Comparing Outcomes of Routine Care for Depression: The Dilemma of Case-Mix Adjustment

Teresa L. Kramer, PhD
Richard B. Evans, PhD
Reid Landes, MS
Michael Mancino, MD
Brenda M. Booth, PhD
G. Richard Smith, MD

Abstract

The purpose of this study was to formulate and test two case-mix models for depression treatment that permit comparisons of patient outcomes across diverse clinical settings. It assessed demographics; eight, diagnostic-specific, case-mix variables; and clinical status at baseline and follow-up for 187 patients. Regressions were performed to test two models for four dependent variables including depression severity and diagnosis. Individual treatment settings were then ranked based on a comparison of actual versus predicted outcomes using regression coefficients and predictor variables. A model inclusive of baseline physical health status and depression severity predicted depression severity, mental health, and physical health functioning at follow-up. A simpler model performed well in predicting depression remission. This study identifies variables to be included in case-mix adjustment models and demonstrates statistical methods to control for differences across settings when comparing depression outcomes.

Introduction

A complex component of health outcomes evaluation involves case-mix adjustment to account for intrinsic differences among patient populations that may affect the outcomes of care. This issue is particularly salient because variations in the rate of improvement may be due to initial patient factors rather than to the quality of care provided. Therefore, report cards, reimbursement strategies, and incentives based on unadjusted outcomes standards have the potential to bias health care systems or providers who treat more severely ill patients.

Address correspondence to Teresa L. Kramer, PhD, Assistant Professor, Centers for Mental Healthcare Research, 5800 W. 10th Street, Suite 605, Little Rock, AR 72204; e-mail: KramerTeresaL@exchange.uams.edu.

Richard B. Evans, PhD, is an Assistant Professor, Division of Biometry, Centers for Mental Healthcare Research.

Reid Landes, MS, is a Research Associate, Division of Biometry, Centers for Mental Healthcare Research.

Michael Mancino, MD, is a Post Doctoral Fellow, Centers for Mental Healthcare Research.

Brenda M. Booth, PhD, is an Associate Director, Centers for Mental Healthcare Research.

G. Richard Smith, MD, is Director, Centers for Mental Healthcare Research.

Journal of Behavioral Health Services & Research, 2001, 28(3), 287–300. © 2001 National Council for Community Behavioral Healthcare.

Methods that account for preexisting patient differences have been developed for medical-surgical treatment of multiple disorders.[1] By comparison, only a few approaches to case-mix adjustment have been proposed in studies of utilization and outcomes for general behavioral health or specific psychiatric disorders. For example, utilization studies have developed models to predict or compare length of stay across hospital settings using diagnosis, comorbidity and severity of medical and psychiatric problems, prior psychiatric history, and selected demographic variables.[2–4] Hendryx and colleagues[5] recently developed and tested case-mix models specifically for generic clinical outcomes in publicly funded mental health outpatient settings. Case-mix variables included demographics, presence of severe diagnoses (major depression, schizophrenia, or bipolar disorder), substance abuse, baseline functional status, quality of life, and satisfaction with services. In their study, unadjusted versus risk-adjusted scores resulted in differentially ranked clinic performance for each of the dependent variables. They concluded that regression techniques can be applied to risk-adjustment models in public mental health outpatient settings.

Although generic models offer promising mechanisms to adjust outcomes for heterogeneous groups of patients, there is evidence to suggest that other variables may be more accurate for predicting diagnostic-specific outcomes. For example, Sherbourne et al[6] found that depressive symptoms decreased between baseline and a 1-year follow-up for patients who were employed, had fewer medical comorbidities, and used avoidance coping styles less frequently. Research also has demonstrated a poorer prognosis in depressed patients with lower baseline functional status and well-being, as assessed by the SF-36.[7]

The goal of this study is to formulate and test two case-mix models for depression treatment that will permit comparisons of patient outcomes from three clinics with relatively diverse patients. As discussed, models of risk should combine clinical judgment and empirical modeling to maximize statistical performance and acceptability.[8] In light of these considerations, the current study approached the dilemma of case mix using a two-step process.

First, variables that were hypothesized to be related to the outcome domains were selected. The case-mix variables included in this analysis were developed specifically for major depressive disorder (MDD) based on empirical review and consensus by an expert panel of researchers and clinicians. These experts convened to develop the Depression Outcomes Module (DOM), an instrument designed to evaluate the process and outcomes of care for MDD.[9] Items or scales addressing these case-mix variables were written and field-tested as part of DOM development. The self-report measures selected as important case-mix variables that would predict depression outcomes include baseline depression severity, as assessed by the Depression Arkansas Scale (D-ARK)[10]; lifetime history of an alcohol or drug disorder; family history of depression or alcoholism; number of previous psychiatric hospitalizations; age of depression onset; number of previous depressive episodes; number of support persons available to the patient; and number of medical comorbidities, as assessed by a self-report checklist of medical disorders. The reliability and validity of the self-report case-mix variables are discussed elsewhere.[10–12] Demographic variables also were included as predictors of outcomes.

Second, empirical models were developed using methods recommended by others.[5,8] Included in the first and simpler model (Model A) were six demographic variables, which may differentiate patients in various treatment settings, and the eight case-mix variables identified by the expert panel and tested in the DOM. A second more complex model (Model B) also was developed employing two other variables hypothesized to be relevant to depression outcomes: the SF-36 physical component summary (PCS) and the SF-36 mental component summary (MCS) scores.[13] These two variables were added to the model to account for patient differences in functional status attributable to comorbid mental or physical problems that may negatively impact outcomes.[5,7]

Data were derived from clinical outcomes management systems (OMSs) established in three treatment settings to evaluate care under customary conditions for patients and provide a systematic assessment of health outcomes to improve the delivery and quality of services.[14] Because providers are increasingly required to provide data on their outcomes of care for multiple stakeholders as well as accrediting agencies, outcomes monitoring has gained widespread popularity. However, the utility of these efforts will be significantly limited if effective case-mix variables and methods are not identified to adjust outcomes and provide an equitable framework for benchmarking.

Methods

Study populations and settings

This study consists of 336 outpatients diagnosed with and treated for MDD in three types of specialty mental health treatment settings. Of these 336 patients, 187 completed baseline and follow-up assessment.

1. Department of Veterans Affairs (VA) clinic—As part of the OMS for MDD in a mental health clinic of a Veterans Affairs Medical Center, 121 patients were assessed. Of these, 53 (51.5%) patients completed both baseline and 3-month follow-up assessments.
2. University clinic—In an adult psychiatry clinic affiliated with a teaching hospital and university medical center, 159 patients were recruited to participate in an OMS for MDD. Of these, 100 (54.3%) patients completed both baseline and 3-month follow-up assessments.
3. Staff model managed care organization (MCO)—Fifty-six patients treated by providers of a staff model clinic managed behavioral health care organization were enrolled in a depression project using an OMS. Both patient baseline and 3-month follow-up data were available for 34 (60.7%) of these patients.

Measures

The DOM monitors the processes and outcomes of care for depression.[9] It contains a patient baseline assessment (PBA), a clinician baseline assessment (CBA), and a patient follow-up assessment (PFA).

The PBA includes the D-ARK, the SF-36,[13,15,16] and items assessing demographics and case mix. The D-ARK is a self-administered, 11-item instrument that quantifies depression severity in the past 4 weeks, based on a scale of 0 to 3, for each of the nine MDD symptoms of the *Diagnostic and Statistical Manual of Mental Disorders-IV* (DSM-IV).[17] (The D-ARK includes one item to assess weight and appetite change and two items to assess suicidality.) The D-ARK is scored on a 0 to 100 scale by summing all items and multiplying by 33.33. Caseness is determined by the presence of depressive symptoms using DSM-IV criteria. Test-retest reliability for severity of depressive symptoms measured 1 week apart is high ($r = 0.87$; $p < .0001$). Item to scale correlations range from .53 to .88 ($p < .0001$). The D-ARK also has demonstrated high internal reliability (Cronbach's alpha = .92; $n = 44$ and .90; $n = 54$), high concurrent validity with the Inventory to Diagnose Depression[18] ($r = .89$, $n = 44$, and $r = .94$; $n = 54$), and a 94% concordance rate of diagnosis of MDD with the Structured Clinical Interview for DSM-III-R.[10,19]

The CBA provides information on patient diagnosis, exclusion criteria, and psychotropic medications (ie, type, dosage, and date prescribed). Other case-mix items relevant to this study in the DOM were derived from existing instruments. The question on previous psychiatric history in the PBA was derived from the National Health Interview Survey.[20] Items regarding social support and medical comorbidity were derived from the Medical Outcomes Study.[21] Two questions concerning a history of dysthymia, five questions concerning comorbidity for substance abuse, and one question regarding age of depression onset were adapted from the Diagnostic Interview Schedule.[22] Family history of mental illness or substance abuse was assessed using an item from the Family History-Research Diagnostic Criteria Interview.[23] Finally, a history of previous depressive episodes was assessed using an item developed by the DOM expert panel. Cognitive interviews were conducted during the pilot study of the instrument; reliability tests were conducting during the validation phase. A full description of the items and their development is available from the authors.

The PCS and MCS were established by combining the mental health and the physical health subscales of the SF-36, which have been described elsewhere.[13] The reliability of the PCS and the MCS is 0.93 and 0.88, respectively.

The PFA provides information on treatment that the patient has received since baseline and the outcomes of care. Disease-specific outcomes include symptom severity and remission, as assessed by the D-ARK. The patient's mental and physical health functioning is evaluated by the SF-36.

Procedure

A system for monitoring depression outcomes was implemented in three treatment settings. For the purposes of this quality improvement initiative, patients diagnosed with depression were asked by their clinician to complete the DOM as part of their baseline assessment and to monitor their outcomes over time. Patients were 18 years of age or older, able to read and write English, cognitively unimpaired, and consented to participate. Exclusion criteria included clinician report of an organic mental disorder, bipolar affective disorder, a psychotic disorder, uncomplicated bereavement, or depression as a known side effect of some previously prescribed medication.

The PBA and CBA were completed at the time of the initial visit. Follow-up occurred at 3 months following baseline at the time of the clinic visit or by mail if the patient had discontinued treatment. Nonresponsive patients were contacted three times to obtain their data.

Statistical methods

Chi square analyses for nominal data and *t* tests for continuous data were conducted to compare patients who completed only baseline assessments with those who completed both baseline and follow-up assessments. Independent variables included demographic and diagnostic-specific, case-mix factors as well as baseline clinical status. Chi square analyses and analyses of variance were conducted comparing patients in the three clinics on the above variables.

The dependent variables used in the risk-adjustment models were depression severity (D-ARK), depression diagnosis (D-ARK), physical health functioning (SF-36 PCS), and mental health functioning (SF-36 MCS) at follow-up. Demographic predictor variables included age, household income categories, marital status (recoded to married or nonmarried), education (recoded to high school education/some college or college degree), gender, and ethnicity (recoded to Caucasian or non-Caucasian). Other predictor variables included the eight, diagnostic-specific, case-mix factors associated with depression outcomes previously discussed and SF-36 PCS and MCS at baseline. It also was hypothesized that age, age of depression onset and baseline depression severity, and SF-36 PCS and MCS would have a curvilinear effect on outcomes, so quadratic effects were added into the model.

Two regression models were constructed for each of the four dependent variables (six linear and two logistic). The simpler model (Model A) included as potential terms the demographic variables contained within the DOM, the eight case-mix variables selected by the expert panel to predict depression, and three of the five aforementioned quadratic variables. The more complex model (Model B) included all the variables from Model A along with the SF-36 PCS and MCS and their quadratics as potential variables in the models.

To develop the models for adjusting outcomes, only patients completing both the baseline and 3-month follow-up assessments were used for model building ($n = 187$). The data were randomly split such that approximately 80% of the data went to a training set and the remaining subjects into a cross-validation set.[24,25] Using the training set ($n = 149$), all subsets of the independent variables in decreasing order of R^2 (or χ^2 score for logistic regressions) were obtained for each of the four dependent variables. The subsets with the highest R^2 (or χ^2 score) and for which all terms were significant at the .10 level were chosen as models. A level of significance at .10 was used to promote the discovery of relationships among the variables. Therefore, the variables most predictive of outcomes were retained in the final models. To guard against overfitting the data, the validation set was used to cross-validate the models. The criterion for cross-validation was general agreement of R^2 or c—the area under the receiver operating characteristic (ROC) curve.

For the linear regressions, in order to construct prediction intervals to compare with clinic observed means, the mean values for all the predictor variables at each clinic were obtained. These values represent the average patient at that particular clinic. The 95% confidence limits for the expected value of the dependent variables (depression severity, MCS, and PCS) for the average patient at each clinic were calculated. The observed means for each clinic were then seen to either fall in or out of that clinic's 95% prediction interval.

For the logistic regression models, the mean and variance of the predicted values of depression diagnosis for each clinic were obtained. These were used to construct 95% prediction intervals for each clinic, which then were compared with the observed mean from their respective clinic.

Results

Baseline assessments were available on 336 eligible patients. Although demographic or clinical data do not exist regarding patients who did not complete the baseline assessment, clinicians reported that they attempted to recruit the majority of their patients and that only 10% to 20% of the patients declined.

Fifty-six percent ($n = 187$) of those patients completing baseline assessments also participated in the 3-month follow-up. Patients with baseline only versus baseline and follow-up were not significantly different on age, race, marital status, sex, or income. However, patients who completed both assessments were better educated, $\chi^2(2, N = 336) = 6.331$, $p < .05$. In addition, those who completed follow-up versus those who did not were not significantly different on baseline depression severity, baseline SF-36 PCS, MCS, or any of the eight case-mix variables.

Patients in the three clinics were significantly different on demographic, case-mix, and clinical variables at baseline. Results and levels of significance are presented in Tables 1 and 2.

Overall, patients in the MCO were more likely to have characteristics associated with better outcomes, while patients in the VA clinic and university clinic varied. For example, patients in the MCO were more likely to be better educated with higher incomes, were less likely to have had a prior depressive episode, and were more likely to have social supports. Although patients in the MCO were significantly *less* impaired on the SF-36 PCS (mean = 50.6, standard error of measure [SEM] = 1.7) than the VA clinic (mean = 35.7, SEM = 1.1), $t(326) = -7.3$, $p < .001$, they were significantly *more* impaired on the SF-36 MCS (mean = 22.8, SEM = 1.3) compared with the VA clinic (mean = 26.6, SEM = 0.8), $t(326) = 2.52$, $p < .05$.

Differences in unadjusted outcomes were found for the three groups on depression severity, $F(2, n = 179) = 5.15$, $p < .001$; the SF-36 MCS, $F(2, n = 180) = 6.23$, $p < .001$; the SF-36 PCS, $F(2, n = 180) = 21.46$, $p < .001$; and the number of patients who no longer met criteria for MDD at follow-up, $\chi^2(2, n = 187) = 10.48$, $p < .01$. If ranked according to these unadjusted outcomes,

Table 1
Baseline demographics

Demographic variables	VA clinic ($n = 121$)	University clinic ($n = 159$)	MCO clinic ($n = 56$)
	Clinics		
Sex*			
Male	108 (89.3)	41 (25.9)	14 (25.0)
Female	13 (10.7)	117 (74.1)	42 (75.0)
Race†			
African American	23 (19.0)	25 (16.1)	4 (7.3)
Caucasian	92 (76.0)	127 (81.9)	50 (90.9)
Other	6 (5.0)	3 (1.9)	1 (1.8)
Marital status‡			
Married	56 (46.3)	63 (39.9)	38 (70.4)
Divorced/separated	53 (43.8)	65 (41.1)	8 (14.8)
Widowed	2 (1.7)	3 (1.9)	0 (0)
Never married	10 (8.3)	27 (17.1)	8 (14.8)
Education§			
High school graduate or less	59 (48.8)	65 (40.9)	14 (25.9)
Some college	47 (38.8)	54 (34.0)	24 (44.4)
College graduate or more	15 (12.4)	40 (25.2)	16 (29.6)
Income‖			
<$20,000	72 (61.0)	80 (52.3)	4 (7.3)
$20,000–$39,999	35 (29.7)	32 (20.9)	23 (41.8)
$40,000–$59,999	8 (6.8)	21 (13.7)	14 (25.5)
>$60,000	3 (2.5)	20 (13.1)	14 (25.5)
Age** (mean, standard deviation)	50.3 (10.9)	40.9 (12.2)	39 (9.6)
Range	24.4–76.8	18.2–78.8	21.8–63.7

VA, Department of Veterans Affairs; MCO, managed care organization
*$\chi^2(2, N = 336) = 124.99, p < .001$
†χ^2 Not significant
‡χ^2 Not significant
§$\chi^2(2, N = 336) = 6.33, p < .042$
‖χ^2 Not significant
**$F(2, N = 336) = 30.4, p < .0001$

patients in the VA clinic would have had consistently poorer outcomes, while patients in the MCO would have consistently demonstrated better outcomes.

Results of the regression analyses (see Tables 3 and 4) indicated that the more complex model (Model B) was better at predicting depression severity, mental health, and physical health functioning than the simpler model (Model A). With regard to depression severity, two terms—baseline depression severity and the SF-36 MCS—accounted for almost 40% of the variance in outcomes. By comparison, the R^2 of the validation set was somewhat less ($R^2 = .20$). In the models for SF-36 MCS at follow-up, R^2 was .12 for the simpler model and .23 for the more complex model (R^2 for the validation model was .27 and .37, respectively). The strongest predictor models were

Table 2
Baseline case mix variables

Case-mix variables	Clinics VA clinic (n = 121)	 University clinic (n = 159)	 MCO clinic (n = 56)
Depression severity (D-ARK; mean, SD)*	55.9 (19.9)	57.4 (18.7)	54.6 (20.7)
Lifetime history of alcoholism/ drug dependency†	64 (52.9%)	62 (38.9%)	18 (32.1%)
Medical comorbidities (mean, SD)‡	4.0 (2.8)	3.4 (2.7)	2.3 (2.4)
Family history of depression or alcoholism§	61 (51.3%)	94 (60.3%)	36 (65.5%)
Previous psychiatric hospitalizations‖	40 (33.3%)	34 (21.7%)	8 (14.5%)
Age of depression onset**			
<12 years	13 (11.0%)	30 (19.5%)	4 (7.4%)
12–18 years	17 (14.4%)	47 (30.5%)	22 (40.7%)
19–35 years	52 (44.1%)	58 (37.7%)	19 (35.2%)
36–64 years	30 (25.4%)	17 (11.0%)	9 (16.7%)
65+ years	6 (5.1%)	2 (1.3%)	0 (0%)
Number of prior episodes††			
0	6 (5.3%)	10 (6.4%)	10 (18.5%)
1–3	21 (18.4%)	35 (22.4%)	11 (20.4%)
4+	46 (40.4%)	60 (38.5%)	28 (51.9%)
Always	41 (36.0%)	51 (32.7%)	5 (9.3%)
Support persons‡‡			
0	57 (47.5%)	38 (24.2%)	10 (18.2%)
1–3	49 (40.8%)	96 (61.1%)	30 (54.5%)
4–6	12 (10.0%)	18 (11.5%)	14 (25.5%)
6+	2 (1.7%)	5 (3.2%)	1 (1.8%)

VA, Department of Veterans Affairs; MCO, managed care organization; D-ARK, Depression Arkansas Scale; UC, university clinic; SD, standard deviation

* F not significant

† $\chi^2(2, N = 336) = 8.57, p < .01$ (VA, UC > MCO)

‡ $F(2, N = 332) = 8.41, p < .001$ (VA, UC > MCO)

§ Not significant

‖ $\chi^2(2, N = 332) = 8.64, p < .01$ (VA > UC > MCO)

** $\chi^2(8, N = 326) = 31.71, p < .001$ (UC > VA > MCO)

†† $\chi^2(6, N = 324) = 20.79, p < .05$ (UC > VA > MCO)

‡‡ $\chi^2(6, N = 332) = 29.07, p < .001$ (UC > VA > MCO)

for the dependent variable of SF-36 PCS ($R^2 = .56$ and .78 for Models A and B, respectively). Again, the validation set resulted in a higher R^2. Models A and B resulted in equivalent c statistics for depression diagnosis at follow-up. The cross-validation resulted in lower, but acceptable, c statistics.

Table 5 provides the 95% prediction intervals for the mean predicted values of the clinics given their case mix, contrasted against the observed mean (in parentheses) for that clinic. As shown, improvement in depression symptoms and diagnosis were well within the expected range for each

Table 3

Regression statistics for Models A and B for depression severity and diagnosis

Variable	Model A		Model B	
	Depression severity coefficient (SE)	Depression diagnosis coefficient (SE)	Depression severity coefficient (SE)	Depression diagnosis coefficient (SE)
Baseline depression	.27 (.08)‖	.02 (.01)‖	.34 (.8)§	.02 (.01)‖
Income	−.30 (.07)§	−.03 (.01)§		
Medical comorbidity	2.50 (.63)§	.17 (.08)‖		
SF-36 physical	†	†	−.83 (.11)‖	
SF-36 physical*	†	†		.001 (.0002)§
Age of onset*		.0003 (.0002)**		
	$R^2 = .29$ (Adj. $R^2 = .28$)	$c = .78$	$R^2 = .39$ (Adj. $R^2 = .38$)	$c = .80$
	(Validation $R^2 = .19$)	(Validation $c = .69$)	(Validation $R^2 = .20$)	(Validation $c = .63$)

SE, standard error

*Quadratic effect

†Not entered in Model A

§$p < .001$

‖$p < .05$

**$p < .10$

Table 4

Regression statistics for Models A and B for mental and physical health

	Model A		Model B	
Variable	**Mental health coefficient (SE)**	**Physical health coefficient (SE)**	**Mental health coefficient (SE)**	**Physical health coefficient (SE)**
Depression severity	−.10 (.05)‡			
Prior psychiatric hospitalization		−.4.13 (1.93)‖		
Income	.14 (.04)‡	.18 (.03)**		
Medical comorbidity		−2.34 (.26)‖		
SF-36 physical	†	†		1.80 (.27)**
SF-36 mental	†	†	−.80 (.45)‡	
Gender		5.58 (1.50)§		
Support				.53 (.25)‖
SF-36 physical*	†	†	.004 (.0008)**	−0.01 (.003)**
Family history				2.40 (1.10)**
Income*				
Education		4.50 (1.63)§		
Age of onset*	−.002 (.0009)‖		−.002 (.0009)‖	
SF-36 mental*	†	†	.020 (.008)§	
	$R^2 = .12$ (Adj. $R^2 = .10$) (Validation $R^2 = .27$)	$R^2 = .56$ (Adj. $R^2 = .54$) (Validation $R^2 = .65$)	$R^2 = .23$ (Adj. $R^2 = .21$) (Validation $R^2 = .37$)	$R^2 = .77$ (Adj. $R^2 = .76$) (Validation $R^2 = .80$)

SE, standard error

* Quadratic effect

† Not entered in Model A

‡ $p < .10$

§ $p < .01$

‖ $p < .05$

** $p < .001$

Table 5

Performance ratings by clinic for Models A and B

	VA clinic predicted range (observed)	University clinic predicted range (observed)	MCO predicted range (observed)
Model A—Patient outcomes			
Depression severity	39.78–47.16 (46.68)	37.42–44.15 (37.79)	27.92–36.12 (33.25)
Depression diagnosis	.58–.83 (.76)	.47–.67 (.51)	.23–.56 (.41)
Mental health	28.80–33.48 (30.36)	32.43–36.40 (35.85)	35.07–40.02 (33.31)†
Physical health	31.94–36.32 (31.89)†	40.18–43.40 (42.61)	46.37–50.46 (50.74)*
Model B—Patient outcomes			
Depression severity	41.90–48.92 (46.68)	36.80–42.98 (37.79)	25.78–33.21 (33.25)*
Depression diagnosis	.55–.81 (.76)	.47–.66 (.51)	.21–.54 (.41)
Mental health	26.00–31.77 (30.36)	30.37–34.84 (35.85)*	32.84–37.84 (33.31)
Physical health	34.37–37.25 (31.89)†	42.23–45.47 (42.61)	49.13–52.20 (50.74)

*Performance is better than predicted

†Performance is worse than predicted

clinic for Model A, resulting in no clinic outliers. In Model B, the MCO performed somewhat better than expected on depression severity outcomes. The VA clinic performed worse than expected on the PCS under Models A and B, while the MCO performed worse than expected on the MCS and better than expected on the PCS under Model A. Therefore, only the VA clinic was identified as an outlier clinic on only PCS under both models.

Discussion

In this study of MDD outcomes, demographic and case-mix variables differentiated patients among the three treatment settings at baseline. If the demographic and case-mix variables had been similar across sites, it could have been assumed that differences in outcomes were related only to variations in the quality of care provided. However, due to intrinsic patient differences, outcomes were adjusted using two empirically derived models, which may vary in applicability and utility, depending on the treatment site.

As hypothesized, the case-mix variables identified by the expert panel moderately predicted depression severity and diagnosis at follow-up. In fact, Model A accounted for 29% of the variance in depression severity with three unique variables (baseline depression severity, number of medical comorbidities, and income). By adding the SF-36 PCS in Model B, an additional 10% of the variance was accounted for without the variables of income and medical comorbidities. However, the additional burden of administering, scoring, and entering data on the SF-36 for all patients may not offset the additional predictive power obtained.

Although different variables were included in Models A and B for depression diagnosis, the *c* statistic was identical. Model A contained four unique variables (baseline depression severity, number of medical comorbidities, income, and the quadratic effect of age at onset) while Model B contained only two variables. Although Model B appears to be more parsimonious, because it only consists of two variables, it requires the additional administration of the SF-36, which would again significantly increase patient burden and clinic costs. Therefore, Model A may be the most appropriate for those clinics wanting to minimize administrative costs when measuring outcomes by presenting a more parsimonious approach for comparing outcomes. Model A uses two variables (in addition to baseline depression severity) that can be measured by asking only two questions of the patient, as opposed to multiple questions on the SF-36. Model A may therefore be more meaningful to clinicians and less burdensome for patients, while fulfilling the delivery system's needs for internal quality improvement.

By comparison, Model B may be particularly appealing when the necessity of predictive power outweighs the costs of measuring more variables, such as when performance determines how a delivery system is ranked on report cards or whether it receives financial incentives. Interestingly, the hypothesized case-mix variables that were thought to predict depression outcomes were not strongly associated with general mental health functioning at follow-up. Using Model A or B accounted for only 12% and 23%, respectively, of the variance in SF-36 outcomes. These findings are consistent with those of Bovasso and colleagues[26] where R^2 ranged from .14 to .21 in predicting outcomes of psychiatric patients on the General Health Questionnaire.[27] In contrast, Hendryx and colleagues[5] developed models for quality of life and functioning with R^2 in the .30 range. It may be that general mental health functioning varies considerably depending on the time period and assessment measure and that diagnostic-specific, case-mix variables are less applicable to more generic outcome variables.

A substantial proportion of the variance was accounted for in the physical functioning domain. The predictive power of the regression was substantially increased over Model A with Model B: 77% of the variance was accounted for by adding baseline physical functioning as well as the quadratic effect of this variable. The surprisingly higher predictive power of Models A and B on the SF-36 PCS when compared with the SF-36 MCS opens speculation that this may be due to the higher stability of physical functioning over a 3-month period, best predicted by baseline status. Thus, Model B appears to be more effective for adjusting outcomes for physical functioning and well-being in depressed patients, although again it is more burdensome with administration of the SF-36 and measurement of the case-mix variables.

By applying the regression models, this study demonstrated how to adjust outcomes to rank performance by various treatment settings. In each case, an expected predicted outcome and its prediction interval for each clinic were calculated, given the case mix of the respective treatment setting. The predicted value was compared to what was actually observed. All three clinics were ranked similarly on their depression outcomes (severity and diagnosis) using the simpler model (Model A) for case-mix adjustment. In other words, patients' depressive symptoms were within the expected range, and the expected proportion of patients still met criteria for a depression diagnosis at the follow-up, allowing for case-mix adjustment. In addition, the VA clinic performed worse than expected only on physical health outcomes, using either Model A or B for adjustment. The reader may recall that the VA clinic patients were significantly worse than patients in the other three clinics at follow-up without adjusting for case-mix differences. This finding would support the use of such case-mix methods to ensure that the playing field is leveled when patients are more ill in one setting.

This study has several limitations. Because the sample size for the MCO is small relative to the other two clinics, the prediction interval for outcomes is greater, which may account for the tendency for its performance to fall within the expected intervals. This accentuates the need for additional studies with much larger sample sizes. Although a larger follow-up rate also would have enhanced the validity of the study, it is important to note that, with the exception of educational level, there were no significant differences between patients who did and did not complete the follow-up assessment. Second, although the variance accounted for in depression outcomes by patient characteristics is consistent with other studies of general mental health outcomes,[5] there is room for improvement in the predictive ability of the models. The cross-validation resulted in an R^2 that was sometimes higher and sometimes lower than the original test set. Though the apparent degree of fit from the training data set will be closer than the true fit, *on average*,[25] it was encouraging to see that the cross-validation set had a higher percentage of the variation in the dependent variable explained on some of the models. When the cross-validated R^2 was higher than the original R^2, outlier subjects (subjects with large residuals) were removed from the training sets and the R^2 recalculated. In those instances, the R^2 of the training models modestly exceeded the R^2 of the validation models. However, given that the sample size was relatively small and no outlier exerted undue influence on the models, all data were retained including the outliers. Interestingly, the selected variables were still predictive in the validation sets, though not as robustly, possibly due to the small size of the validation set.

In addition, future research is needed to determine whether the predictive ability of the models can be enhanced by additional variables, such as those provided by instrumental variables that reduce selection bias (eg, financial or geographic barriers to care).[28–31] Finally, there is considerable diversity in the populations of the three clinics. Comparisons across VA clinics, MCOs, or hospital-based psychiatric clinics only may significantly alter the models. This present study represents an extreme example of case-mix adjustment, but demonstrates that case-mix adjustment using diagnostic-specific variables and regression techniques can work across very diverse settings.

As discussed by Hendryx and colleagues,[5] risk-adjustment models should include patient factors that are related to outcomes, that explain sufficient variance in outcomes to prevent adverse selection, and that are available and inexpensive. Moreover, they suggest that original models should account for a significant portion of the variance and should be cross-validated on other data sets. The current study incorporated these guidelines, providing a similar approach to case-mix adjustment for depression that is effective in comparing outcomes. This approach resulted in two models with a reduced set of variables that can be tested in other populations.

Implications for Behavioral Health Services

More systems are moving toward performance-based contracting and financial incentives. This requires careful selection of outcome measures and identification of appropriate case-mix variables and methods to compare outcomes across different providers or settings. Otherwise, payers, regulators, and other stakeholders have no idea whether poor outcomes are related to invalid or unreliable measurement tools, treatment of a sicker patient population, or inappropriate or low-quality care.

This study demonstrates that case-mix models can be developed to compare depression outcomes across clinical settings. Not surprisingly, baseline depression severity was a stable predictor of depression severity and diagnosis at the 3-month follow-up. In addition, medical comorbidity, as assessed by patient report of number of medical problems or a more complex measure such as the SF-36, contributed to poorer depression outcomes. Income and age of onset were inconsistent predictors. By comparison, general mental health outcomes in depressed patients, as assessed by the SF-36, were not well predicted by the model, suggesting a need for additional case-mix variables in this domain. Although physical health outcomes are not usually of primary concern in behavioral health settings, they can best be predicted by baseline physical health status, which may have more relevance in primary care settings where providers are addressing both aspects of the patient's well-being.

Once the appropriate variables were identified through expert opinion and empirical validation, clinics were ranked on their depression outcomes. Without these adjustments, the VA clinic consistently performed worse than the other two clinics, but when expected versus observed outcomes were compared, ranking improved. The relative performance of the other two clinics in the outcomes domains also changed, providing strong evidence for the use of risk-adjustment models in outcomes management systems.

Although this article offers two examples of model development with specified outcome variables, there are multiple strategies that can be employed to adjust outcomes across settings. Two critical steps in this process are (1) agreement on what outcomes and potentially relevant case-mix variables to measure and when, and (2) data sharing across large behavioral health settings. By collaborating in such a manner, providers and treatment agencies can begin to build models for a number of behavioral health disorders across age groups (children and adolescents versus adults), payer sources (private versus public), and treatment modalities (inpatient versus outpatient). Future work should be devoted to improving these models by validating them on other samples of depressed patients and applying the statistical methods to other diagnoses. In addition, investigators need to determine what factors differentiate high-performing clinics from others and techniques for transferring these best practices to other settings. These efforts will contribute significantly to our understanding of risk and provide the necessary tools for national benchmarking and standards.

Acknowledgments

This research was supported, in part, by an Independent Scientist Award from National Institute of Drug Abuse (NIDA) to Dr Booth (K02 DA00346), by a National Institute of Mental Health (NIMH) Centers for Mental Health Services Research grant to Dr Smith (P50-MH48197), and by a T-32 grant from NIMH to Dr Mancino (MH20024-01).

References

1. Keeler EB, Kahn KL, Draper D, et al. Changes in sickness at admission following the introduction of the prospective payment system. *Journal of the American Medical Association.* 1997;264:1962–1968.
2. Fortney JC, Booth BM, Smith GR. Variation among VA hospitals in length of stay for treatment of depression. *Psychiatric Services.* 1996;47:608–613.
3. Hodges K, Wong MM. Use of the child and adolescent functional assessment scale to predict service utilization and cost. *The Journal of Mental Health Administration.* 1997;24:278–290.
4. Horn SD, Chambers AF, Sharkey PD, et al. Psychiatric severity of illness: a case mix study. *Medical Care.* 1989;27:69–84.
5. Hendryx MS, Dyck DG, Srebnik D. Risk-adjusted outcome models for public mental health outpatient programs. *Health Services Research.* 1999;34:171–195.
6. Sherbourne CD, Hays RD, Wells KB. Personal and psychosocial risk factors for physical and mental health outcomes and course of depression among depressed patients. *Journal of Consulting and Clinical Psychology.* 1995;63:345–355.
7. Wells KB, Burnam MA, Rogers W, et al. The course of depression in adult outpatients: results from the Medical Outcomes Study. *Archives of General Psychiatry.* 1992;49:788–794.
8. Iezzoni LI. *Risk Adjustment for Measuring Health Care Outcomes.* 2nd ed. Melrose Park, IL: Health Administration Press; 1997.
9. Smith GR, Burnam A, Burns B, et al. *Major Depression Outcomes Module: User's Manual.* Little Rock, AR: University of Arkansas for Medical Sciences; 1994.
10. Kramer TL, Smith GR. Tools to improve the detection and treatment of depression in primary care. In: Maruish MEP, ed. *Handbook of Psychological Assessment in Primary Care Settings.* Manway, NJ: Lawrence Erlbaum Associates; 1999.
11. Rost K, Smith GR, Burnam MA, et al. Measuring the outcomes of care for mental health problems: the case of depressive disorder. *Medical Care.* 1992;30(5, suppl):MS266–MS273.
12. Rost K, Williams C, Wherry J, et al. The process and outcomes of care for major depression in rural family practice settings. *Journal of Rural Health.* 1995;11:114–121.
13. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): conceptual framework and item selection. *Medical Care.* 1992;30:473–483.
14. Ellwood PM. Shattuck lecturee—Outcomes management: a technology of patient experience. *New England Journal of Medicine.* 1988;318:1549–1556.
15. Ware JE, Kosinski M, Bayliss MS, et al. Comparison of methods for the scoring and statistical analysis of the SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. *Medical Care.* 1995;33:AS264–AS279.
16. Ware JE, Snow KK, Kosinski M. *SF-36 Health Survey Manual and Interpretation Guide.* Boston: New England Medical Center, Health Institute; 1993.
17. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.* Washington, DC: American Psychiatric Association; 1994.
18. Zimmerman M, Coryell W, Corenthal C, et al. A self-report scale to diagnose major depressive disorder. *Archives of General Psychiatry.* 1986;43:1076–1081.
19. Spitzer RL, Williams JBW, Gibbon M, et al. The structured clinical interview for DSM-III-R (SCID) 1: history, rationale, and description. *Archives of General Psychiatry.* 1992;49:624–629.
20. National Center for Health Statistics. *Current Estimates from the US National Health Interview Survey.* Washington, DC: Government Printing Office; 1989.
21. Tarlov AR, Ware JE, Greenfield S, et al. The Medical Outcomes Study: an application of methods for evaluating the results of medical care. *Journal of the American Medical Association.* 1989;262:925–930.
22. Robins LN, Helzer JE, Croughan J, et al. National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics and validity. *Archives of General Psychiatry.* 1981;38:381–389.
23. Andreasen NC, Endicott J, Spitzer RL. The family history method using diagnostic criteria. *Archives of General Psychiatry.* 1977;34:1229–1235.
24. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis.* Boca Raton, FL: Chapman & Hall/CRC; 2000:46–47.
25. Mosteller F, Tukey JW. *Data Analysis and Regression: A Second Course in Statistics.* Reading, MA: Addison-Wesley; 1977:37.
26. Bovasso GB, Eaton WW, Armenian HR. The long-term outcomes of mental health treatment in a population-based study. *Journal of Consulting and Clinical Psychology.* 1999;67:529–538.
27. Goldberg DP. *The Detection of Psychiatric Illness by Questionnaire.* London: Oxford University Press; 1972.
28. Sun GW, Shook TL, Kay GL. Inappropriate use of bi-variable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology.* 1996;49:907–916.
29. Barnow BS, Cain GG, Goldberger AS. Issues in the analysis of selectivity bias. *Evaluation Study Review Annual.* 1980;5:43–59.
30. Breen R. *Regression Models: Censored, Sample Selected or Truncated Data.* Thousand Oaks, CA: Sage; 1996.
31. Moffitt R. Program evaluation with non-experimental data. *Evaluation Review.* 1991;15:291–314.