score Range	n (%)
PHQ-9 (n=148)	6.82	5.63	0–24	–
Mild depression (5–9)				45 (30.41)
Moderate depression (10–14)				20 (13.51)
Moderately severe depression (15–19)				13 (8.78)
Severe depression (≥20)				5 (3.38)
HADS-D (n=162)	4.39	3.75	0–15	–
Possible cases (8–10)				16 (9.88)
Probable cases (≥11)				16 (9.88)
CES-D (n=148)	14.16	11.96	0–49	–
Mild depression (16–26)				31 (20.95)
Major depression (≥27)				23 (15.54)
DASS-D (n=154)	3.67	4.07	0–19	–
Mild depression (5–6)				17 (11.04)
Moderate depression (7–10)				14 (9.09)
Severe depression (11–13)				7 (4.55)
Extremely severe depression (≥14)				7 (4.55)
BDI-II (n=137)	12.47	10.77	0–48	–
Mild depression (14–19)				22 (16.06)
Moderate depression (20–28)				15 (10.95)
Severe depression (29–63)				13 (9.49)

Abbreviations: BDI-II, Beck Depression Inventory-II; CES-D, Centre for Epidemiologic Studies Depression Scale; DASS-D, Depression Anxiety Stress Scales-Depression subscale; HADS-D, Hospital Anxiety and Depression Scale-Depression subscale; PHQ-9, Patient Health Questionnaire-9.

somatic items were also purposefully excluded from the DASS. The DASS-21 is scored as three 7-item subscales: depression, anxiety, and stress (maximum subscale score = 21). Only the Depression subscale (DASS-D) was included in this analysis.

BDI-II: The BDI-II is a 21-item scale that was initially developed to assess the efficacy of psychoanalytically oriented psychotherapy in depressed individuals. When completing this scale, participants are asked to rate each item on a variable 4-point response scale (0–3). The BDI has established psychometric properties,[23] and scores on this scale range from 0 to 63.[11]

Participants also completed the major depressive episode module of the SCID. Patients were rated on whether they met the diagnostic criteria for a current major depressive episode in the past month (present vs absent). These interviews were conducted by 2 trained registered psychologists experienced in the diagnosis of depression. The SCID interviewers were blinded to participants' responses on the scales.

Data Analysis

The Rasch Unidimensional Measurement Models (RUMM) software version 2030[24] was first used (partial credit model) to create 5 subtests corresponding to each of the 5 scales to examine (1) overall fit statistics, (2) dimensionality (Objective 1), (3) the person-item map (Objective 2), and (4) test characteristic curves (Objective 3). Rasch analysis is a rigorous psychometric approach increasingly used to obtain an in-depth understanding of a scale's measurement properties[25,26] and allow for the identification of measurement issues (eg, item bias, misfit, response format) not easily detected by classical test theory approaches.[25] The Rasch measurement model assumes that the probability of a participant endorsing an item is a logistic function of the relative difference between the person's level of, for example, depression and the level of depression represented by an item.

Fit Analyses

Fit analyses examine the extent to which the 5 scales (or subtests) correspond to the Rasch measurement model. First, the chi-square probability and the summary fit residual standard deviations (SDs) for items and persons were examined. Good fit is indicated by a nonsignificant (using Bonferroni-adjusted alpha value) chi-square and fit residual SDs of less than 1.5.[27] Second, individual item and person-fit residual values were inspected to identify values outside the range of ± 2.5. Third, differential item functioning (DIF) was examined by conducting analysis of variance (using Bonferroni-adjusted alpha level) of the standardized response residuals for each subtest across each level of the person factors and class interval (ie, at different levels of depression). DIF can

occur when different groups within the sample, despite equal levels of depression, respond in a different manner to a scale.[28] Person factors considered in these analyses were age and sex. Last, local dependency and dimensionality were examined. To identify local dependency, the residual correlation matrix was examined and pairs of subtests with correlations exceeding 0.3 were taken to indicate dependency. To examine the dimensionality of the scales (or subtests), principal component analysis of the residuals was performed to identify the 2 subgroups of subtests that showed the most difference from one another. Differences between person estimates (location values) derived from these 2 subgroup of subtests were then compared using a series of *t* tests. If more than 5% of these *t* tests were significant, the scale was deemed multidimensional.[28]

Person-Item Map

The person-item map (Figure 1) was examined to compare the range of severity of depression captured by the 5 scales, the relative difficulty of endorsing the scales, and the distribution of transition points. Person-item maps plot the scales' score thresholds (transitions between scores) against the level of the trait being measured (ie, depression). The left side of a person-item map displays the distribution of respondents along the Rasch calibrated metric scale of the trait being measured (indicated by x's in Figure 1). This is referred to as the *location* value. Cases at the bottom of the map have the lowest location value, representing low/no depression, whereas cases at the top have higher severity levels of depression. The right side of the map displays the score thresholds. For example, the value of CES-D.02 represents the transition between a score of 1 to 2.

Test Characteristic Curves

Test equating analyses (or co-calibration), using test characteristic curves (Figure 2), were conducted to co-calibrate the scores across the 5 scales. Test char-

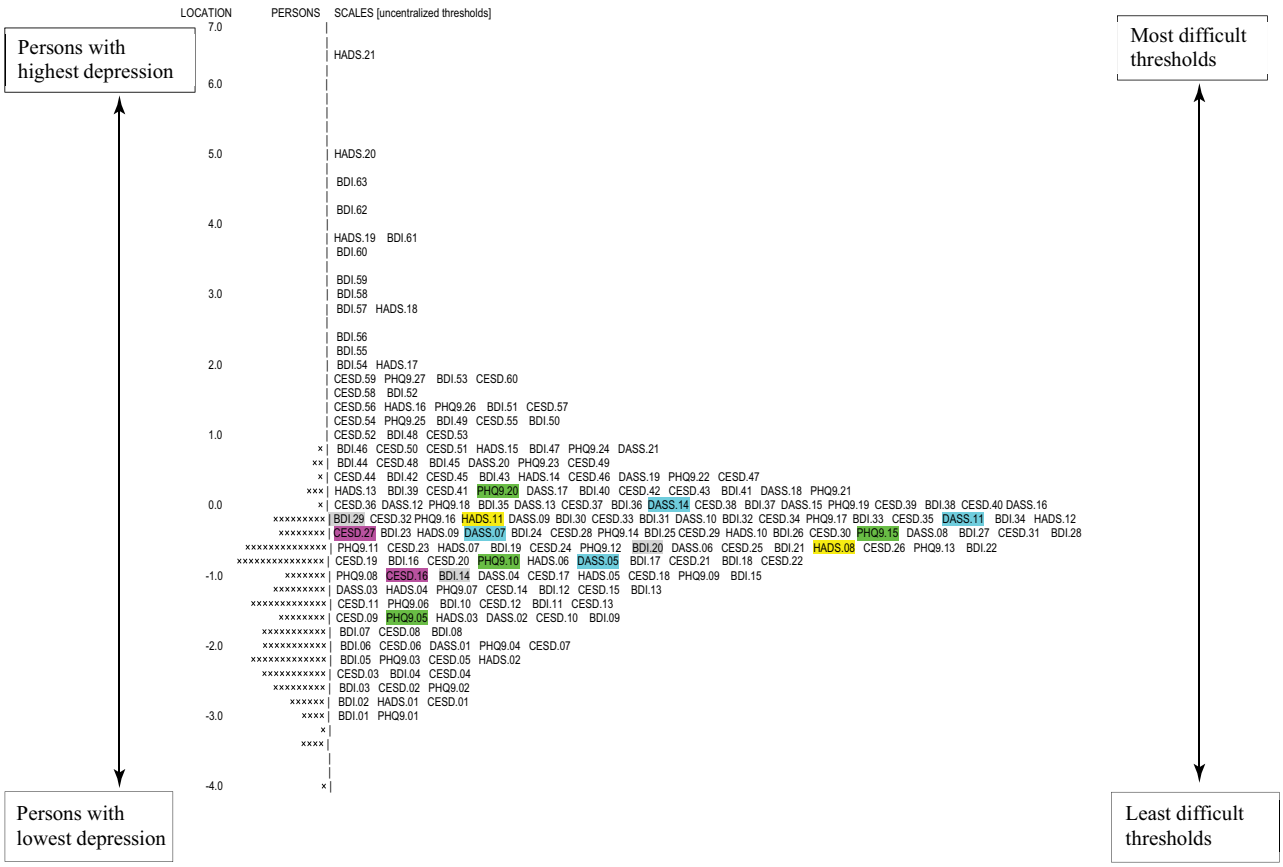

Figure 1 Item map showing location values for the Hospital Anxiety and Depression Scale-Depression subscale (HADS-D), the Centre for Epidemiologic Studies Depression Scale (CES-D), the Beck Depression Inventory-II (BDI-II), the Depression Anxiety Stress Scales–Depression subscale (DASS-D), and the Patient Health Questionnaire-9 (PHQ-9). The x denotes persons.

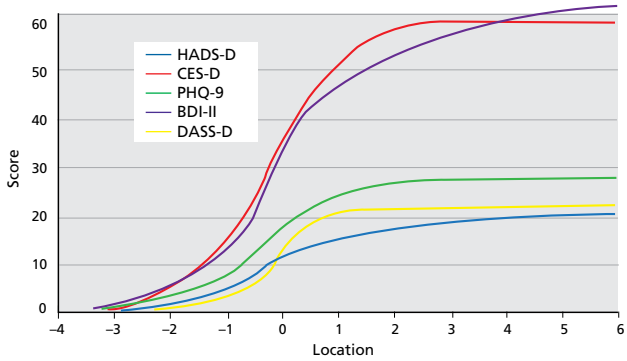

Figure 2 Equating of scores across scales. The x axis is the location on the logit scale or level of depression.
Abbreviations: BDI-II, Beck Depression Inventory-II; CES-D, Centre for Epidemiologic Studies Depression Scale; DASS-D, Depression Anxiety Stress Scales-Depression subscale; HADS-D, Hospital Anxiety and Depression Scale-Depression subscale; PHQ-9, Patient Health Questionnaire-9.

acteristic curves plot participants' scores on a scale (y axis) against their ability or location (x axis) on the measured trait (ie, depression) for each scale (or subtest). RUMM2030 expresses participants' ability (ie, depression severity) on a common logit scale. From this, it is possible to determine the scores on each scale that represent an equivalent level of depression severity. The process is analogous to equating kilograms to pounds by converting measurement from one scale to the other.

Cutoff Scores for Criterion Determined Major Depression

Receiver operating characteristic (ROC) curves were plotted using the Statistical Package for Social Sciences (IBM SPSS Statistics V22.0; Armonk, NY) to examine the ability of each scale to detect cases of major depression as identified by the SCID. For each curve, the chosen cutoff score offered the best compromise (or best balance) for both sensitivity and specificity. The area under the curve estimate was also used as an indicator of the overall accuracy of the scales to identify cases of depression (values 0.70–0.80 = acceptable; 0.81–0.90 = good; and 0.91–1.00 = excellent). The chosen cutoff scores were further examined for whether they also offered the best balance for positive predictive value (PPV) and negative predictive value (NPV).

Results

Sample

Initially, 322 patients consented to study participation, with 168 completing study scales and 162 providing sufficient data for this analysis. Table 2 presents participant demographic and illness variables. Also, the mean scores for the 5 depression scales are summarized in Table 1. The prevalence of depression according to the SCID was 14.2%.

Table 2 Patients' Characteristics

Demographics	n[a] (%)
Sex (N=161)	
Female	107 (66.46)
Age, y	
≤55	56 (34.57)
56–65	61 (37.65)
≥66	45 (27.78)
Education (N=161)	
Primary school	96 (59.63)
Secondary school	26 (16.15)
College	39 (24.22)
Marital status (N=161)	
Married or living as married	113 (70.19)
Widowed	13 (8.07)
Divorced/Separated	27 (16.77)
Never married	8 (4.97)
Primary cancer type	
Breast	71 (43.83)
Hematologic	25 (15.43)
Colorectal	20 (12.35)
Lung	14 (8.64)
Genitourinary	11 (6.79)
Other	21 (12.96)
Stage (N=110)	
Early	20 (18.18)
Regional 2	15 (13.64)
Regional 3	18 (16.36)
Advanced	27 (24.55)
Unsure	30 (27.27)
Time since diagnosis, mo (N=161)	
≤12	59 (36.65)
13–24	30 (18.63)
25–48	30 (18.63)
≥49	42 (26.09)

[a]N=162, unless specified otherwise.

Fit Analyses

Analysis confirmed the subtests' fit to the Rasch model expectations, as indicated by a nonsignificant item trait interaction chi-square statistic (χ^2 = 6.22; degrees of freedom = 10; P=.80). Although the summary item fit residual SD (1.84) exceeded the accepted value of 1.5, the individual items fit residuals were all less than 2.5. The summary person fit residual SD was 0.96. No DIF and no local dependency were present. There was no evidence of multidimensionality with a series of independent t tests comparing person estimates from the subtests identified using principal component analysis of the

residuals, indicating that only 1.94% of tests were statistically different.

Person-Item Maps

As illustrated in the person-item map (Figure 1), the HADS-D and BDI-II have distribution of scores spanning the widest location values from low to high depression severity. Hence, the HADS-D and BDI-II have the widest measurement range. Conversely, the DASS-D had the narrowest measurement range. The CES-D and PHQ-9 had a comparable measurement range. At lower levels of depression, the HADS-D, CES-D, BDI-II, and PHQ-9 were the easiest to endorse, whereas at higher levels of depression, the HADS-D and BDI-II were the most difficult to endorse. The distribution of the transition points around the cutoff scores reveals that the mild cutoff score on the PHQ-9 was the easiest to attain, which means the PHQ-9 identified more cases of mild depression than the CES-D, BDI-II, and DASS-D (Table 1). Conversely, the cutoff score on the PHQ-9 for severe depression is the highest on the construct.

Co-Calibrate the Scales

Co-calibration of scales' scores is illustrated in Figure 2 and detailed in Table 3. The results show that the cutoff scores across the scales were not equivalent. The mild cutoff score of the PHQ-9 was equivalent to scores below the mild cutoff scores of the CES-D, BDI-II, and DASS-D. However, the mild cutoff scores of the CES-D and BDI-II were at a comparable level. The DASS-D mild cutoff score was found to be approximately equivalent to the moderate cutoff scores on the PHQ-9 and BDI-II.

The HADS-D possible cutoff score was approximately equivalent to those for major, moderately severe, and moderate depression on the CES-D, PHQ-9, and DASS-D, respectively. This explains the relatively lower prevalence of possible depression according to the HADS-D (Table 1). Similarly, the HADS-D probable case cutoff score was equivalent to the severe cutoff score on the DASS-D.

ROC Curve Analysis

As detailed in Table 4, the best balance of sensitivity and specificity was identified for the following scores: PHQ-9 of 9 or greater, HADS-D of 7 or greater, CES-D of 22 or greater, BDI-II of 16 or greater, and DASS-D of 6 or greater. These cutoff scores also offered the best balance between PPV and NPV for the PHQ-9, HADS-D, CES-D, and DASS-D. However, the BDI-II cutoff score was one point higher (≥17).

Discussion

To understand the extent to which variations in estimates of depression prevalence are a measurement artefact, this study examined the measurement range and cutoff scores of the PHQ-9, HADS-D, CES-D, DASS-D, and BDI-II. Five major findings warrant attention. First, when considered as subtests, the 5 scales were found to fit Rasch measurement model expectations and measured the common underlying construct of depression. This finding is comparable to those reported by Covic et al.[29,30]

Second, the scales were found to differ considerably in the range of depression severity measured,

Table 3 Equivalent Scores Across Scales

PHQ-9	HADS-D	CES-D	BDI-II	DASS-D
5 (mild)	2.7	9.9	8.7	1.7
7.8	4.4	**16 (mild)**	13.2	3.3
8.4	4.8	17.1	**14 (mild)**	3.6
10 (moderate)	6	20.5	16.6	4.5
10.9	6.7	22.3	18.3	**5 (mild)**
11.8	7.4	24.1	**20 (moderate)**	5.6
12.6	**8 (possible)**	25.6	21.6	6.1
13.3	8.5	**27 (major)**	23.2	6.6
13.7	8.9	27.9	24.2	**7 (moderate)**
15 (moderately severe)	9.8	30.7	27.7	8.4
15.5	10.1	31.7	**29 (severe)**	9
16.7	**11 (probable)**	34.5	32.5	10.9
16.8		34.6	32.7	**11 (severe)**
18.4	12	38.5	36.8	**14 (extremely severe)**
20 (severe)	12.9	42.2	40.1	16.7

Abbreviations: BDI-II, Beck Depression Inventory-II; CES-D, Centre for Epidemiologic Studies Depression Scale; DASS-D, Depression Anxiety Stress Scales–Depression subscale; HADS-D, Hospital Anxiety and Depression Scale–Depression subscale; PHQ-9, Patient Health Questionnaire–9.

Table 4 PPV, NPV, Sensitivity, and Specificity, Percentages for Selected Cutoff Scores on the Various Scales

Cutoff Score	PPV (95% CI)	NPV (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
PHQ-9				
≥5 (mild)	22.9 (14.7–33.7)	96.9 (88.2–99.5)	90.5 (68.2–98.3)	49.2 (40.2–58.2)
≥9[a,b]	40.0 (26.1–55.6)	97.1 (91.0–99.2)	86.7 (62.6–96.2)	78.6 (70.2–85.2)
≥10 (moderate)	39.5 (24.5–56.5)	94.5 (87.9–97.7)	71.4 (47.7–87.8)	81.7 (73.7–87.8)
≥15 (moderately severe)	61.1 (36.1–81.7)	92.2 (85.8–96.0)	52.4 (30.3–73.6)	94.4 (88.5–97.5)
≥20 (severe)	80.0 (29.9–98.9)	88.0 (81.3–92.7)	19.0 (6.3–42.6)	99.2 (95.0–1.00)
HADS-D				
≥7[a,b]	42.5 (27.4–59.0)	95.0 (89.1–98.0)	73.9 (51.3–88.9)	83.3 (75.8–88.9)
≥8 (possible)	46.9 (29.5–65.0)	93.8 (87.8–97.1)	65.2 (42.8–82.8)	87.7 (80.7–92.5)
≥11 (probable)	50.0 (25.5–74.5)	89.7 (83.2–93.9)	34.8 (17.2–57.2)	94.2 (88.5–97.3)
CES-D				
≥16 (mild)	33.3 (21.5–47.6)	96.8 (90.2–99.2)	85.7 (62.6–96.2)	71.4 (62.6–78.9)
≥22[a,b]	41.9 (27.4–57.8)	97.1 (91.2–99.3)	85.7 (62.6–96.2)	80.2 (71.9–86.5)
≥27 (major)	60.9 (38.8–79.5)	94.4 (88.3–97.5)	66.7 (43.1–84.5)	92.9 (86.5–96.5)
BDI-II				
≥14 (mild)	34.0 (21.6–48.9)	95.3 (87.9–98.5)	81.0 (57.4–93.7)	71.3 (62.0–79.2)
≥16[a]	38.1 (24.0–54.3)	94.7 (87.5–98.0)	76.2 (52.5–90.9)	77.4 (68.5–84.4)
≥17[b]	41.7 (26.0–59.1)	94.0 (86.9–97.5)	71.4 (47.7–87.8)	81.7 (73.2–88.1)
≥20 (moderate)	50.0 (31.1–68.9)	93.5 (86.6–97.1)	66.7 (43.1–84.5)	87.8 (80.1–92.9)
≥29 (severe)	76.9 (46.0–93.8)	91.1 (84.2–95.2)	47.6 (26.4–69.7)	97.3 (92.0–99.3)
DASS-D				
≥5 (mild)	40.0 (26.1–55.6)	97.2 (91.5–99.3)	85.7 (62.6–96.2)	79.5 (71.4–85.9)
≥6[a,b]	44.4 (28.3–61.7)	95.7 (89.8–98.4)	76.2 (52.5–90.9)	84.8 (77.3–90.3)
≥7 (moderate)	55.6 (35.6–74.0)	95.2 (89.5–98.0)	71.4 (47.7–87.8)	90.9 (84.3–95.0)
≥11 (severe)	64.3 (35.6–86.0)	91.4 (85.1–95.3)	42.9 (22.6–65.6)	96.2 (90.9–98.6)
≥14 (extremely severe)	83.3 (36.5–99.1)	89.1 (82.7–93.5)	23.8 (9.1–47.5)	99.2 (95.2–1.00)

Abbreviations: BDI-II, Beck Depression Inventory-II; CES-D, Centre for Epidemiologic Studies Depression Scale; DASS-D, Depression Anxiety Stress Scales–Depression subscale; HADS-D, Hospital Anxiety and Depression Scale–Depression subscale; NPV, negative predictive value; PHQ-9, Patient Health Questionnaire–9; PPV, positive predictive value.
[a]Cutoff score with best balance between sensitivity and specificity.
[b]Cutoff score with best balance between PPV and NPV.

with the BDI-II and HADS-D having the widest measurement range. In their analysis of the CES-D and BDI-II among adolescents, using item response theory, Olino et al[31] also found the BDI-II to have the broadest measurement range. The present study adds that the HADS-D and BDI-II have a comparable measurement range, but the HADS-D was even more useful in assessing depression in the high-severity range. Given their wide measurement range, the HADS-D and BDI-II might be best suited for measuring intervention response among patients with clinical levels of depression. Conversely, the CES-D, PHQ-9, and DASS-D, because of their narrower measurement range and concentration of items at the lower depression severity range, might be more appropriate for identifying depression in samples in which the expected severity is lower.

Third, the published cutoff scores on the scales did not represent equivalent depression severity. These findings confirm that variation in prevalence of depression is likely to be attributable to the scale used. Most strikingly, at the mild and moderate end of the spectrum, the PHQ-9 was the easiest to endorse. This finding was further supported by the co-calibration

analysis, whereby the mild cutoff score of the PHQ-9 was equivalent to scores below the mild cutoff scores of the CES-D, BDI-II, and DASS-D. The ROC curve analyses complemented these findings by identifying that the optimal cutoff score of 9 on the PHQ-9 is closer to the moderate range, which is equivalent to the mild range of other scales. Thekkumpurath et al[32] compared the PHQ-9 against clinical interviews and found that among patients with cancer, the optimal cutoff score for the PHQ-9 was 8. In other clinical contexts, the optimal cutoff score of the PHQ-9 has been found to range from 9 to 12.[33]

Fourth, severity labels attributed to cutoff scores might be misleading. This was most obvious for the HADS-D's possible label, because the co-calibration analyses revealed that this cutoff score was comparable to depression in the major or moderate range as captured by the CES-D and DASS-D, respectively. Similarly, the HADS-D probable cutoff score was equivalent to the severe cutoff score on the DASS-D. This finding is comparable with that reported by Covic et al[30] among individuals with rheumatoid arthritis. This means that researchers and clinicians can have confidence that patients identified as possible cases by the HADS-D have sufficient symptoms to warrant further referral. This conclusion is further supported by comparing the HADS-D with the SCID, whereby the optimal HADS-D cutoff score was 7. Other ROC analyses have suggested that a score as low as 5[34,35] on the HADS-D among patients with cancer results in the best trade-off between sensitivity and specificity.

Fifth, based on the analyses conducted, some clear recommendations can be put forward regarding the future use of these scales. For the PHQ-9, its narrow measurement range and high sensitivity and NPV support its particular utility as a screening measure (rather than a case-finding tool). The scale with the highest specificity and PPV was the DASS-D, suggesting that it is one of the most appropriate scales for case finding. The HADS-D followed the DASS-D in terms of its suitability as a case-finding instrument. However, the CES-D had the best trade-off across NPV, PPV, sensitivity, and specificity, which, combined with its broad measurement range, suggests that it might be the scale to favor overall. This finding corroborates those of the Luckett et al[14] review. The overall accuracy of the BDI-II was the lowest, and therefore it is least recommended scale for future use.

Strengths and Limitations

The strengths of the study are the inclusion of a mixed group of oncology outpatients and the inclusion of SCID as a goal standard to examine the accuracy statistics of the cutoff scores. The use of Rasch analysis allowed for the co-calibration of the scales on a common metric. One limitation was that more women than men were included; however, no DIF on sex was detected. Another limitation is that recruitment occurred in one center, which affects the generalizability of the findings. Also, findings need to be corroborated in a larger sample and across other illness contexts.

Conclusions

Comparison of 5 depression scales commonly used for patients with cancer provided evidence that at least some of the variability in estimates of prevalence of depression is caused by measurement artefact. The labels of mild, moderate, and severe depression attributed to the cutoff scores across these scales were not equivalent. Overall, a score in the mild to moderate depression range on the PHQ-9 represented a lower level of depression severity than on the other scales. Conversely, the HADS-D possible case cutoff score was found to represent a much higher level of depression according to the other scales. These findings have direct impact in terms of interpreting prevalence estimates across studies and the selection of a scale in managing patients in cancer care. Furthermore, study findings can guide researchers and clinicians in choosing the scale most fitting for their context, including which cutoff score to favor.

Acknowledgements

The authors wish to express their sincere thanks to the patients who provided the survey data. They are also grateful to the research staff, Kylie Harris, Stacey Hosking, and Georgia Carr, for recruiting and interviewing participants.

References

1. Clinical Practice Guidelines for the Psychosocial Care of Adults with Cancer. Available at: http://www.nhmrc.gov.au/_files_nhmrc/file/publications/synopses/cp90.pdf. Accessed September 21, 2015.
2. Carlson LE, Angen M, Cullum J, et al. High levels of untreated distress and fatigue in cancer patients. Br J Cancer 2004;90:2297–2304.

3. Miller K, Massie MJ. Depressive disorders. In: Holland J, Breitbart W, Jacobsen PB, et al, eds. Psycho-Oncology. New York, NY: Oxford University Press; 2010:311–318.
4. Mitchell AJ, Chan M, Bhatti H, et al. Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies. Lancet Oncol 2011;12:160–174.
5. Walker J, Holm Hansen C, Martin P, et al. Prevalence of depression in adults with cancer: a systematic review. Ann Oncol 2013;24:895–900.
6. Mitchell AJ, Meader N, Davies E, et al. Meta-analysis of screening and case finding tools for depression in cancer: evidence based recommendations for clinical practice on behalf of the Depression in Cancer Care consensus group. J Affect Disord 2012;140:149–160.
7. Rowland J, Massie MJ. Breast cancer. In: Holland J, Breitbart W, Jacobsen P, eds. Psycho-Oncology. Oxford: Oxford University Press; 2010:177–186.
8. Li M, Boquiren V, Lo C et al. Depression and anxiety in supportive oncology. In: Davis M, Feyer P, Ortner P, et al, eds. Supportive Oncology. Philadelphia, PA: Elsevier; 2011:528–540.
9. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiat Scand 1983;67:361–370.
10. Centre for Epidemiologic Studies. Centre for Epidemiologic Studies Depression Scale (CES-D). Rockville: National Institute of Mental Health; 1970.
11. Beck AT, Brown GK, Steer RA. Manual for the Beck Depression Inventory-II. San Antonio, TX: The Psychological Corporation; 1996.
12. Lovibond PF, Lovibond SH. Manual for the Depression Anxiety & Stress Scales. 2nd ed. Sydney, Australia: Psychology Foundation; 1995.
13. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606–613.
14. Luckett T, Butow PN, King MT, et al. A review and recommendations for optimal outcome measures of anxiety, depression and general distress in studies evaluating psychosocial interventions for English-speaking adults with heterogeneous cancer diagnoses. Support Care Cancer 2010;18:1241–1262.
15. Clover K, Carter G, Mackinnon A, Adams C. Is my patient suffering clinically significant emotional distress? Demonstration of a probabilities approach to evaluating algorithms for screening for distress. Support Care Cancer 2009;17:1455–1462.
16. Clover K, Leigh Carter G, et al. Concurrent validity of the PSYCH-6, a very short scale for detecting anxiety and depression, among oncology outpatients. Aust N Z J Psychiatry 2009;43:682–688.
17. Carter G, Britton B, Clover K, et al. Effectiveness of QUICATOUCH: a computerised touch screen evaluation for pain and distress in ambulatory oncology patients in Newcastle, Australia. Psychooncology 2012;21:1149–1157.
18. Bjelland I, Dahl A, Haug T, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale: an updated literature review. J Psychosom Res 2002;52:69–77.
19. Lambert S, Pallant JF, Boyes A, et al. A Rasch analysis of the Hospital Anxiety and Depression Scale (HADS) among cancer survivors. Psychol Assess 2013;25:379–390.
20. Zich JM, Attkisson CC, Greenfield TK. Screening for depression in primary care clinics: the CES-D and the BDI. Int J Psychiatry Med 1990;20:259–277.
21. Hann D, Winter K, Jacobsen P. Measurement of depressive symptoms in cancer patients: evaluation of the Center for Epidemiological Studies Depression Scale (CES-D). J Psychosom Res 1999;46:437–443.
22. Vodermaier A, Linden W, Siu C. Screening for emotional distress in cancer patients: a systematic review of assessment instruments. J Natl Cancer Inst 2009;101:1464–1488.
23. Wang YP, Gorenstein C. Assessment of depression in medical patients: a systematic review of the utility of the Beck Depression Inventory-II. Clinics (Sao Paulo) 2013;68:1274–1287.
24. Andrich D, Lyne A, Sheridan B, Luo G. RUMM2030. Perth, Western Australia: RUMM Laboratory; 2010.
25. Hagquist C, Bruce M, Gustavsson J. Using the Rasch model in nursing research: an introduction and illustrative example. Int J Nurs Stud 2009;46:380–393.
26. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007;46:1–18.
27. Shea TL, Tennant A, Pallant JF. Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). BMC Psychiatry 2009;9:21.
28. Tennant A, Conaghan PG. The Rasch Measurement Model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007;57:1358–1362.
29. Covic T, Cumming SR, Pallant JF, et al. Depression and anxiety in patients with rheumatoid arthritis: prevalence rates based on a comparison of the Depression, Anxiety and Stress Scale (DASS) and the Hospital, Anxiety and Depression Scale (HADS). BMC Psychiatry 2012;12:6.
30. Covic T, Pallant JF, Tennant A, et al. Variability in depression prevalence in early rheumatoid arthritis: a comparison of the CES-D and HAD-D Scales. BMC Musculoskelet Disord 2009;10:18.
31. Olino TM, Yu L, Klein DN, et al. Measuring depression using item response theory: an examination of three measures of depressive symptomatology. Int J Methods Psychiatr Res 2012;21:76–85.
32. Thekkumpurath P, Walker J, Butcher I, et al. Screening for major depression in cancer outpatients: the diagnostic accuracy of the 9-item patient health questionnaire. Cancer 2011;117:218–227.
33. 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;22:1596–1602.
34. Singer S, Kuhnt S, Gotze H, et al. Hospital anxiety and depression scale cutoff scores for cancer patients in acute care. Br J Cancer 2009;100:908–912.
35. Katz MR, Kopek N, Waldron J, et al. Screening for depression in head and neck cancer. Psychooncology 2004;13:269–280