Predictive Modeling of Total Healthcare Costs Using Pharmacy Claims Data: A Comparison of Alternative Econometric Cost Modeling Techniques

Author(s): Christopher A. Powers, Christina M. Meyer, M. Christopher Roebuck and Baze Vaziri

Source: *Medical Care*, Vol. 43, No. 11 (Nov., 2005), pp. 1065-1072
Published by: Lippincott Williams & Wilkins
Stable URL: http://www.jstor.org/stable/3768184
Accessed: 06-07-2018 15:50 UTC

Predictive Modeling of Total Healthcare Costs Using Pharmacy Claims Data

A Comparison of Alternative Econometric Cost Modeling Techniques

Christopher A. Powers, PharmD, Christina M. Meyer, MHS, M. Christopher Roebuck, MBA, and Baze Vaziri, MBA

Objective: We sought to evaluate several statistical modeling approaches in predicting prospective total annual health costs (medical plus pharmacy) of health plan participants using Pharmacy Health Dimensions (PHD), a pharmacy claims-based risk index.

Methods: We undertook a 2-year (baseline year/follow-up year) longitudinal analysis of integrated medical and pharmacy claims. Included were plan participants younger than 65 years of age with continuous medical and pharmacy coverage (n = 344,832). PHD drug categories, age, gender, and pharmacy costs were derived across the baseline year. Annual total health costs were calculated for each plan participant in follow-up year. Models examined included ordinary least squares (OLS) regression, log-transformed OLS regression with smearing estimator, and 3 two-part models using OLS regression, log-OLS regression with smearing estimator, and generalized linear modeling (GLM), respectively. A 10% random sample was withheld for model validation, which was assessed via adjusted r^2, mean absolute prediction error, specificity, and positive predictive value.

Results: Most PHD drug categories were significant independent predictors of total costs. Among models tested, the OLS model had the lowest mean absolute prediction error and highest adjusted r^2. The log-OLS and 2-part log-OLS models did not predict costs accurately as the result of issues of log-scale heteroscedasticity. The 2-part model using GLM had lower adjusted r^2 but similar performance in other assessment measures compared with the OLS or 2-part OLS models.

Conclusion: The PHD system derived solely from pharmacy claims data can be used to predict future total health costs. Using PHD with a simple OLS model may provide similar predictive accuracy in comparison to more advanced econometric models.

Key Words: predictive modeling, risk assessment, two-part model, log model, generalized linear model, pharmacy

(*Med Care* 2005;43: 1065–1072)

From Caremark, Hunt Valley, Maryland.

Reprints: Christina M. Meyer, MHS, 11311 McCormick Road, Executive Plaza II, Suite 230, Hunt Valley, MD 21031. E-mail: christina.meyer@caremark.com.

ISSN: 0025-7079/05/4311-1065

Predictive modeling increasingly is used to forecast the expected health services utilization or costs of plan participants for managed care organizations. Data can be obtained from various sources, including medical and pharmacy insurance claims, health risk-assessment surveys, or laboratory data. These models can be used to risk-adjust provider payments and proactively identify high-risk plan participants for disease-management programs.

Several predictive models of total healthcare costs are available commercially. Some examples include adjusted clinical groups (ACGs), chronic illness and disability payment system (CDPS), diagnostic cost groups (DCGs), RxGroups, RxRisk, and episode risk groups (ERGs).[1] Most models use medical insurance claims to generate main predictor variables and often optionally incorporate pharmacy data. Although few models predict risk based solely on pharmacy claims data,[2–13] models of this type may offer some advantages over those using medical claims data. Pharmacy data generally are timelier and less costly than medical data because nearly all prescriptions are adjudicated in real-time using automated electronic databases. Pharmacy data also tend to contain fewer coding errors.[14–16] Conversely, pharmacy claims data may not always portray an accurate clinical picture because some prescribed medications go unfilled and some medications have multiple indications from which disease status must be inferred.

Pharmacy Health Dimensions (PHD) is a pharmacy-based risk index that was developed by expanding upon previously published drug classification algorithms.[2,3,5,8] PHD uses pharmacy claims data to predict prospective total annual healthcare costs (medical plus pharmacy costs) of participants. Prior research has found the PHD to be effective at predicting hospitalizations and total healthcare costs over the course of 3 years and also is correlated with health-related quality of life.[17–19]

Much research has been conducted on the statistical methodologies for modeling healthcare cost data.[20–26] However, the extant literature on predictive models of healthcare costs have not fully incorporated these suggestions. Most previous pharmacy-based models have relied on ordinary least squares (OLS) regression; the Chronic Disease Score,[3] Pediatric Chronic Disease Score,[5] RxRisk Model,[6] RxRisk-V,[10,11] Medicaid-Rx model,[12] Pharmacy Cost Groups,[7] and RxGroups[9] all use this

approach. However, healthcare cost data typically have mixed distributions that contain a disproportionate mass of observations at zero (ie, nonusers) and a right-skewed distribution for users. Two-part modeling has been used to address this issue of a mixed distribution of healthcare cost data.[21,22,25] This econometric approach first estimates the probability of incurring any cost and subsequently models the level of cost conditional on having incurred any cost.

Because healthcare cost data tend to be highly right-skewed, transformation of the dependent variable often is used.[20,23] However, parameter estimates from OLS regression of the commonly used log-transformed dependent variable are on the log scale. Meaningful interpretation of these coefficients requires that they be retransformed to the original scale (eg, dollars or units of health services utilization) using the smearing estimator to correct for retransformation bias.[23,24] This process can become quite complex in cases involving log-scale heteroscedasticity, the nonconstant error variance across observations.

The use of generalized linear modeling (GLM) also has been proposed as a method for dealing with the problematic distributional properties of healthcare cost and utilization data.[20,21,25] GLM can be used to obtain predicted values that maintain the original scale of the data even when a transformation might otherwise be necessary because GLM relates the expected value of the response to the linear predictor via a monotonic link function and accounts for nonconstant variance by explicitly modeling the response probability distribution.[21]

The objective of this study was to compare various types of cost modeling approaches for predictive modeling of total healthcare costs using pharmacy claims data. We examined results from empirical specifications using several OLS regression and 2-part modeling techniques. Finally, we tested the validity of these predictive models using a separate sample.

METHODS

Pharmacy Health Dimensions (PHD)

The clinical content of PHD has been presented elsewhere.[18] In general, PHD categorizes a year of pharmacy claims into 51 distinct implied disease states. Several algorithms for classification of chronic disease from pharmacy claims have been published.[2,3,5,8] However, published algorithms do not include newly available therapies or changes in the treatment of some chronic diseases. In response, we created a new comprehensive classification based on the existing literature and other updates provided by an internal clinical panel of physicians, pharmacists, and epidemiologists. Categories included from the original Chronic Disease Score published by Von Korff were asthma, steroid dependency (formerly asthma and/or rheumatism), cancer, cardiovascular disease, diabetes, epilepsy, glaucoma, gout, high cholesterol, hypertension, migraine, Parkinson disease, respiratory illness, rheumatoid arthritis, tuberculosis, and ulcers.[2] A follow-up study by Clark et al[3] appended the categories of AIDS/HIV, anxiety and depression (which we combined), Crohn's disease, cystic fibrosis, end-stage renal disease, liver failure, chronic pain, psychosis, renal failure, and ulcerative colitis. The pediatric Chronic Disease Score version developed by Fishman added the categories of attention deficit hyperactivity disorder, eczema, growth hormone, and immunodeficiency.[5] A separate classification system proposed by Malone et al[8] contributed categories for allergic rhinitis, antiarrhythmics, benign prostatic hyperplasia, transplant, and psoriasis.

We augmented these previous classifications with the following categories: anorexics, dementia, hepatitis C, lifestyle drugs, menopause, manic depression, multiple sclerosis, osteoporosis, overactive bladder, and peripheral vascular disease. Ultimately, each plan participant receives a binary yes/no indicator for each of the 51 categories.

Data

Integrated medical and pharmacy insurance claims data from a >600,000-participant state employer in the southern United States were examined. Benefits were administered by a commercial medical carrier (offering preferred provider organization and health maintenance organization plans) and a pharmacy benefit manager. A portion of the enrollees were covered under the State Children's' Health Insurance Program (SCHIP) and were removed from the analysis. The data included a 1-year baseline period (5/01/01–4/30/02) and a 1-year follow-up period (5/01/02–04/30/03), with claims paid through October 31, 2003. Plan participants were included in the analyses if they were younger than 65 years of age at the end of the study period and had continuous medical and pharmacy coverage during the 2-year study period. Those 65 years of age and older were excluded to avoid including plan participants who had supplemental medical coverage with Medicare, which may have not been captured in the data set.

Using data from the baseline period, the 51 PHD dichotomous variables were derived, as well as 12 participant age/gender indicators and total baseline pharmacy cost as measured by the ingredient cost paid for medications, which is the cost before any fees charged by the pharmacy or any coinsurance or copayments paid by the plan participant. Annual total costs (pharmacy plus medical) were calculated for each participant for the follow-up period. For medical costs, the allowed amounts (as opposed to the charge to or payment by the insurer) were used, which were the contracted rates of reimbursement for medical services established by the plan. Costs before participant cost-sharing were used to minimize any effects of changes in plan design between the study periods. For plan participants without any claims who were determined to be eligible for services through verification of eligibility records a cost of 0 was applied.

Statistical Analyses

A split-sample design was used, in which a 10% random sample of the total study population was withheld for model validation. Descriptive statistics, including the Student t test, the χ^2 test, and the Wilcoxon rank sum test (a nonparametric test for comparisons of samples having nonnormal distributions), were used to test for differences in subject characteristics between training and validation samples.

To predict participants' total health costs in the follow-up year as a function of baseline information, several multivariate econometric approaches were explored. The first model used OLS in estimating total costs on 4 sets of regressors: (1) a vector of age/gender indicators only; (2) age/gender indicators along with a baseline continuous measure of pharmacy cost; (3) age/gender indicators, continuous baseline pharmacy cost, and the vector of PHD binary variables; and (4) age/gender indicators, categorical baseline pharmacy cost (0 to <80%, 80% to <95%, 95% to <99%, and 99% to 100%; chosen to group plan participants with the highest pharmacy expenditures and the greatest variability in costs), and the vector of PHD binary variables. The independent variables included in all subsequent models were those of this fourth set.

Next, OLS regression using log-transformed follow-up total cost was examined. Following the common, albeit improper, practice, a value of $1 was added to all follow-up year total costs observations to avoid undefined solutions for the log of zero. The smearing estimator developed by Duan[24] was used to retransform predictions from the log-scale to the desired original scale (dollars). However, the null hypothesis of constant variance after log transformation was rejected ($P < 0.0001$), using the Breusch-Pagan/Cook-Weisberg test for heteroscedasticity.[27,28] We also retransformed using a separate smearing factor for each decile of the predicted values in an attempt to improve the retransformation of log-scale predictions in the presence of heteroscedasticity.[26]

Three separate 2-part models also were examined; all of which used logistic regression to assess the probability of incurring any cost for the first part of the model with various scenarios tested for the best second part model. In the first 2-part model, OLS regression was used in the second part to obtain a prediction of the level of cost conditional upon incurring any expense. In the second 2-part model, a log-OLS regression model retransformed via the smearing factor (as described above) was used in the second part. The third and final 2-part model used a GLM with a gamma response probability distribution and a log link function for the second part of the model. A gamma response probability distribution was chosen for the GLM based on the results of the modified Park test on the raw scale residuals as proposed by Manning and Mullahy.[20,29]

To assess the performance of each model, we included the following the statistical evaluation criteria: predicted mean, predicted median, and adjusted r^2 for the sample in which the model was estimated (training sample). For the validation sample, we also estimated the predicted mean and median, and to make relative comparisons of model performance between different types of models, we derived the mean absolute prediction error and adjusted r^2.[6] The mean absolute prediction error was calculated by first subtracting each plan participant's actual follow-up year total costs from their predicted total costs and taking the absolute value of this difference, and then the mean was calculated for these absolute differences. The adjusted r^2 was calculated by squaring the correlation coefficient (r) between the actual and predicted values and then adjusting for the number of observations (n) and variables (k) in each model.

Predictive model performance was further assessed by categorizing actual and predicted costs using 4 different percentile thresholds (<50%, ≥90%, ≥95%, and ≥99%) and determining the positive predictive value (PPV) and specificity of the models for each of these cost groups. Sensitivity was examined but not presented because it is equivalent to PPV when comparing by equal percentiles of actual and predicted. These measures represent the accuracy of the model in classifying individual plan participants into low-cost and high-cost groups (ie, how well the models predict the bottom 50% and top 10%, 5%, and 1% of participants), as opposed to describing how much variation in the population can be explained by the model. All statistical analyses were conducted using SAS, version 8.2 and STATA, release 8.2.[30,31]

RESULTS

Plan Participant Characteristics

The characteristics of plan participants in the training and validation samples are presented in Table 1. There were no significant differences in the characteristics of plan participants between samples except for a small, yet statistically significant difference in the proportion of females aged 35 to 44 years (11.7% in the training sample vs. 11.3% in the validation sample; $P = 0.02$). The mean age of plan participants was approximately 40 years in both the training and validation samples and both groups were similarly composed of a larger proportion of females (61.0% in training sample and 60.5% in validation sample). The mean pharmacy costs in the baseline year were $640 (median, $166) for the training sample and $639 (median, $168) for the validation sample. The distribution of actual follow-up total costs was highly right-skewed and contained the typical mass of observations at zero.

OLS Model

Among the 4 OLS models, the 2 models containing age/gender, prior annual pharmacy cost, and PHD categories (models 3 and 4) had improved performance over the OLS age/gender model (model 1) and the OLS age/gender plus prior annual pharmacy cost model (model 2) as demonstrated by increases in the adjusted r^2 values shown in Table 2. Removing prior annual pharmacy costs from the final model resulted in a reduction of the training model adjusted r^2 to 0.124 (not shown in Table 2). The addition of PHD categories also resulted in increases in the adjusted r^2 values and decreases in mean absolute prediction errors for the validation sample.

In the training sample, the actual mean total cost for the follow-up year was $3198 (median, $749) and the mean predicted total cost was $3198 (median, $2215) from the final OLS model. Likewise, for the validation sample, the actual mean total cost for the follow-up year was $3145 (median, $736) and the mean predicted cost in this sample was $3221 (median, $2215).

TABLE 1. Characteristics of Plan Participants Included in the Cost Modeling Study using Pharmacy Health Dimensions

	Training Sample (n = 310,413)	Validation Sample (n = 34,419)	*P*
Age (years), mean ± SD*	39.8 ± 16.8	39.8 ± 16.8	0.88
Proportion of females, n (%)†	189,273 (61.0)	20,817 (60.5)	0.08
Gender/age categories†			
Male			
Age 0–11 yr, n (%)	13,601 (4.4)	1522 (4.4)	0.73
Age 12–17 yr, n (%)	12,359 (4.0)	1399 (4.1)	0.45
Age 18–34 yr, n (%)	21,193 (6.8)	2413 (7.0)	0.20
Age 35–44 yr, n (%)	19,169 (6.2)	2177 (6.3)	0.27
Age 45–54 yr, n (%)	30,473 (9.8)	3368 (9.8)	0.85
Age 55–64 yr, n (%)	24,345 (7.8)	2723 (7.9)	0.65
Female			
Age 0–11 yr, n (%)	12,840 (4.1)	1364 (4.0)	0.12
Age 12–17 yr, n (%)	11,835 (3.8)	1350 (3.9)	0.31
Age 18–34 yr, n (%)	33,230 (10.7)	3677 (10.7)	0.90
Age 35–44 yr, n (%)	36,291 (11.7)	3873 (11.3)	0.02
Age 45–54 yr, n (%)	56,771 (18.3)	6291 (18.3)	0.96
Age 55–64 yr, n (%)	38,306 (12.3)	4262 (12.4)	0.82
Year 1 pharmacy cost, mean ± SD (median)‡	$640 ± 1586 ($166)	$639 ± 1338 ($168)	0.61
Year 2 total costs, mean ± SD (median)‡	$3198 ± 10,344 ($749)	$3145 ± 9970 ($736)	0.46

*P value calculated using t test.
†P value calculated using χ^2 test.
‡P value calculated using non-parametric Wilcoxon rank-sum test.

TABLE 2. Actual and Predicted Follow-up Total Costs (Untruncated) and Selected Statistics for Training and Validation Samples

	Training Sample (n = 310,413)				Validation Sample (n = 34,419)				Mean Absolute Prediction Error
	Median	Mean	Standard Deviation	Adjusted r^2	Median	Mean	Standard Deviation	Adjusted r^2	
Actual follow-up year total costs	$749	$3198	$10,344	—	$736	$3145	$9970	—	—
Ordinary least squares (OLS) model predicted total costs									
Model 1: age/gender	$3099	$3198	$1559	0.023	$3099	$3195	$1564	0.025	$3541
Model 2: age/gender/prior prescription cost*	$2556	$3198	$3082	0.089	$2558	$3195	$2719	0.085	$3156
Model 3: age/gender/prior prescription cost/ PHD categories*	$2047	$3198	$3790	0.134	$2036	$3219	$3835	0.114	$3074
Model 4: age/gender/prior prescription cost/ PHD categories†	$2215	$3198	$3760	0.132	$2215	$3221	$3843	0.111	$3088
Log-transformed OLS model,‡ predicted total costs									
Single smearing estimator	$2330	$17,068	$59,664	0.065	$2330	$17,590	$65,432	0.043	$16,095
Ten smearing estimators (by deciles)	$2225	$3584	$7791	0.063	$2206	$3649	$8564	0.042	$3497
Two-part model: logistic/OLS,‡ predicted total costs	$2172	$3198	$3760	0.132	$2057	$3221	$3845	0.111	$3088
Two-part model: logistic/log transformed OLS,‡ predicted total costs									
Single smearing estimator	$1817	$5099	$11,948	0.095	$1813	$5225	$13,762	0.059	$4614
Ten smearing estimators (by deciles)	$2153	$3382	$5675	0.095	$2143	$3444	$6517	0.057	$3292
Two-part model: logistic/generalized linear model (GLM),‡§ predicted total costs	$2090	$3348	$6723	0.090	$2070	$3456	$10,447	0.032	$3296

*Continuous previous prescription cost variable.
†Categorical previous prescription cost variable using the following percentile groups: 0 to <80, 80 to <95, 95 to <99, 99 to 100.
‡Using same independent variable model structure as OLS model 4.
§Generalized linear model using a gamma response probability distribution with log link function.

Nearly all of the 51 PHD categories were significant independent predictors of total health costs. PHD categories that were not significant included acne, psoriasis, eczema, growth hormone, immunodeficiency, attention deficit hyperactivity disorder, menopause, osteoporosis, manic depression, and lifestyle.

Log-OLS Model

Table 2 also presents the summary statistics from the log-transformed OLS model using the single and decile smearing factors. Compared with the untransformed OLS model, training model adjusted r^2 values were lower for the log-transformed model using the single smearing factor and using the smearing factor by deciles (0.132 vs. 0.065 and 0.063, respectively) and this trend was repeated for the validation sample. Predicted total costs for the follow-up year were highly distorted when retransforming with a single smearing factor, having a mean of $17,068 (median, $2330) in the training sample; this finding also was similar in the validation sample. As previously discussed, this result is not surprising given the nonconstant variance in the error structure of this model, which remained even after log transformation. Although not obvious from the adjusted r^2, retransforming by deciles resulted in improved predictions as measured by the mean absolute prediction error as compared with the single smearing estimator approach ($3497 vs. $16,095, respectively). Examination of residual plots also indicated that the error term on the log-scale was heteroscedastic by several PHD disease categories simultaneously.

Logistic/OLS Two-Part Model

In the training sample, 46,720 plan participants (15.1%) had a total cost of 0 in the follow-up year and, thus, use of a 2-part model seemed appropriate. Summary statistics for the 2-part model with an OLS second part, presented in Table 2, suggest that there was no improvement in performance in comparison to the simple OLS model. This can be seen by the similar adjusted r^2 values from both samples and also by the equal mean absolute predicted errors from the validation sample.

Logistic/Log-OLS Two-Part Model

There were increases in the adjusted r^2 values for the training and validation samples when using a 2-part model with a log-transformed second part compared with the simple log-transformed OLS model, using single or decile smearing estimators, as Table 2 shows. Although adjusted r^2 values were again similar whether using the single or decile smearing approach, the mean absolute predicted error was lower when using the latter retransformation.

Logistic/GLM Two-Part Model

Results from the 2-part model with a GLM second part revealed a lower adjusted r^2 in comparison to the OLS model or the 2-part OLS model (see Table 2). However, the differences in the mean absolute prediction error were less pronounced ($3296 vs. $3088 and $3088, respectively). Adjusted r^2 values and mean absolute prediction error from the 2-part model using GLM were similar to the 2-part log-OLS model retransformed by deciles.

Predicting Low- and High-Cost Plan Participants

The PPV and specificity of study models for correctly identifying low (bottom 50%) and high cost (top 10%, 5%, and 1%) plan participants are presented in Table 3. The age/gender OLS model did not correctly predict any plan participants above the top 5th percentile. Adding PHD categories (OLS Models 3 and 4) increased the PPV and specificity for all top cost groups in comparison to the more parsimonious OLS models (Models 1 and 2) in both the training and validation samples.

Comparing the different modeling approaches using PHD categories revealed that all models had similar performance at predicting low-cost plan participants (bottom 50th percentile PPV range 74.9% to 76.2%, for training sample). For identifying high-cost individuals, all models had similar specificity at each of the top percentile groups. The OLS model had higher PPV for identifying high cost participants compared with the log-transformed OLS model (irrespective of retransformation method), especially for the highest of high cost participants (top 1st percentile PPV 14.1% vs. 10.7%, respectively for the training sample). Using a 2-part OLS model versus a simple OLS model did not greatly affect the accuracy of predicting participants into the top 10th, 5th, or 1st percentiles of cost as demonstrated by similar PPVs between the models at these levels. However, slightly higher PPVs were observed for the 2-part log-OLS model compared with the simple log-OLS model. When comparing the 2-part model using OLS to the 2-part model using GLM, there were only slight differences in the PPVs when predicting the top cost participants.

DISCUSSION

Although general concurrent risk assessment and case-mix adjustment have been important considerations in managed care, the ability to accurately project future expenditures of individual participants in a health plan recently has gained much attention. In addition, the ability to do so in a timely fashion with scant resources and limited plan participant information is of great interest. This study investigated several advanced statistical approaches for modeling healthcare costs using only data contained in pharmacy claims as predictors.

Prospective modeling of plan participants' total annual health services costs using PHD with a simple OLS model demonstrated equivalence and, in some cases superiority, to the variety of advanced econometric models examined in this study, when comparing adjusted r^2 values, absolute mean prediction error, and accuracy (ie, PPV) of predicting high-cost plan participants. Although the use of OLS on a log-transformed dependent variable or GLM for the second part of a 2-part model have been suggested for healthcare cost data, neither of these techniques resulted in improvements in performance for this study population.

Other researchers have previously used pharmacy data to predict total healthcare costs and utilization.[2–13] Clark et al revised the Chronic Disease Score, originally presented by Von Korff et al, by estimating empirically derived weights

TABLE 3. Positive Predictive Value and Specificity of Predicted Costs at Various Cost Groupings for the Training and Validation Samples*

	Training Sample (n = 310,413)				Validation Sample (n = 34,419)			
	Bottom 50% (n = 155,208)	Top 10% (n = 31,041)	Top 5% (n = 15,522)	Top 1% (n = 3105)	Bottom 50% (n = 17,211)	Top 10% (n = 3443)	Top 5% (n = 1722)	Top 1% (n = 345)
Ordinary least squares (OLS)								
Model 1: age/gender								
Positive predictive value	67.2%	16.4%	—	—	66.4%	17.1%	—	—
Specificity	66.0%	92.7%	—	—	67.5%	92.7%	—	—
Model 2: age/gender/prior prescription cost†								
Positive predictive value	74.6%	36.8%	27.5%	9.9%	74.7%	36.6%	27.0%	8.9%
Specificity	74.6%	93.0%	96.2%	99.1%	74.7%	92.9%	96.2%	99.1%
Model 3: age/gender/prior prescription cost/PHD categories†								
Positive predictive value	75.4%	38.8%	30.0%	14.4%	75.4%	39.5%	29.6%	13.8%
Specificity	75.4%	93.2%	96.3%	99.1%	75.4%	93.3%	96.3%	99.1%
Model 4: age/gender/prior prescription cost/PHD categories‡								
Positive predictive value	74.9%	38.9%	29.8%	14.1%	75.1%	39.6%	29.4%	14.4%
Specificity	74.9%	93.2%	96.3%	99.1%	75.3%	93.3%	96.3%	99.1%
Log-transformed OLS§								
Single smearing estimator								
Positive predictive value	75.6%	37.1%	27.2%	10.7%	75.9%	37.8%	27.3%	11.0%
Specificity	75.6%	93.0%	96.2%	99.1%	75.9%	93.1%	96.2%	99.1%
Ten smearing estimators (by deciles)								
Positive predictive value	75.4%	36.3%	27.2%	10.7%	75.5%	37.4%	27.3%	11.0%
Specificity	75.4%	92.9%	96.2%	99.1%	75.5%	93.0%	96.2%	99.1%
Two-part: logistic/OLS§								
Positive predictive value	75.7%	38.7%	29.8%	14.0%	75.9%	39.6%	29.5%	14.4%
Specificity	75.6%	93.2%	96.3%	99.1%	75.9%	93.3%	96.3%	99.1%
Two-part: logistic/log transformed OLS§								
Single smearing estimator								
Positive predictive value	76.2%	38.0%	28.4%	12.2%	76.4%	38.7%	27.9%	11.5%
Specificity	76.2%	93.1%	96.2%	99.1%	76.4%	93.2%	96.2%	99.1%
Ten smearing estimators (by deciles)								
Positive predictive value	75.5%	37.6%	28.4%	12.2%	75.9%	38.1%	27.9%	11.5%
Specificity	75.5%	93.1%	96.2%	99.1%	75.9%	93.1%	96.2%	99.1%
Two-Part: logistic/Generalized Linear Model§								
Positive predictive value	75.7%	38.4%	29.3%	14.0%	75.9%	39.7%	28.7%	13.5%
Specificity	75.7%	93.2%	96.3%	99.1%	75.9%	93.3%	96.2%	99.1%

*Dashes (—) indicate that no plan participants were predicted to have costs at the specified cost level.
†Continuous previous prescription cost variable.
‡Categorical previous prescription cost variable using the following percentile groups: 0 to <80, 80 to <95, 95 to <99, 99 to 100.
§Using same independent variable model structure as OLS Model 4.

using an OLS model and thus were able to predict cost outcomes, including total costs.[2,3] Using age, gender, and their medication classification variables, the Clark et al predictive model explained 10% of the variation in future total costs (ie, $r^2 = 0.10$). By comparison, the model presented in this article, which includes age, gender, and the PHD medication classification variables was able to explain 12% of the variation in future total costs. However, caution must be used when directly comparing these results since the models were constructed in different populations using different time periods (ie, Clark et al used 6 months of data, whereas a full year of data was used in this study).

Several OLS prospective cost models using diagnosis information from medical claims data recently have been

reviewed in a research study sponsored by the Society of Actuaries.[1] Results using offered weights and no data truncation showed that the CDPS model had an r^2 of 0.103 with a mean absolute prediction error of 2299 (equal to 1.03 times the actual mean of 2232), and the DCGs had an r^2 of 0.143 with a mean absolute prediction error of 2187 (0.98 times actual mean). ERGs, which uses both medical and pharmacy claims information, was found to have an r^2 of 0.146 and a mean absolute prediction error of 2082 (0.93 times actual mean). By comparison, the OLS-based PHD model examined in the present study demonstrated comparable performance to these models using medical data, with an r^2 of 0.132 and a mean absolute prediction error of 3088 (0.98 times actual mean of 3145).

Since the seminal work by Von Korff et al, the use of pharmacy data for risk assessment and predicting future healthcare costs and utilization has become more commonplace and is becoming accepted as methodologically sound.[4,6,7,9,13] The reliability and validity of using pharmacy data to predict costs and utilization has also been demonstrated in other distinct populations, including Veterans Affairs, Medicaid, and pediatrics.[5,8,10–12] However, in all of these previous pharmacy-based studies, the analytical modeling methods presented have been limited to OLS regression.

Determination of the appropriate modeling approach for a particular population is rarely an easy task, particularly with healthcare data. Suggestions from health economists are widely available[20–26] but are not always used in practice. Manning and Mullahy offer a relatively straightforward recommendation to first examine residuals to determine whether to use an OLS-based model or a GLM approach (using the Park test to guide selection of the response probability distribution).[20,29] However, they also caution that if the log-scale residuals are heteroscedastic with respect to the independent variables, then the OLS findings will be biased unless retransformed using a correction which incorporates the log scale variance function, or alternatively using a GLM if modeling this variance function is not easily accomplished. In our analysis, we were not able to model the variance function for the log-scale error due to its complexity. Additionally, we found that the log-scale residuals were extremely heavy tailed (leptokurtotic), a finding that Manning and Mullahy state can lead to appreciable losses in precision when modeling using GLM.

A key distinction of the present study is the assortment of approaches to the econometric modeling of total healthcare cost data from pharmacy claims, employed to observe the predictive performance of each and ultimately select the superior model. Additionally, no data in the present study was truncated to remove the effect of outliers. Although such a task would likely have improved the overall fit of the models, the important ability to predict future total costs for the highest cost plan participants that often contribute a disproportionate amount of cost to the total population was preserved. The study also extends the literature by using comparison metrics other than r^2, namely PPV and specificity, which allow for a more intuitive interpretation of the performance of the model in predicting high cost participants and demonstrates relevance for using such models for grouping plan participants based on predicted costs for a variety of applications such as intervention strategies. Finally, the relatively recent data analyzed in this study accommodates treatment with newer agents that were not available when prior predictive models were published.

Although the models were estimated in a relatively large population and included a randomly withheld validation sample, this study was limited to 1 state employer health plan. Parameter estimates were found to be nonsignificant for some of the PHD drug categories for rare, short-lived, and/or episodic conditions. Furthermore, the study excluded elderly plan participants and those who may have died or disenrolled during the study period because of the 2-year continuous eligibility inclusion criterion. Consequently, the findings of this study may not be generalizable to all insured populations (eg, Medicaid or Medicare).

The acquisition, cleaning, and use of medical cost data are often prohibitive and costly for health plans and employers. Given these constraints, the PHD medication classification system derived solely from pharmacy insurance claims data is a viable alternative for plan sponsors who want to predict future total costs of their plan participants. While it was found that a simple OLS regression model may provide similar predictive accuracy in comparison to more advanced econometric approaches when modeling using PHD in this population, it is our recommendation that researchers carefully examine and compare the available modeling approaches before settling on a simple OLS model.

ACKNOWLEDGMENTS

The authors would like to recognize the contributions of Peggy Pierson, Christopher White, and Elliott Gerstner for data support; Dr. Michael Rushnak, Dr. Jeff Kalmanowicz, Dr. Aaron Eaton, and Dr. Deborah Cooper for clinical review; Dr. Olga Parsons for methodological review; and 3 anonymous reviewers for their insightful comments and suggestions.

REFERENCES

1. Cumming RB, Knutsin D, Cameron BA, Derrick B. A comparative analysis of claims-based methods of health risk assessment for commercial populations. Society of Actuaries. May 24, 2002. Available at: http://www.soa.org/ccm/content/areas-of-practice/special-interest-sections/health/health-section-sponsored-research. Accessed August 18, 2005.
2. Von Korff M, Wagner EH, Saunders K. A chronic disease score from automated pharmacy data. *J Clin Epidemiol.* 1992;45:197–203.
3. Clark DO, Von Korff M, Saunders K, et al. A Chronic Disease Score with empirically derived weights. *Med Care.* 1995;33:783–795.
4. Johnson RE, Hornbrook MC, Nichols GA. Replicating the Chronic Disease Score (CDS) from automated pharmacy data. *J Clin Epidemiol.* 1994;47:1191–1199.
5. Fishman PA, Shay DK. Development and estimation of a Pediatric Chronic Disease Score using automated pharmacy data. *Med Care.* 1999;37:874–883.
6. Fishman PA, Goodman MJ, Hornbrook MC, et al. Risk adjustment using automated ambulatory pharmacy data: the RxRisk Model. *Med Care.* 2003;41:84–99.
7. Lamers LM. Pharmacy Cost Groups. A risk-adjuster for capitation payments based on the use of prescribed drugs. *Med Care.* 1999;37:824–830.
8. Malone DC, Billups SJ, Valuck RJ, et al. Development of a chronic

disease indicator score using Veterans Affairs Medical Center medication database. *J Clin Epidemiol.* 1999;52:551–557.
9. Zhao Y, Ellis RP, Ash AS, et al. Measuring population health risks using inpatient diagnosis and outpatient pharmacy data. *Health Serv Res.* 2001;36:180–193.
10. Sales AE, Liu CF, Sloan KL, et al. Predicting costs of care using a pharmacy-based measure risk adjustment in a veteran population. *Med Care.* 2003;41:753–760.
11. Sloan KL, Sales AE, Liu CF, et al. Construction and characteristics of the RxRisk-V: a VA-adapted pharmacy-based case-mix instrument. *Med Care.* 2003;41:761–774.
12. Gilmer T, Kronick R, Fishman PA, et al. The Medicaid Rx model: pharmacy-based risk adjustment for public programs. *Med Care.* 2001; 39:1188–1202.
13. Putnam KG, Buist DSM, Fishman PA, et al. Chronic Disease Score as a predictor of hospitalization. *Epidemiology.* 2002;13:340–346.
14. Lewis NJ, Patwell JT, Breisacher BA. The role of insurance claims databases in drug therapy outcomes research. *Pharmacoeconomics.* 1993;4:323–330.
15. Levy AR, O'Brien BJ, Sellors C, et al. Coding accuracy of administrative drug claims in the Ontario Drug Benefit database. *Can J Clin Pharmacol.* 2003;10:67–71.
16. Thomas M, Cleland J, Price D. Database studies in asthma pharmacoeconomics: uses, limitations and quality markers. *Expert Opin Pharmacother.* 2003;4:351–358.
17. Powers CA, Meyer CM, Cooper D, et al. Predicting one and two-year risk of hospitalization using Patient Health Dimensions, a pharmacy-based risk index (abstract for material presented at the 16th Annual Meeting of the Academy of Managed Care Pharmacy). *J Manage Care Pharmacy.* 2004;10:200.
18. Meyer CM, Cooper D, Kalmanowicz J, et al. Three year predictive model of medical cost risk and methodological issues related to an expanded pharmacy claims risk index (abstract for material presented at the 8th Annual Meeting of the International Society for Pharmacoeconomics and Outcomes Research). *Value Health.* 2003;6:212.
19. Meyer CM, Liberman J, Kalmanowicz J, et al. Predicting variation in health-related quality of life (HRQOL) based on a pharmacy claims health risk index in adult asthmatics. Presented at Academy Health Meeting, Nashville, TN, 2003.
20. Manning WG, Mullahy J. Estimating log models: to transform or not to transform. *J Health Econ.* 2001;20:461–494.
21. Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. *J Health Econ.* 1999;18:153–171.
22. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health economics. *J Health Econ.* 1998;17:247–281.
23. Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *J Health Econ.* 1998;17:283–295.
24. Duan N. Smearing estimate: a nonparametric retransformation method. *J Am Stat Assoc.* 1983;78:605–610.
25. Veazie PJ, Manning WG, Kane RL. Improving risk adjustment for Medicare capitated reimbursement using nonlinear models. *Med Care.* 2003;41:741–752.
26. Buntin MB, Zaslavsky AM. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *J Health Econ.* 2004;23:525–542.
27. Breusch TS, Pagan AR. A simple test for heteroscedasticity and random coefficient variation. *Econometrica.* 1979;47:1287–1294.
28. Cook RD, Weisberg S. Diagnostics for heteroscedasticity in regression. *Biometrika.* 1983;71:1–10.
29. Park R. Estimation with heteroscedastic error terms. *Econometrica.* 1966;38:888.
30. SAS Version 8.02. Cary, NC: SAS Institute Inc.; 2001.
31. Stata Statistical Software. Release 8.0. College Station, TX: Stata Corp LP; 2004.