Assessment of Outcome-Based Measures of Depression Care Quality in Veterans Health Administration Facilities

Paul N. Pfeiffer, MD
Kara Zivin, PhD
Avinash Hosanagar, MD
Vanessa Panaite, PhD
Dara Ganoczy, PhD
H. Myra Kim, ScD
Timothy Hofer, MD
John D. Piette, PhD

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11414-022-09813-4.

Address correspondence to Paul N. Pfeiffer, VA Center for Clinical Management Research, VA Ann Arbor Healthcare System, Ann Arbor, MI, USA; Mental Health Service, VA Ann Arbor Healthcare System, Ann Arbor, MI, USA; University of Michigan Medical School, Ann Arbor, MI, USA. ppfeiffe@medumich.edu.

Kara Zivin, VA Center for Clinical Management Research, VA Ann Arbor Healthcare System, Ann Arbor, MI, USA; School of Public Health, University of Michigan, Ann Arbor, MI, USA; University of Michigan Medical School, Ann Arbor, MI, USA. ppfeiffe@medumich.edu.

Dara Ganoczy, VA Center for Clinical Management Research, VA Ann Arbor Healthcare System, Ann Arbor, MI, USA. ppfeiffe@medumich.edu.

H. Myra Kim, VA Center for Clinical Management Research, VA Ann Arbor Healthcare System, Ann Arbor, MI, USA;University of Michigan Consulting for Statistics, Computing, and Analytics Research, Ann Arbor, MI, USA. ppfeiffe@medumich.edu.

Timothy Hofer, VA Center for Clinical Management Research, VA Ann Arbor Healthcare System, Ann Arbor, MI, USA. ppfeiffe@medumich.edu.

John D. Piette, VA Center for Clinical Management Research, VA Ann Arbor Healthcare System, Ann Arbor, MI, USA; University of Michigan Medical School, Ann Arbor, MI, USA; School of Public Health, University of Michigan, Ann Arbor, MI, USA. ppfeiffe@medumich.edu.

Avinash Hosanagar, Mental Health Service, VA Ann Arbor Healthcare System, Ann Arbor, MI, USA; University of Michigan Medical School, Ann Arbor, MI, USA.

Timothy Hofer, University of Michigan Medical School, Ann Arbor, MI, USA.

Vanessa Panaite, James A. Haley Veterans' Hospital, Tampa, FL, USA.

The Journal of Behavioral Health Services & Research, 2023. 49–67 © 2022, National Council for Mental Wellbeing. DOI 10.1007/s11414-022-09813-4

Abstract

To inform the potential use of patient-reported depression symptom outcomes as measures of care quality, this study collected and analyzed longitudinal Patient Health Questionnaire (PHQ9) scores among 1,638 patients who screened positive for major depression according to a PHQ9≥10 across 29 Department of Veterans Affairs facilities. The study found baseline PHQ9, prior mental health visits, physical functioning, and treatment expectancy were consistently associated with subsequent PHQ9 outcomes. No facilities outperformed any others on PHQ9 scores at the 6-month primary endpoint, and the corresponding intra-class coefficient was≤.01 for the entire sample (n = 1,214) and 0.03 for the subgroup of patients with new depression episodes (n = 629). Measures of antidepressant receipt, psychotherapy, or treatment intensification were not associated with 6-month PHQ9 scores. PHQ9 outcomes are therefore unlikely to be useful as quality indicators for VA healthcare facilities due to low inter-facility variation, and new care process measures are needed to inform care for patients with chronic depression prevalent in this sample.

Introduction

Health system assessments of depression care quality typically consist of process measures, such as adequacy of antidepressant medication prescribed or psychotherapy visits completed.[1–4] Performance measures based on patient-reported outcomes (e.g., improvement in symptoms) are potentially important additions to care quality assessment because they reflect the realized effectiveness of implemented evidence-based treatments.[5] Performance measures based on depression symptom outcomes have been developed by the National Committee for Quality Assurance (NCQA) and the International Consortium for Health Outcomes Measurement (ICHOM), and health systems may consider adopting these or similar measures to inform quality improvement initiatives.[6,7]

Within large health systems, performance measures may be used to compare quality across facilities in addition to understanding system-wide performance. For patient-reported outcome performance measures (PRO-PMs) to be useful for identifying quality improvement opportunities across practice sites, it is important to establish whether there are meaningful differences in outcomes between sites that are due to care quality and not due to chance or differences in patient populations.[8] Though findings are not consistent, studies show patient characteristics such as age, gender, race, and socioeconomic status predict treatment response via mechanisms that may be independent of the quality or adequacy of the treatment received.[9,10] If these characteristics differ at a population level (e.g., an older vs. younger patient population), PRO-PM scores may need to be adjusted for these characteristics (referred to as case-mix adjustment).[11,12] Kramer et al.[12] evaluated the optimal case-mix adjustment factors for depression outcomes among outpatients with depression and identified that baseline depression severity and physical functioning were associated with depression outcomes. However, this prior study did not use the Patient Health Questionnaire (PHQ9)[13] to measure depression symptoms (as recommended for depression PRO-PMs) or assess performance across multiple sites within a single health system.

Treatment with antidepressant medications and certain psychotherapies improves depression symptoms in clinical trials,[14] and measures of adequate receipt of these treatments were associated with improved outcomes in studies of quality improvement interventions (e.g., collaborative care management.[15–17] However, these findings may not generalize to routine care where

the relationships between process measures reflecting treatment receipt and outcome measures are less well established. Identifying process measures that impact PRO-PM performance could provide more actionable targets for quality managers and providers.

This study conducted a longitudinal survey of depression outcomes among US Department of Veterans Affairs (VA) patients receiving care in Midwestern facilities to address existing knowledge gaps and inform the potential adoption of depression PRO-PMs. Although VA providers collect PHQ9 depression outcomes as part of routine care, this study measured PHQ9 outcomes independently using largely automated systems to avoid potential bias and imprecision from clinically administered measures and missing outcomes for patients who drop out of care. The study's key aims were to (1) identify baseline patient characteristics that best predict subsequent depression outcomes (e.g., case-mix variables); (2) after adjusting for case-mix, assess variation in depression outcomes across facilities; and (3) assess whether various measures of depression care processes are associated with case-mix adjusted symptom outcomes in routine care.

Methods

Setting and sample

This study recruited patients who accessed care across 29 VA Healthcare facilities in a Midwestern Veterans Integrated Service Network between June 2017 and October 2019. The sampling strategy was designed to recruit similar numbers of patients across these facilities, regardless of size, to improve cross-site analyses. On a weekly basis, the study team identified patients from electronic medical record data with a clinical diagnosis of depression (ICD-10 codes: F32.0x, F32.1x, F32.2x, F32.4x, F32.8x, F32.9x, F33.0x, F33.1x, F33.2x, F33.9x, F34.1x, F43.21, F43.23), a PHQ2 score > 2, or a new prescription for an antidepressant medication at a primary care or mental health provider visit during the past week. Patients without recent provider documentation of current depression symptoms and those with a diagnosis of bipolar disorder, schizophrenia/schizoaffective disorder, or a neurocognitive disorder such as dementia were excluded. While some depression quality measures (e.g., those focused on antidepressant treatment adequacy) only include patients with new episodes of care, others include all patients with a depression diagnosis regardless of when treatment initiated (i.e., new and prevalent cases).[1,7] Sampling was evenly stratified between patients with new vs. prevalent episodes of depression to assess whether this distinction is associated with differences in outcomes and the impact of including new vs. prevalent patients on PRO-PMs. New episodes were conservatively defined as no depression diagnosis or positive 2-item PHQ screen (PHQ2) in the past year and no antidepressant treatment in the prior 6 months.[18] Study staff contacted and screened potentially eligible patients with the PHQ9. Those with a score of 10 or more (which has 88% sensitivity and specificity for detecting a major depressive episode)[13] could complete the remaining baseline assessment and follow-up assessments at 6 weeks, 3 months, 6 months, and 12 months. The VA Ann Arbor Healthcare System's Institutional Review Board approved human subjects' involvement.

Survey administration

Participants completed all survey assessments, including the initial PHQ9 screening, by self-report via the participant's choice of telephone interactive voice response (IVR), web-based survey accessed via a short message service (SMS) text message link, or mailed paper surveys.

The IVR and SMS systems conducted automated follow-up assessments supplemented by study staff reminder calls.

Measures

The primary outcome for the study was depression symptom severity at 6 months according to the continuous PHQ9 score. The PHQ9 has good sensitivity and specificity for identifying a major depressive episode, has a range from 0 to 27, and is the measure the NCQA and ICHOM depression PRO-PMs use. [1,6,13,19] The PHQ9 was used to also construct 3 PRO-PMs from dichotomized patient outcomes: 5-point reduction in PHQ9 from baseline to follow-up (i.e., improvement), 50% reduction in PHQ9 from baseline to follow-up (i.e., response), and PHQ9 of 5 or less at follow-up (i.e., remission). These measure definitions were chosen from outcome assessments of collaborative care interventions for depression in primary care settings.[15]

The baseline survey assessed patient characteristics associated with depression outcomes, and, when possible, survey items were aligned with those recommended by the ICHOM.[6,9,10] To minimize response burden, single-item assessments were used unless otherwise noted to measure the following: race, ethnicity, marital status, people living in the home, education, employment, financial distress, Medical Outcomes Study Social Support scale (4 items), [20] social distress (4 items),[21] age of onset of first depressive episode (before or after age 18), number of lifetime depressive episodes, duration of current depressive episode (greater or less than 2 years), current antidepressant use and duration, current receipt of psychotherapy, depression treatment expectancy, anxiety score (0 to 100 with 100 most severe), pain score (0 to 100 with 100 most severe), PROMIS physical functioning scale (4 items),[22] and general health.[23] Additional variables extracted from the medical record included age, gender, service-connected disability, Elixhauser comorbidity score[24] (modified to remove mental health and substance use disorders), substance use disorders, anxiety disorders, pain disorders, new treatment episode, and number of prior mental health visits. Follow-up assessments included the PHQ9. Mode of survey completion (IVR, SMS, or paper) and an indicator for survey completion during the COVID-19 pandemic (applicable to 7.7% of 6-month and 27.2% of 12-month surveys) were also included. Quality of care provided by a facility might have influenced some baseline variables, such as prior treatment and treatment expectancy; however, these variables were included given the important baseline patient characteristics they may represent separate from care quality, and unadjusted analyses were conducted without these included.

Medical record data was used to construct three care process measures for patient-level analyses. Adequate antidepressant treatment was defined as receipt of at least an 84-day supply of antidepressant medication over the 114 days prior to the 3-month assessment among those receiving any antidepressant, consistent with a NCQA measure.[1] Psychotherapy treatment adequacy was defined as receipt of at least 3 psychotherapy visits in the 84 days prior to the 3-month assessment among those receiving any psychotherapy. At least three psychotherapy visits have been used in other studies to define minimally adequate treatment in the VA.[4,25] The third measure represented an exploratory measure of treatment intensification, defined as whether a patient who had not experienced a 5-point improvement in PHQ9 score from baseline to 6 weeks received a new antidepressant medication, an increase in antidepressant medication dose, a depression augmentation agent, or initiation of psychotherapy.

Analyses

To identify a parsimonious set of case-mix adjustment variables, a backward stepwise variable selection process was used. The purpose of removing variables was to reduce the risk of over-fitting

the models and to inform health systems regarding which variables to prioritize for the purpose of case-mix adjustment. The initial model included all baseline variables (including baseline PHQ9 score and an indicator for new episodes) and exempted predictors of follow-up assessment completion from removal to account for missing data assuming missingness at random. Using linear regression models, variables that improved the Akaike information criteria (AIC) the most were removed, one at a time, until the AIC no longer improved with removal of any subsequent variables. A hierarchical model was used to determine coefficients for the retained variables with 6-month continuous PHQ9 scores as the outcome and facilities as random intercepts. Intra-class correlation (ICC) coefficients were used to describe the variation in outcomes at the facility-level relative to patient-level variation within facilities by dividing the facility level variance by the total variance. ICCs were calculated for PHQ9 scores at 6 months using both unadjusted and adjusted hierarchical models and for the subgroup of patients with new episodes of depression. Two sets of sensitivity analyses were used to assess whether the results are robust to changes in the primary model. In the first set, the analyses were repeated using 3- and 12-month PHQ9 scores as outcomes to assess sensitivity to duration of follow-up. In the second set, analyses were repeated using 6-month dichotomous outcomes of improvement, response, and remission based on PHQ9 scores to assess sensitivity to alternative outcome definitions. Visual comparisons were used to assess for outlying performing facilities using (1) the unadjusted mean change in PHQ9 scores at 6 months (calculated as 6 months minus baseline) along with their 95% confidence intervals and (2) the random intercepts for each facility in the final adjusted model of 6-month PHQ scores. Finally, each of the 3 measures of depression care processes were included in each of the prior hierarchical models. The sample size was designed to provide adequate precision to estimate the intraclass correlation coefficient.[26]

Results

Enrollment, follow-up completion, and sample characteristics

Study staff screened 17,433 patients with medical record indicators of depression, yielding 10,666 (61%) eligible following chart review. Of these, 5,138 could not be reached by phone, 2,652 refused, 131 proved ineligible, and 2,745 consented to participate. Of those consented, 2,390 (87%) completed the baseline PHQ-9 screen and 1,638 (69%) of baseline completers had a PHQ-9 score ≥ 10 and became eligible for follow-up assessments. Participants completed 1,224 (75%) assessments at 6-weeks, 1,295 (79%) at 3 months, 1,214 (74%) at 6 months, and 1,106 (68%) at 12 months. Older age and SMS survey method were positively associated with completion of the 3-, 6-, and 12-month assessments; additionally, greater educational attainment was positively associated with 3-month assessment completion, and a substance use disorder diagnosis and degree of social distress were negatively associated with 12-month assessment completion.

The sample completing the 6-month assessments ($N = 1,214$) had a mean age of 52 (SD = 15) years and was 20% female, 80% White, 11% Black, 4% Hispanic, 0.7% Asian American or Pacific Islander, and 5% multiracial or other race. Only 7% of participants described their current depressive episode as their first, and 24% indicated their current depressive episode as less than 2 years duration. Frequency of comorbid diagnoses included 43% for anxiety disorders, 43% for PTSD, 18% for a substance use disorders, and 72% for a pain diagnosis. Although 52% of patients were considered new episodes of depression via medical record screening (e.g., no depression diagnosis in past year, no antidepressant in past 6 months), most participants (88%) had previously seen a VA mental health provider (i.e., for mental health diagnoses other than depression for those with new depression episodes) in the past year. The mean number of mental health visits was 9.6 (SD = 14.2) and for those with any visits it was 11.0 (SD = 14.6). In the year prior to their baseline, 70% of

participants received antidepressant treatment, and 63% received psychotherapy according to medical record data. According to survey responses, 73% of participants were taking an antidepressant at baseline. Only 7% of participants expected their treatment to be very successful, 30% expected treatment to be moderately successful, 47% expected treatment to be somewhat successful, and 17% expected their treatment to not at all be successful (see Table 1).

Depression outcomes and case-mix adjustment variables

Participants had a mean PHQ9 score of 16.2 (SD = 4.4) at baseline, 14.4 (SD = 5.7) at 3 months, 13.8 (SD = 5.9) at 6 months, and 13.8 (SD = 6.2) at 12 months. At 3 months, 27.4% of participants had a 5-point improvement in PHQ9 score, 12.4% had a 50% improvement, and 6.0% had a score of 5 or less. At 6 months, these figures were 30.6%, 14.7%, and 8.2% and at 12 months were 32.5%, 16.7%, and 9.4%, respectively.

Age and mode of survey completion predicted follow-up survey completion at 6 month and were excluded from variable reduction. Following removal of 17 variables, statistically significant predictors of lower PHQ9 scores at 6 months were female gender, less than 2-year duration of current depressive episode, depression onset before age 18, substance use disorder diagnosis, and expectancy that treatment would be very successful (Table 2). Predictors of greater PHQ9 scores at 6 months were greater baseline PHQ9, separated marital status, worse physical functioning, anxiety rating, and number of past-year mental health visits. Across models of 3-month and 12-month continuous outcomes and dichotomous 6-month outcomes (5-point improvement, 50% improvement, and PHQ9 less than 5) (Table 3), no variable was a significant predictor across all models. Baseline PHQ9 score, physical functioning, total prior mental health visits, and expectancy that treatment would be very successful were significant in 4 of the 5 sensitivity models, while depression episode duration of less than 2 years and expectancy that treatment would be moderately successful were significant in 3 of the 5. Female gender was associated with lower PHQ9 scores at 3 and 12 months but not with greater likelihood of 5-point improvement, response, or remission at 6 months.

Depression outcomes and treatment facility

In the unadjusted model of continuous PHQ9 outcomes at 6 months, the ICC was 0.01 showing little variation across facilities relative to within facilities. In the model including baseline case-mix adjustment variables, the ICC was < 0.01. Sensitivity analyses of 3 month and 12 months PHQ scores and those using dichotomous outcome definitions at 6 months did not find any ICC above 0.01. In subgroup analysis of patients with new episodes of depression, the ICC for 6-month continuous outcomes was 0.03. Mean change in unadjusted PHQ9 outcomes at 6 months by facility are depicted in Fig. 1 with overlapping 95% confidence intervals for the true means. In adjusted models, the random intercepts for facility also had overlapping confidence intervals with no outliers (please see electronic supplementary material).

Depression outcomes and care processes

Seventy-one percent of participants with any antidepressant received an adequate 84-day supply, 54% with any psychotherapy received at least 3 sessions, and 32% of those whose PHQ9 did not improve by 5 points at 6 weeks received intensification of treatment. None of these indicators significantly predicted continuous 6-month PHQ9 outcomes when added separately to unadjusted or fully adjusted patient-level models. In the sensitivity analyses, at least 3 sessions of psychotherapy yielded higher odds of remission at 6 months (OR 2.70; 95% CI: 1.10, 6.64; $p = 0.03$).

Table 1
Sample Characteristics ($N = 1,214$)

	N	%
Age, M (SD)	51.5	(14.9)
Female	237	19.5
Race/ethnicity		
White	928	80.1
Black	126	10.9
Hispanic/Latino	40	3.5
Asian	8	0.7
Multiracial/Other	57	4.9
Education		
HS or less	276	23.4
Some college	601	50.9
College grad	175	14.8
Graduate school	129	10.9
Marital status		
Married	628	52.0
Never married	162	13.4
Divorced	347	28.7
Separated	48	4.0
Widowed	23	1.9
Living arrangement		
Lives with spouse/partner	673	56.8
Lives with other family/friends	234	19.7
Lives alone	277	23.4
Lives in long-term care facility	2	0.2
Employment		
Employed full time	320	27.3

Table 1
(continued)

	N	%
Employed part time	89	7.6
Seeking employment	78	6.7
Not working by choice	260	22.2
Disabled due to mental health	209	17.9
Disabled other	215	18.4
Financial worries		
Nearly every day	555	47.0
More than half days	248	21.0
Several days	273	23.1
Not at all	104	8.8
Social support (range: 4 to 20), M (SD)	12.3	(4.4)
Social distress (range: 4 to 16), M (SD)	10.9	(3.4)
Depression < 2 years	272	23.9
Depression onset		
After 18	699	60.1
Before 18	211	18.1
Do not remember	253	21.8
Number of episodes		
Several prior	603	52.1
One prior	109	9.4
1st episode	86	7.4
Not episodic	360	31.1
Service-connected disability	904	74.5
Physical functioning (range: 4 to 20), M (SD)	10.6	(4.1)
Elixhauser comorbidity score, M (SD)	1.3	(4.5)

Table 1
(continued)

	N	%
General health		
Poor	267	22.4
Fair	530	44.4
Good	318	26.7
Very good	61	5.1
Excellent	17	1.4
Anxiety diagnosis	520	42.8
SUD diagnosis	213	17.6
PTSD diagnosis	523	43.1
Pain diagnosis	870	71.7
Pain rating (range: 0 to 100), M (SD)	50.9	(29.6)
Anxiety rating (range: 0 to 100), M (SD)	60.7	(27.2)
Total MH visits, M (SD)	9.6	(14.2)
Treatment expectancy		
Not at all successful	184	16.5
Somewhat successful	522	46.7
Moderately successful	339	30.3
Very successful	73	6.5
New depression episode (medical record)	629	51.8
Past-year treatment (medical record)		
Psychotherapy	765	63.0
Antidepressant medication	853	0.3
Antidepressant use (survey)		
More than 6 months	524	44.1
3 to 6 months	58	4.9

Table 1
(continued)

	N	%
1 to 3 months	291	24.5
Not taking	315	26.5
Survey completion mode		
Interactive voice response	454	37.4
SMS and web survey	631	52.0
Paper	129	10.6

Table 2
Predictors of PHQ9 scores following variable reduction

		6 months ($n=994$)		3 months ($n=1,022$)	12 months ($n=886$)
Variable		Beta	95% CI	Beta	Beta
Intercept		5.22**	(2.70, 7.73)	5.44**	8.69**
Baseline PHQ		0.38**	(0.30, 0.47)	0.46**	0.42**
Female		−1.42**	(−2.27, −0.57)	−1.08**	−1.53**
Marital status	Married	Reference			
	Never Married	0.64	(−0.42, 1.70)		
	Divorced	−0.11	(−0.92, 0.70)		
	Separated	2.62**	(0.91, 4.33)		
	Widowed	0.15	(−2.28, 2.57)		
Employment	Working full time	Reference			
	Working part time	0.33	(−1.03, 1.69)	0.19	0.04
	Seeking employment	0.79	(−0.60, 2.17)	0.25	−0.21
	Not working by choice	−0.37	(−1.47, 0.73)	−0.81	−1.39*
	Disabled, mental health	0.76	(-0.28, 1.80)	1.10*	0.37
	Disabled, other	-0.87	(-1.99, 0.25)	-1.28*	-0.99
Financial worries	Nearly every day	Reference			
	More than half the days			−0.33	−0.89
	Several days			−0.66	−1.08*
	Not at all			−1.69**	−1.81*
Depression < 2 years		−1.87**	(−2.66, −1.07)	−0.92*	−1.53**
Depression onset	After 18	Reference			
	Before 18	−1.19**	(−2.07, −0.31)	−0.76	−0.39
	Do not remember	0.77	(−0.08, 1.61)	0.52	0.75
Physical functioning		0.19**	(0.09, 0.28)	0.09	0.12*
Substance use disorder		-0.89*	(−1.78, −0.01)	−0.77	−0.94
Posttraumatic stress disorder					0.78*

Table 2
(continued)

Variable		6 months ($n=994$) Beta	95% CI	3 months ($n=1,022$) Beta	12 months ($n=886$) Beta
Anxiety rating (0 to 100)		0.03**	(0.01, 0.04)	0.01*	
Pain rating (0 to 100)				0.02**	0.02*
Mental health visits		0.03*	(0.01, 0.05)	0.03*	0.03*
Treatment expectancy	Not at all successful	Reference			
	Somewhat successful	−0.48	(−1.45, 0.49)	−0.59	−1.28*
	Moderately successful	−0.95	(−2.00, 0.10)	−1.18*	−2.73**
	Very successful	−2.79**	(−4.38, −1.20)	−1.04	−3.04**

Empty cells indicate the variable was not included in the final model. *$p<.05$, **$p<.01$. The following variables were included as covariates in some of the models but had no statistically significant associations with any of the outcomes: age, method of survey administration, education, social support, social distress, number of prior depressive episodes, pain diagnosis, and living arrangement

Table 3

Models of dichotomous 6-month outcomes

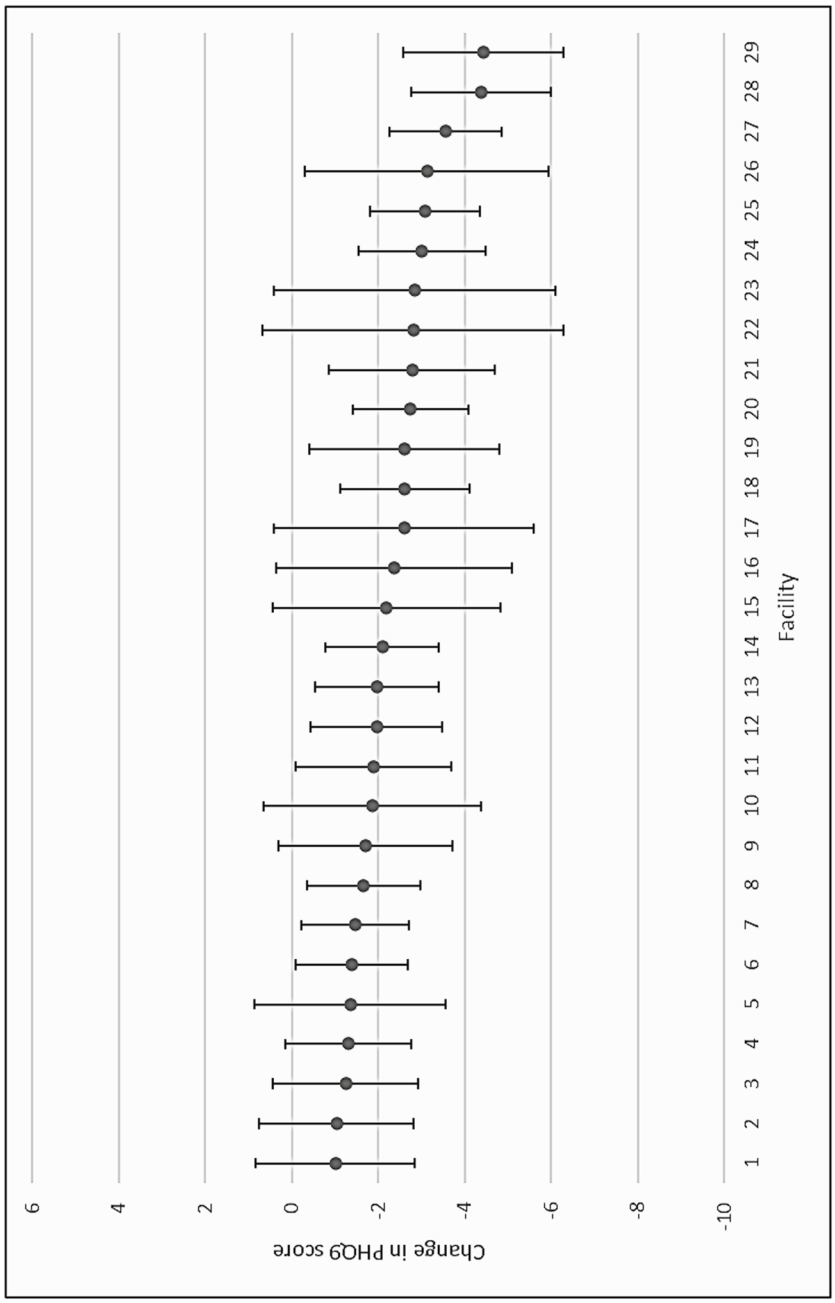
	5-point improvement ($n=1024$)		50% improvement ($n=975$)		PHQ9 5 or less ($n=978$)	
	OR	95% CI	OR	95% CI	OR	95% CI
Intercept	0.02**	(0.01, 0.08)	0.23	(0.05, 1.07)	3.78	(0.32, 44.9)
Baseline PHQ	1.21**	(1.16, 1.26)	1.06**	(1.02, 1.12)	0.95	(0.88, 1.02)
Marital status						
Married (Reference)						
Never married	0.98	(0.61, 1.56)				
Divorced	1.24	(0.87, 1.76)				
Separated	0.34*	(0.13, 0.90)				
Widowed	1.81	(0.65, 5.04)				
Education						
High school or less (Reference)						
Some college	1.61*	(1.11, 2.33)			1.60	(0.83, 3.05)
College grad	1.00	(0.60, 1.66)			0.62	(0.22, 1.69)
Graduate school	1.19	(0.69, 2.05)			1.05	(0.41, 2.70)
General health						
Poor (Reference)						
Fair			0.53*	(0.31, 0.90)	0.41*	(0.20, 0.88)
Good			0.76	(0.42, 1.37)	0.45	(0.20, 1.04)
Very good			0.31*	(0.11, 0.88)	0.24*	(0.06, 0.86)
Excellent			0.59	(0.14, 2.44)	0.31	(0.05, 1.77)
Depression < 2 years	1.85**	(1.32, 2.61)	1.33	(0.86, 2.06)	1.59	(0.91, 2.77)
Depression onset						
After 18 (Reference)						
Before 18			1.84**	(1.17, 2.89)	1.17	(0.62, 2.22)
Do not remember			0.76	(0.44, 1.31)	0.64	(0.30, 1.35)
Physical functioning	0.92**	(0.89, 0.96)	0.94*	(0.88, 0.99)	0.85**	(0.78, 0.93)

Table 3
(continued)

	5-point improvement ($n=1024$)		50% improvement ($n=975$)		PHQ9 5 or less ($n=978$)	
Anxiety rating (0 to 100)	0.99**	(0.99, 1.00)			1.00	(0.99, 1.01)
Mental health visits	0.98**	(0.97, 0.99)	0.97**	(0.95, 0.99)	0.97	(0.94, 1.00)
Treatment expectancy						
Not at all successful (Reference)						
Somewhat successful	1.73*	(1.10, 2.73)	1.35	(0.72, 2.54)	0.71	(0.30, 1.67)
Moderately successful	2.17**	(1.32, 3.55)	1.90	(0.97, 3.70)	0.81	(0.34, 1.96)
Very successful	3.25**	(1.62, 6.51)	4.61**	(2.02, 10.56)	3.29*	(1.19, 9.11)
Survey method						
IVR (Reference)						
SMS	0.77	(0.55, 1.08)	0.72	(0.47, 1.09)	0.60	(0.34, 1.05)
Paper	0.88	(0.52, 1.48)	0.36*	(0.14, 0.92)	0.07*	(0.01, 0.56)
COVID era			2.37**	(1.29, 4.35)	3.15**	(1.46, 6.78)

*$p<.05$, **$p<.01$

Note: Empty cells reflect the variable was not included in the model following variable reduction. The following variables were included as covariates in some of the models but had no statistically significant associations with any of the outcomes: age, gender, social support, social distress, financial stress, pain rating, pain diagnosis, service-connected disability, Elixhauser comorbidity score, new depression episode, anxiety diagnosis, PTSD diagnosis, substance use disorder diagnosis, and antidepressant use



confidence intervals

Fig. 1

Unadjusted mean change in PHQ9 score from baseline to 6 months by VA facility including 95% confidence intervals

Discussion

This study found depression outcomes were primarily influenced by baseline patient characteristics. Indicators of more severe or treatment-resistant depression, specifically greater baseline PHQ9, duration of current depressive episode more than 2 years, and greater number of prior mental health visits, consistently predicted worse subsequent depression outcomes. Physical functioning also consistently predicted outcomes, consistent with Kramer et al.'s prior study of depression case-mix adjustment and other work investigating the relationship between medical illness and depression.[12,27] Treatment expectancy has been shown to influence outcomes in clinical trials for depression.[28,29] The results of this study extend these findings by demonstrating that treatment expectancy also predicts PHQ9 outcomes in routine care. These findings support the ICHOM approach of including physical functioning and treatment expectancy among other depression case-mix variables. However, unlike baseline PHQ9 scores and prior mental health visits, physical functioning and treatment expectancy are not often collected or contained within existing medical records, and the costs of collecting these additional measures may only prove worthwhile if substantial differences in these characteristics exist across planned comparison settings.

This study found that only a minimal amount of variation in depression outcomes is explained by the facility in which patients received their care and no individual facility outperformed the others. These findings suggest that at least in the VA healthcare system, performance measures based on depression symptoms are unlikely to be useful for comparing care quality across facilities. Consistency in outcomes across facilities could be due to similar patterns of care delivery, although several aspects of depression care vary across VA facilities, such as the propensity to provide psychotherapy vs. antidepressant medications and the propensity for patients to be treated by an integrated primary care mental health provider.[30,31] Since this study used facility as the unit of analysis, clinically significant differences in quality and resultant outcomes may exist within individual clinics or care teams yet manifest on average as small differences at the facility level. Depression PRO-PMs in the VA health system may need to focus on identifying quality improvement opportunities within particular clinics (e.g., primary care) or teams rather than VA facilities as a whole. The minimal impact of treatment setting on depression outcomes could have resulted from the substantial degree to which baseline patient characteristics and unobserved factors present before or during treatment (e.g., patients' life events) determine depression outcomes.[32,33]

Study findings suggest PRO-PMs may more reliably detect differences in outcomes across facilities when restricted to patients with new depressive episodes. Our criteria for identifying new patients (e.g., no depression diagnosis or positive screen in 1 year, no antidepressant in 6 months) did not screen out patients with chronic untreated depression, depression treatment outside the VA, or prior VA mental health use for other diagnoses. Refining the criteria for new patients could further improve PRO-PMs for new patients; however, performance measures will still be needed for chronic and treatment-resistant depression in the VA given the prevalence of these conditions.

While there was little variation in PHQ9 outcomes across facilities, these findings do not inform use of the PHQ9 with individual patients as part of measurement-based care. Although the study did not measure use of PHQ9 scores by individual clinicians, if use by clinicians is consistent across facilities (i.e., consistent high or low-level use), then use would not drive differences in outcomes across facilities despite potential effectiveness with individual patients. It is possible more efficient depression outcome measures (e.g., a single-item assessment of mood)[34] may be sufficient for comparing quality across facilities; however, in settings, where the PHQ9 is routinely collected for patient care, using alternative outcome measures may only add to patient and health system burden.

This study found none of the process measures of adequate antidepressant medication treatment, adequate psychotherapy, or treatment intensification were associated with depression outcomes in the primary analyses. Psychotherapy receipt predicted remission at 6 months in a sensitivity analysis, though the wide confidence interval suggests remaining cautious about this finding. These

primarily null findings could be due to treatment selection bias mitigating treatment effects, such that patients with more severe or persistent symptoms may be more likely to seek and receive adequate treatment or treatment intensification; conversely, patients tend to stop depression treatment once symptom improvement has been achieved.[35,36] Although this study's adjusted models included baseline symptom severity, this may not account for symptom and functional improvements that drive treatment decisions over time.

Although the studied process measures appropriately reflect receipt of effective evidence-based treatments, limitations to the degree to which they capture fidelity to the interventions used in clinical trials could also explain the largely null associations. Measures of antidepressant treatment adequacy use pharmacy prescription fills and not actual ingestion of medication or appropriate dosing, and psychotherapy procedure codes do not ensure the therapist utilized a specific evidence-based psychotherapy protocol. Future research should explore depression care process measures that more accurately assess fidelity, are not reliant on patient treatment adherence, or reflect shared decision-making and patient preferences when treatment does not meet criteria for adequacy.[37,38]

The degree of treatment resistance (i.e., continued symptoms despite prior treatment) in the study population may also explain the lack of associations between process measures and outcomes. In the STAR*D trial of sequential antidepressant treatments, remission rates decreased to 13.7% and 13.0%, respectively, by the third and fourth antidepressant trial rates, and rates of relapse in these groups were high (> 50%) within 12 months.[39] This study's remission rates of < 10% across time points despite subsequent treatment are largely consistent with the patients in STAR*D who did not respond to initial treatments and likely reflect a patient population with more treatment-resistant depression.

This study of VA patients with a high proportion of chronic depression (episode duration > 2 years) receiving care in the Midwest may not generalize to other treatment populations or to VA patients with new-onset depression or residing in other regions of the USA. Generalizability was strengthened by recruiting patients across 29 different VA facilities, and this study avoided referral bias by using medical records to identify participants. The inability to include data from patients who refused any study participation or who were unable to be contacted may bias the results. However, among enrolled participants eligible for follow-up assessments, this study adjusted for characteristics that predicted follow-up completion and had an adequate 74% of participants' follow-up at 6 months. Since the study sample was designed to include a mix of patients with new and existing depression diagnoses evenly distributed across facilities, descriptive characteristics of the sample (e.g., age) may not represent all patients with depression in the VA.

Implications for Behavioral Health

Depression outcome-based quality measures generated using automated methods do not appear reliable for assessing differences in care quality between VA facilities that treat patients with predominantly chronic depression. The lack of association between patient outcomes and measures of antidepressant or psychotherapy use suggests that current process measures may not adequately capture provider fidelity to evidence-based practices or are confounded by treatment nonadherence when patients' symptoms improve. Based on strong associations observed at the individual level, if depression outcomes are used to compare clinics or care teams, outcomes should be adjusted for baseline patient characteristics including depression severity, duration of depression, prior specialty mental health service use, treatment expectancy, and physical functioning. Depression care quality improvement efforts in the VA and related research should focus on identifying and improving care for treatment-resistant depression, given the high prevalence of chronic depression and limited symptom improvement that was observed.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. *Antidepressant medication management.* Washington, DC: National Committee for Quality Assurance (NCQA), 2021. Available at: https://www.ncqa.org/hedis/measures/antidepressant-medication-management/. Accessed 01 December, 2021.
2. Chermack ST, Zivin K, Valenstein M, et al. The prevalence and predictors of mental health treatment services in a national sample of depressed veterans. *Medical Care* 2008; 46(8):13-20. https://doi.org/10.1097/MLR.0b013e318178eb08.
3. Fullerton CA, Busch AB, Normand SL, et al. Ten-year trends in quality of care and spending for depression: 1996 through 2005. *Archives Of General Psychiatry* 2011; 68(12):1218-1226. https://doi.org/10.1001/archgenpsychiatry.2011.146.
4. Farmer MM, Rubenstein LV, Sherbourne CD, et al. Depression quality of care: Measuring quality over time using VA electronic medical record data. *Journal of General Internal Medicine* 2016; 31(1):36-45. https://doi.org/10.1007/s11606-015-3563-4.
5. Donabedian A: The quality of care- how can it be assessed. *Journal of the American Medical Association* 1988; 260(12):1743-1748. https://doi.org/10.1001/jama.1988.03410120089033.
6. Obbarius A, van Maasakkers L, Baer L, et al. Standardization of health outcomes assessment for depression and anxiety: Recommendations from the ICHOM depression and anxiety working group. *Quality of Life Research* 2017; 26(12):3211-25. https://doi.org/10.1007/s11136-017-1659-5.
7. *HEDIS Depression Measures Specified for Electronic Clinical Data Systems*. Washington, DC: National Committee for Quality Assurance (NCQA), 2021. Available at: https://www.ncqa.org/hedis/the-future-of-hedis/hedis-depression-measures-specified-for-electronic-clinical-data/. Accessed on 01 December, 2021.
8. Deutsch R, Smith L, Gage B, et al. Patient-reported outcomes in performance measurement: commissioned paper on pro-based performance measures for healthcare accountable entities., 2012. In Washington, DC: *National Quality Forum,* 2012, pp. 1–46.
9. Carter GC, Cantrell RA, Zarotsky V, et al. Comprehensive review of factors implicated in the heterogeneity of response in depression. *Depression and Anxiety* 2012; 29(4):340-54. https://doi.org/10.1002/da.21918.
10. Van HL, Schoevers RA, Dekker J. Predicting the outcome of antidepressants and psychotherapy for depression: A qualitative, systematic review. *Harvard Review of Psychiatry* 2008; 16(4):225-34. https://doi.org/10.1080/10673220802277938.
11. *Patient Reported Outcomes (PROs) in Performance Measurement.* National Quality Forum website, 2013. Available online at: https://www.qualityforum.org/publications/2012/12/patient-reported_outcomes_in_performance_measurement.aspx. Accessed 1 December, 2021.
12. Kramer TL, Evans RB, Landes R, et al. Comparing outcomes of routine care for depression: The dilemma of case-mix adjustment. *Journal of Behavioral Health Services & Research* 2001;28(3):287-300. https://doi.org/10.1007/BF02287245.
13. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine* 2001; 16(9):606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.
14. Cuijpers P, van Straten A, van Oppen P, et al. Are psychological and pharmacologic interventions equally effective in the treatment of adult depressive disorders? A meta-analysis of comparative studies. *The Journal of Clinical Psychiatry* 2008; 69(11). https://doi.org/10.4088/JCP.v69n1102.
15. Unutzer J, Carlo AC, Arao R, et al. Variation in the effectiveness of collaborative care for depression: Does it matter where you get your care? *Health Affairs* 2020; 39(11):1943-1950. https://doi.org/10.1377/hlthaff.2019.01714.
16. Gilbody S, Bower P, Fletcher J, et al. Collaborative care for depression: A cumulative meta-analysis and review of longer-term outcomes. *Archives of Internal Medicine* 2006; 166(21):2314-2321. https://doi.org/10.1001/archinte.166.21.2314.
17. Schoenbaum M, Unützer J, McCaffrey D, et al. The effects of primary care depression treatment on patients' clinical status and employment. *Health Services Research* 2002; 37(5):1145-1158. https://doi.org/10.1111/1475-6773.01086.
18. Lowe B, Kroenke K, Grafe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *Journal of Psychosomatic Research* 2005; 58(2):163-171. https://doi.org/10.1016/j.jpsychores.2004.09.006.
19. Kroenke K, Spitzer RL. The PHQ-9: A new depression diagnostic and severity measure. Psychiatric Annals 2002; 32(9):509-515. https://doi.org/10.3928/0048-5713-20020901-06.
20. Gjesfjeld CD, Greeno CG, Kim KH. A confirmatory factor analysis of an abbreviated social support instrument: The MOS-SSS. *Research on Social Work Practice* 2008; 18(3):231-237. https://doi.org/10.1177/1049731507309830.
21. Teo AR, Choi H, Valenstein M. Social relationships and depression: Ten-year follow-up from a nationally representative study. *Plos One* 2013; 8(4). https://doi.org/10.1371/journal.pone.0062396.
22. Jensen RE, Potosky AL, Reeve BB, et al. Validation of the PROMIS physical function measures in a diverse US population-based cohort of cancer patients. *Quality of Life Research* 2015; 24(10):2333-2344. https://doi.org/10.1007/s11136-015-0992-9.
23. Bowling A. Just one question: If one question works, why ask several? *Journal of Epidemiology and Community Health* 2005; 59(5):342-345. https://doi.org/10.1136/jech.2004.021204.
24. van Walraven C, Austin PC, Jennings A, et al. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med Care* 2009; 47(6):626-33. https://doi.org/10.1097/MLR.0b013e31819432e5.
25. Levine DS, McCarthy JF, Cornwell B, et al. Primary care-mental health integration in the VA health system: Associations between provider staffing and quality of depression care. *Psychiatric Services* 2017; 68(5):476-481. https://doi.org/10.1176/appi.ps.201600186.

26. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistic in Medicine* 2012; 31(29):3972-3981. https://doi.org/10.1002/sim.5466.
27. Katon WJ. Clinical and health services relationships between major depression, depressive symptoms, and general medical illness. *Biological Psychiatry* 2003; 54(3):216-26. https://doi.org/10.1016/S0006-3223(03)00273-7.
28. Rutherford BR, Marcus SM, Wang P, et al. A randomized, prospective pilot study of patient expectancy and antidepressant outcome. *Psychological Medicine* 2013; 43(5):975-982. https://doi.org/10.1017/S0033291712001882.
29. Visla A, Constantino MJ, Newkirk K, et al. The relation between outcome expectation, therapeutic alliance, and outcome among depressed patients in group cognitive-behavioral therapy. *Psychotherapy Research* 2018; 28(3):446-56. https://doi.org/10.1080/10503307.2016.1218089.
30. Leung LB, Yoon J, Rubenstein LV, et al. Changing patterns of mental health care use: The role of integrated mental health services in veteran affairs primary care. *Journal of the American Board of Family Medicine* 2018; 31(1):38-48. https://doi.org/10.3122/jabfm.2018.01.170157.
31. Pfeiffer PN, Glass J, Austin K, et al. Impact of distance and facility of initial diagnosis on depression treatment. *Health Services Research* 2011; 46(3):768-786. https://doi.org/10.1111/j.1475-6773.2010.01228.x.
32. Schindler A, Hiller W, Witthoft M. What predicts outcome, response, and drop-out in CBT of depressive adults? A naturalistic study. *Behavioural and Cognitive Psychotherapy* 2013; 41(3):365-370. https://doi.org/10.1017/S1352465812001063.
33. Disabato DJ, Kashdan TB, Short JL, et al. What predicts positive life events that influence the course of depression? A longitudinal examination of gratitude and meaning in life. *Cognitive Therapy and Research* 2017; 41(3):444-458. https://doi.org/10.1007/s10608-016-9785-x.
34. Killgore WD. The visual analogue mood scale: Can a single-item scale accurately classify depressive mood state? *Psychological reports* 1999; 85(3 Pt 2):1238-43. https://doi.org/10.2466/pr0.1999.85.3f.1238.
35. Demyttenaere K, Enzlin P, Dewe W, et al. Compliance with antidepressants in a primary care setting, 1: Beyond lack of efficacy and adverse events. *Journal of Clinical Psychiatry* 2001; 62:30-33.
36. Lee AA, Sripada RK, Hale AC, et al. Psychotherapy and depressive symptom trajectories among VA patients: Comparing dose-effect and good-enough level models. *Journal of Consulting and Clinical Psychology* 2021; 89(5):379-92. https://doi.org/10.1037/ccp0000645.
37. Atkins D. The next generation of clinical performance measures. *Journal of General Internal Medicine* 2016; 31(1):3-5. https://doi.org/10.1007/s11606-015-3575-0.
38. Cole S, Reims K, Kershner L, et al. Improving care for depression: performance measures, outcomes and insights from the health disparities collaboratives. *Journal of Health Care for the Poor and Underserved* 2012; 23(3):154-73. https://doi.org/10.1353/hpu.2012.0138.
39. Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *American Journal of Psychiatry* 2006; 163(11):1905-1917. https://doi.org/10.1176/ajp.2006.163.11.1905.