Neuropsychiatric Disease and Treatment

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Psychometric evaluation and linking of the PHQ-9, QIDS-C, and VQIDS-C in a real-world population with major depressive disorder

Emily OC Palmer[1], Sheryl Ker[2], Miguel E Rentería[3], Thomas Carmody[4], A John Rush[5,6]

[1]Holmusk Europe Ltd, London, UK; [2]KKT Technologies, Pte. Ltd, Singapore; [3]QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia; [4]Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX, USA; [5]Duke University School of Medicine, Duke University School of Medicine, Durham, NC, USA; [6]Clinical sciences, Duke-National University of Singapore, Singapore

Correspondence: Emily OC Palmer, Holmusk Europe Ltd, 414 Linen Hall, 162-168 Regent St, London, W1B 5TE, UK, Email emily.palmer@holmusk.com

Purpose: Major depressive disorder (MDD) is a leading cause of disability worldwide. An accurate assessment of depressive symptomology is crucial for clinical management and research. This study assessed the convergent validity, reliability, and total scale score interconversion across the 9-item Patient Health Questionnaire (PHQ-9) self-report, the 16-item Quick Inventory of Depressive Symptomatology-clinician report (QIDS-C) (two widely used clinical ratings) and the 5-item Very Brief Quick Inventory of Depressive Symptoms-clinician report (VQIDS-C), which evaluate the core features of MDD.

Patients and Methods: This study leveraged electronic health record (EHR)-derived, de-identified data from the NeuroBlu Database (Version 23R1), a longitudinal behavioural health real-world platform. Classical Test Theory (CTT) and Item Response Theory (IRT) analyses were used to evaluate the reliability, validity of, and conversions between the scales. The Test Information Function (TIF) was calculated for each scale, with greater test information reflecting higher precision and reliability in measuring depressive symptomology. IRT was also used to generate conversion tables so that total scores on each scale could be compared to the other.

Results: The study sample (n = 2,156) had an average age of 36.4 years (standard deviation [SD] = 13.0) and 59.7% were female. The mean depression scores for the PHQ-9, QIDS-C, and VQIDS-C were 12.9 (SD = 6.6), 12.0 (SD = 4.9), and 6.18 (SD = 3.2), respectively. The Cronbach's alpha coefficients for PHQ-9, QIDS-C, and VQIDS-C were 0.9, 0.8, and 0.7, respectively, suggesting acceptable internal consistency. PHQ-9 (TIF = 30.3) demonstrated the best assessment of depressive symptomology, followed by QIDS-C (TIF = 25.8) and VQIDS-C (TIF = 17.7).

Conclusion: Overall, PHQ-9, QIDS-C, and VQIDS-C appear to be reliable and convertible measures of MDD symptomology within a US-based adult population in a real-world clinical setting.

Keywords: psychometrics, linking, patient health questionnaire, quick inventory of depressive symptoms, depression, real-world data

Introduction

Major depressive disorder (MDD) is a leading cause of disability worldwide, with an estimated lifetime prevalence in the United States (US) of 20%.[1] Accurate and reliable assessment of MDD is crucial for effective treatment and management. The Patient Health Questionnaire-9 (PHQ-9; self-report) and the clinician rated Quick Inventory of Depressive Symptomatology (QIDS-C)[2] are widely used and validated tools to assess the prevalence and severity of depressive symptoms in patients with MDD. These scales are used both in clinical research and practice. The PHQ-9 assesses the prevalence over the past two weeks of the nine criterion depressive symptom domains that define a major depressive episode based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-5).[3] It is a self-administered questionnaire that can be completed in a few minutes, making it a convenient tool for clinicians and patients. The PHQ-9 has demonstrated good psychometric properties in various populations, including adults[4,5] and adolescents,[6] and has been translated into multiple languages.[7–13]

The QIDS-C consists of 16 items that cover the same nine DSM-5 criterion symptom domains; it has clinician report and self-report versions.[2,14] The QIDS-C has demonstrated good psychometric properties, including high internal consistency,

Received: 13 October 2023
Accepted: 28 February 2024
Published: 26 March 2024

test-retest reliability, and strong convergent validity with other depression measures.[15] In recent years, the 5-item Very Quick Inventory of Depressive Symptomatology (VQIDS-C)[16] was developed as a more convenient method for rapid clinical assessment and is effective at evaluating the core features of depression. The self-report version of the QIDS and PHQ-9 have been compared in Asian primary care settings and shown good internal consistency.[17] However, these measures have not been evaluated together in a large real-world US cohort. Previous research has demonstrated the value of linking depression scales such as the PHQ-9 and Hamilton Rating Scale for Depression (HAM-D)[18] and Montgomery-Åsberg Depression Rating Scale (MÅDRS), Self Rated Scale (SRS), PHQ-9 and The Beck Depression Inventory-II (BDI-II).[19] Given the high prevalence of MDD and the need for an accurate and reliable assessment of this disorder, validating the PHQ-9, QIDS-C, and VQIDS-C using real-world data is paramount. In this study we aimed to answer the following research questions: 1) How do the individual symptom domains perform in relation to the overall trait when evaluating the QIDS-C, VQIDS-C, and PHQ-9 in the same patients?; and 2) Can conversion tables be developed to estimate total scores among the three rating scales?

Methods

Participants

This study leveraged EHR derived, de-identified data from the NeuroBlu Database (Version 23R1) a longitudinal behavioural health real-world database comprising both structured and unstructured patient-level clinical data.[20] At the time of this study, the NeuroBlu Database included data from over 30 behavioural health centres across the US spanning over 20 years, consisting of data from over 1.4 million patients.

Patients were selected from the NeuroBlu Database if they had at least one recorded non-psychotic MDD diagnosis and did not have a record of schizophrenia, schizoaffective disorder, and bipolar disorder at any time (Supplemental Table 1). Patients were also required to have a record within the EHR of both the QIDS-C and the PHQ-9 (after MDD diagnosis) on the same day.

Institutional review board (IRB) approval of the study protocol, including a waiver of Health Insurance Portability and Accountability Act authorization, was obtained prior to study conduct, and covers data originating from all sites represented. Approval was granted by the WCG IRB. (The Holmusk Real-World Evidence Parent Protocol; IRB registration number 1–1470336-1; Protocol ID HolmuskRWE_1.0).

Measures

Demographic characteristics and clinical features were based on documentation at the time of entry into the EHR (sex, race, ethnicity) or baseline (the first instance in which QIDS-C and PHQ-9 are recorded on the same day).

PHQ-9 is a 9-item self-report instrument designed to assess the prevalence of depressive symptoms in patients over the last two weeks. The items on the scale correspond to the nine diagnostic criteria for major depressive disorder from DSM-5 (eg depressed mood, low interest, sleep disturbance, fatigue, appetite changes, etc.).[3] The prevalence of each symptom item is scored on a four-point Likert scale ranging from 0 (not at all) to 3 (nearly every day). The total score ranges from 0 to 27, with higher scores indicating greater presence of depressive symptoms.

The QIDS-C is a 16-item clinician-rated scale developed to measure the severity of depressive symptoms in both research and clinical settings.[2,14] The 16 items cover the same nine DSM-5 criterion symptom domains of a major depressive episode as the PHQ-9. Each item is rated on a four-point Likert scale, ranging from 0 to 3 with responses largely aimed at defining the overall severity of the symptom over the prior 7 days. Items 1–4, 6–9, and 15–16 assess sleep, appetite/weight, and psychomotor symptoms, respectively. For these three domains, only the highest score is counted towards the total QIDS-C score. The total score, therefore, ranges from 0 to 27, with higher scores indicating more severe symptoms of depression.

The VQIDS-C consists of five items selected from the longer 16-item QIDS-C[16,21,22] to reflect core depressive symptoms (sad mood, self-outlook, involvement, fatigue, and psychomotor changes) modelled after the 6-item Hamilton rating scale for depression.[23] The brevity of the scale allows for smart phone use. VQIDS-C scores range from 0 to 15, with higher scores indicating more severe depressive symptoms. This tool is sensitive to changes in adults[16] and has been validated through linking to other depression scales.[22]

Removal of Invalid Data

When linking psychometric scales, it is essential to ensure that the conversion system relies on valid data, the removal of invalid data prior to linking scales is a practice recommended by previous studies.[24,25] The validity of data can be assessed using the similarity of scores among patients on comparable scales. For instance, if a patient's overall score on the PHQ-9 indicates severe depressive symptomatology, a similar score should be reflected on the QIDS-C. Invalid data was therefore eliminated by plotting the total PHQ-9 score against the total QIDS-C score for each patient. The patients with a more than 10-point difference between the scores, which is indicative of a two-severity category difference were then removed, this process ensures the reliability and accuracy of the link between the scales. This threshold was based on the clinical opinion that it is unlikely that a patient would move two severity categories in the space of a week, as the PHQ-9 captures a two-week period compared to the QIDS-C which is 7 days.

Data Analysis

Descriptive statistics were reported for demographic characteristics and baseline clinical variables. Continuous variables were summarised using mean with standard deviation (SD) or median with interquartile range (IQR), as appropriate. Categorical and ordinal variables were summarised using frequencies and percentages.

Classical Test Theory (CTT) was used to evaluate the internal consistency of the PHQ-9, QIDS-C, and VQIDS-C. CTT tests the reliability and validity of a scale based on its items. CTT assumes that each observed score is a combination of an underlying true score representing the latent trait (depressive symptomology) and unsystematic error. Cronbach's alpha and domain-total correlations were calculated to determine the extent to which domains on the scales were interrelated and consistently measured the underlying latent trait (depressive symptomology). Values of Cronbach's alpha range from 0 to 1, with values greater than 0.70 considered acceptable. For domain-total correlation: values greater than 0.3 indicate that a domain discriminates well.

Item Response Theory (IRT)[26] is a statistical framework that models the relationship between individual test domains and an underlying latent trait or construct. Unlike CTT, it does not assume that all domains contribute equally to the underlying latent trait or construct, but it does assume the unidimensionality of each scale. Unidimensionality was assessed using parallel analysis.[27] A graded response IRT model was used in the current study as previously described.[28]

IRT estimates discrimination (a) and difficulty (b) parameters for each domain in each scale. The discrimination parameter measures how effectively a domain can differentiate between individuals with different levels of depression. The difficulty parameters represent the level of depression at which an individual has a particular probability of endorsing a specific response category for a domain. In this context, b0 indicates the depression level at which a subject would be equally likely to score either "0" compared with "1", "2", or "3". Similarly, b1 represents an equal likelihood of scoring "0" or "1" versus "2" or "3". Lastly, b2 indicates an equal likelihood of scoring "0", "1", or "2" versus "3". IRT was also used to calculate the Test Information Function (TIF) which is calculated by the addition of the information of every domain in a scale. The information refers to the precision at which the latent trait can be estimated by each domain.

IRT methods as described by previous literature[29,30] were used to link the total scores for each scale. An individual's IRT latent trait score was estimated by multiplication of the item response category characteristic curve (IRCCC) corresponding to patients' response patterns to represent the posterior distribution, and an average of the distribution was taken. This IRT latent trait score measures depressive symptomology in standard deviation units where 0 represents the average. This IRT latent trait score is then averaged for every individual with a specific summary score total for each scale, this is then linked to a corresponding summary score for a different scale depending on which score has the closest IRT latent trait score, creating a conversion table.

The linking of scales was then validated using multiple methods (Correlation, Root Mean Square Error (RMSE), and Kappa) to compare the score estimated by the conversion with the actual score reported for that patient. Firstly, the correlation coefficient between the estimated scores and the actual score reported for that patient was calculated, 0.7–0.9 represents a high correlation. Secondly, RMSE, which represents the average difference between the actual and estimated scores was calculated. As RMSE is scale-dependent, the interpretation differs by scale, but a lower RMSE is better. Finally, the Kappa was calculated to represent the level of agreement between the estimated severity category (none, mild, moderate, severe, very severe) and the actual score severity category. Kappa ranges from −1 to 1, where 1 indicates perfect agreement, 0 indicates agreement due to chance, 0.21–0.40 represents a fair agreement and 0.41–0.60 represents a moderate agreement.[31]

Data Privacy

The data ingestion and aggregation process comply with HIPAA guidelines, including applying the Safe Harbor definition for de-identified data.[32]

Software

The following programs were used for the various analysis: ltm version 1.2-0 for the IRT graded response model, mirt version 1.38.1 for the confirmatory factor analysis, psych version 2.3.3 for Cohen's kappa and exploratory factor analysis and paran version 1.5.2 for the parallel analysis.

Results

Demographics

Of the 2,231 study eligible patients, 75 had at least a 10-point difference between scores and were removed bringing the final cohort to 2,156. A comparison of demographic features between the final cohort and those that were excluded is shown in Supplemental Tables 2 and 3.

Table 1 shows the features of the analytic sample (n = 2,156) with a mean age of 36.4 years and included nearly 60% females. Most patients were white or black or African American. Nearly a third endorsed Latino or Hispanic ethnicity. Overall depression severity was moderate based on the PHQ-9 and QIDS-C.

Table 1 Demographic characteristics at first entry into the electronic health record and measurements after first MDD diagnosis

Demographic characteristic	**Cohort** **N = 2156**
Age (years), mean ± SD	36.4 ± 13.0
Age (years), n (%)	
18–34	1120 (52.0)
35–49	627 (29.1)
50–64	365 (16.9)
≥65	44 (2.0)
Sex, n (%)	
Female	1288 (59.7)
Male	868 (40.3)
Race, n (%)	
White	1337 (62.0)
Black or African American	688 (31.9)
Unknown	81 (3.8)
Other[a]	47 (2.2)
Multiracial	3(0.1)

(*Continued*)

Table 1 (Continued).

Demographic characteristic	**Cohort** **N = 2156**
Ethnicity, n (%)	
Not Hispanic or Latino	1302 (60.4)
Hispanic or Latino	641 (29.7)
Unknown	213 (9.9)
Measurements, n (%)	
PHQ-9 [Mean ± SD]	12.9 (6.6)
None (0–4)	251 (11.6)
Mild (5–9)	513 (23.8)
Moderate (10–14)	472 (21.9)
Moderately severe (15–20)	509 (23.6)
Severe (20–27)	411 (19.1)
QIDS-C [Mean ± SD]	12.0 (4.9)
None (0–5)	234 (10.9)
Mild (6–10)	585 (27.1)
Moderate (11–15)	846 (39.2)
Severe (16–20)	409 (19.0)
Very Severe (21–27)	82 (3.8)
VQIDS-C [Mean ± SD]	6.2 (3.2)
None (0–2)	171 (7.9)
Mild (3–5)	758 (35.2)
Moderate (6–9)	677 (31.4)
Severe (9–12)	443 (20.5)
Very Severe (13–15)	107 (5.0)

Notes: [a]Other category includes Asian, Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Vietnamese, Chinese, Filipino and Korean.

Abbreviations: PHQ-9, Patient Health Questionnaire-9; QIDS-C, clinician rated Quick Inventory of Depressive Symptomatology; SD, standard deviation; VQIDS-C, 5-item Very Quick Inventory of Depressive Symptomatology.

Classical Test Theory

CTT was used to evaluate the psychometric properties of three scales: PHQ-9, QIDS-C, and VQIDS-C. The internal consistency of the scales was examined using Cronbach's alpha and domain-to-total score correlation, the coefficients for PHQ-9, QIDS-C, and VQIDS-C were 0.9, 0.8, and 0.7, respectively, with the domain-to-total score correlation ranging from 0.3 to 0.8 (Supplemental Table 4). All three scales therefore demonstrated acceptable internal consistency.

Unidimensionality

Parallel analysis was used to assess the unidimensionality of PHQ-9, QIDS-C, and VQIDS-C. For all three scales, the first eigenvalues from the sample data were much larger than that of the randomly generated data (PHQ-9 – 3.4 vs 0.1, QIDS-C – 2.4 vs 0.1 and VQIDS-C – 1.6 vs 0.1). The second eigenvalues from the sample data were much lower than the first values and were similar to or lower than those of the randomly generated data (PHQ-9 – 0.1 vs 0.1, QIDS-C – 0.1 vs 0.1 and VQIDS-C - −0.003 vs 0.1). These results provide evidence of unidimensionality.

Item Response Theory

The IRT parameter estimates for the questions within each scale are shown in Table 2. Questions 1, 2, and 4 of the PHQ-9 pertaining to little interest or pleasure, feeling down, depressed, or hopeless, and feeling tired or having little energy were found to be the most sensitive in distinguishing depression. Question 9 of the PHQ-9, relating to suicidal thoughts, required higher levels of depression to be endorsed. Question 13 of the QIDS-C assessed general interest and was the most sensitive at discriminating depression, this was also found for the VQIDS-C. Both question 12 relating to suicidal

Table 2 Item Response Theory parameter estimates for PHQ-9, QIDS-C, and VQIDS-C (N = 2,156)

	Domain Descriptor	Discrimination	Difficulty		
		A	B0	B1	B2
PHQ-9					
1	Little interest or pleasure in doing things	2.16	−1.06	−0.02	0.82
2	Feeling down, depressed, or hopeless	2.14	−1.63	−0.32	0.57
3	Trouble falling or staying asleep, or sleeping too much	1.33	−1.67	−0.57	0.23
4	Feeling tired or having little energy	1.95	−1.31	−0.18	0.69
5	Poor appetite or overeating	1.49	−0.70	0.18	1.09
6	Feeling bad about yourself – or that you are a failure	1.86	−0.95	−0.04	0.63
7	Trouble concentrating	1.69	−1.00	0.03	0.82
8	Moving or speaking so slowly that other people could have noticed. Or the opposite — being so fidgety or restless. That you have been moving around a lot more than usual	1.27	−0.20	0.91	1.74
9	Thoughts that you would be better off dead, or of hurting yourself	1.17	0.69	2.06	3.03
QIDS-C					
1234	Falling Asleep/Sleep During the Night/Waking Up Too Early/Sleeping Too Much:	0.91	−3.10	−1.94	−0.80
5	Feeling sad	1.56	−2.07	−0.42	1.26
6789	Decreased Appetite Increased Appetite Decreased Weight (Within the Last Two Weeks)/Increased Weight (Within the Last Two Weeks)	0.55	−1.61	1.03	2.57
10	Concentration/decision making	1.53	−1.37	0.10	2.06
11	View of myself	1.62	−0.79	0.32	0.85
12	Thoughts of death or suicide	1.06	0.85	2.85	5.03
13	General interest	1.98	−0.96	0.27	1.40
14	Energy level	1.63	−1.10	0.45	2.16
1516	Feeling slowed down/ Feeling restless	0.88	−0.55	1.82	4.17

(*Continued*)

Table 2 (Continued).

	Domain Descriptor	Discrimination	Difficulty		
		A	B0	B1	B2
VQIDS-C					
5	Feeling Sad	1.45	−2.15	−0.44	1.31
11	View of myself	1.50	−0.83	0.33	0.88
13	General interest	2.21	−0.92	0.25	1.34
14	Energy level	1.82	−1.04	0.43	2.05
15	Feeling slowed down/ Feeling restless	0.86	0.83	2.81	5.28

Abbreviations: PHQ-9, Patient Health Questionnaire-9; QIDS-C, clinician rated Quick Inventory of Depressive Symptomatology; VQIDS-C, 5-item Very Quick Inventory of Depressive Symptomatology.

thoughts, and questions 15 and 16, pertaining to psychomotor symptoms, required higher depression for endorsement. The TIF for PHQ-9, QIDS-C, and VQIDS-C were 30.3, 25.8, and 17.8, respectively, suggesting that PHQ-9 provided the most accurate measure of depressive symptomatology (Figure 1). Table 3 and Figure 2 links the scores between PHQ-9 and QIDS-C, and VQIDS-C. The validation of the estimated scores compared with the true scores is shown in Table 4.

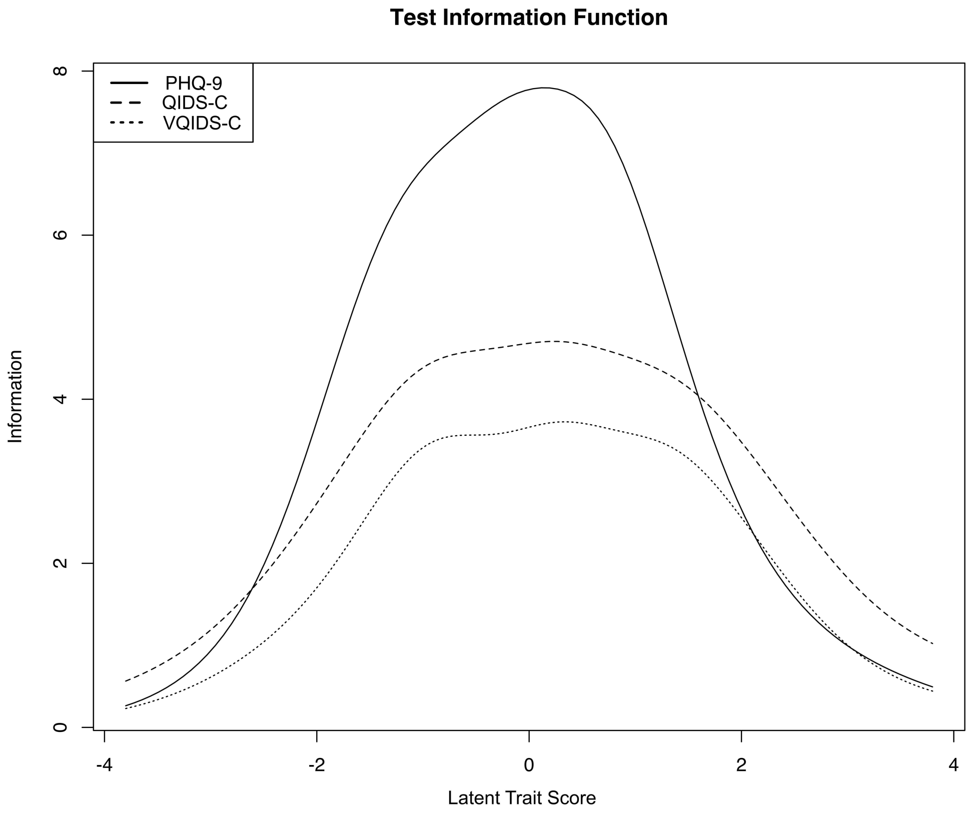

Figure 1 Test Information Functions for the PHQ-9, QIDS-C, and VQIDS-C.
Abbreviations: PHQ-9, Patient Health Questionnaire-9; QIDS-C, clinician rated Quick Inventory of Depressive Symptomatology; VQIDS-C, v5-item Very Quick Inventory of Depressive Symptomatology.

Table 3 Linking of the Summary Total Scores of PHQ-9, QIDS-C, and VQIDS-C

PHQ-9	Equivalent QIDS-C
0	1
1	2
2	4
3	4
4	5
5	6
6	7
7	7
8	8
9	9
10	10
11	11
12	12
13	12
14	13
15	13
16	14
17	15
18	16
19	17
20	17
21	18
22	19
23	20
24	21
25	22
26	23
27	24
PHQ-9	**Equivalent VQIDS-C**
0	0
1	0
2	1

(*Continued*)

Table 3 (Continued).

3	1
4	2
5	2
6	3
7	3
8	4
9	4
10	5
11	5
12	6
13	6
14	6
15	7
16	8
17	8
18	9
19	9
20	10
21	10
22	11
23	12
24	12
25	13
26	14
27	15
QIDS-C	**Equivalent PHQ-9**
0	NA
1	0
2	1
3	2
4	2
5	4
6	5
7	6

(*Continued*)

Table 3 (Continued).

8	8
9	9
10	11
11	12
12	13
13	15
14	16
15	18
16	19
17	20
18	21
19	22
20	23
21	24
22	25
23	26
24	27
25	27
26	NA
27	27
QIDS-C	**Equivalent VQIDS-C**
0	NA
1	0
2	0
3	1
4	1
5	2
6	2
7	3
8	3
9	4
10	5
11	5

(*Continued*)

Table 3 (Continued).

12	6
13	7
14	8
15	8
16	9
17	9
18	10
19	11
20	11
21	12
22	13
23	14
24	15
25	15
26	NA
27	15
VQIDS-C	**Equivalent QIDS-C**
0	2
1	4
2	6
3	7
4	9
5	10
6	12
7	13
8	15
9	16
10	18
11	19
12	21
13	22
14	23
15	24

(*Continued*)

Table 3 (Continued).

VQIDS	Equivalent PHQ-9
0	1
1	2
2	5
3	7
4	9
5	11
6	13
7	15
8	17
9	19
10	21
11	23
12	24
13	25
14	26
15	27

Abbreviations: PHQ-9, Patient Health Questionnaire-9; QIDS-C, clinician rated Quick Inventory of Depressive Symptomatology; VQIDS-C, 5-item Very Quick Inventory of Depressive Symptomatology.

Discussion

The current study assessed the convergent validity, reliability, and interconversion of the PHQ-9, QIDS-C, and VQIDS-C in a large real-world cohort of adults with a non-psychotic MDD diagnosis. Results showed that all the scales were unidimensional and demonstrated good internal consistency. This is consistent with findings from two previous analyses in adult populations, the first in 400 Singaporean primary care patients (Cronbach's alpha: PHQ-9 = 0.87 and QIDS-SR = 0.79)[17] and the second in 297 inpatients in China with a diagnosis of depression (Cronbach's alpha: PHQ-9 = 0.88 and QIDS-SR = 0.83).[33] Our findings are also consistent with a recent evaluation of PHQ-9, QIDS-SR, and VQIDS-SR in adolescents with depression (Cronbach's alpha: PHQ-A (adapted for adolescents) = 0.86, QIDS-SR = 0.80 and VQIDS-SR = 0.76).[34]

All the scale items related to "lack of interest in things" and "feeling tired" had a high level of discrimination. This finding reflects previous research that found "lack of interest in things" was the most sensitive at distinguishing levels of depression for PHQ-9, and "feeling tired" was the most sensitive for QIDS-SR and VQIDS-SR.[32] For the PHQ-9, "feelings of sadness" also highly discriminated levels of depression. This item was less prominent in the QIDS-C and VQIDS-C, with the "view of self" as a highly discriminating factor for these scales. Items related to "suicidality" or "psychomotor symptoms" required a higher degree of depression to endorse.

In alignment with our findings, the symptoms associated with MDD exhibit a notable convergence across various dimensions. This convergence is observed regardless of the assessment method, whether through prevalence assessment or evaluation of symptom severity, and irrespective of whether self-reported (PHQ-9) or assessed by clinician rating (QIDS-C). Specifically, the symptoms harmonize around five pivotal dimensions, encompassing mood, interest, concentration, energy, and self-view. In contrast, psychomotor changes and suicidal ideation demonstrate weaker

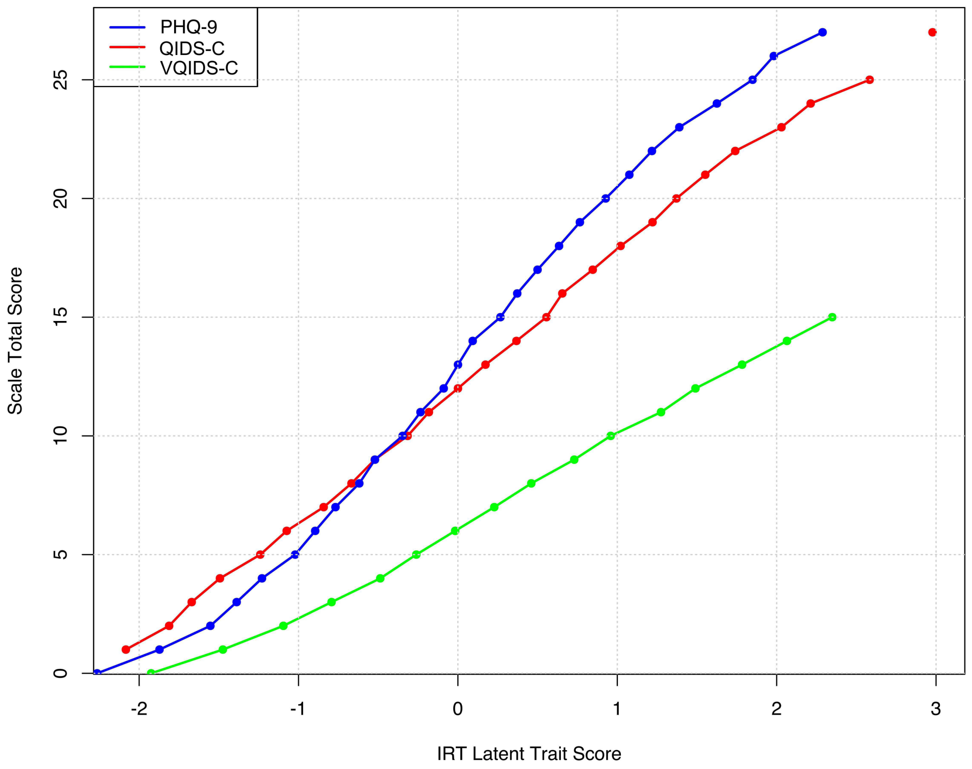

Figure 2 Linking the total PHQ-9, QIDS-C, and VQIDS-C scores based on the IRT latent trait score.
Abbreviations: IRT, Item Response Theory; PHQ-9, Patient Health Questionnaire-9; QIDS-C, clinician rated Quick Inventory of Depressive Symptomatology; VQIDS-C, 5-item Very Quick Inventory of Depressive Symptomatology.

associations, as evidenced by their relatively lower correlation coefficients. The literature supports the idea that suicidal ideation is not universally present in all patients,[35] and its inclusion as a marker of overall depression severity could be questioned due to its cross-diagnostic nature and sensitivity to contextual variables.[36] Similarly, research highlights that psychomotor changes necessitate a higher level of severity for recognition.[37,38] Particularly intriguing is the inclusion of four out of the five core dimensions within the VQIDS-C. This observation prompts consideration for the replacement of the psychomotor aspect derived from the six-item Hamilton with the dimension of concentration, thereby contributing to a more comprehensive representation of the core features of depression.

Table 4 Validation of estimated scores vs actual reported scores

Scale	Correlation	RMSE	Kappa (CI)
QIDS-C (derived from PHQ-9)	0.80	3.17	0.36 (0.34–0.39)
QIDS-C (derived from VQIDS-C)	0.91	2.09	0.55 (0.52–0.58)
PHQ-9 (derived from QIDS-C)	0.81	4.01	0.39 (0.37–0.42)
PHQ-9 (derived from VQIDS-C)	0.76	4.49	0.31 (0.29–0.34)
VQIDS-C (derived from PHQ-9)	0.75	2.39	0.35 (0.32–0.38)
VQIDS-C (derived from QIDS-C)	0.91	1.37	0.56 (0.54–0.59)

Abbreviations: CI, confidence interval; PHQ-9, Patient Health Questionnaire-9; QIDS-C, clinician rated Quick Inventory of Depressive Symptomatology; RMSE, Root Mean Square Error; VQIDS-C, 5-item Very Quick Inventory of Depressive Symptomatology.

This study represents the first effort to link PHQ-9, QIDS-C, and VQIDS-C in an adult population using real-world EHR data, resulting in valuable conversion tables. It is noteworthy to highlight that the conversion from QIDS-C and VQIDS-C to the PHQ-9 appears to exhibit lower reliability compared to the conversion of PHQ-9 scores to QIDS-C and VQIDS-C scores. This observation is underscored by the higher RMSE of 4 in the former case. However, a closer examination of the TIF supports this outcome and suggests that the slightly increased RMSE aligns with the underlying psychometric structure. Notably, our findings also underscore the viability of employing the abbreviated VQIDS-C if there is insufficient time to perform the full QIDS-C, given its considerably reduced administration time while still providing a reasonably accurate approximation of the QIDS-C total score. In addition, the viability of the VQIDS-C conversion to provide an approximation of the QIDS-C could be used to overcome issues of data missingness within EHR based datasets to increase the measures available for the analysis of real-world data. Clinicians and researchers can utilize the conversion tables to replace missing scores of patients with the corresponding scores, thereby maximizing the available data for patients or specific groups of interest. This enables more comprehensive descriptive and inferential analyses, enhancing the depth of understanding and insights gained from the data. It is important to utilize our conversion tables while keeping in mind that the agreement between any two instruments, irrespective of the statistical method used, can be influenced by variations in the population being studied.[39]

There are several limitations of the current study. First, due to the real-world nature of the data source, the administration of the scales was not randomized, so that order effects were not controlled for. Order effects can manifest as primacy effects (early items having a more significant impact) or recency effects (later items having a greater impact).[40] The current data source omits details on the timing of the administration of the scales, however, if these data were available the impact of such order effects could have been more thoroughly investigated. Additionally, the real-world data used in this study were derived from community mental health centres and thus findings may not be generalizable to other settings, such as primary care or population studies. Second, the scales being equated were completed from different perspectives. The PHQ-9 was a patient self-report, whereas the QIDS-C was clinician reported. Third, there was a lack of complete scale range across all scale scores. For example, no patients scored 0 or 26 on the QIDS-C. Fourth, there are inherent differences in the scales such as the period that the scales correspond to, PHQ-9 – 2 weeks and QIDS-C – 7 days or severity versus prevalence. Fifth, limitations of IRT include sample dependence, assumption that item responses are independent, and the assumption that item parameters are invariant across time. These factors should be carefully considered when interpreting the current findings. Sixth, the current study considers only scales recorded at one moment in time, however, previous research has demonstrated the value of linking improvement or worsening from a baseline score.[41] Future research using the scales in this study could consider such a longitudinal perspective. Seventh, although the removal of patients with a 10-point difference between PHQ-9 and QIDS-C was a study requirement, this criterion may have disregarded the discrepancy between the subjective and objective experience of depressive symptomatology. Indeed, previous research has shown that patient and clinician reported depression scale do not always demonstrate agreement,[42–44] and highlight the importance of capturing both viewpoints. Finally, the generalizability and applicability of the conversion tables should be considered. This sample was sourced from specialty mental health settings, where higher levels of comorbidity and more severe depression might be found compared with primary care settings or epidemiological samples.

Conclusion

PHQ-9, QIDS-C, and VQIDS-C were equated using real-world data from the EHRs of patients diagnosed with non-psychotic MDD. Establishing a link between these scales has the potential to open up new opportunities for measurement integration, research collaboration, and improved patient care. For clinicians and researchers, the conversion tables developed in this study may have several uses, including to address missing data in a patient's clinical history and to harness data from multiple sources to produce larger more representative cohorts for research studies. Despite the potential benefits of linking these scales it is important not to overlook the value of capturing multiple dimensions when measuring psychiatric constructs such as depression. We have demonstrated that variation in the perspective of the individual completing the scale, in addition to prevalence and severity of symptoms may deferentially contribute to depressive symptomology.

Abbreviations

MDD, Major depressive disorder; PHQ-9, Patient Health Questionnaire-9; QIDS-C - Quick Inventory of Depressive Symptomatology-Clinician Report; DSM-IV, Diagnostic and Statistical Manual of Mental Disorders; ICD-10, International Classification of Disease – Version 10; VQIDS-C, 5-item Very Quick Inventory of Depressive Symptomatology- Clinician Report; EHR, Electronic Health Record; CTT, Classical Test Theory; IRT, Item Response Theory; IQR, Interquartile Range; SD, Standard Deviation; RMSE, Root Mean Square Error; IRCCC, Item Response Category Characteristic Curve; MÅDRS, Montgomery-Åsberg Depression Rating Scale; SRS, Self Rated Scale; BDI-II), The Beck Depression Inventory-II; HAM-D, Hamilton Rating Scale for Depression.

Data Sharing Statement

The data supporting this study originate with Holmusk Technologies, Inc. These de-identified data may be made available upon request and are subject to license agreement with Holmusk. Interested parties should contact publications@holmusk.com to determine licensing terms.

Ethics Approval

This study was conducted in accordance with the 1964 Declaration of Helsinki and its subsequent amendments. Institutional review board (IRB) approval of the study protocol, including a waiver of HIPAA authorization, was obtained prior to study conduct, and covers data originating from all sites represented. Approval was granted by the WCG IRB. (The Holmusk Real-World Evidence Parent Protocol; IRB registration number 1-1470336-1; Protocol ID HolmuskRWE_1.0).

Acknowledgments

Cody Patton (Holmusk Technologies, Inc) provided editorial support.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work was funded by Holmusk Technologies, Inc.

Disclosure

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Emily OC Palmer reports a relationship with Holmusk Europe, Ltd. that includes: employment. Sheryl Ker reports a relationship with KKT Technologies, Pte. Ltd. that includes: employment. At the time the study was conducted Miguel E Rentería reports a relationship with KKT Technologies, Pte. Ltd. that includes: employment. A. John Rush has received consulting fees from Compass Inc., Curbstone Consultant LLC, Emmes Corp., Evecxia Therapeutics, Inc., Holmusk Technologies, Inc., ICON, PLC, Johnson and Johnson (Janssen), John Peter Smith Foundation, Liva-Nova, MindStreet, Inc., Neurocrine Biosciences Inc., Otsuka-US, Singapore Ministry of Health; speaking fees from Liva-Nova, Johnson and Johnson (Janssen); and royalties from Wolters Kluwer Health, Guilford Press and the University of Texas Southwestern Medical Center, Dallas, TX (for the Inventory of Depressive Symptoms and its derivatives). He is also named co-inventor on two patents: US Patent No. 7,795,033: Methods to Predict the Outcome of Treatment with Antidepressant Medication, Inventors: McMahon FJ, Laje G, Manji H, Rush AJ, Paddock S, Wilson AS; and US Patent No. 7,906,283: Methods to Identify Patients at Risk of Developing Adverse Events During Treatment with Antidepressant Medication, Inventors: McMahon FJ, Laje G, Manji H, Rush AJ, Paddock S. Thomas Carmody has received consulting fees from Alkermes, Inc. and Holmusk Technologies, Inc. The authors report no other conflicts of interest in this work.

References

1. Hasin DS, Sarvet AL, Meyers JL, et al. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry.* 2018;75(4):336–346. doi:10.1001/jamapsychiatry.2017.4602
2. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry.* 2003;54(5):573–583. doi:10.1016/s0006-3223(02)01866-8
3. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing; 2013.
4. 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. doi:10.1046/j.1525-1497.2001.016009606.x
5. Martin A, Rief W, Klaiberg A, Braehler E. Validity of the brief patient health questionnaire mood scale (PHQ-9) in the general population. *Gen Hosp Psychiatry.* 2006;28(1):71–77. doi:10.1016/j.genhosppsych.2005.07.003
6. Allgaier A-K, Pietsch K, Frühe B, Sigl-Glöckner J, Schulte-Körne G. screening for depression in adolescents: validity of the patient health questionnaire in pediatric care. *Depression and Anxiety.* 2012;29(10):906–913. doi:10.1002/da.21971
7. Arthurs E, Steele RJ, Hudson M, Baron M, Thombs BD, Group CSR. Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning. *PLoS One*. 2012;7(12):e52028. doi:10.1371/journal.pone.0052028
8. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry.* 2008;8(1):1–7. doi:10.1186/1471-244X-8-46
9. Miller AP, Espinosa da Silva C, Ziegel L, et al. Construct validity and internal consistency of the patient health questionnaire-9 (PHQ-9) depression screening measure translated into two Ugandan languages. *Psychiatry Res Communications*. 2021;1(2):100002. doi:10.1016/j.psycom.2021.100002
10. Muramatsu K, Miyaoka H, Kamijima K, et al. Performance of the Japanese version of the patient health questionnaire-9 (J-PHQ-9) for depression in primary care. *Gen Hosp Psychiatry.* 2018;52:64–69. doi:10.1016/j.genhosppsych.2018.03.007
11. Sawaya H, Atoui M, Hamadeh A, Zeinoun P, Nahas Z. Adaptation and initial validation of the patient health questionnaire–9 (PHQ-9) and the generalized anxiety disorder–7 questionnaire (GAD-7) in an Arabic speaking Lebanese psychiatric outpatient sample. *Psychiatry Res*. 2016;239:245–252. doi:10.1016/j.psychres.2016.03.030
12. Wang W, Bian Q, Zhao Y, et al. Reliability and validity of the Chinese version of the patient health questionnaire (PHQ-9) in the general population. *Gen Hosp Psychiatry.* 2014;36(5):539–544. doi:10.1016/j.genhosppsych.2014.05.021
13. Woldetensay YK, Belachew T, Tesfaye M, et al. Validation of the patient health questionnaire (PHQ-9) as a screening tool for depression in pregnant women: afaan oromo version. *PLoS One*. 2018;13(2):e0191782. doi:10.1371/journal.pone.0191782
14. Reilly TJ, MacGillivray SA, Reid IC, Cameron IM. Psychometric properties of the 16-item quick inventory of depressive Symptomatology: a systematic review and meta-analysis. *J Psychiatr Res*. 2015;60:132–140. doi:10.1016/j.jpsychires.2014.09.008
15. Trivedi MH, Rush AJ, Ibrahim HM, et al. The inventory of depressive symptomatology, clinician rating (IDS-C) and self-report (IDS-SR), and the quick inventory of depressive symptomatology, clinician rating (QIDS-C) and self-report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychol Med*. 2004;34(1):73–82. doi:10.1017/s0033291703001107
16. De La Garza N, John Rush A, Grannemann BD, Trivedi MH. Toward a very brief self-report to assess the core symptoms of depression (VQIDS-SR(5)). *Acta Psychiatr Scand*. 2017;135(6):548–553. doi:10.1111/acps.12720
17. Sung SC, Low CCH, Fung DSS, Chan YH. Screening for major and minor depression in a multiethnic sample of A sian primary care patients: a comparison of the nine-item patient health questionnaire (PHQ-9) and the 16-item quick inventory of depressive symptomatology–self-report (QIDS-SR16). *Asia-Pacific Psychiatry.* 2013;5(4):249–258. doi:10.1111/appy.12101
18. Ma S, Yang J, Yang B, et al. The patient health questionnaire-9 vs. the Hamilton rating scale for depression in assessing major depressive disorder. *Front Psychiatry.* 2021;12:747139. doi:10.3389/fpsyt.2021.74713
19. Hawley CJ, Gale TM, Smith PS, et al. Equations for converting scores between depression scales (MÅDRS, SRS, PHQ-9 and BDI-II): good statistical, but weak idiographic, validity. *Hum Psychopharmacol*. 2013;28(6):544–551. doi:10.1002/hup.2341
20. Patel R, Wee SN, Ramaswamy R, et al. NeuroBlu, an electronic health record (EHR) trusted research environment (TRE) to support mental healthcare analytics with real-world data. *BMJ Open*. 2022;12(4):e057227. doi:10.1136/bmjopen-2021-057227
21. Bernstein IH, Rush AJ, Carmody TJ, Woo A, Trivedi MH. Clinical vs. self-report versions of the quick inventory of depressive symptomatology in a public sector sample. *J Psychiatr Res*. 2007;41(3):239–246. doi:10.1016/j.jpsychires.2006.04.001
22. Rush AJ, Madia ND, Carmody T, Trivedi MH. Psychometric and clinical evaluation of the clinician (vqids-c(5)) and self-report (vqids-sr(5)) versions of the very quick inventory of depressive symptoms. *Neuropsychiatr Dis Treat*. 2022;18:289–302. doi:10.2147/ndt.S342457
23. Kyle PR, Lemming OM, Timmerby N, Søndergaard S, Andreasson K, Bech P. The validity of the different versions of the Hamilton depression scale in separating remission rates of placebo and antidepressants in clinical trials of major depression. *J Clin Psychopharmacol*. 2016;36(5):453–456. doi:10.1097/jcp.0000000000000557
24. Huang X-J, Ma H-Y, Wang X-M, Zhong J, Sheng D-F, Xu M-Z. Equating the PHQ-9 and GAD-7 to the hads depression and anxiety subscales in patients with major depressive disorder. *J Affective Disorders*. 2022;311:327–335. doi:10.1016/j.jad.2022.05.079
25. Velozo CA, Byers KL, Wang Y-C, Joseph BR. Translating measures across the continuum of care: using rasch analysis to create a crosswalk between the functional independence measure and the minimum data set. *J Rehabil Res Dev.* 2007;44(3):467. doi:10.1682/JRRD.2006.06.0068
26. Toland MD. Practical guide to conducting an item response theory analysis. *J Early Adolesc*. 2013;34(1):120–151. doi:10.1177/0272431613511332
27. Humphreys LG, Montanelli RG. An investigation of the parallel analysis criterion for determining the number of common factors. *Multivar Behav Res*. 1975;10(2):193–205. doi:10.1207/s15327906mbr1002_5
28. Samejima F. *Graded Response Model of the Latent Trait Theory and Tailored Testing*. Washington DC: US Government Printing Office; 1976:5–17.

29. Orlando M, Sherbourne CD, Thissen D. *Summed-Score Linking Using Item Response Theory: Application to Depression Measurement*. US: American Psychological Association; 2000:354–359.
30. Thissen D, Pommerich M, Billeaud K, Williams VS. Item response theory for scores on tests including polytomous items with ordered responses. *Appl Psychol Meas*. 1995;19(1):39–49. doi:10.1177/014662169501900105
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174. doi:10.2307/2529310
32. US Department of Health and Human Services. Health information privacy. Available from: https://www.hhs.gov/hipaa/index.html. Accessed October 11, 2023
33. Feng Y, Huang W, Tian TF, et al. The psychometric properties of the quick inventory of depressive symptomatology-self-report (QIDS-SR) and the Patient health questionnaire-9 (PHQ-9) in depressed inpatients in China. *Psychiatry Res*. 2016;243:92–96. doi:10.1016/j.psychres.2016.06.021
34. Nandy K, Rush AJ, Carmody T. Reducing the clinical-research gap: a comparison of depressive symptom self-reported measures in psychiatric outpatient youth. *J Clin Psych*. 2023;85. doi:10.4088/JCP.23m14861
35. Bostwick JM, Pankratz VS. Affective disorders and suicide risk: a reexamination. *Am J Psychiatry*. 2000;157(12):1925–1932. doi:10.1176/appi.ajp.157.12.1925
36. Fergusson DM, Woodward LJ, Horwood LJ. Risk factors and life processes associated with the onset of suicidal behaviour during adolescence and early adulthood. *Psychol Med*. 2000;30(1):23–39. doi:10.1017/s003329179900135x
37. Parker G. Defining melancholia: the primacy of psychomotor disturbance. *Acta Psychiatr Scand*. 2007;115(s433):21–30. doi:10.1111/j.1600-0447.2007.00959.x
38. Parker G, Bassett D, Outhred T, et al. Defining melancholia: a core mood disorder. *Bipolar Disorders*. 2017;19(3):235–237. doi:10.1111/bdi.12501
39. Dorans NJ. Linking scores from multiple health outcome instruments. *Qual Life Res*. 2007;16:85–94. doi:10.1007/s11136-006-9155-3
40. Bradley JV. Complete counterbalancing of immediate sequential effects in a Latin square design. *J Am Stat Assoc*. 1958;53(282):525–528. doi:10.1080/01621459.1958.10501456
41. Leucht S, Rothe P, Davis JM, Engel RR. Equipercentile linking of the BPRS and the PANSS. *Eur Neuropsychopharmacol*. 2013;23(8):956–959. doi:10.1016/j.euroneuro.2012.11.004
42. Sakurai H, Suzuki T, Yoshimura K, Mimura M, Uchida H. Predicting relapse with individual residual symptoms in major depressive disorder: a reanalysis of the STAR*D data. *Psychopharmacology*. 2017;234(16):2453–2461. doi:10.1007/s00213-017-4634-5
43. Zimmerman M, Martinez JA, Attiullah N, et al. Why do some depressed outpatients who are in remission according to the Hamilton depression rating scale not consider themselves to be in remission? *J Clin Psychiatry*. 2012;73(6):790–795. doi:10.4088/JCP.11m07203
44. Tada M, Uchida H, Suzuki T, Abe T, Pollock BG, Mimura M. Baseline difference between patients' and clinicians' rated illness severity scores and subsequent outcomes in major depressive disorder: analysis of the sequenced treatment alternatives to relieve depression data. *J Clin Psychopharmacol*. 2014;34(3):297–302. doi:10.1097/JCP.0000000000000112