Measurement Matters - Ascertaining Response to Depression Treatment in Primary Care

Andrew D. Carlo

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Public Health

University of Washington

2019

Committee:

Jürgen Unützer

Michelle Garrison

Program Authorized to Offer Degree:

Health Services

University of Washington

Abstract

Measurement Matters - Ascertaining Response to Depression Treatment in Primary Care

Andrew D. Carlo

Chair of the Supervisory Committee:
Jürgen Unützer
Departments of Global Health & Health Services

Introduction/Background: Depression is a common behavioral health problem that has been linked to elevated disability, morbidity, mortality and medical spending. Although there are numerous evidence-based treatments for depression in various settings, the accuracy and reliability of symptom measurement remain elusive, making it difficult to ascertain clinical progress. There are inherent flaws in the longitudinal use of validated symptom scales, such as the Patient Health Questionairre-9 (PHQ-9), and the literature is divided on how to interpret changes in patients' scores over time. Different metrics have been used to denote depression "response" or "remission," but they have had minimal empirical validation. Further, it remains unclear how changes in rating scale scores impact patient-centered outcomes (PCOs), such as quality of life and social connectivity.

Methods: This investigation was conducted in two phases. The first was via a secondary analysis of data from the Improving Mood–Promoting Access to Collaborative Treatment (IMPACT) trial, an 1800-participant RCT of Collaborative Care for the treatment of depression in primary care. In IMPACT, all participants were surveyed and interviewed with a variety of evidence-based instruments at baseline and periodically. PHQ-9 scores were also tracked for the treatment group (n=906). Baseline and follow-up surveys for all participants included, among other instruments, the Health Information National Trends Survey (HINTS), the Strengths Self-Efficacy Scale (SSES), a self-reported Quality of Life measure

and the Sheehan Disability Scale (SDS). For the treatment group only, we calculated response rates and analyzed the average associated change in general health, social functioning, quality of life and degree of disability (with simple linear regression significance testing) across nine PHQ-9 depression "response" and "remission" metrics and three intervals of time (3, 6 and 12 months). Pair-wise tests for equivalence in mean differences in outcome for all combinations of single-component PHQ-9 depression "response" and "remission" metrics were also conducted using ANOVA with specified linear contrasts. The second phase was analyzing data collected by the University of Washington's Advancing Integrated Mental Health Solutions (AIMS) center from 42 health care organizations serving more than 11,000 patients in real-world collaborative care implementations nationwide. Using the same nine PHQ-9 "response" and "remission" metrics, we ranked all 42 health care organizations with direct sample means and a multilevel logistic regression model with empirical Bayes predictions. For both methods, correlations between all pair-wise rank orders tests for significance were conducted using the Spearman's rank-order correlation coefficient.

Results: Using PHQ-9 data from IMPACT, choice of metric had a substantial impact on depression treatment response rates, with 3- 6- and 12-month rates ranging from 32.3-76.7%, 37.9-82.2% and 42.1-84.9%, respectively. Further, PCOs such as general health, social functioning, quality of life and degree of disability change all demonstrated statistically significant improvement with depression response or remission. Three metrics were top-performing with respect to associated PCO differences- (1) $\geq$50% decrease from baseline, (2) $\geq$ 50% decrease from baseline AND score of <10 and (3) Score <5 (remission). However, when combinations of the five single-component metrics were tested for equivalence, most pair-wise comparisons were found to be statistically insignificant. Remission (score < 5) was more consistently different from the single-component response metrics. Using real-world AIMS

data, organization-level rankings differed across depression response and remission metrics, though most were highly correlated. Of all metrics, remission was least positively correlated with the others.

Conclusions: Our findings demonstrate that choice of depression response metric substantially impacts observed response rates and that PCOs tend to improve with depression. However, our comparative associations between depression response/remission metrics and patient-centered outcomes were mixed overall and we are consequently unable to suggest an optimal metric. Organization-level rankings for depression response/remission vary depending on choice of metric, but their rankings are highly correlated. Given the increasing global focus on population health and measurement-based care, future research should prioritize the determination of an optimal, pragmatic metric for depression outcome ascertainment.

Background:

Depression is a common behavioral health problem that affects between five and ten percent of adult patients seen in the primary care setting[1]. Previous research has clearly demonstrated that depression is a leading cause of disability, lower quality of life, diminished productivity and reduced employment rates globally[2]. In the United States, a recent study estimated that the total economic burden of major depressive disorder (MDD) has reached at least $210.5 billion per year, a 21.5% increase from 2005[3].

Although plentiful evidence-based treatments have been validated for the treatment of depression in primary care, hospital and specialty mental health settings, it remains challenging to accurately and reliability measure outcome progress. This is for a variety of reasons: (1) depression symptoms are capricious (often improving or worsening rapidly with or without improvement[4,5]), (2) there are inherent flaws in the longitudinal use of validated symptom scales (primarily due to a lack of demonstrated unidimensionality or measurement invariance[6]), (3) the literature is divided on how to interpret changes in patients' scores over time, and (4) it remains unclear exactly how changes in rating scale scores impact the outcomes most important to patients (e.g., quality of life and social connectivity). Nevertheless, outcome ascertainment for depression treatment is becoming increasingly important as measurement-based care, population health and value-based payment become priorities for health organizations nationwide[7]. There is a consequent growing need for efficient, pragmatic, and scalable ways to assess treatment progress.

Of the numerous validated rating scales for depression (e.g., the Hamilton Depression Rating Scale (HAM-D)[8], the Beck Depression Inventory (BDI)[9], the Symptom Checklist–20 (SCL-20)[10], and the Patient Health Questionairre-9 (PHQ-9))[11], the PHQ-9 has consistently been one of the most used and

validated in primary care, specialty behavioral health and research settings[12–14]. As a result, it will be the instrument of choice for this investigation. The PHQ-9 has nine items that are each scored from zero to three, for a maximum score of 27. See **Table 1** for recommended interpretations of PHQ-9 scores.

Table 1: PHQ-9 Score Ranges and Interpretations[15]

Score	Depression Severity	Comment
0-4	Minimal or None	Monitor; may not require treatment
5-9	Mild	Use clinical judgment (symptom duration, functional impairment) to determine necessity of treatment
10-14	Moderate	
15-19	Moderately Severe	Warrants active treatment with psychotherapy, medications, or combination
20-27	Severe	

Oftentimes, the PHQ-9 and other scales are used as a way to ascertain the severity of baseline depression symptoms and to track patients' progress over time with treatment. In the literature, treatment is usually quantified using the terms "response" (or "partial response") and "remission." Although not ubiquitously defined in this way, "remission" (on the PHQ-9 scale) is often defined as achieving a score of less than five (for a patient with a previous score in a category suggestive of depression symptoms)[11,16]. The literature is much more divided on treatment "response," with many different definitions being described across studies[17,18]. Some include a single term, while others are compound (including more than one term). Definitions also vary in whether or not they specify a minimum baseline PHQ-9 score[17].

Table 2: Depression Response and Remission Metrics			
Depression Response	**Metrics with a single term from the published literature**	**1**	Reduction from baseline PHQ-9 of ≥ 50%[16]
		2	Absolute reduction from baseline PHQ-9 of ≥ 5 points[14]
		3	PHQ-9 score of < 10[11,16]
	Compound metrics from the published literature	**4**	Reduction from baseline PHQ-9 of ≥ 50% OR PHQ-9 score of < 10[19]
		5	Reduction from baseline PHQ-9 of ≥ 50% AND PHQ-9 score of < 10[11,15,17]
		6	Absolute reduction from baseline PHQ-9 score of ≥ 5 points AND PHQ-9 score of < 10[17,20]
	Additional potential single term and compound metrics	**7**	Absolute reduction from baseline PHQ-9 of ≥ 10 points
		8	Reduction from baseline PHQ-9 of ≥ 50% AND Absolute reduction from baseline PHQ-9 score of ≥ 5 points
Depression Remission		**9**	PHQ-9 score of < 5[11,15,17]

Each of the above metrics in **Table 2** has potential strengths and weaknesses for assessment of depression treatment, but no known studies to date have determined which of these best correlates with clinical improvement in patient-centered outcomes, such as self-reported health, social connectivity and disability burden[21]. One seminal study assessed the construct validity of the PHQ-9 using the 20-item Short-Form General Health Survey (SF-20), self-reported counts of sick days and clinic visits, and symptom-related difficulty[15]. Findings demonstrated that higher PHQ-9 scores were associated with substantially lower functional status on all six SF-20 sub-scales[15]. This data, of note, was cross-sectional and did not assess response to depression treatment over time. Another study secondarily analyzed randomized controlled trial (RCT) data to determine the minimal clinically important difference (MCID) for the PHQ-9, estimating that it was between 2.59 and 4.78[14]. This led the authors to conclude that PHQ-9 score changes of five or greater reflect a clinically relevant change in individuals undergoing depression treatment.

One final study used data from a 114-person collaborative care RCT[17,22] to compare the initial depression response metric proposed by Kroenke and colleagues[15] (decrease in baseline PHQ-9 score by 50% AND a score of less than 10) to structured interviews and three other depression metrics derived from previous studies on the concepts of reliable and clinically significant change[23]. In general, all measured metrics were found to have good agreement (kappa>0.60)[17]. The authors also reported that metrics combining multiplicative terms (50% change) or absolute terms (five-point change or greater) with the requirement of a score less than ten tended to classify the same patients as improved or not improved[17]. At the same time, they suggested that the absolute change metric may be preferable for multiple reasons, one of which is that it has been validated as the MCID in previous studies[17]. These findings, however, are limited by the study's relatively small sample size and availability of PHQ-9 scores at only two time points - baseline and three months.

In this investigation, the authors aim to contribute to the published literature by comparing the associations of eight depression response metrics and one depression response metric with clinical improvement in patient-centered outcomes (PCOs) over a twelve-month observation period **(Table 3)**. The authors secondarily analyze data from the multi-center IMPACT study[24], a 2002 RCT with 1,801 participants that studied later-life depression outcomes for patients treated with the Collaborative Care Model (CoCM). Additionally, the authors leverage longitudinal PHQ-9 data from 11,000 patients across nine states collected by the University of Washington's Advancing Integrated Mental Health Solutions (AIMS) Center[25] to analyze the extent to which different depression response and remission metrics influence organization-level treatment outcome rank order.

Table 3: Patient-Centered Outcomes (Dependent Variables) and Corresponding Questions from IMPACT Interview					
Variable	**Outcome**	**Question**	**Response Type**	**Levels**	**Source**
1	General Health	In general, would you say your health is:	Categorical	Excellent, Very Good, Good, Fair, Poor, Don't know, Refuse	SF-36[26]
2	Strengths and Resources	Over the past two weeks, how often have you made an effort to keep in touch with relatives or friends?	Categorical	Not at all, Occasionally, Sometimes, Fairly Often, Very Often, Don't know, Refuse	Self-Efficacy Questions[27]
3	Quality of Life	Please rate your quality of life during the past month on a scale from 0 to 10.	Continuous	0-10, Don't know, Refuse	IMPACT[24]
4	Sheehan Disability	To what extent has your health interfered with your work, including paid work or work around the house, in the past month?	Continuous	0 (Not at all) - 10 (Unable to carry on any activities), Don't know, Refuse	Sheehan Disability Scale[28]
	Sheehan Disability	To what extent has your health interfered with your family life, in the past month?	Continuous	0 (Not at all) - 10 (Unable to carry on any activities), Don't know, Refuse	Sheehan Disability Scale[28]
	Sheehan Disability	To what extent has your health interfered with your social life or relationships with others outside of your family, in the past month?	Continuous	0 (Not at all) - 10 (Unable to carry on any activities), Don't know, Refuse	Sheehan Disability Scale[28]

Methods:

IMPACT Analyses:

Data Source:

Conducted between 1999 and 2001, the Improving Mood–Promoting Access to Collaborative Treatment (IMPACT) trial is the largest RCT of CoCM to date, consisting of 1801 participants from a

sample of 7 sites representing 8 diverse health organizations[24]. All study patients were treated in primary care and were randomized to usual care or CoCM, which is an integrated care approach where psychiatric consultants work with a team of mental health and primary care providers to treat common mental health and addiction problems. Assessment interviews were completed on all randomized patients at baseline and at 3, 6, and 12 months for depression, depression treatments, satisfaction with care, functional impairment, quality of life and a variety of other outcomes[24]. The SCL-20 was the study's instrument of choice for depression outcome measurement, with response being defined as a decrease of greater than or equal to 50% from baseline. In addition to the SCL-20 and other outcomes measured as part of ongoing study interviews by research staff, intervention arm subjects (n=906) were assessed longitudinally using the PHQ-9. These assessments were conducted by clinical staff as part of clinical treatment and were recorded in the Care Management Tracking System (CMTS)[29], a web-based registry that was originally developed for IMPACT[30] and has since been extensively used to track depression outcomes in diverse primary care clinics implementing CoCM[31–39].

Study Population:

The population for this secondary data analysis will be those in the intervention arm of the IMPACT trial (n=906). This is because these subjects have baseline and follow-up PHQ-9 scores (in addition to the SCL-20), while the control arm had SCL-20 scores only. Since the PHQ-9 has become the standard instrument for depression assessment in primary care, results will have greater public health impact if based on this instrument.

Independent Variables:

PHQ-9 assessments were conducted by clinical staff and recorded in the CMTS patient registry as part of ongoing care. Unlike SCL-20 scores and other outcomes, which were collected at specific time periods by research staff, PHQ-9 scores were not obtained on a defined timeline. Participants with only one PHQ-9 score were excluded from the analysis. To build the data frame for this investigation, the authors used the date of each patient's first PHQ-9 score as time zero. PHQ-9 assessments closest to three, six and twelve months (within ±45 days) were used as scores for each respective time point. In this way, participants' longitudinal PHQ-9 data were adapted into a data frame with scores at baseline, three, six and twelve months. For each of the nine different response and remission metrics listed above in **Table 2**, binary variables (1/0) were created to indicate whether or not participants responded to depression treatment between baseline and three, six and twelve months. Percentage response was calculated for each metric.

Dependent Variables:

The four dependent variables in this investigation were derived from four different questions or survey instruments recorded longitudinally in the IMPACT trial. See **Table 3** for details on relevant questions, as well as their levels and sources. Three questions from the Sheehan Disability Scale[28] were summed together to create a composite disability metric. For each of the four dependent variables, differences in scores were calculated between baseline and three, six and twelve months for each participant.

Analyses:

Using simple linear regression models with no covariate adjustment, the difference-in-differences in outcomes (with p-values) were determined for each combination of PHQ-9 response or remission

metric and dependent variable at each time interval (baseline to three, six and twelve months). For example, the difference-in-difference in reported quality of life between baseline and three months was calculated between participants who did and did not have a ≥ 50% decrease in PHQ-9 score over that same time period. Means of the difference-in-differences were calculated for each patient-centered outcome across the three time periods for a total of twelve outcome-specific time points. Each depression response or remission metric was color-coded based on whether it was above or below the average (red for below, green for above).

To determine whether associated changes in dependent variables were significantly different among metrics, pair-wise tests for equivalence were conducted using ANOVA with specified linear contrasts for all combinations of the four single-component response metrics and the remission metric. Due to sample size limitations, these tests were not conducted on the multi-component (compound) metrics. See **Figure 1** for visualization of the four quadrants created by pair-wise metric comparisons and the null hypothesis for each AVOVA test.

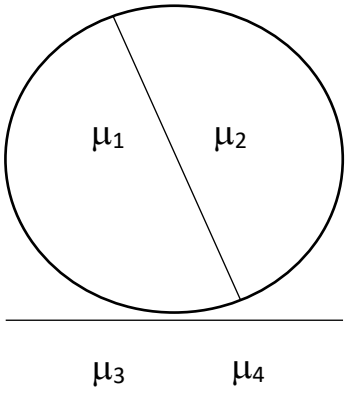


Figure 1: Visualization of Pair-Wise Metric Comparison between Metrics A and B

μ_1 = Metric A (1), Metric B (0)
μ_2 = Metric A (1), Metric B (1)
μ_3 = Metric A (0), Metric B (0)
μ_4 = Metric A (0), Metric B (1)

$$H_o: \frac{P_1\mu_1 + P_2\mu_2}{P_1 + P_2} - \frac{P_3\mu_3 + P_4\mu_4}{P_3 + P_4} = \frac{P_2\mu_2 + P_4\mu_4}{P_2 + P_4} - \frac{P_1\mu_1 + P_3\mu_3}{P_1 + P_3}$$

μ_j = Outcome
Pj = Proportion of participants meeting condition

All calculations were conducted with R base code[40].

AIMS Analysis:

Data Source:

In the years following IMPACT, the UW Department of Psychiatry and Behavioral Sciences founded the Advancing Integrated Mental Health Solutions (AIMS) Center[25] to offer implementation support and ongoing assistance for practices implementing CoCM. Care managers in all participating clinics used CMTS to record contacts with patients and to track PHQ-9 depression scores over time. With the consent of participating organizations, the UW Department of Psychiatry and Behavioral Sciences compiled a dataset of 11,303 adult patients with depressive symptoms who were treated in one of 144 primary care clinics and had depression outcomes tracked between 2008 and 2018. The dataset includes all patients 18 years of age or older who had at least two documented PHQ-9 assessment scores - one at baseline and one or more within the following twelve months. Analysis of this de-identified dataset was granted exemption status by the University of Washington Institutional Review Board (ID # STUDY00005907).

Study Population: All organizations in the dataset had at least one clinic implementing CoCM. Within organizations, all clinics with at least ten CoCM patients enrolled were included. At each outcome time point, a clinic's data was included if at least one patient had a recorded PHQ-9 at that time period. For example, one clinic could have 35 total patients enrolled in CoCM and 25, 14 and zero recorded PHQ-9 outcomes at three, six and twelve months respectively. This clinic would therefore be included in the denominator for the three and six-month outcomes, but not for the twelve-month outcome.

Independent Variables: For this analysis, one depression remission and eight depression response metrics were used as the independent variables, including all metrics from **Table 2**.

Dependent Variables: None

Analyses: All 42 organizations in the sample were ranked using two different methods according to their improvement rates across six depression metrics. First, organization-level ranks were calculated using a non-parametric sample means approach. Next, ranks were calculated using an unadjusted multilevel logistic regression model to account for the clustering of clinics within organizations and empirical Bayes predictions. For each method (sample means and multilevel logistic regression), correlations between all pair-wise rank orders tests for significance were conducted using the Spearman's rank-order correlation coefficient.

Results:

IMPACT Data

Depression Response and Remission:

IMPACT had 1812 patients, 906 of which were randomized to the treatment and control groups. Only the treatment group patients (n=906) were eligible to have PHQ-9 scores measured and, therefore, only this cohort was analyzed as a part of this investigation. Of the 906 treatment group participants, 871 had an initial PHQ-9 value, 869 had at least one follow-up PHQ-9 value and 838 had at least one PHQ-9 score within 45 days of 3-months, 6-months or 12-months. As seen in **Table 4**, 830, 796 and 701 participants had PHQ-9 scores specifically within 45 days of three, six and twelve months, respectively.

Response and remission rates varied considerably across metrics for all time points. At three months, response rates ranged from 32.3% to 76.7%, while remission was 45.7%. Six- and twelve-month values

showed incremental increases across all metrics **(see Table 4)**. Of the six metrics for response that have been described previously in the literature, three-month rates ranged from 58.0% to 76.7%. One proposed metric for response (absolute decrease in PHQ- score of ≥10) was actually met by a lower percentage of participants than that of remission, indicating how comparatively rare it is to have a change of that magnitude.

Table 4: Depression Response and Remission Rates After 3, 6 and 12 Months Across 9 Different Metrics (n=869)

		Described in Previous Literature?	3-month response (n=830)	6-month response (n=796)	12-month response (n=701)
1	**≥ 50% decrease from baseline**	Yes	60.6%	64.9%	70.3%
2	**Absolute decrease of ≥ 5**	Yes	66.5%	70.9%	74.9%
3	**Score of <10**	Yes	75.5%	80.9%	84.2%
4	**≥ 50% decrease from baseline OR score of <10**	Yes	76.7%	82.2%	84.9%
5	**≥ 50% decrease from baseline AND score of <10**	Yes	59.4%	63.7%	69.6%
6	**Absolute decrease of ≥5 AND score <10**	Yes	58.0%	63.4%	68.3%
7	**Absolute decrease of ≥10**	No	32.3%	37.9%	42.1%
8	**≥ 50% decrease from baseline AND absolute decrease of ≥ 5**	No	55.4%	60.2%	65.0%
9	**Score of < 5**	Yes	45.7%	50.0%	58.8%

Patient-Centered Outcome (PCO) Measures:

As shown in **Table 5**, mean General Health scores stayed relatively constant over time, but ultimately demonstrated a small decrease between baseline and twelve months. On average, Strengths and Resources and Quality of life increased over the course of the study period. All three Sheehan Disability items examined in this investigation demonstrated mean decreases between baseline and twelve

months **(Table 6)**. The counts of participants with missing (Don't know or Refused) patient-centered outcome data remained the same across all measures.

Table 5: Distribution of Scores for Patient-Centered Outcome Measures (n=838) - GH, SR and QOL										
Outcome	**Level**	**Description**	**Baseline**		**3-Months**		**6-Months**		**12-Months**	
			n	**%**	**n**	**%**	**n**	**%**	**n**	**%**
General Health	1	Excellent	48	5.73	41	4.89	28	3.34	32	3.82
	2	Very Good	137	16.3	177	21.1	167	19.9	186	22.2
	3	Good	295	35.2	300	35.8	305	36.4	295	35.2
	4	Fair	255	30.4	233	27.8	248	29.6	222	26.5
	5	Poor	103	12.3	85	10.1	81	9.67	74	8.83
	8/9	DK/REF	0	0	2	0.239	9	1.07	29	3.46
	Mean		3.27		3.17		3.23		3.15	
Strengths and Resources	0	Not at all	74	8.83	33	3.94	40	4.77	32	3.82
	1	Occasionally	180	21.5	141	16.8	143	17.1	115	13.7
	2	Sometimes	118	14.1	116	13.8	118	14.1	122	14.6
	3	Fairly Often	241	28.8	243	29.0	277	33.1	254	30.3
	4	Very Often	225	26.9	303	36.2	251	30.0	286	34.1
	8/9	DK/REF	0	0	2	0.239	9	1.07	29	3.46
	Mean		2.43		2.77		2.67		2.80	
Quality of Life	1	Death	14	1.67	4	0.477	7	0.835	6	0.716
	2	2	26	3.10	20	2.39	23	2.74	17	2.03
	3	3	54	6.44	40	4.77	27	3.22	29	3.46
	4	4	96	11.5	58	6.92	53	6.32	42	5.01
	5	5	277	33.1	165	19.7	171	20.4	137	16.4
	6	6	122	14.6	111	13.3	117	14.0	85	10.1
	7	7	104	12.4	154	18.4	166	19.8	150	17.9
	8	8	86	10.3	158	18.9	165	19.7	201	24.0
	9	9	20	2.39	72	8.6	55	6.56	77	9.19
	10	Perfect	18	2.15	35	4.18	29	3.46	51	6.09
	98/99	DK/REF	0	0	2	0.239	9	1.07	29	3.46
	Mean		5.38		6.25		6.24		6.61	

Table 6: Distribution of Scores for Composite Patient-Centered Outcome Measure (n=838) - SD										
Outcome	**Level**	**Description**	**Baseline**		**3-Months**		**6-Months**		**12-Months**	
			n	**%**	**n**	**%**	**n**	**%**	**n**	**%**
SD 1 - Work	0	Not at all	91	10.9	139	16.6	124	14.8	154	18.4
	1	1	10	1.19	38	4.54	42	5.01	56	6.68
	2	2	26	3.10	61	7.28	64	7.64	72	8.59
	3	3	62	7.40	91	10.9	76	9.07	69	8.23
	4	4	55	6.56	50	5.97	61	7.28	51	6.09
	5	5	212	25.3	171	20.4	181	21.6	137	16.3
	6	6	81	9.67	65	7.76	66	7.88	57	6.80
	7	7	105	12.5	69	8.23	81	9.67	70	8.35
	8	8	110	13.1	91	10.9	70	8.35	86	10.3
	9	9	43	5.13	34	4.06	30	3.58	30	3.58
	10	Unable to Carry on Any Activities	43	5.13	27	3.22	34	4.06	27	3.22
	98/99	DK/REF	0	0	2	0.239	9	1.07	29	3.46
	Mean		5.31		4.38		4.44		4.15	
SD 2 - Family Life and Home Responsibilities	0	Not at all	237	28.3	281	33.5	290	34.6	304	36.3
	1	1	18	2.15	54	6.44	45	5.37	45	5.37
	2	2	48	5.73	67	8.00	57	6.80	71	8.47
	3	3	45	5.37	57	6.80	65	7.76	64	7.64
	4	4	42	5.01	56	6.68	41	4.89	34	4.06
	5	5	182	21.7	110	13.1	113	13.5	97	11.6
	6	6	56	6.68	46	5.49	50	5.97	46	5.49
	7	7	76	9.07	56	6.68	49	5.85	45	5.37
	8	8	73	8.71	68	8.12	60	7.16	62	7.40
	9	9	28	3.34	13	1.55	30	3.58	18	2.15
	10	Unable to Carry on Any Activities	33	3.94	28	3.34	29	3.46	23	2.75
	98/99	DK/REF	0	0	2	0.239	9	1.07	29	3.46
	Mean		4.01		3.28		3.34		3.06	
SD 3 - Social Life	0	Not at all	201	24.0	229	27.3	237	28.3	256	30.5
	1	1	17	2.03	47	5.61	35	4.18	45	5.37
	2	2	37	4.42	77	9.19	57	6.80	63	7.52
	3	3	48	5.73	66	7.88	71	8.47	79	9.43
	4	4	48	5.73	67	8.00	45	5.37	46	5.49
	5	5	148	17.7	99	11.8	120	14.3	99	11.8
	6	6	71	8.47	59	7.04	58	6.92	57	6.80

	7	7	71	8.47	56	6.68	58	6.92	50	5.97
	8	8	82	9.79	78	9.31	63	7.52	54	6.44
	9	9	40	4.77	22	2.63	30	3.58	30	3.58
	10	Unable to Carry on Any Activities	75	8.95	36	4.30	55	6.56	30	3.58
	98/99	DK/REF	0	0	2	0.239	9	1.07	29	3.46
	Mean		4.60		3.70		3.88		3.44	

Associations Between Depression Response and Other Health Patient-Centered Outcomes (PCOs):

For most depression response metrics and depression remission across all time periods, the mean differences in other health outcome scores between responders/remitters and non-responders/non-remitters were statistically significant **(Table 7)**.

Overall, differences in General Health between responders and non-responders (across metrics) ranged from -0.138 to -0.307, with the three-month scores being lower than those at baseline in all cases (an indication of better self-reported General Health). These differences largely attenuated at six months (ranging from -0.0855 to -0.204), before improving (often beyond three-month differences) at twelve months (ranging from -0.178 to -0.284). Strengths and Resources scores increased for all metrics across all time periods (an indication of improved social connectivity). Differences were larger overall between baseline and six months (ranging from 0.228 to 0.443) than between baseline and three months (ranging from 0.120 to 0.300). At twelve months, differences in Strengths and Recourses scores either attenuated mildly from six-month differences or continued to improve (ranging from 0.308 to 0.566). Differences in Quality of life scores, overall, improved incrementally across the study period. Between baseline and three months, differences ranged from 0.652 to 1.16. Corresponding ranges for six and twelve months were 0.764 to 1.25 and 1.12 to 1.71, respectively. Sheehan Disability

score differences were negative for all metrics across all time periods (indicative of reduced self-reported disability). In most cases, the magnitude of decrease improved over time, with three-, six- and twelve-month differences ranging from -1.85 to -3.25, -2.57 to -4.08 and -3.94 to -5.07, respectively.

Metric one (fifty percent improvement in PHQ-9 score from baseline) had the highest number of mean differences with a greater than the mean score (11/12). This was followed by metrics five and nine (both with 8/12) and metrics three, four and seven (all with 7/12). The lowest performing metric was number two, with only one mean difference with a greater than the mean score. This was followed by metrics six and eight (both with 2/12). When three-month depression response/remission were compared to outcomes at six and twelve months, differences were attenuated at most metrics and time periods **(Table S1 in the supplement)**.

Table 7: Mean Difference in Patient-Centered Outcome Measures at 3, 6 and 12 Months Between Depression Responders and Non-Responders Across 6 Different Depression Response Metrics

		3-month outcomes (n=830)				6-month outcomes (n=796)				12-month outcomes (n=701)			
		GH	SR	QOL	SD	GH	SR	QOL	SD	GH	SR	QOL	SD
1	≥50% decrease from baseline	-0.239 (***)	0.205 (*)	0.950 (***)	-2.68 (***)	-0.190 (**)	0.423 (***)	1.07 (***)	-3.51 (***)	-0.284 (***)	0.512 (***)	1.45 (***)	-4.52 (***)
2	Absolute decrease of ≥ 5	-0.138	0.191	0.652 (***)	-1.85 (**)	-0.130	0.332 (**)	0.764 (***)	-2.57 (***)	-0.229 (**)	0.308 (*)	1.47 (***)	-4.16 (***)
3	Score of <10	-0.250 (**)	0.120	1.16 (***)	-2.00 (**)	-0.0855	0.397 (**)	1.14 (***)	-3.04 (***)	-0.187	0.469 (**)	1.57 (***)	-4.85 (***)
4	≥ 50% decrease from baseline OR score of <10	-0.233 (**)	0.163	1.15 (***)	-2.37 (***)	-0.132	0.416 (**)	1.25 (***)	-3.60 (***)	-0.244 (*)	0.446 (**)	1.62 (***)	-5.07 (***)
5	≥ 50% decrease from baseline AND score of <10	-0.256 (***)	0.175	0.984 (***)	-2.42 (***)	-0.161 (*)	0.418 (***)	1.02 (***)	-3.22 (***)	-0.250 (**)	0.530 (***)	1.44 (***)	-4.44 (***)
6	Absolute decrease of ≥5 AND score <10	-0.214 (**)	0.193	0.837 (***)	-2.21 (***)	-0.128	0.436 (***)	1.04 (***)	-3.08 (***)	-0.178 (*)	0.400 (***)	1.37 (***)	-3.94 (***)
7	Absolute decrease of ≥10	-0.307 (***)	0.300 (**)	0.736 (***)	-3.25 (***)	-0.204 (**)	0.443 (***)	0.666 (***)	-3.61 (***)	-0.282 (***)	0.441 (***)	1.12 (***)	-3.99 (***)
8	≥50% decrease from baseline AND absolute decrease of ≥5	-0.207 (**)	0.229 (*)	0.857 (***)	-2.34 (***)	-0.148 (*)	0.438 (***)	0.955 (***)	-3.29 (***)	-0.216 (**)	0.451 (***)	1.40 (***)	-4.29 (***)
9	Score of <5	-0.287 (***)	0.216 (*)	0.890 (***)	-3.00 (***)	-0.186 (**)	0.228 (*)	0.867 (***)	-4.08 (***)	-0.276 (***)	0.566 (***)	1.71 (***)	-3.98 (***)
Mean		-0.237	0.199	0.913	-2.46	-0.152	0.392	0.975	-3.33	-0.238	0.458	1.46	-4.36

p<0.05 (*); p<0.01 (**); p<0.001 (***)
GH - General Health; SR - Strengths & Resources; QOL - Quality of Life; SD - Sheehan Disability
Green - > mean score; Yellow - At mean score; Red - < mean score

Tests for Equivalency Across Single-Component Metrics:

When all combinations of single-component metrics at all time periods were tested for equivalence in mean differences using ANOVA with specified linear contrasts, metrics were often found to not be statistically different. At three months **(Table 8)**, statistically significant differences were found for at least two patient-centered health outcomes in three pair-wise comparisons. Two of these were between metrics on the same continua – (1) score <10 (metric three) and score <5 (metric nine) and (2) absolute decrease of ≥ 5 (metric two) and absolute decrease of ≥ 10 (metric seven). The third was between and absolute decrease of ≥ 5 (metric two) and score <5 (metric nine). Two other pair-wise comparisons had one statistically significant patient-centered health outcome difference.

Table 8: F-Tests for Pair-Wise Tests for Equivalence in 3-Month Mean Differences in Outcomes for All Combinations of Single-Component PHQ-9 Depression "Response" Metrics

	Score of < 5				Absolute decrease of ≥ 5				Score of <10				Absolute decrease of ≥ 10			
	GH	**SR**	**QOL**	**SD**	**GH**	**SR**	**QOL**	**SD**	**GH**	**SR**	**QOL**	**SD**	**GH**	**SR**	**QOL**	**SD**
≥ 50% decrease from baseline	1.02	0.060	0.010	1.33	2.82	0.280	4.25 (*)	2.17	0.335	0.446	0.305	1.10	0.449	1.16	1.70	1.09
Score of < 5					4.95 (*)	0.611	2.89	4.94 (*)	1.50	1.54	0.246	5.81 (*)	0.005	0.852	1.98	0.194
Absolute decrease of ≥ 5									0.147	1.08	2.06	0.597	5.34 (*)	1.02	0.172	4.80 (*)
Score of <10													1.98	2.60	0.804	4.44 (*)

p<0.05 (*); p<0.01 (**); p<0.001 (***)
GH - General Health; SR - Strengths & Resources; QOL - Quality of Life; SD - Sheehan Disability

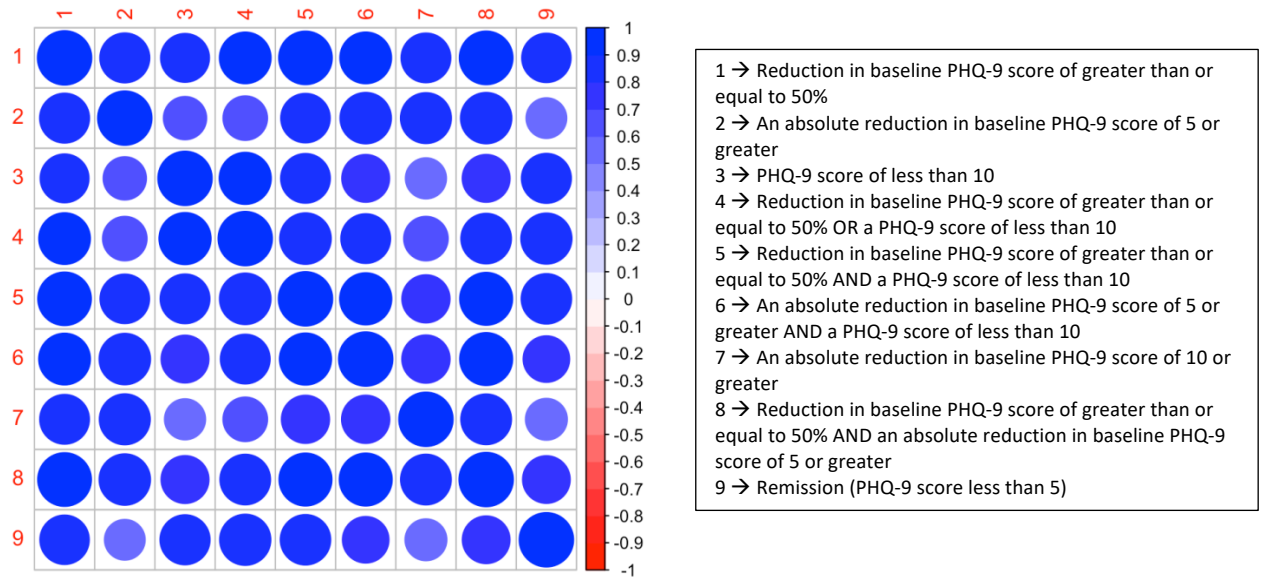
Similar, albeit not identical findings were noted at six and twelve months - see **Tables S2 and S3** in the supplement for details. At six months, statistically significant differences were found for at least two patient-centered health outcomes in one comparison, while five other comparisons had one statistically significant difference. At twelve months, statistically significant differences were found for at least two patient-centered health outcomes in two comparison, while two other comparisons had one statistically significant difference.

AIMS Data

Rank Order Comparisons Across Metrics:

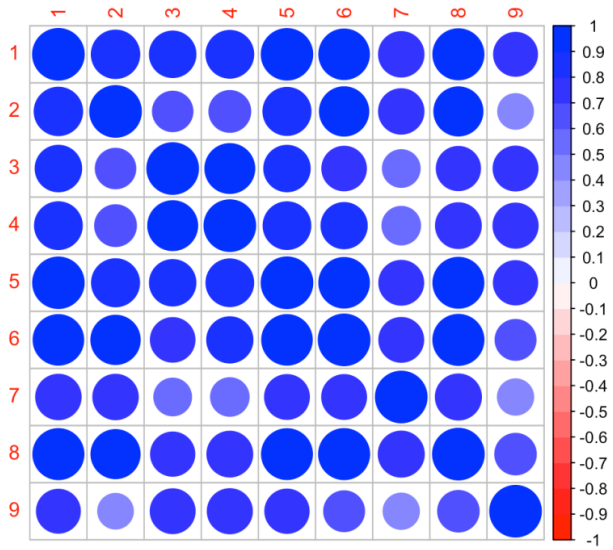
The full AIMS data sample consisted of 145 clinics and 42 organizations. A total of 135 clinics met the inclusion criteria of having a total collaborative care panel size of at least ten. At three, six and twelve months, 135, 130 and 113 clinics met inclusion criteria by having at least one PHQ-9 outcome, respectively. Organization-level rank orders derived from a multilevel logistic regression model and direct sample means at three, six and twelve months across all nine metrics (eight response and one remission) are visible in **Tables S4, S5, S6, S7, S8 and S9** in the supplement. Correlation matrices calculated using Spearman's rank-order correlation coefficient for model-based and sample mean rankings for six and twelve months are displayed in **Figures 2, 3, 4 and 5** below. Corresponding figures for three months are visible as **Figures S1 and S2** in the supplement.

Figure 2: Multilevel Logistic Regression Model - 6 Months:



At six months with the multilevel logistic regression model and empirical Bayes predictions, all metrics were positively correlated with the others. Metrics one, five, six and eight were most positively correlated with others, while two, three, four, seven and nine were less highly correlated (though still substantially correlated).

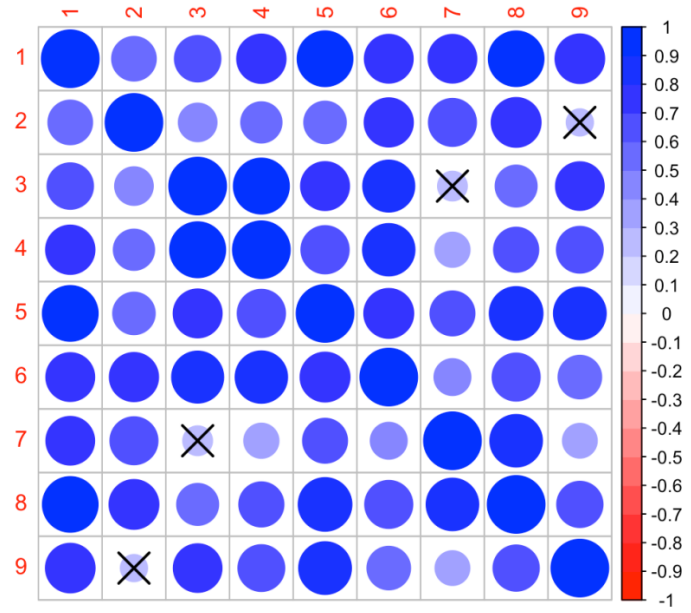
Figure 3: Multilevel Logistic Regression Model - 12 Months:



1 → Reduction in baseline PHQ-9 score of greater than or equal to 50%
2 → An absolute reduction in baseline PHQ-9 score of 5 or greater
3 → PHQ-9 score of less than 10
4 → Reduction in baseline PHQ-9 score of greater than or equal to 50% OR a PHQ-9 score of less than 10
5 → Reduction in baseline PHQ-9 score of greater than or equal to 50% AND a PHQ-9 score of less than 10
6 → An absolute reduction in baseline PHQ-9 score of 5 or greater AND a PHQ-9 score of less than 10
7 → An absolute reduction in baseline PHQ-9 score of 10 or greater
8 → Reduction in baseline PHQ-9 score of greater than or equal to 50% AND an absolute reduction in baseline PHQ-9 score of 5 or greater
9 → Remission (PHQ-9 score less than 5)

At six months with the multilevel logistic regression model and empirical Bayes predictions, correlations were slightly weaker overall than those at six months. Patterns in correlation, however, were quite similar, with metrics one, five, six and eight being the most highly correlated. Metric nine (remission) was least correlated with the others.

Figure 4: Sample Means - 6 Months*:

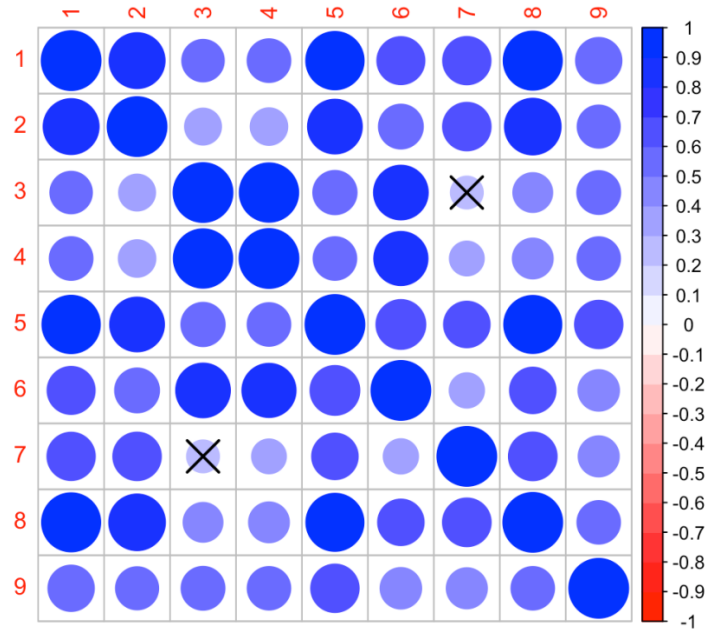


1 → Reduction in baseline PHQ-9 score of greater than or equal to 50%
2 → An absolute reduction in baseline PHQ-9 score of 5 or greater
3 → PHQ-9 score of less than 10
4 → Reduction in baseline PHQ-9 score of greater than or equal to 50% OR a PHQ-9 score of less than 10
5 → Reduction in baseline PHQ-9 score of greater than or equal to 50% AND a PHQ-9 score of less than 10
6 → An absolute reduction in baseline PHQ-9 score of 5 or greater AND a PHQ-9 score of less than 10
7 → An absolute reduction in baseline PHQ-9 score of 10 or greater
8 → Reduction in baseline PHQ-9 score of greater than or equal to 50% AND an absolute reduction in baseline PHQ-9 score of 5 or greater
9 → Remission (PHQ-9 score less than 5)

*An "X" denotes a lack of statistical significance at the $p<0.05$ level

At six months using a sample means approach, correlations were weaker than those derived from the multilevel logistic regression model. Metrics seven and nine were the least positively correlated with other metrics, while one, four, five, six and eight were the most. Metric two was positively correlated, but to a lesser extent than other metrics. Two comparisons were statistically insignificant.

Figure 5: Sample Means - 12 Months*:



1 → Reduction in baseline PHQ-9 score of greater than or equal to 50%
2 → An absolute reduction in baseline PHQ-9 score of 5 or greater
3 → PHQ-9 score of less than 10
4 → Reduction in baseline PHQ-9 score of greater than or equal to 50% OR a PHQ-9 score of less than 10
5 → Reduction in baseline PHQ-9 score of greater than or equal to 50% AND a PHQ-9 score of less than 10
6 → An absolute reduction in baseline PHQ-9 score of 5 or greater AND a PHQ-9 score of less than 10
7 → An absolute reduction in baseline PHQ-9 score of 10 or greater
8 → Reduction in baseline PHQ-9 score of greater than or equal to 50% AND an absolute reduction in baseline PHQ-9 score of 5 or greater
9 → Remission (PHQ-9 score less than 5)

*An "X" denotes a lack of statistical significance at the $p<0.05$ level

Using a sample means approach at twelve months, metrics were weaklier correlated overall than the sample means approach at six months. However, only one comparison was statistically insignificant. Strength of correlation was comparable across metrics, with the exceptions of metrics seven and nine (which were weaklier correlated). No negative correlations were noted.

Discussion:

IMPACT:

Depression Response and Remission Rates

Findings demonstrated that the various response metrics were associated with significantly different response rates, with three- six- and twelve-month rates ranging from 32.3-76.7%, 37.9-82.2% and 42.1-84.9%, respectively. Depression remission rates at three-, six- and twelve-months were 45.7%, 50.0% and 58.8%, respectively. Of note, remission rates at all time periods were higher than the corresponding response rates using metric seven (absolute reduction in baseline PHQ-9 score of 10 or greater), indicating that this latter metric is challenging to achieve.

Our findings suggest that an organization's choice of depression response metric could alter the perceived effectiveness of a treatment program and that response rates should only be compared across organizations using the same metric. Of note, we do not propose that PHQ-9 response rates (using any metric) from this investigation serve as benchmarks for real-world or community collaborative care implementations, as they are derived from a large RCT and likely have poor external validity. Further, our depression response and remission rates are substantially different than those reported with SCL-20 measurements in the IMPACT trial[24]. Response in IMPACT was defined as a reduction in SCL-20 score of greater than or equal to 50% (similar to metric one using the PHQ-9 in this investigation); rates in the intervention group at three-, six- and twelve months were 31.8%, 49.34% and 44.67%, respectively[24]. In IMPACT, remission was defined as an SCL-20 score of less than 0.5 and rates at three-, six- and twelve months were found to be 15.8%, 30.1% and 25.0%, respectively[24].

These results are consistent with prior studies comparing the SCL-20 and PHQ-9 treatment response rates with IMPACT data, all of which have found that changes in PHQ-9 scores are typically greater than concomitant changes in SCL-20 scores[11,14,15]. The reasons for this remain minimally understood. One possible contributing factor from the IMPACT trial was differences in the administration of the two instruments. The PHQ-9 was administered by nurses as part of clinical care, while the SCL-20 was managed by trained research assistants. Although this could have caused intervention arm patients to report fewer depression symptoms on the PHQ-9 (biasing the results in favor of improved response and remission rates), this has never been demonstrated empirically[11]. However, since depression symptom instruments (e.g., PHQ-9) are typically administered by clinical staff in real-world collaborative care implementations, the impact of this workflow should be assessed in future research.

Changes in Patient-Centered Outcomes (PCOs) Across Depression Treatment Metrics

Our findings demonstrated that all four patient-centered outcomes (PCOs) improved for all nine metrics across all time periods. For Strengths and Resources, Quality of Life and Sheehan Disability, the magnitude of improvement increased incrementally over time. For General Health, the improvement attenuated at six months (relative to three months), before rebounding at twelve months. These findings are not unexpected, as previous research has extensively shown that patient-centered outcomes (such as those assessed in this investigation) tend to improve as depression symptoms improve[41–43].

When comparing overall differences in patient-centered outcomes across time periods and metrics, we found that metrics one (fifty percent improvement in PHQ-9 score from baseline), five (fifty percent improvement in PHQ-9 score from baseline AND a score of less than 10) and nine (PHQ-9 score less

than five - remission) were the strongest performers. Although not particularly surprising, this finding is significant in that it corroborates the response metric from IMPACT[24] (metric one), the original PHQ-9 response metric (metric five)[11,15,17] and the almost universally adopted PHQ-9 remission metric (metric nine)[11,15,17]. It is noteworthy that metric four (fifty percent improvement in PHQ-9 score from baseline OR a score of less than 10) was not far behind. This suggests that changing "And" to "Or" (the only difference between metrics four and five) may not lead to a notable difference in patient centered outcomes, despite being associated with substantial differences in response rates **(Table 4)**. Conversely, metric seven (not known to be used in any research or real-world setting) was an almost equally high performer while be associated with substantially lower response rates. This once again highlights the importance of comparing real-world depression improvement rates across organizations using the same metric.

Previous research has shown (albeit with fewer metrics) that different metrics tend to classify similar patients as improved or not improved with respect to depression symptoms[17]. Although our results do not contradict that assertion, we do find that some metrics may be more consistently associated with improvements in patient-centered outcomes than others.

Tests for Equivalency Across Single-Component Metrics:

Although power limitations precluded us from testing for equivalency across all metrics, we were able to compare the single-component metrics (one, two, three, seven and nine). Overall, findings demonstrated that the single-component metrics most significantly different from one another across the three time periods (based on pair-wise comparisons) were the following: (1) score <10 (metric three) and score <5 (metric nine), (2) absolute decrease of ≥ 5 (metric two) and absolute decrease of $\geq$

10 (metric seven) and (3) an absolute decrease of ≥ 5 (metric two) and score <5 (metric nine). It is unsurprising that the first and second are different from one another, as they are on the same continuum (we would expect a score of less than five to be associated with more improved patient-centered outcomes than a score less than ten). The third significant difference is a notable finding, as these metrics are commonly used for response (an absolute decrease of ≥ 5)[14] and remission (score <5)[11,15,17]. Since we would expect response and remission to be significantly different from one another, this provides some supporting evidence for the constructs of the aforementioned metrics.

At the same time, we see little in the way of evidence for significant differences between metrics one, seven and nine. These were all top-performing metrics with regard to their associated differences in patient-centered outcomes. One interpretation of this lack of significant differences could be that these metrics are in their own "league" - indistinguishable from one another, but different than the others. However, this is not supported by the data, as there is also little evidence for difference between metrics one and two, the latter of which was the lowest performing metric. Overall, this suggests that the single-component metrics are largely not statistically distinguishable in pair-wise comparisons. This does not, however, negate the analysis and comparisons conducted above **(Table 7)** and does not prevent comparisons of metrics in aggregate across the three time periods. Further, it does not suggest anything about the multi-component metrics, as these could not be directly examined using this statistical method.

AIMS Data

Although choice of depression response or remission metric had a visible impact on organization-level rank-order, our findings demonstrated that most rank-orders were highly correlated with one another.

Using a multilevel logistic regression model with empirical Bayes predictions, rank-orders based on metrics one, five, six and eight were most highly correlated, while those based on metric nine were least correlated. Results were similar with a sample means approach (no statistical model); metrics one, four, five, six and eight were the most correlated, while nine was least correlated.

Among the top three metrics determined by associated changes in patient-centered outcomes (one, five and nine), metrics one and five had rank orders that were consistently highly correlated. This provides some further evidence in support of these metrics - both are associated with favorable changes in patient-centered outcomes and both rank organizations similarly. The last of the top-three metrics, score of less than five (remission), was more weakly (though still positively) correlated with all of the other metrics. Although the remission metric was not notably superior to metrics one or five with regard to associated patient-centered outcome changes, it was found to rank organizations differently. It also had the most statistically significant differences in the aforementioned pair-wise comparisons. These findings collectively suggest that the remission metric is exceptional to some extent. This is logically intuitive, and one would expect ranking based on remission to be substantially different from ranking based on any response metric.

Limitations

The findings in this investigation have a number of limitations. First, our preferential use of the PHQ-9 limited our IMPACT analysis to 900 participants (instead of the full study population of 1800). However, we believe that this choice makes research more relevant to clinicians, researchers and policymakers, who most often use the PHQ-9. Additionally, our IMPACT sample consisted entirely of treatment arm participants, with control arm participants being omitted (due to lack of measured PHQ-9 scores). At

the same time, an argument can be made that this is not a substantial limitation, as patients do not often have ongoing PHQ-9 score ascertainment when they are not engaged in care of some kind. Our comparative analysis of depression response and remission metrics was limited to those with a single component due to statistical power constraints. Because of this, we could not provide pair-wise comparisons for four of our nine metrics. The real-world AIMS Center dataset was limited by numerous instances of missing data, including follow-up PHQ-9 scores and demographic information. However, our sample of 145 clinics and 42 organizations across multiple states remains one of the largest real-world collaborative care implementation datasets to date.

Conclusions:

Our findings demonstrate that choice of depression response metric substantially impacts observed response rates and that PCOs tend to improve with depression. However, our comparative associations between depression response/remission metrics and patient-centered outcomes were mixed overall and we are consequently unable to suggest an optimal metric. Organization-level rankings for depression response/remission vary depending on choice of metric, but their rankings are highly correlated. Given the increasing global focus on population health and measurement-based care, future research should prioritize the determination of an optimal, pragmatic metric for depression outcome ascertainment.

References:

1. Unützer J, Park M. Strategies to Improve the Management of Depression in Primary Care. *Prim Care - Clin Off Pract*. 2012;39(2):415-431. doi:10.1016/j.pop.2012.03.010
2. Schoenbaum M, Unützer J, McCaffrey D, Duan N, Sherbourne C, Wells KB. The effects of primary care depression treatment on patients' clinical status and employment. *Health Serv Res*. 2002;37(5):1145-1158. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed5&NEWS=N&AN=2002434074.
3. Greenberg PE, Fournier AA, Sisitsky T, Pike CT, Kessler RC. The economic burden of adults with

major depressive disorder in the United States (2005 and 2010). *J Clin Psychiatry*. 2015;76(2):155-162. doi:10.4088/JCP.14m09298
4. Cuijpers P. The Challenges of Improving Treatments for Depression. *JAMA*. 2018;86(2):320-333. doi:10.1001/jama.2018.17824
5. Levkovitz Y, Tedeschini E, Papakostas GI. Efficacy of Antidepressants for Dysthymia. *J Clin Psychiatry*. 2011;72(04):509-514. doi:10.4088/JCP.09m05949blu
6. Fried EI, van Borkulo CD, Epskamp S, Schoevers RA, Tuerlinckx F, Borsboom D. Measuring depression over time...or not? lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychol Assess*. 2016;28(11):1354-1367. doi:10.1037/pas0000275
7. Song Z, Navathe AS, Emanuel EJ, Volpp KG. Incorporating value into physician payment and patient cost sharing. *Am J Manag Care*. 2018;24(3):126-128.
8. Hamilton M. A Rating Scale for Depression. *J Neurol Neurosurg Psychiatry*. 1960;23(1):56-62. doi:10.1136/jnnp.23.1.56
9. Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clin Psychol Rev*. 1988;8(1):77-100. doi:10.1016/0272-7358(88)90050-5
10. Williams JW, Stellato CP, Cornell J, Barrett JE. The 13- and 20-Item Hopkins Symptom Checklist Depression Scale: Psychometric Properties in Primary Care Patients with Minor Depression or Dysthymia. *Int J Psychiatry Med*. 2005;34(1):37-50. doi:10.2190/u1b0-nkwc-568v-4mak
11. Kroenke K, Spitzer RL. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatr Ann*. 2002;32:509-616.
12. Arroll B, Goodyear-Smith F, Crengle S, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med*. 2010;8(4):348-353. doi:10.1370/afm.1139
13. SAMHSA-HRSA Center for Integrated Health Solutions. Screening Tools.
14. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care*. 2004;42(12):1194-1201. doi:10.1159/000317133
15. 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:10.1097/01.MLR.0000093487.78664.3C
16. Katzelnick DJ, Duffy FF, Chung H, Regier DA, Rae DS, Trivedi MH. Depression Outcomes in Psychiatric Clinical Practice: Using a Self-Rated Measure of Depression Severity. *Psychiatr Serv*. 2011;62(8):929-935. doi:10.1176/ps.62.8.pss6208_0929
17. McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: A comparison of methods. *J Affect Disord*. 2010;127(1-3):122-129. doi:10.1016/j.jad.2010.04.030
18. Evans C, Margison F, Barkham M. The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evid Based Ment Health*. 1998;1(3):70-72. doi:10.1136/ebmh.1.3.70
19. Ziring JP, Black J, Gogia S, Stine N, Chokshi DA. It's Time to Rethink How We Measure Remission from Depression. *NEJM Catal*. 2016. https://catalyst.nejm.org/rethink-measure-depression-remission/.
20. New York State Office of Mental Health (OMH). *New York State Medicaid Collaborative Care Provider Certification Packet*.; 2014. http://uwaims.org/nyscci/files/NYSMedicaidCollaborativeCareDepressionProgramRFA.pdf.

21. Rush AJ, Thase ME. Improving Depression Outcome by Patient-Centered Medical Management. *Am J Psychiatry*. 2018;(December):appi.ajp.2018.1. doi:10.1176/appi.ajp.2018.18040398
22. Richards DA, Lovell K, Gilbody S, et al. Collaborative care for depression in UK primary care: A randomized controlled trial. *Psychol Med*. 2008;38(2):279-287. doi:10.1017/S0033291707001365
23. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol*. 1991;59(1):12-19. doi:10.1109/APCC.1999.820481
24. Unützer J, Katon W, Callahan CM, et al. Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. *JAMA*. 2002;288(22):2836-2845. http://www.ncbi.nlm.nih.gov/pubmed/12472325.
25. University of Washington AIMS Center. AIMS Center - Advancing Integrated Mental Health Solutions. https://aims.uw.edu. Published 2019. Accessed May 23, 2019.
26. Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. *Health Econ*. 1993;2(3):217-227. http://www.ncbi.nlm.nih.gov/pubmed/8275167.
27. Von Korff M, Ormel J, Katon W, Lin EH. Disability and depression among high utilizers of health care. A longitudinal analysis. *Arch Gen Psychiatry*. 1992;49(2):91-100. doi:10.1001/archpsyc.1992.01820020011002
28. Leon AC, Olfson M, Portera L, Farber L, Sheehan D V. Assessing Psychiatric Impairment in Primary Care with the Sheehan Disability Scale. *Int J Psychiatry Med*. 1997;27(2):93-105. doi:10.2190/T8EM-C8YH-373N-1UWD
29. University of Washington AIMS Center. Care Management Tracking System (CMTS). https://aims.uw.edu/resource-library/care-management-tracking-system-cmts. Published 2019. Accessed April 15, 2019.
30. Unützer J, Choi Y, Cook IA, Oishi S. A web-based data management system to improve care for depression in a multicenter clinical trial. *Psychiatr Serv*. 2002;53(6):671-673, 678. doi:10.1176/ps.53.6.671
31. Bauer AM, Chan YF, Huang H, Vannoy S, Unützer J. Characteristics, management, and depression outcomes of primary care patients who endorse thoughts of death or suicide on the PHQ-9. *J Gen Intern Med*. 2013;28(3):363-369. doi:10.1007/s11606-012-2194-2
32. Cerimele JM, Chan YF, Chwastiak LA, Unützer J. Pain in primary care patients with bipolar disorder. *Gen Hosp Psychiatry*. 2014;36(2):228-229. doi:10.1016/j.genhosppsych.2013.11.004
33. Chan YF, Huang H, Bradley K, Unützer J. Referral for substance abuse treatment and depression improvement among patients with co-occurring disorders seeking behavioral health services in primary care. *J Subst Abuse Treat*. 2014;46(2):106-112. doi:10.1016/j.jsat.2013.08.016
34. Chan Y-F, Huang H, Sieu N, Unützer J. Substance Screening and Referral for Substance Abuse Treatment in an Integrated Mental Health Care Program. *Psychiatr Serv*. 2013;64(1):88-90. doi:10.1176/appi.ps.201200082
35. Huang H, Bauer AM, Wasse JK, et al. Care managers' experiences in a collaborative care program for high risk mothers with depression. *Psychosomatics*. 2013;54(3):272-276. doi:10.1016/j.psym.2012.07.011
36. Huang H, Chan YF, Katon W, et al. Variations in depression care and outcomes among high-risk mothers from different racial/ethnic groups. *Fam Pract*. 2012;29(4):394-400. doi:10.1093/fampra/cmr108
37. Ratzliff ADH, Ni K, Chan Y-F, Park M, Unützer J. A Collaborative Care Approach to Depression Treatment for Asian Americans. *Psychiatr Serv*. 2013;64(5):487-490.

doi:10.1176/appi.ps.001742012
38. Cerimele JM, Chan Y-F, Chwastiak LA, Avery M, Katon W, Unützer J. Bipolar Disorder in Primary Care: Clinical Characteristics of 740 Primary Care Patients With Bipolar Disorder. *Psychiatr Serv*. 2014;65(8):1041-1046. doi:10.1176/appi.ps.201300374
39. Unützer J, Chan YF, Hafer E, et al. Quality improvement with pay-for-performance incentives in integrated behavioral health care. *Am J Public Health*. 2012;102(6):E41-E45. doi:10.2105/AJPH.2011.300555
40. The R Foundation. The R Project for Statistical Computing.
41. Lin EH, VonKorff M, Russo J, et al. Can depression treatment in primary care reduce disability? A stepped care approach. *Arch Fam Med*. 2011;9(10):1052-1058. http://www.ncbi.nlm.nih.gov/pubmed/11115207.
42. Lin EHB, Katon W, Von Korff M, et al. Effect of improving depression care on pain and functional outcomes among older adults with arthritis: a randomized controlled trial. *JAMA*. 2003;290(18):2428-2429. doi:10.1001/jama.290.18.2428
43. Kroenke K, Spitzer RL. The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatr Ann*. 2002;32(9):509-515. doi:10.3928/0048-5713-20020901-06