A Predictive Model of Hospitalization Risk Among Disabled Medicaid Enrollees

John F. McAna, PhD; Albert G. Crawford, PhD; Benjamin W. Novinger, MS; Jaan Sidorov, MD; Franklin M. Din, DMD; Vittorio Maio, PharmD; Daniel Z. Louis, MS; and Neil I. Goldfarb, BA

Objectives: To identify Medicaid patients, based on 1 year of administrative data, who were at high risk of admission to a hospital in the next year, and who were most likely to benefit from outreach and targeted interventions.

Study Design: Observational cohort study for predictive modeling.

Methods: Claims, enrollment, and eligibility data for 2007 from a state Medicaid program were used to provide the independent variables for a logistic regression model to predict inpatient stays in 2008 for fully covered, continuously enrolled, disabled members. The model was developed using a 50% random sample from the state and was validated against the other 50%. Further validation was carried out by applying the parameters from the model to data from a second state's disabled Medicaid population.

Results: The strongest predictors in the model developed from the first 50% sample were over age 65 years, inpatient stay(s) in 2007, and higher Charlson Comorbidity Index scores. The areas under the receiver operating characteristic curve for the model based on the 50% state sample and its application to the 2 other samples ranged from 0.79 to 0.81. Models developed independently for all 3 samples were as high as 0.86. The results show a consistent trend of more accurate prediction of hospitalization with increasing risk score.

Conclusions: This is a fairly robust method for targeting Medicaid members with a high probability of future avoidable hospitalizations for possible case management or other interventions. Comparison with a second state's Medicaid program provides additional evidence for the usefulness of the model.

Am J Manag Care. 2013;19(5):e166-e174

For author information and disclosures, see end of text.

According to a report from the Kaiser Commission on Medicaid and the Uninsured, in 2011, due to the recession, there were 2 major factors driving state needs to control costs in their Medicaid programs: reduced state budgets and increased enrollment in programs. However, despite these economic pressures, a survey of Medicaid directors conducted by the commission showed a commitment to assuring access to high-quality care delivered in the most efficient manner possible.[1] States want to improve the care and well-being of participants in their Medicaid programs and at the same time bring some control to rising program costs.

One method of meeting this commitment would be to identify those segments of the Medicaid population accounting for disproportionate percentages of the costs and within those groups to identify the highest risk members and intervene early in order to avoid unnecessary high-cost care. In 2003, the elderly and disabled constituted approximately 25% of the Medicaid population. However, they accounted for about 70% of Medicaid spending that year: 43% by people with disabilities and 26% by the elderly. Only 30% of the spending was accrued by the remaining 75% of the Medicaid population.[2]

Predictive modeling may assist Medicaid plans in identifying program participants at highest risk of future health problems. According to Knutson and Bella,[3] "Predictive models are data-driven, decision-support tools that estimate an individual's future potential healthcare costs and/or opportunities for care management." Predictors can be derived from administrative data. This was an approach taken by Billings et al[4] when developing a predictive model for the National Health Service in England to identify patients at high risk for rehospitalization. Claims and enrollment data are readily available to payers and can be used in models to target specific groups of interest and to provide risk scores for individuals. These scores could then be provided to case/care/disease managers to help them more readily identify those in need of their services.

A randomized trial conducted by Wennberg et al[5] has shown that a targeted care management program can be successful in reducing medical costs and hospitalizations. Billings and Mijanovich[6] showed that care management for chronic disease Medicaid patients who had been hospitalized could be cost-effective and could improve the health of this population. Of importance to Medicaid plans, they showed that existing data resources can

In this article
Take-Away Points / e167
Published as a Web exclusive
www.ajmc.com

be used to predict patients at greatest risk of future hospital readmissions within 12 months of an index admission.

These studies also stress the need for developing models and care management plans specific to Medicaid populations. Eligibility requirements, such as low income and/or disability, and other factors typically associated with this population (eg, homelessness, substance use, or low educational achievement) distinguish them from those populations typically covered by commercial plans and their vendors.

Take-Away Points

Predictive models are powerful tools that can be used to estimate future healthcare costs and opportunities for interventions for individuals.

- Administrative data can be successfully used to identify individuals for care management.
- This study provides a robust method for developing a predictive model to identify these individuals.
- The model is based on available data; most of the derived variables are relatively easy to generate from the data; and risk scores, either developed on a proprietary basis or open source, are easily incorporated into the model.

Hospitalizations are known to be high-cost events and are easily identifiable and categorized from claims and encounter data. It is also well known from the literature that patients with chronic diseases and multiple comorbidities are at high risk for hospitalization or rehospitalization.[7] This situation should hold true regardless of which state Medicaid plan is under study. For these reasons, developing a model predictive of hospitalization for patients with chronic diseases and multiple comorbidities would provide the best opportunity for targeting patients for case/care management that could reduce avoidable costs and be generalizable across states.

In this article, we describe the development of a model to predict hospitalizations among enrollees identified as disabled in a state Medicaid program. Its purpose was to identify Medicaid patients, based on 1 year of administrative data, at high risk of admission to a hospital in the next year and most likely to benefit from outreach and targeted interventions. Previous studies have examined a similar population, but specific to readmissions,[6] or have looked at specific diagnoses.[8] Applying the model to Medicaid data from a second state supports the generalizability of the model to other programs in other states.

METHODS

Claims and Enrolment Data

Data for a 2-year period (2007 and 2008) were extracted from a data mart containing a subset of membership and claims information for all Medicaid enrollees in 1 state Medicaid program. (The state was in the southern part of the United States; however, contractual agreements prevent the authors from identifying the actual state studied.) Data extracted for 2007 (measurement year) were used to derive the predictors for the model; outcomes were derived from 2008 (prediction year) data.

The claims experience of the eligible members included information from all measurement year claims/encounters. Inpatient, outpatient, professional, and pharmacy claims were used to obtain predictors based on utilization and diagnosis. Eligibility and enrollment files were used to establish study eligibility and to provide demographic predictor variables.

Study Population

The Medicaid population is composed of numerous subpopulations defined by the state's eligibility categories and benefits structures. It would be inappropriate to develop 1 model based on the entire eligible population. Each group has its own characteristics, risk factors, and outcomes.

The population chosen for this study included disabled enrollees who were fully covered by Medicaid and were continuously enrolled for both measurement year and prediction year. Members were identified as disabled if they were enrolled in one of the aid categories defined by the state for the disabled. The disabled were chosen because they comprised a large portion of the enrolled Medicaid members for the state and were more likely than most other enrollees to be continuously enrolled for at least 2 years. In 2009, for the state under study, the disabled comprised 19.2% of the Medicaid population and accounted for 44.4% of the costs. Fully covered, continuously enrolled members were chosen because of the need for the most complete claims picture possible. Those enrollees with full Medicaid coverage were chosen to reduce loss of information due to incomplete Medicare claims data from the states involved.

The Model

Logistic regression was used to provide predicted probabilities for the occurrence of an inpatient hospital stay for individuals. The regression was performed in Stata and a stepwise process was used for including variables in the model (P was set at .05 for inclusion). The model was specified in a prospective manner. Demographic, utilization, diagnosis, and prescription drug data for 2007 were used to predict hospitalizations in 2008. The coefficients for the most powerful (ie, statistically significant) variables were used.

The dependent variable was the occurrence of an inpatient stay in the prediction year. Inpatient stays were identified by the occurrence of a valid, paid inpatient claim. Admissions

due to major trauma or pregnancy were omitted, as these were felt to be unpredictable from the data and less amenable to intervention. Trauma that could be treated outpatient or through an emergency department (ED) was not addressed in this study. The independent variables used in the model included the following:

1. Inpatient stay(s) in the measurement year
2. Total length of stay (in days)
3. Primary/preventive office visit in the measurement year
4. Gender
5. Race
6. Age
7. Charlson Comorbidity Index
8. Chronic disease score
9. Mental health diagnosis in the measurement year (substance abuse)
10. Mental health diagnosis in the measurement year (other than substance abuse)
11. ED visit in the measurement year
12. Polypharmacy (8 or more different drugs prescribed in the measurement year)
13. The disease categories included in the chronic disease score (cystic fibrosis, end-stage renal disease [ESRD], human immunodeficiency virus [HIV], anxiety and tension, asthma, bipolar disorder, cardiac disease, coronary/peripheral vascular, depression, diabetes, epilepsy, gastric acid disorder, glaucoma, heart disease/hypertension, hyperlipidemia, hypertension, inflammatory bowel disease, liver failure, malignancies, pain, pain with inflammation, Parkinson's disease, psychotic illness, renal disease, rheumatoid arthritis, thyroid disorder, transplant, and tuberculosis)

The model was developed by using a 50% sample of claims data for 1 state. The model was validated in 2 ways. It was tested against the other 50% of the eligible population and also against a second state (in the Midwest) to evaluate its generalizability. Stepwise logistic regressions were run separately for each of the samples and the results were compared. Variables were included and excluded based on their significance level in relation to the other variables included in the model. Because the chronic disease score (CDS) and Charlson Comorbidity Index (CCI) and individual disease categories were included in the model, multicollinearity was a possibility. The standard errors of the parameter estimates were examined to determine if multicollinearity existed. Multicollinearity was a possibility if any of the standard errors were large (ie, over 2). All of the standard errors were much lower than 2.

The performance of the model was evaluated using the receiver operating characteristics (ROC) curve. Based on data from the first 12-month period, the model also assigned scores reflecting each member's risk of hospitalization in the second 12-month period.

Disease Classification and Severity Adjustment

Because of the wide range of conditions included under the heading of "disabled," adjustment for severity and comorbidity were needed in the model. The CCI[9,10] and the CDS[11] were used to provide this adjustment. The CCI is a widely used index based on the *International Classification of Diseases, Ninth Revision, Clinical Modification* (*ICD-9-CM*) codes. Originally developed to predict mortality, it has been shown to be generally useful in controlling for comorbidities for other purposes. Cox proportional hazards modeling was used in determining its contents and weighting scheme. The Statistical Analysis System (SAS) code used for the CCI score was based on the work of Hude Quan et al.[10] They based their work on the Deyo version of the CCI,[9] which adapted the original CCI for use with administrative databases, updating the codes used to reflect changes in the CCI during the intervening period. It was felt that this version of the CCI score was the most up-to-date. For the purposes of this study, diagnoses were derived from all inpatient, outpatient, and professional claims and encounters. The CDS is based on current medication use. It was created by a panel of health professionals using a pharmaceutical database to reach consensus decisions as to which classes of medications should be included in the score and how they should be weighted to correspond to disease complexity and severity.[5] The Clark version[11] of the CDS was chosen based on the findings of a study by Putnam et al[12] that showed the version performed somewhat better than the others at predicting hospitalization. Each individual disease category was also used in the model to identify those with the most significant impact on admission.

RESULTS

The mean age, sex, and racial composition of the 3 study populations are presented in **Table 1**. There are only minimal differences between the two 50% samples from the first state. However, there are obvious demographic differences between the 2 states. The second state has substantially smaller percentages of disabled members 21 years and younger, and of members identified as black. Members in

Table 1. Descriptive Statistics: Disabled Members Continuously Enrolled 2 Years

		State 1: Year 1		State 1: Year 2		State 2: Year 2	
		N = 42,382	%	N = 41,592	%	N = 71,431	%
Sex	Female	22,687	53.5	22,425	53.9	38,515	53.9
	Male	19,695	46.5	19,167	46.1	32,916	46.1
Age group, y	0-21	15,061	35.5	14,721	35.4	10,193	14.3
	22-45	12,850	30.3	12,461	30.0	37,520	52.5
	46-64	13,202	31.2	13,189	31.7	23,718	33.2
	65+	1269	3.0	1221	2.9	0	0
Race	Black	22,027	52.0	21,618	52.0	9881	13.8
	White	16,078	37.9	15,834	38.1	59,654	83.5
	Other	4277	10.1	4140	10.0	1896	2.7
Inpatient admissions, measurement year	0	34,767	82.0	34,060	81.9	63,554	89.0
	1-2	6507	15.4	6341	15.3	6366	8.9
	3+	1108	2.6	1191	2.9	1511	2.1
Inpatient LOS, measurement year	0	35,002	82.6	34,278	82.4	63,515	88.9
	1-3	3181	7.5	3128	7.5	2535	3.6
	4+	4199	9.9	4186	10.1	5381	7.5
Inpatient admits, prediction year	No	36,127	85.2	35,522	85.4	65,711	92.0
	Yes	6255	14.8	6070	14.6	5720	8.0
Primary care visits, measurement year	0	8699	20.5	8324	20.0	10,354	14.5
	1-3	12,772	30.1	12,416	29.9	19,846	27.8
	4+	20,911	49.3	20,852	50.1	41,231	57.7
Charlson Comorbidity Index	0	23,997	56.6	23,373	56.2	27,931	39.1
	1-2	11,583	27.3	11,318	27.2	20,858	29.2
	3+	6802	16.1	6901	16.6	22,642	31.7
Chronic disease score	0-299	12,830	30.3	12,498	30.1	6612	9.3
	300-1299	10,318	24.4	10,031	24.1	37,104	51.9
	1300-3399	9493	22.4	9435	22.7	12,285	17.2
	3400+	9741	23.0	9628	23.2	15,430	21.6
Alcohol/substance abuse dx, measurement year	No	41,246	97.3	40,540	97.5	68,529	95.9
	Yes	1136	2.7	1052	2.5	2902	4.1
Mental health dx (not substance abuse), measurement year	No	29,195	68.9	28,624	68.8	36,531	51.1
	Yes	13,187	31.1	12,968	31.2	34,900	48.9
ED visit, measurement year	No	24,522	57.9	23,860	57.4	39,895	55.9
	Yes	17,860	42.1	17,732	42.6	31,536	44.2
Polypharmacy	No	25,080	59.2	24,409	58.7	46,962	65.7
	Yes	17,302	40.8	17,183	41.3	24,469	34.3

dx indicates diagnosis; ED, emergency department; LOS, length of stay.

the second state were also more likely to have higher CDS and CCI scores, less likely to have an inpatient stay, more likely to have a primary care visit, and more likely to have a mental health diagnosis for something other than substance abuse.

In general, the variables selected for the stepwise logistic regression provided strong predictive results for the disabled, continuously enrolled populations examined. The strongest predictors in the model developed from the first 50% sample were over age 65 years, inpatient stays in the measurement year, and higher CCI scores (Table 2). Table 3 shows the percentages of correctly predicted outcomes at 9 levels of predictive probability (risk score). The higher the risk score is, the fewer the number of predicted hospital-

Table 2. Stepwise Logistic Regression: Hospitalization in Prediction Year by Selected Predictors in Measurement Year: Disabled Members Continuously Enrolled 2 Years. Model Based on 50% Random Sample of State 1 Membership

Predictors		Odds Ratio	SE	*z*	*P*>|*z*|	[95% CI]	
Age, y	22-45 vs 0-21	1.48	0.07	8.58	.000	1.35	1.61
	46-64 vs 0-21	1.74	0.08	11.7	.000	1.58	1.91
	65+ vs 0-21	2.83	0.24	12.54	.000	2.41	3.34
Race	White vs black	1.09	0.03	2.61	.009	1.02	1.16
IP stays, measurement year	1-2 vs 0	2.42	0.11	18.95	.000	2.21	2.65
	3+ vs 0	7.45	0.55	27.2	.000	6.45	8.61
IP LOS, measurement year	1-3 days vs 0	0.85	0.05	–2.88	.004	0.76	0.95
ED visits in measurement year	Yes vs no	1.45	0.05	10.82	.000	1.36	1.56
Primary care visits, measurement year	1-3 vs 0	1.13	0.06	2.07	.038	1.01	1.26
	4+ vs 0	1.38	0.08	5.63	.000	1.23	1.54
Polypharmacy	Yes vs no	1.29	0.06	5.6	.000	1.18	1.41
Charlson Comorbidity Index	1-2 vs 0	1.69	0.07	13.29	.000	1.57	1.83
	3+ vs 0	2.46	0.11	19.47	.000	2.25	2.69
Disease categories							
Pain	Yes vs no	1.20	0.04	4.89	.000	1.11	1.28
Cardiac disease	Yes vs no	1.34	0.06	6.2	.000	1.22	1.47
Psychotic illness	Yes vs no	0.81	0.04	–4.2	.000	0.73	0.89
Cystic fibrosis	Yes vs no	2.59	0.51	4.79	.000	1.75	3.82
Rheumatoid arthritis	Yes vs no	1.21	0.05	4.59	.000	1.12	1.32
Renal disease	Yes vs no	2.04	0.35	4.12	.000	1.45	2.87
Hyperlipidemia	Yes vs no	0.81	0.04	–4.73	.000	0.75	0.89
Epilepsy	Yes vs no	1.16	0.05	3.73	.000	1.07	1.25
Gastric acid disease	Yes vs no	1.12	0.04	3.1	.002	1.04	1.21
Malignancies	Yes vs no	1.33	0.12	3.19	.001	1.12	1.59
HIV	Yes vs no	0.59	0.10	–3.05	.002	0.42	0.83
Heart disease/hypertension	Yes vs no	1.13	0.05	3	.003	1.04	1.22
Mental health: not substance abuse	Yes vs no	0.89	0.03	–2.99	.003	0.83	0.96

CI indicates confidence interval; ED, emergency department; HIV, human immunodeficiency virus; IP, inpatient; LOS, length of stay; SE, standard error.

izations, but the higher percentage of accurately predicted hospitalizations.

These ROC findings were similar whether the coefficients from the 50% randomized sample were used across the different populations, or the stepwise logistic regressions were applied separately for each population. Results were generally better if the stepwise logistic regression was run separately for each population (Table 4). The areas under the ROC for the model based on the 50% state sample and its application to the 2 other samples ranged from 0.79 to 0.81. When separate models were developed for each of the 3 samples, the areas ranged from 0.79 to 0.86.

DISCUSSION

This study shows that administrative data can be used to identify individuals for care management. The greater the risk score, the more likely a hospitalization occurred and the less likely a false positive was identified. The model can also be used to identify those factors most likely to be associated with high risk of hospitalization. The model identifies factors significantly associated with the outcome, and the coding in the analytic file used allowed identification of those risk factors for the specific individuals identified as high risk.

Table 3. Predictions of Hospitalization vs No Hospitalization, Including Total Percentage Correctly Classified

Predicted Probability ≥	Hospitalization in Prediction Year	Predicted Probability <	No Hospitalization in Prediction Year	% Correctly Classified
State 1: Test set				
0.1	5074 = 26.6% (19,068)	0.1	22,133 = 94.9% (23,314)	64.2%
0.2	3597 = 36.9% (9740)	0.2	29,984 = 91.9% (32,642)	79.2%
0.3	2565 = 47.1% (5448)	0.3	33,244 = 90.0% (36,934)	84.5%
0.4	1853 = 54.6% (3395)	0.4	34,585 = 88.7% (38,987)	86.0%
0.5	1271 = 62.2% (2042)	0.5	35,356 = 87.6% (40,340)	86.4%
0.6	724 = 68.5% (1057)	0.6	35,794 = 86.6% (41,325)	86.2%
0.7	456 = 74.3% (614)	0.7	35,969 = 86.1% (41,768)	86.0%
0.8	175 = 79.9% (219)	0.8	36,083 = 85.6% (42,163)	85.6%
0.9	8 = 88.9% (9)	0.9	36,126 = 85.3% (42,373)	85.3%
State 1: Validation set				
0.1	4981 = 26.4 % (18,880)	0.1	21,623 = 95.2% (22,712)	64.0%
0.2	3648 = 36.7% (9933)	0.2	29,237 = 92.3% (31,659)	79.1%
0.3	2599 = 46.4% (5599)	0.3	32,522 = 90.4% (35,993)	84.4%
0.4	1904 = 54.6% (3488)	0.4	33,938 = 89.1% (38,104)	86.2%
0.5	1295 = 62.1% (2087)	0.5	34,730 = 87.9% (39,505)	86.6%
0.6	773 = 69.0% (1121)	0.6	35,174 = 86.9% (40,471)	86.4%
0.7	504 = 74.3% (678)	0.7	35,348 = 86.4% (40,914)	86.2%
0.8	187 = 79.2% (236)	0.8	35,473 = 85.8% (41,356)	85.7%
0.9	11 = 68.8% (16)	0.9	35,517 = 85.4% (41,576)	85.4%
State 2: Validation set				
0.1	5058 = 12.6% (40,271)	0.1	30,498 = 97.9% (31,160)	49.8%
0.2	3875 = 23.9% (16,217)	0.2	53,369 = 96.7% (55,214)	80.1%
0.3	2842 = 37.5% (7579)	0.3	60,974 = 95.5% (63,852)	89.3%
0.4	2111 = 45.0% (4687)	0.4	63,135 = 94.6% (66,744)	91.3%
0.5	1524 = 52.8% (2889)	0.5	64,346 = 93.9% (68,542)	92.2%
0.6	1002 = 60.9% (1646)	0.6	65,067 = 93.2% (69,785)	92.5%
0.7	679 = 65.6% (1035)	0.7	65,355 = 92.8% (70,396)	92.4%
0.8	297 = 70.7% (420)	0.8	65,588 = 92.4% (71,011)	92.2%
0.9	23 = 82.1% (28)	0.9	65,706 = 92.0% (71,403)	92.0%

Given the need to control costs and provide efficient and effective services to state Medicaid populations, this study provides a robust method for developing a predictive model and targeting individuals who could become high-cost members and who could potentially benefit from some type of case/care/disease management intervention.

There are a number of strengths to this methodology. The model is based on already available data; no additional data collection is needed. Most of the derived variables are relatively easy to generate from the data once the appropriate raw data are extracted. Risk scores, either developed on a proprietary basis or open source, are easily incorporated into the model.

The model is fairly robust across samples. Similar C (area under the ROC curve) statistics are achieved across samples using the same set of coefficients. An area under the ROC curve of 0.8 is considered fair to good, and this model consistently achieves that level for the disabled populations.

The logistic regression produces a risk score (predicted probability of an inpatient stay) for each member subject. High-risk individuals can be reviewed for the variable flags present in their records (eg, all disease/condition variables are

Table 4. Odds Ratios (ORs) From Individual Stepwise Logistic Regressions Performed Separately for Each Sample (only ORs for significant independent variables included)

Predictor variable	Category	State 1: Test set	State 1: Validation set	State 2: Validation set
Age, y	22-45 vs 0-21	1.48	1.46	
	46-64 vs 0-21	1.74	1.93	1.21
	65+ vs 0-21	2.83	3.26	
Race	White vs black	1.09		0.71
Gender	Male vs female		0.94	
IP stays, measurement year	1-2 vs 0	2.42	1.46	2.36
	3+ vs 0	7.45	3.98	5.90
IP LOS, measurement year	1-3 days vs 0	0.85	1.47	
	4+ days vs 0		1.76	1.23
ED visits, measurement year	Yes vs no	1.45	1.56	1.27
Primary care visits, measurement year	1-3 vs 0	1.13		
	4+ vs 0	1.38	1.24	
Polypharmacy	Yes vs no	1.29	1.27	1.69
Charlson Comorbidity Index	1-2 vs 0	1.69	1.75	1.48
	3+ vs 0	2.46	2.28	1.56
Chronic disease score	300-1299 vs 0-299			0.36
Disease categories				
Pain	Yes vs no	1.20	1.22	1.39
Cardiac disease	Yes vs no	1.34	1.22	1.56
Psychotic illness	Yes vs no	0.81	0.86	
Cystic fibrosis	Yes vs no	2.59	1.84	2.01
Rheumatoid arthritis	Yes vs no	1.21	1.20	1.16
Renal disease	Yes vs no	2.04	2.08	
Hyperlipidemia	Yes vs no	0.81	0.82	0.83
Epilepsy	Yes vs no	1.16	1.16	1.19
Gastric acid disease	Yes vs no	1.12	1.14	1.16
Malignancies	Yes vs no	1.33	1.42	1.44
HIV	Yes vs no	0.59		
Heart disease/hypertension	Yes vs no	1.13		1.11
Anxiety and tension	Yes vs no		1.16	
Liver failure	Yes vs no		1.64	1.31
Coronary/peripheral vascular	Yes vs no		1.30	1.21
Parkinson's disease	Yes vs no		0.80	
Gout	Yes vs no		1.31	
Inflammatory bowel disease	Yes vs no		1.75	1.46
Glaucoma	Yes vs no		0.79	
Asthma	Yes vs no			1.26
Hypertension	Yes vs no			1.13
Diabetes	Yes vs no			1.18
ESRD	Yes vs no			1.67
Depression	Yes vs no			1.10
Mental health: not substance abuse	Yes vs no	0.89	0.84	0.65

ED indicates emergency department; ESRD, end-stage renal disease; HIV, human immunodeficiency virus; IP, inpatient; LOS, length of stay.

dichotomized [0 = no, 1 = yes]) that account for their high scores. At prediction probabilities (risk scores) of 0.4 or better, 86% or more of the members are classified correctly as to whether or not they had inpatient stays in the prediction year. The model can be developed and run using most available/common statistical packages, eg, Stata, SAS.

There are a number of potential limitations of the model. The complexity of most Medicaid plans makes pulling the correct/necessary data from the databases the most complicated part of using the model. Also, the structure of Medicaid plans varies from state to state. How different programs are defined and how eligibility is decided are not consistent. Variations such as these account for some of the demographic differences. For example, the age distributions for the 2 states studied are different. In 1 state, approximately 3% of the study population was 65 years or older, and in the other state, none of the study subjects were over 65 years of age. The presence of seniors over age 65 years in 1 state may be due to Medicare eligibility requirements. Each state may want to use this methodology to target different segments of their eligible population. One of the purposes of this project was to develop an easily generalizable method for creating risk scores that could be used in different states.

Another limitation to this study is the use of stepwise logistic regression. This technique has a number of problems (eg, overstated R^2 values, understated P values, exacerbated collinearity problems).[13] However, a process for modeling was needed that used generally available software (SAS, Stata, SPSS, etc), was feasible and intuitive for use by non-statisticians, and could be utilized by different programs in different states.

A large role in the modeling is played by the chronic disease score and its disease categories. This score is highly dependent on the completeness and accuracy of the prescription data available. This limits the model's practical use to subpopulations from Medicaid plans for which fairly complete prescription data are available (eg, dual eligibles may not be an appropriate subgroup).

Most of the independent variables cannot distinguish between appropriate and inappropriate treatment, limiting the model's ability to identify actionable situations revealed by the patient data. In some cases, better results for the model are achieved by running the stepwise regression for the particular subpopulation of interest rather than applying coefficients developed with another subpopulation (eg, the results using the second state's findings suggest that, using the same variables, the stepwise logistic should be run separately for each state/population of interest).

The results obtained from this project are promising, but are, at best, preliminary. Further study is needed in several areas. Narrowing the focus of the dependent variable, avoidable hospitalizations, to a more targeted group of diagnoses (eg, those ambulatory care sensitive–conditions), and including more independent variables targeting behaviors that can be impacted by care management, and studying those with the largest associations with the outcome could provide practical, implementable results for more immediate use by the different plans. Also, it would be important to evaluate the model for shorter follow-up periods. The earlier individuals can be identified for intervention, the better the chances of avoiding unnecessary hospitalizations. There is also a need to examine potential interactions among the variables used in this study. The variable used to examine previous hospitalizations, given its high level of significance, should also be reexamined. Stratifying the analysis by number of previous hospitalizations could provide important information for targeting individuals for intervention.

The statistical model itself should also be further revised and evaluated. As stated above, there are a number of known problems with the stepwise approach to regression, and results from other variable selection procedures should be compared with those obtained with the methods used in this study. Whether or not to use an automated variable selection procedure should also be reviewed.

Further work is also needed in delineating the subpopulations most amenable to interventions, and in testing the model in more states and assessing how well it works in states with very different demographic patterns. Also, the model should be evaluated to determine whether or not it has any practical value for use with other outcomes (eg, ED visits, observational stays).

Author Affiliations: From Thomas Jefferson University (JFM, AGC, VM, DZL), Philadelphia, PA; Greater Philadelphia Business Coalition on Health (NIG), Philadelphia, PA; HP Enterprises (BWN, JS, FMD), Camp Hill, PA.

Funding Source: This research was supported by a contract with HP Enterprise Services. Decisions on inclusion or exclusion of material and the decisions on where or whether to publish were made solely by the authors. HP provides data processing services to a number of state Medicaid plans, and provided access to 2 state databases for the purposes of this study.

Author Disclosures: Mr Novinger and Mr Din report employment with HP Enterprises. The authors (JFM, AGC, JS, VM, DZL, NIG) report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (JFM, AGC, BWN, JS, VM, DZL, NIG); acquisition of data (BWN, JS, NIG); analysis and interpretation of data (JFM, AGC, VM, DZL, NIG); drafting of the manuscript (JFM, AGC, JS, NIG); critical revision of the manuscript for important intellectual content (JFM, AGC, BWN, VM, NIG); statistical analysis (JFM, AGC); provision of study materials or patients (FMD); obtaining funding (FMD, NIG); administrative, technical, or logistic support (JFM, AGC, BWN, FMD, DZL); and supervision (JFM, AGC, BWN, FMD, NIG).

Address correspondence to: John F. McAna, PhD, Jefferson School of Population Health, 901 Walnut St, 10th Fl, Philadelphia, PA 19107. E-mail: john.mcana@jefferson.edu.

REFERENCES

1. Smith, VK, Gifford K, Ellis E, Rudowitz R, Snyder L. Moving Ahead Amid Fiscal Challenges: A Look at Medicaid Spending, Coverage and Policy Trends. Results from a 50-State Medicaid Budget Survey for State Fiscal Years 2011 and 2012. Washington, DC: Kaiser Commission on Medicaid and the Uninsured; 2011.

2. Stanton MW, Rutherford MK. The high concentration of U.S. health care expenditures. Rockville, MD: Agency for Healthcare Research and Quality; 2005. Research in Action Issue 19. AHRQ Pub. No. 06-0060.

3. Knutson D, Bella M, Llanos K. Predictive Modeling: A Guide for State Medicaid Purchasers. Center for Health Care Strategies, Inc; August 2009.

4. Billings J, Dixon J, Mijanovich T, Wennberg D. Case finding for patients at risk for readmission to hospital: development of algorithm to identify high risk patients. *BMJ*. 2006;333(7563):327.

5. Wennberg DE, Marr A, Lang L, O'Malley S, Bennett G. A randomized trial of a telephone care-management strategy. *N Engl J Med*. 2010;363(13):1245-1255.

6. Billings J, Mijanovich T. Improving the management of care for high-cost Medicaid patients. *Health Affairs*. 2007;26(6):1643-1655.

7. Agency for Healthcare Research and Quality. Research in Action, issue #19. The high concentration of U.S. health care expenditures, June 2006.

8. Hollenbeak CS, Chirumbole M, Novinger B, Sidorov J, Din F. Predictive models for diabetes patients in Medicaid. *Popul Health Manag*. 2011;14(5):239-242.

9. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol*. 1992;45(6):613-619.

10. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43(11):1130-1139.

11. Clark DO, Von Korff M, Saunders K, Baluch WM, Simon GE. A chronic disease score with empirically derived weights. *Med Care*. 1995;33:783-795.

12. Putnam KG, Buist DS, Fishman P, et al. Chronic Disease Score as a predictor of hospitalization. *Epidemiology*. 2002;13(3):340-346.

13. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer-Verlag; 2002. ■