Methods Section

Risk-Adjusted Outcome Models for Public Mental Health Outpatient Programs

Michael S. Hendryx, Dennis G. Dyck, and Debra Srebnik

Objective. To develop and test risk-adjustment outcome models in publicly funded mental health outpatient settings. We developed prospective risk models that used demographic and diagnostic variables; client-reported functioning, satisfaction, and quality of life; and case manager clinical ratings to predict subsequent client functional status, health-related quality of life, and satisfaction with services.

Data Sources/Study Setting. Data collected from 289 adult clients at five- and ten-month intervals, from six community mental health agencies in Washington state located primarily in suburban and rural areas. Data sources included client self-report, case manager ratings, and management information system data.

Study Design. Model specifications were tested using prospective linear regression analyses. Models were validated in a separate sample and comparative agency performance examined.

Principal Findings. Presence of severe diagnoses, substance abuse, client age, and baseline functional status and quality of life were predictive of mental health outcomes. Unadjusted versus risk-adjusted scores resulted in differently ranked agency performance.

Conclusions. Risk-adjusted functional status and patient satisfaction outcome models can be developed for public mental health outpatient programs. Research is needed to improve the predictive accuracy of the outcome models developed in this study, and to develop techniques for use in applied settings. The finding that risk adjustment changes comparative agency performance has important consequences for quality monitoring and improvement. Issues in public mental health risk adjustment are discussed, including static versus dynamic risk models, utilization versus outcome models, choice and timing of measures, and access and quality improvement incentives.

Key Words. Risk adjustment, mental health, outpatient programs

The purpose of this article is to develop and test outcomes risk-adjustment models for public mental health outpatient treatment programs. "Outcomes risk" refers to the probability of a poor outcome that clients bring to a treatment episode: the higher a client's pre-intervention risk status, the lower the expected outcome to the episode of care, all other things being equal. The reason for developing risk-adjustment models where treatment outcomes are the dependent variables is to enable public treatment agencies and state mental health authorities to improve the quality of care. Risk-adjustment models can contribute to quality improvement by enabling outcomes to be compared fairly across agencies, by providing outcome data for state mental health authorities (SMHAs) to use in imposing performance-based financial consequences on provider agencies, and by providing agencies with incentives to improve access for patients at the highest severity levels. SMHAs are responsible for making providers accountable for outcomes of care delivered to publicly supported consumers. Such accountability is fair only if it can be defined in risk-adjusted terms.

We address the statistical qualities of risk-adjustment models in a sample of mental health outpatients treated in public agencies in Washington state. The results demonstrate that regression techniques developed to conduct risk-adjustment in other health areas (e.g., Iezzoni 1994; Hornbrook and Goodman 1996) can be applied to public mental health outpatient settings. The results suggest the types of variables to be included in mental health outpatient risk-adjustment outcome models. The results also offer preliminary evidence that using adjusted versus unadjusted outcomes will lead to different conclusions about the comparative performance of mental health treatment agencies. The Discussion section of the article develops the applications of these models for quality improvement purposes and outlines risk-adjustment issues in public mental health.

BACKGROUND

Providing medical care services is an expensive responsibility of state and county governments. In 1990, state governments funded or operated 2,859

Address correspondence and requests for reprints to Michael S. Hendryx, Ph.D., Associate Professor, The Washington Institute for Mental Illness Research and Training, Washington State University, 601 West First Ave., Spokane, WA 99201. Dennis G. Dyck, Ph.D. is Professor and Director, The Washington Institute for Mental Illness Research and Training, Washington State University, and Debra Srebnik, Ph.D. is Assistant Professor, Dept. of Psychiatry and Behavioral Sciences, University of Washington. This article, submitted to *Health Services Research* on August 29, 1997, was revised and accepted for publication on June 4, 1998.

specialty mental health organizations (Lutterman 1994) and controlled an estimated $14 billion in expenditures in the specialty mental health sector (Frank and McGuire 1996). In Washington state these expenditures totaled over $106 million (Lutterman and Hollen 1992). In response to high costs and limited federal support, states are implementing capitated payment contracts, global budgets, and other managed care strategies for publicly funded (largely Medicaid) mental health care (Anderson et al. 1996; Brach 1995; Essock and Goldman 1995; Masland et al. 1996). In Washington state, outpatient services are funded via capitated contracts between the state and 14 regional support networks. At the time of this study inpatient care had not yet been incorporated into these contracts; such incorporation is now taking place. The state public mental health system provided treatment to over 90,000 outpatients in 1997.

State mental health authorities (SMHAs) have the responsibility to ensure that publicly funded treatment agencies deliver appropriate and effective services and that clients with complex needs have access to services. States are rapidly developing outcomes measuring systems for outpatient mental health treatment, and are using outcomes to establish standards, disseminate performance reports, and reward best performers (National Association of State Mental Health Program Directors [NASMHPD] 1998). However, SMHAs cannot rely on unadjusted outcome indicators that are affected by variables outside of the mental health system. To establish an outcome standard that does not take into account differences in prior severity might encourage agencies to select less ill clients and to underserve clients who are more ill. From a consumer protection perspective, risk adjustment is needed to improve access and quality of care for those persons with greater healthcare needs. From a treatment agency perspective, risk-adjustment models can be used to identify and evaluate quality improvement efforts.

Risk-adjustment techniques have been relatively well developed in certain areas of medical care, such as inpatient ICU care (Knaus, Wagner, Draper, et al. 1991), and predicting Medicare costs (Gruenberg, Kaganova, and Hornbrook 1996; Hornbrook and Goodman 1996). Studies have also been conducted to predict inpatient psychiatric or substance abuse readmission (e.g., Peterson, Swindle, Phibbs, et al. 1994) and to predict utilization in order to set risk-adjusted capitation rates in the public mental health sector (Brach 1995; Masland et al. 1996). Predicting utilization to set payment rates is problematic to the extent that rates provide an incentive to maintain the severity levels of mental health clients who are receiving long-term treatment, because a healthier client population would reduce payment levels in the next payment cycle. This is one reason why incentives for quality improvement must also be incorporated into payment rates. This issue will be reexamined in the Discussion section of the article. SMHAs are beginning to incorporate risk-adjustment of outcomes: Massachusetts tracks functional status and well-being stratified by baseline scores in its Medicaid reporting requirements (Brach 1995), and Indiana reports 90-day and one-year improvement rates stratified by three baseline functioning groups. However, regression-based risk-adjustment models that predict outcomes have not been addressed for public outpatient programs that treat persons with serious and persistent mental illness.

RISK-ADJUSTMENT MODEL DEVELOPMENT AND CRITERIA

As noted by Hornbrook and Goodman (1996), risk-adjustment models should possess a number of attributes. They should incorporate client characteristics that are related to health status and other relevant outcomes, including pre-intervention health status; they should explain sufficient variance to reduce incentives for adverse selection of patients who are not as ill; and they should use available and inexpensive data. Risk-adjustment models should be stable and not take advantage of chance variation in data; this may be accomplished by developing prospective models that predict outcomes at time t from variables at time $t-1$ (Newhouse, Beeuwkes, and Chapman 1997), and by developing models based on multiple runs on different samples (Hornbrook and Goodman 1996). Models derived from a particular sample should also be validated in new samples.

In addition, data sources should be sufficiently robust so that the variables cannot be deliberately manipulated. Although the collection of client self-report data has been criticized for this reason (Newhouse, Beeuwkes, and Chapman 1997), reliance on data reported by treatment providers is also open to this possibility. Our approach is to rely, therefore, on data reported from multiple sources, including clients, treatment providers, and on agency management information system (MIS) data collected by treatment providers for other purposes.

Selection of relevant outcomes for public mental health outpatients, and selection of the appropriate candidate predictors, is an important but difficult issue. Stakeholder groups, including mental health consumers, providers, and administrators, often debate what constitutes service goals and outcome priorities (Nelson 1994). Our selection of outcomes for the care of persons with serious mental illness was based on the conceptual model of Rosenblatt and Attkisson (1993) and on input from a multi-stakeholder advisory committee composed of consumers, family members, providers, and regional and state administrators. Rosenblatt and Attkisson (1993) proposed four outcome domains: clinical status (psychopathology and symptomatology), functional status (the ability to fulfill social and role functions), life satisfaction (including satisfaction with services but also indicators of well-being or happiness), and welfare and safety (which has also been termed quality of life and concerns basic and fundamental needs). There was general agreement among the stakeholders that each of these domains needed to be included. Consequently, our choice of survey tools, as described under Methods, was intended to capture these domains. Appropriate predictors are those variables that relate to the outcome domains. These variables might include age, gender, race, and diagnosis to the extent that they are demographic proxy measures for physiological health, population cohort, social roles, stress, discrimination, gender-related genetic factors, or other influences. Health status is a candidate predictor (Hornbrook and Goodman 1995). Because our focus is on a population that often receives ongoing care for serious and persistent mental illness, taking into account the baseline or prior level of an outcome domain may also be important; such measures may capture some of the predictable individual differences in outcomes. Finally, if the outcome domains themselves are interrelated, a baseline score in one domain might be predictive of an outcome in another domain. As hypothetical examples, if service satisfaction influences treatment motivation and compliance, satisfaction might be a predictor of later functional outcomes; and a better functional capacity might improve one's chances for a better subsequent quality of life.

We employed risk-adjustment methods as developed and recommended by others, especially Hornbrook and Goodman (1996); Gruenberg, Kaganova, and Hornbrook (1996); Newhouse, Beeuwkes, and Chapman (1997); and Iezzoni and colleagues (Iezzoni 1994), and extended these methods to multiple relevant outcomes in the public sector outpatient mental health field. For confidence in developing accurate risk-adjustment models for mental health outpatient programs, we needed to ensure that the models satisfy a number of criteria:

1. Original models should account for significant portions of outcome variance; model content validity is improved by capturing more of the predictable outcome variance.
2. Validation models should be accurate; average prediction errors

found when the models are applied to a new sample should be small and nonsignificant.

3. The variance of predicted values should approximate the observed variance in a validation sample, indicating that the models predict different outcomes for different individuals.
4. The predicted values should provide a good fit to the observed distribution in a validation sample, measured by the size and significance of adjusted R^2s and by intercepts close to zero.
5. Prediction errors should be uncorrelated with policy-relevant groups in a validation sample, including groups defined by age, gender, and race, indicating that the models are not biased in terms of these variables.
6. Models should have a practical effect as measured by the relative performance of agencies using unadjusted versus adjusted outcomes.
7. Models should satisfy these criteria independently for each relevant outcome.

METHODS

STUDY POPULATION AND SETTING

The data for this study come from adult (18 years of age and older) outpatients treated in one of six publicly funded community mental health agencies in Washington state in 1995–1997. The original purpose of the study was to evaluate the possible impact on outcome and efficiency variables of a new state law that eliminated some procedural rules in place of developing an outcomes system. The mental health agencies are located in primarily rural and suburban areas. Two cohorts of clients were followed over time. The cohorts were constructed by attempting to survey all clients who entered a given treatment agency for a scheduled appointment over a selected period of time; the period of data collection at each agency reflected the agency's total caseload. Clients were approached by other clients hired and trained for the study and were asked to complete the survey instrument, and about 60 percent of clients thus approached agreed to participate. The two cohorts entered the study at different times: Cohort 1 during May–August 1995 and Cohort 2 during March–June 1996. Each cohort was measured at three time points approximately five months apart. Models are developed from Cohort 1 and validated in Cohort 2; validation results presented later in this article are thus based on a different sample and a different time period than those on which the models were constructed.

The initial sample sizes were 229 for Cohort 1 and 218 for Cohort 2, or 447 total. Complete cases numbered 289 at the first follow-up (Time 2) and 260 at the second follow-up (Time 3), representing 65 percent of the original sample size at first follow-up and 58 percent at second follow-up. Cases were lost to follow-up primarily because subjects dropped out of treatment or terminated treatment, but in some instances (about 23 percent of the lost cases) subjects were dropped because the case manager failed to complete clinical status ratings. The final samples thus included only those in treatment through each of the follow-up periods, and are less representative of short-term clients. Most analyses are based on the sample with completed baseline and first follow-up measures, where loss to follow-up was less severe than when the second follow-up sample was included; therefore, descriptive characteristics of the sample are based on this Time 1–Time 2 group.

The average age of the sample was 45.1 years, and 58 percent were women. Eighty-nine percent were white; because of small numbers of clients in specific ethnic categories we were forced to code the race category as white or nonwhite. MIS data indicated that 49 percent had a primary diagnosis that we classified as severe: schizophrenia, major depression, or bipolar disorder. The sample was dominated by long-term clients; 79 percent of them had been receiving services in the state public mental health system for at least three to four years, and 41 percent had received some form of treatment services in each of the prior 14 calendar quarters.

We analyzed the representativeness of the completed sample at Time 2 relative to cases lost to follow-up. Complete cases did not differ from cases lost to follow-up on gender, race (white or nonwhite), prevalence of the three diagnosis groups, or baseline scores on the outcome domains. One group difference was significant: the average age of completed cases was younger (mean 45.1), relative to that of lost cases (mean 48.6; $t = -2.07$, df = 446, $p < .04$). We also compared the sample to the entire treatment population at these agencies along demographic and diagnostic variables. The sample was not significantly different from the population of mental health clients at the treatment agencies in terms of gender, race, or most diagnostic groups. However, the sample relative to the population was older (the population mean was 34.9, $t = -4.00$, df = 4360, $p < .0001$), and was more likely to have a schizophrenia diagnosis (27 percent of the sample versus 13 percent of the population had this diagnosis, $\chi^2 = 39.02$, df = 1, $p < .001$).

MEASURES

The intent in selecting the measures was to produce a brief yet sound and comprehensive survey packet. As noted in the introductory sections of this article, the measures were selected with input from a study advisory committee that included providers, administrators, and consumers. The measures were also pilot-tested by members of a consumer advisory committee. We wanted the measures to reflect multiple outcome domains important to mental health clients (Rosenblatt and Attkisson 1993; McGlynn 1996): satisfaction with services, quality of life, functional capacities, and clinical status. The self-report and clinician-rated measures have been found to be psychometrically reliable and valid, and clients and treatment staff could complete them in brief periods of time (Srebnik, Hendryx, Stevenson, et al. 1997). (The client and case manager survey instruments are available from the authors upon request.)

The survey relied primarily on previously developed and psychometrically tested instruments. The survey included the eight-item Client Satisfaction Questionnaire (CSQ) (Nguyen, Attkisson, and Stegner 1983) and the SF-12 (McHorney, Ware, and Raczek 1993). Consistent with published guidelines (McGee et al. 1996), we used SF-12 items to create a three-item Activities of Daily Living (ADL) scale, and a three-item Role scale. The Role scale was formed to represent the extent to which physical or emotional problems interfered with work, daily, or social behaviors. A four-item Mental Health Status scale was also formed from the items in the Mental Health and the Energy SF-12 scales. Seven items were drawn from the Lehman Quality of Life Interview (Lehman 1991) to measure social functioning and safety. Items from a California public mental health survey (Veit 1995) measured goals in four areas: work, school, training, and volunteering; we computed the difference between actual and desired activity in each area, resulting in four items we labeled goal satisfaction. We developed two items to assess client involvement in treatment. We also developed two items to assess whether treatment was appropriate to the client's age and ethnic/cultural background. One item asked clients to rate their perceived safety at the mental health center. Four items assessed client skills in managing their own stress and symptoms. Two items assessed whether clients had been victimized by violent or nonviolent crime. Two items assessed living conditions (whether the client had enough food and money and whether he or she wanted to move from a current residence or remain there).

Case managers completed the 4-D Classification Scale (Comtois, Ries, and Armstrong 1994), which includes four seven-point scales (scored 0 through 6) that assess symptomatology, functioning, substance abuse, and treatment compliance. Based on the intercorrelations among the items, we calculated a mean from three items that was used as our indicator of clinical status (we labeled this three-item indicator the 3-D) and used the remaining item, on degree of substance abuse, as a separate predictor.

We performed a maximum likelihood factor analysis with oblique rotation to examine the composition of the survey items relative to the conceptual outcome domains of Rosenblatt and Attkisson (1993). We used an oblique rotation because the outcome domains were significantly intercorrelated. The factor analysis results suggested three outcome domains consistent with the conceptual framework: satisfaction, functioning, and quality of life. The factor eigenvalues were 7.76, 3.30, and 1.11, respectively. When we tried to force a four-factor solution, the analysis failed due to commonalities greater than one. Particular scales used to represent each factor are listed by bullet further on. The clinical domain, which we had hoped to capture with the 4-D, was not present. In order to represent each domain with one variable, scale scores (e.g., the CSQ) were standardized to mean = 10 and standard deviation = 1. A domain score was then calculated as the mean of the standardized scores, with higher scores being favorable. The resulting outcome scores were normally distributed, with small standard deviations and little skewness. Each domain was used as one of our outcome indicators. From the MIS data we used client gender, age, race, and primary diagnosis.

The following lists summarize the dependent and independent variables used in the risk-adjustment models.

Dependent Variables

- Functional Status domain (role, ADLs, mental health, skills, and social functioning)
- Quality of Life domain (crime, safety, living conditions, and the 3-D)
- Satisfaction with Services domain (CSQ, involvement, appropriateness, and safety at the mental health center)

Independent Variables

- Sex
- Race
- Age
- Presence of severe primary diagnosis: major depression, schizophrenia, or bipolar disorder

- Baseline levels of substance abuse as rated by case managers
- Baseline functional status
- Baseline quality of life
- Baseline satisfaction with services

ANALYSES

Estimation Techniques

The first steps involved the use of linear regression (Schwartz and Ash 1994) to identify models for each outcome variable, using Time 1 data to predict outcomes measured at Time 2. Using data from Cohort 1, we conducted a series of regression analyses for each dependent variable. The first regression analysis included only demographic variables as predictors, the second added diagnosis and substance abuse to demographics, the third used only the baseline rating of the matching outcome domain, and the fourth used all predictors. The purpose of this approach was to examine the ability of the various models to contribute explained outcome variance as measured by adjusted R^2, as used in other risk-adjustment studies (Gruenberg, Kaganova, and Hornbrook 1996).

However, a risk in model identification lies in overfitting the model by taking advantage of chance variation in the data. To address this risk, we conducted a fifth regression analysis based on results of a stability analysis. The stability analysis was conducted by drawing 25 independent repeated samples with replacement (Gruenburg, Kaganova, and Hornbrook 1996). Each sample consisted of a randomly selected 50 percent of the entire two-cohort sample. There is no objective rule for retaining stable predictors; we retained only predictors that were significant at $p < .20$ in at least 10 of the 25 test samples. We refer to this specification as the reduced set model. The size of the adjusted R^2 values of the five model formulations is our test of Criterion 1.

Model Performance Tests

After examining the model identification results, the strongest models for each dependent variable were validated on a separate sample. This was done by using the intercepts and coefficients obtained from the Cohort 1 models to predict Cohort 2, Time 2 outcomes. The validity of the models was examined by testing the degree of average prediction error (Criterion 2), and by testing the variance of predicted values relative to the variance of observed values (Criterion 3). Validity was also tested by regressing the predicted outcomes on actual outcomes; the adjusted R^2s should be significant and intercepts should be close to zero (Criterion 4). We also tested the models by examining the correlation between predicted errors and demographic categories that included age, gender, and race (correlations should be nonsignificant to meet Criterion 5). Finally, we computed at the agency level a mean predicted to observed ratio for each Cohort 2–Time 2 outcome domain, and compared the relative ranks of the agencies when using these ratios versus using the unadjusted observed outcome scores (Criterion 6). Criterion 7 means that the foregoing tests should be satisfied independently for each outcome, indicating that multiple relevant outcomes can be successfully represented by outcome-specific risk-adjustment models.

As a final test of the time stability of the models, we used the intercept and coefficient terms identified in the Cohort 1 reduced set model to predict second follow-up or Time 3 outcomes in Cohort 2. This is the only use of the Time 3 data and is used as an exploratory test of the degree to which predictive strength may decrease over time.

RESULTS

MODEL IDENTIFICATION

Tables 1–3 show the results of the regression analyses conducted on the Cohort 1 sample, one table per outcome. The first model, which includes only demographics, failed to account for significant variance in two of three outcomes; only greater age predicted better ratings of quality of life. Although this finding appears counterintuitive for physical health outcomes, it reflects the well-known observation that positive symptoms associated with mental illnesses such as schizophrenia frequently abate with age. The addition of diagnosis and substance abuse in the second model made a significant improvement only to the functioning outcome. The third model included only the baseline indicator of each respective outcome and shows a strong relationship between baseline and follow-up score in all three outcomes. Although the improvements were modest, the comprehensive and/or reduced set models increased the adjusted R^2 for all three outcomes. Because the baseline, comprehensive, and reduced set models were the only ones to satisfy Criterion 1 (that models should account for significant portions of outcome

Table 1: Regression Statistics for Alternative Risk-Adjustment Models in Predicting Time 2 Functioning

Baseline Variables	*Demographic*		*Diagnosis and Substance Abuse*		*Baseline Only*		*Comprehensive*		*Reduced Set*	
	Coefficient	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*
Age	.0035*	.0031	−.0001*	.0044	–	–	−.0016*	.0038	–	–
Race	.0853*	.1918	.0798*	.1834	–	–	.1771*	.1562	–	–
Gender	.0702*	.1137	.0219*	.1301	–	–	−.0158*	.1103	–	–
Major depression	–	–	−.6130**	.3001	–	–	−.5151**	.2547	–	–
Schizophrenia	–	–	.4700***	.1344	–	–	.4089***	.1190	.3209***	.0976
Bipolar disorder	–	–	.2123*	.1738	–	–	.1230*	.1481	–	–
Substance abuse	–	–	.0383*	.0431	–	–	.0444*	.0372	–	–
Functioning	–	–	–	–	.6044***	.0718	.4744***	.0847	.5416***	.0730
Satisfaction	–	–	–	–	–	–	−.0171*	.0761		
Quality of life	–	–	–	–	–	–	.2284	.0869	.1468**	.0744
Intercept	9.78	0.34	9.64	0.45	4.06	0.72	2.78	1.17	3.12	0.89
Adjusted R^2	−.01	–	.11	–	.31	–	.37	–	.35	–
F-ratio	0.47*	–	3.16***	–	70.81***	–	7.89***	–	29.06***	–

* ns; ** $p < .05$; *** $p < .01$.

Table 2: Regression Statistics for Alternative Risk-Adjustment Models in Predicting Time 2 Satisfaction

Baseline Variables	*Demographic*		*Diagnosis and Substance Abuse*		*Baseline Only*		*Comprehensive*		*Reduced Set*	
	Coefficient	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*
Age	.0080***	.0040	.0083*	.0057	–	–	.0085**	.0047	.0069*	.0043
Race	.0277*	.2453	.0043*	.2375	–	–	.0055*	.1930	–	–
Gender	.0992*	.1450	.2310*	.1679	–	–	.1633*	.1358	–	–
Major depression	–	–	.0892*	.3887	–	–	–.0241*	.3147	–	–
Schizophrenia	–	–	.1064*	.1735	–	–	.0209*	.1465	–	–
Bipolar disorder	–	–	.1228*	.2248	–	–	.0008*	.1828	–	–
Substance abuse	–	–	.1428*	.0558	–	–	.0836**	.0459	.0726**	.0427
Functioning	–	–	–	–	–	–	.2497***	.1046	.2548****	.0988
Satisfaction	–	–	–	–	.7224****	.0760	.6215****	.0940	.6280****	.0892
Quality of life	–	–	–	–	–	–	.0286*	.1074	–	–
Intercept	9.39	0.44	8.49	0.58	2.77	0.76	–0.06	1.44	0.49	1.18
Adjusted R^2	.007	–	.03	–	.36	–	.38	–	.39	–
F-ratio	1.36*	–	1.51*	–	90.28****	–	8.04****	–	16.27****	–

* ns; ** $p < .10$; *** $p < .05$; **** $p < .01$.

Table 3: Regression Statistics for Alternative Risk-Adjustment Models in Predicting Time 2 Quality of Life

Baseline Variables	*Demographic*		*Diagnosis and Substance Abuse*		*Baseline Only*		*Comprehensive*		*Reduced Set*	
	Coefficient	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*	*Coefficient*	*Std. Error*
Age	.0091****	.0026	.0094**	.0039	–	–	.0061**	.0034	.0048***	.0021
Race	.0024*	.1539	.0316*	.1610	–	–	.1478*	.1391	–	–
Gender	–.0047*	.0933	.0138*	.1150	–	–	–.0457*	.0991	–	–
Major depression	–	–	–.3315*	.2496	–	–	–.2834*	.2151	–	–
Schizophrenia	–	–	–.1032*	.1200	–	–	–.0157*	.1077	–	–
Bipolar disorder	–	–	.0768*	.1557	–	–	.0672*	.1347	–	–
Substance abuse	–	–	.0357*	.0384	–	–	.0382*	.0333	–	–
Functioning	–	–	–	–	–	–	.1336**	.0769	.1307***	.0605
Satisfaction	–	–	–	–	–	–	.0019*	.0662	–	–
Quality of life	–	–	–	–	.5108****	.0593	.4515****	.0786	.4520****	.0616
Intercept	9.61	0.28	9.42	0.39	4.93	0.60	3.61	1.05	3.98	0.75
Adjusted R^2	.06	–	.05	–	.31	–	.31	–	.34	–
F-ratio	4.55****	–	1.86**	–	74.11****	–	6.22****	–	29.13****	–

* ns; ** $p < .10$; *** $p < .05$; **** $p < .01$.

variance) consistently across the three outcomes, only these three models were carried forward to the model validation analysis.

The particular independent variables that predicted an outcome varied across dependent variables. The three demographic indicators were rarely stable predictors, although age was retained in two of the three reduced set models. Specifically, age was positively associated with higher satisfaction with services and higher quality of life. The presence of a schizophrenia diagnosis was retained in one model. Ratings of baseline status were retained in every model; better prior scores predicted better outcomes. The baseline measure of functioning was also retained in the satisfaction and quality of life models, and the baseline measure of quality of life was retained in the functioning model.

TIME 3 PREDICTION

Using a 10-month follow-up period still resulted in significant adjusted R^2s in the reduced set models (.31, .28, and .23 for functioning, satisfaction, and quality of life, respectively). Although significant, the strength of the prediction was reduced as time elapsed.

MODEL VALIDATION

We validated the models by using the regression intercepts and coefficients derived from Cohort 1 data to predict Cohort 2–Time 2 outcomes. This was done for the baseline, comprehensive, and reduced set models. Results are summarized in Table 4. In all cases the predicted mean fell within the 95 percent confidence interval of the observed mean. This result satisfied Criterion 2 (that differences between observed and predicted means should be nonsignificant). Predicted variances were significantly smaller than observed, but the extent to which these differences could be reduced is unclear; our conclusion is that Criterion 3 was partially met but that the models require additional refinement to increase their prediction accuracy.

Criterion 4 was met, in that adjusted R^2s were significant in all nine models, and intercepts were usually not different from zero. Criterion 5 was largely met as only one statistically significant correlation was found between prediction errors and client age, gender, and nonwhite race. (A correlation between gender and residual errors in the comprehensive model of functioning was the exception.)

In the prediction of functioning, the adjusted R^2 of the reduced set model was the same as the baseline model, and the absolute magnitude of average prediction error was also the same. However, the intercept was smaller in the

Table 4: Model Validation Results

	Functioning			*Satisfaction*			*Quality of Life*		
	Baseline	*Comprehensive*	*Reduced Set*	*Baseline*	*Comprehensive*	*Reduced Set*	*Baseline*	*Comprehensive*	*Reduced Set*
Predicted mean	9.906	9.868	9.894	9.987	9.934	9.955	9.956	9.976	9.968
Observed mean	9.900	9.900	9.900	10.009	10.009	10.009	9.991	9.991	9.991
Prediction error	.006*	−.032*	−.006*	−.022*	−.075*	−.054*	−.035*	−.015*	−.023*
Maximum error	−1.437	1.395	−1.576	−2.701	−2.667	−2.660	2.171	1.941	2.092
Minimum error	−.006	−.008	.007	.002	.019	−.002	−.010	−.002	−.001
Observed variance	.653	.653	.653	.794	.794	.794	.521	.521	.521
Predicted variance	.371	.291	.359	.179	.258	.227	.153	.144	.171
F-ratio	1.76**	2.24**	1.83**	4.45**	3.08**	3.50**	3.40**	3.61**	3.05**
Intercept	−2.483**	0.357*	−1.413*	2.592*	2.109*	2.629**	−1.086*	1.515*	−0.083*
Adjusted R^2	.56***	.45***	.56***	.19***	.26***	.23***	.31***	.29***	.34***
Gender error	.14*	.20**	.09*	−.03*	−.09*	−.02*	.07*	.15*	.09*
Race error	−.10*	−.17*	−.11*	−.03*	.03*	−.01*	.02*	−.05*	.04*
Age error	−.02*	−.03*	−.09*	.09*	.04*	.02*	.03*	.02*	−.06*

* ns; ** $p < .05$; *** $p < .0001$.

reduced set model compared to the intercept in the baseline model. In the prediction of satisfaction, both the comprehensive and reduced set models were slightly superior to the baseline model; the comprehensive model had the highest adjusted R^2, the smallest F-ratio, and the smallest intercept. In the prediction of quality of life, the reduced set model was slightly superior to the other two, as evidenced by a higher adjusted R^2, and smaller intercept and F-ratio.

COMPARING AGENCY PERFORMANCE

As a preliminary test of Criterion 6, the impact of the risk-adjustment models, we calculated for each agency an observed to expected ratio for each outcome. In this ratio, the observed score was the actual mean outcome domain score for Cohort 2 at Time 2, and the expected score was the corresponding predicted mean at Time 2 based on the regression equations. We then ranked the six agencies first using the observed unadjusted outcome scores and then using the observed to expected ratios, and we visually compared the ranks. This was done for the baseline, comprehensive, and reduced set models. The results, summarized in Table 5, show that different ranks result when using adjusted versus unadjusted performance scores. In a few instances, the difference in ranks was large; for example, from the reduced set model, agency D ranked second on unadjusted quality of life but fifth on adjusted quality of life, and agency C ranked sixth on unadjusted functioning but second on adjusted functioning. Differences in ranks between adjusted and unadjusted outcomes existed in all three model specifications, although they were not always the same differences. Finally, by comparing the ranked observed to expected ratios across outcomes, Table 5 demonstrates that an agency's ranked risk-adjusted performance on one outcome did not agree with its ranks on other outcomes. Note also that Criterion 7 is met to the extent that outcome-specific risk-adjusted models satisfied Criteria 1 through 6 consistently for all three outcome domains.

DISCUSSION

Our conclusion is that the models offer a promising start in risk-adjusting mental health outpatient outcomes in public treatment agencies. Each of the seven model criteria was satisfied partially or wholly. The presence of a severe diagnosis or of substance abuse, prior scores in the same domain, prior scores in other domains, and client age were significant predictors of one

Table 5: Unadjusted Cohort 2, Time 2 Outcome Ranks Compared to Ranked Observed-to-Expected (O/E) Ratios

Agency	*Unadjusted rank*	*Baseline O/E rank*	*Reduced Set O/E rank*	*Comprehensive O/E rank*
Functioning				
A	1	1	1	1
B	3	4	4	3
C	6	5	2	5
D	2	2	3	4
E	5	6	6	6
F	4	2	5	2
Satisfaction				
A	1	1	1	1
B	4	5	3	4
C	6	4	4	3
D	3	3	5	5
E	2	2	2	2
F	5	6	6	6
Quality of Life				
A	3	2	3	2
B	5	4	4	4
C	6	5	6	6
D	2	6	5	5
E	4	3	2	3
F	1	1	1	1

or more behavioral health outcomes. Although significant outcome variance is strongly predicted by prior levels of the same variable (e.g., baseline functioning predicts follow-up functioning), other variables contribute to prediction as well, albeit modestly, and improve the overall strength of the models. There is nothing undesirable about the fact that different predictors are significant for different outcomes; research on inpatient medical care outcomes is consistent with this (DesHarnais, McMahon, and Wroblewski 1991) and confirms the importance of developing outcome-specific risk-adjustment models.

The reduction in prediction associated with a longer time interval between the collection of predictors and outcomes indicates that shorter times (five months in this study) may be preferable in forecasting outcomes. There are, however, cost trade-offs that must be considered, as well as judgments about the appropriate time frames for particular treatments or illnesses. More frequent measurements are more expensive and reduce the time period in which treatment agencies have the opportunity to affect client status. The benefits of increased statistical prediction, higher follow-up rates, and quicker outcome data must be weighed against these costs. The timing and the number of repeated measurements taken on mentally ill treatment populations is an important topic to consider as risk-adjustment methods are developed.

The results have practical implications in efforts to develop contracts between governments and mental health treatment agencies, and in efforts to improve quality. The finding that unadjusted and adjusted ranks lead to different conclusions regarding comparative agency performance suggests that SMHAs should consider implementing risk-adjustment models as they develop performance contracts that incorporate information from outcomes monitoring systems. We recommend that SMHAs first examine the impact of unadjusted versus adjusted comparative performance, using their own outcomes systems, to confirm whether adjusted outcomes make a difference. To the extent that risk-adjusted outcomes do change comparative agency performance, failure to risk-adjust might lead to serious mistakes by SMHAs in their decisions regarding the agencies with which to contract, which agencies to reward or punish with performance bonuses, the agencies on which to expend audit resources, and other uses of outcomes performance data.

Using risk-adjustment models requires that the models be fair with respect to the age, gender, and race distributions of treatment agencies. Models developed in one setting should not be biased against a particular group of clients if they are applied in new settings that have different age, gender, or race distributions. This is the purpose of testing the models by comparing the correlations of these groups to model errors. In this study, correlated errors were rarely present even when the only predictor was the baseline domain score. However, these variables should continue to be included in risk models as a safeguard to protect agencies that serve vulnerable groups from models that do not reflect their performance fairly. Initial development of models should also attend to adequate representation of all groups to which models will ultimately be applied.

Limitations and Next Steps

The finding that the model specification (baseline versus comprehensive versus reduced set) results in differently ranked performance indicates that the form of the model that is used for applications is important. Unfortunately, the question of which model specification to use has not been definitively answered by this study. However, work can proceed to improve the range and quality of predictor variables. This may be investigated by including diagnostic and functional status measures with greater sensitivity and specificity, and by using a stronger instrument to capture clinical status. Eventually, a reduced set approach can identify the most efficient set of predictors from a larger pool. As validation models are improved, the intercept terms will approach zero and the slopes will approach one, indicating that the outcomes are more accurately predicted at the low and high ends of the distribution.

Efforts can also be made to capture more fully the Rosenblatt and Attkisson (1993) outcome domains. Functional outcomes that include the ability to work and maintain independent residential status would be important to include, for example, as would life satisfaction indicators that go beyond satisfaction with services. Because this study was limited to persons in treatment through the follow-up periods, future research may investigate successful or unsuccessful termination of treatment as an outcome, including incarcerations, hospitalizations, or loss to follow-up, as well as successful resolution. The models should also be tested and confirmed, and comparative ranks examined, in larger, urban, and more ethnically diverse samples. In addition, our results are limited to adult clients, and risk-adjustment models for children will likely have different predictors and different relevant outcome measures.

CONCLUSIONS: ISSUES IN PUBLIC MENTAL HEALTH RISK-ADJUSTED OUTCOME MODELS

Static Condition or Dynamic Change

In using risk-adjusted models an important distinction exists between models that represent risk at a static point in time and models that represent change or improvement over time. Both types of models have a particular advantage, and both must be combined in a comprehensive risk-adjustment approach. The static model, which addresses risk at a baseline period, may be used to help set fair capitated reimbursement rates. Such models have the potential to encourage access for the most severely ill if rates provide appropriate financial compensation for doing so. It may even be possible to set rates that will encourage providers actively to seek out the most severely ill people and so to reduce the probability of adverse selection. However, in a population of seriously mentally ill clients, where ongoing and long-term treatment is common, a "baseline" measure may be an arbitrary point in time that represents clients at various lengths of time in treatment. If static models alone are used, agencies have no incentive to improve outcomes if such improvement will make the agency appear to have a less severely ill population at the next payment cycle. Therefore, incentives must also exist to improve the quality of care by tracking the outcome status of a sample relative to the earlier profile of that same sample: individual clients must be tracked over time. The tools to accomplish the static and dynamic functions of risk adjustment are, respectively, the expected outcome score for the agency at a baseline period (i.e., the denominator of the observed to expected ratio), and the observed to expected ratio itself at the follow-up, to capture favorable treatment progress. If it is clear to agency providers and administration that an SMHA has enacted both incentives, they may conclude that the most logical strategy is to improve access for the most severely ill clients and engage in their best efforts to improve outcomes.

Measurement: Choice of Outcomes, Predictors, and Timing

The study reported here was only partially successful in representing the Rosenblatt and Attkisson (1993) domains. Although baseline measures of the outcomes of interest are relevant to include given the ongoing nature of treatment in this population, work remains to be done to improve the predictive strength further over that afforded simply by prior measures of the same variable. As suggested, both conceptually and by our model results, a single global outcome measure for mental health may not be appropriate: predictors and agency performance both vary by outcome. Setting appropriate incentives then becomes complicated by the number of different outcomes and by the importance that various stakeholders attach to the various domains. We made no attempt to weight the relative importance of the domains; such an attempt would have introduced an additional level of complexity and was premature. However, assigning weight to the domains remains an important problem for future research on risk adjustment for mental health outcomes.

The timing and number of repeated measures is also a multifaceted issue. Collecting initial measures of predictors as a client begins a treatment episode would provide a true baseline and an assessment unconfounded with earlier treatment. As an SMHA implements a comprehensive risk-adjustment approach, it may wish to conduct such measures as new clients enter services over time, and indeed, some SMHAs routinely collect baseline functioning and clinical measures on all new clients. However, limiting data collection to only new clients ignores the many clients with long-term psychiatric disabilities who have received ongoing care from the agency. It may be necessary to form a baseline observation period from all (or a sample of) new clients that began treatment over a given interval plus a sample of already enrolled clients. As the model matures and all ongoing clients have been assessed, it may be possible to convert to baselines of only new clients. Of course, such an approach would need to be coordinated so that the various agencies could conduct the data collection in a comparable way. From each baseline sample, follow-up measures could be done on the same individuals at a later point in time (e.g., at six months or at several six-month intervals) and at the time of discharge or termination of treatment.

Using Models to Create Access and Quality Improvement Incentives

It was suggested earlier that risk-adjustment can be used for improving access and quality of care. An important task for agencies and SMHAs will be to adopt risk-adjustment models that can accomplish both objectives. For example, nontechnical reports can be generated to provide comparative static risk assessments, using the expected outcome scores, and to provide dynamic assessments using observed to expected ratios. Observed to expected ratios are easy to interpret, as numbers greater than one are favorable and numbers less than one are unfavorable. The next step would be to convert these multiple outcome indicators to actual payment figures and performance bonuses or sanctions. Setting performance bonuses or sanctions based on dynamic results might be done by an SMHA in collaboration with providers, family members, and consumer advocates, who collectively would decide on an amount (such as a percentage of the capitation rate) and a trigger for the bonus (such as the number of required outcome indicators in which performance exceeds expected results, based on their consensual weighting of the various outcomes).

Predicting Utilization and Predicting Outcomes

Static models may be particularly relevant for the prediction of utilization, with capitation rates set according to baseline predictions of future use. To the extent that predictions of worse outcomes and the prediction of heavy utilization identify the same clients, such models can achieve their inherent potential for improving access. The contribution of outcome prediction to the static model may be to refine payments to the extent that heavy users of the services do not coincide with the most severely ill clients. The most severely ill clients may not always be the heaviest users because they may be more likely to leave treatment prematurely, more likely to die via suicide (Flechtner, Wolf, and Preibe 1997), and more likely to be in poor adherence in keeping scheduled services appointments.

ACKNOWLEDGMENTS

This research was supported by the Washington State Department of Social and Health Services, Mental Health Division. The authors thank John Whitbeck, Ph.D. for his guidance and critical comments during the course of this study, and the anonymous reviewers of a previous draft of this article for their valuable insights and contributions.

REFERENCES

Anderson, D. F., J. L. Berlant, D. Mauch, and W. R. Maloney. 1996. "Managed Behavioral Health Care Services." In *The Managed Health Care Handbook, 3rd Edition,* edited by P. R. Kongstvedt. Gaithersburg, MD: Aspen.

Brach, C. 1995. *Designing Capitation Projects for Persons with Severe Mental Illness: A Policy Guide for State and Local Officials.* Technical Assistance Monograph 3. Mental Health Policy Resource Center, Technical Assistance Collaborative, Boston.

Comtois, K. D., R. Ries, and H. E. Armstrong. 1994. "Case Manager Ratings of the Clinical Status of Dually Diagnosed Outpatients." *Hospital and Community Psychiatry* 45 (6): 568–73.

DesHarnais, S., L. F. McMahon, and R. Wroblewski. 1991. "Measuring Outcomes of Hospital Care Using Multiple Risk-adjusted Indexes." *Health Services Research* 26 (4): 425–45.

Essock, S. M., and H. H. Goldman. 1995. "States' Embrace of Managed Mental Health Care." *Health Affairs* 14 (3): 34–44.

Flechtner, K., T. Wolf, and S. Preibe. 1997. "Suicide Rates in a Community Psychiatric Service System." *Nervenarzt* 68 (7): 566–73.

Frank, R. G., and T. G. McGuire. 1996. "Introduction to the Economics of Mental Health Payment Systems." In *Mental Health Services: A Public Health Perspective,* edited by B. L. Levin and J. Petrila, pp. 22–37. New York: Oxford University Press.

Gruenberg, L., E. Kaganova, and M. C. Hornbrook. 1996. "Improving the AAPCC with Health-Status Measures from the MCBS." *Health Care Financing Review* 17 (3): 59–75.

Hornbrook, M. C., and M. J. Goodman. 1996. "Chronic Disease, Functional Health Status, and Demographics: A Multi-Dimensional Approach to Risk Adjustment." *Health Services Research* 31 (3): 283–307.

———. 1995. "Assessing Relative Health Plan Risk with the RAND-36 Health Survey." *Inquiry* 32 (1): 56–74.

Iezzoni, L. I., ed. 1994. *Risk Adjustment for Measuring Health Care Outcomes.* Chicago: Health Administration Press.

Knaus, W. A., D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano, and F. E. Harrell. 1991. "The APACHE III Prognostic System: Risk Prediction of Hospital Mortality for Critically Ill Hospitalized Adults." *Chest* 100 (6): 1619–36.

Lehman, A. 1991. *Quality of Life Interview: Core Version.* University of Maryland, Center for Mental Health Services Research.

Lutterman, T. C. 1994. "The State Mental Health Agency Profile System." In *Mental Health, United States, 1994,* edited by R. W. Manderscheid and M. A. Sonnenschein, pp. 165–87. Department of Health and Human Services, Pub. No. (SMA) 94-3000. Washington DC: Government Printing Office.

Lutterman, T. C., and V. L. Hollen. 1992. "Change in State Mental Health Agency Revenues and Expenditures Between Fiscal Years 1981 and 1990." In *Mental Health, United States, 1992,* edited by R. W. Manderscheid and M. A. Sonnenschein, pp. 163–207. Department of Health and Human Services, Pub. No. (SMA) 92-1942. Washington DC: Government Printing Office.

Masland, M. C., G. Piccagli, L. Snowden, and B. J. Cuffel. 1996. "Planning and Implementation of Capitated Mental Health Programs in the Public Sector." *Evaluation and Program Planning* 19 (3): 253–62.

McGee, J., N. Goldfield, K. Riley, and J. Morton. 1996. *Collecting Information from Health Care Consumers: A Resource Manual of Tested Questionnaires and Practical Advice.* Gaithersburg, MD: Aspen Pub.

McGlynn, E. A. 1996. "Setting the Context for Measuring Patient Outcomes." In *Using Client Outcomes Information to Improve Mental Health and Substance Abuse Treatment: New Directions for Mental Health Services,* edited by D. M. Steinwachs, L. M. Flynn, G. S. Norquist, and E. A. Skinner, pp. 19–32. San Francisco: Jossey-Bass.

McHorney, C., J. Ware, and A. Raczek. 1993. "The MOS 36-Item SF-36 II: Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs." *Medical Care* 31 (3): 247–63.

National Association of State Mental Health Program Directors (NASMHPD). 1998. National Technical Assistance Center for State Mental Health Planning. *http://www/nasmhpd.org/ntac/.*

Nelson, G. 1994. "The Development of a Mental Health Coalition: A Case Study." *American Journal of Community Psychology* 22 (2): 229–55.

Newhouse, J. P., M. Beeuwkes, and J. D. Chapman. 1997. *Risk Adjustment and Medicare.* New York: The Commonwealth Fund, Program on Medicare's Future.

Nguyen, T., C. Attkisson, and B. Stegner. 1983. "Assessment of Patient Satisfaction: Development and Refinement of a Service Evaluation Questionnaire." *Evaluation and Program Planning* 6 (3): 299–314.

Peterson, K. A., R. W. Swindle, C. S. Phibbs, B. Recine, and R. H. Moos. 1994. "Determinants of Readmission Following Inpatient Substance Abuse Treatment: A National Study of VA Programs." *Medical Care* 32 (6): 535–50.

Rosenblatt, A., and C. C. Attkisson. 1993. "Assessing Outcomes for Sufferers of

Severe Mental Disorder: A Conceptual Framework and Review." *Evaluation and Program Planning* 16 (4): 347–63.

Schwartz, M., and A. S. Ash. 1994. "Evaluating the Performance of Risk Adjustment Methods: Continuous Measures." In *Risk Adjustment for Measuring Health Care Outcomes,* edited by L. I. Iezzoni, pp. 287–312. Chicago: Health Administration Press.

Srebnik, D., M. S. Hendryx, J. Stevenson, S. Caverly, D. G. Dyck, and A. M. Cauce. 1997. "Development of Outcome Indicators for Monitoring the Quality of Public Mental Health Care." *Psychiatric Services* 48 (7): 903–909.

Veit, S. 1995. California Mental Health Performance Outcome Project. Report for the California State Department of Mental Health.