© Published 2016. This article is a U.S. Government work and is in the public domain in the USA
DOI: 10.1111/1475-6773.12454
METHODS ARTICLE

Risk Adjustment Tools for Learning Health Systems: A Comparison of DxCG and CMS-HCC V21

Todd H. Wagner, Anjali Upadhyay, Elizabeth Cowgill, Theodore Stefos, Eileen Moran, Steven M. Asch, and Peter Almenoff

Objective. To compare risk scores computed by DxCG (Verisk) and Centers for Medicare and Medicaid Services (CMS) V21.

Research Design. Analysis of administrative data from the Department of Veterans Affairs (VA) for fiscal years 2010 and 2011.

Study Design. We regressed total annual VA costs on predicted risk scores. Model fit was judged by *R*-squared, root mean squared error, mean absolute error, and Hosmer–Lemeshow goodness-of-fit tests. Recalibrated models were tested using split samples with pharmacy data.

Data Collection. We created six analytical files: a random sample ($n = 2$ million), high cost users ($n = 261,487$), users over age 75 ($n = 644,524$), mental health and substance use users ($n = 830,832$), multimorbid users ($n = 817,951$), and low-risk users ($n = 78,032$).

Principal Findings. The DxCG Medicaid with pharmacy risk score yielded substantial gains in fit over the V21 model. Recalibrating the V21 model using VA pharmacy data-generated risk scores with similar fit statistics to the DxCG risk scores.

Conclusions. Although the CMS V21 and DxCG prospective risk scores were similar, the DxCG model with pharmacy data offered improved fit over V21. However, health care systems, such as the VA, can recalibrate the V21 model with additional variables to develop a tailored risk score that compares favorably to the DxCG models.

Key Words. Risk adjustment, cost, performance measurement, health economics

Large clinical databases provide a foundation for policy makers and health system managers to track performance metrics across health care systems and over time. Such tracking is necessary to build learning health care systems (Bohmer 2011; Smith et al. 2012). However, to conclude that one hospital has higher mortality rates than another, or that one physician group is more efficient than another, analysts must adjust for inherent differences in the patients'

underlying clinical needs. Numerous risk adjustment techniques have been developed over the past two decades (Schone and Brown 2013), some proprietary and some in the public domain (Pope et al. 2004). Most use International Classification of Diseases, ninth revision (ICD-9) codes to classify patients into homogenous clinical categories, such as diabetes without complications. Some systems only produce these clinical categories, while others use these categories to produce a risk score, which reduces the multidimensional problem with many condition categories and facilitates risk adjustment in health services research. Choosing among these risk adjustment techniques is challenging, yet head-to-head comparisons are quite rare, particularly for proprietary software.

The Department of Veterans Affairs (VA) uses risk adjustment while producing performance metrics for its hospital systems. For example, VA's hospital efficiency metric requires detailed information on patient risk and input prices (Gao et al. 2011). Historically, VA used DxCG, supported by Verisk, which is a commonly used proprietary risk adjustment systems for cost data. We sought to compare DxCG to the CMS V21 system (hereafter referred to as V21), which was developed to adjust capitated payments for Medicare Advantage plans (Pope et al. 2011). For the comparison, we use data from the VA. We set forth two aims. First, we compared age and sex adjusted risk scores computed by DxCG and V21. Second, as many health systems have rich clinical data on their members, in Aim 2 we created a new risk score based on the V21 and additional information, including pharmacy data, and compared this new recalibrated risk score to the DxCG. This second aim investigates the value gained by tailoring and recalibrating the V21 risk score relative to DxCG.

Address correspondence to Todd H. Wagner, Ph.D., VA Palo Alto Health Economics Resource Center (HERC) and Department of Health Research and Policy, Stanford University, 795 Willow Road, 152-MPD, Menlo Park, CA 94025; e-mail: todd.wagner@va.gov. Todd H. Wagner, Ph.D., is also with the Center for Innovation to Implementation, VA Palo Alto, Menlo Park, CA; Department of Health Research and Policy, Stanford University, Palo Alto, CA. Anjali Upadhyay, M.S., and Elizabeth Cowgill, M.P.H., are with the Health Economics Resource Center (HERC), VA Palo Alto, Menlo Park, CA. Theodore Stefanos, Ph.D., and Eileen Moran, M.H.A., are with the VHA Office of Productivity, Efficiency & Staffing, Bedford, MA. Steven M. Asch, M.D., M.P.H., is with the Center for Innovation to Implementation, VA Palo Alto, Menlo Park, CA; Division of General Medical Disciplines, Stanford University, Palo Alto, CA. Peter Almenoff, M.D., is with the Department of Veterans Affairs, Operational Analytics and Reporting, Office of Informatics and Analytics; Department of Veterans Affairs, Office of Secretary; Department of Veterans Affairs, Center of Innovation; Vijay Babu Rayudu Endowed Chair in Patient Safety, University of Missouri-Kansas City, Kansas City, MO.

METHODS

Data and Variables

We extracted utilization and diagnostic data from national databases (National Patient Care Database and Patient Treatment Files) maintained by VA for fiscal years (FY) 2010 and 2011. Cost data were obtained from the Decision Support System (DSS) and the HERC Average Cost Datasets (Phibbs et al. 2003; Wagner, Chen, and Barnett 2003; Yu et al. 2003). For care purchased by the VA from non-VA providers, we extracted paid claims from the Fee Basis datasets.

One challenge in choosing among risk adjustment techniques is that systems developed from commercial providers or for Medicare may not translate to other populations. Therefore, we created six analytical files reflecting veterans who used the VA in FY10 and FY11: (1) 2 million VA users, (2) veterans age 75 and over, (3) high-cost veterans, defined as those veterans in the top 5 percent of annual VA costs, (4) veterans with a mental health condition and/or substance use disorder (MH-SUD sample), (5) veterans with multimorbidity, and (6) low-risk veterans. Samples 2, 4, and 5 were 50 percent random samples from all VA users. Patients in the high cost sample were identified in the HERC Average Cost data, which adjust for geographic wage variation. The multimorbidity sample was defined as any veteran having conditions affecting two or more body systems as measured in the AHRQ Chronic Condition Indicators (Agency for Healthcare Quality and Research 2013). The low-risk group was focused on identifying patients who were relatively healthy. Therefore, we selected veterans whose health utilization, as tracked in diagnostic codes, affected one or fewer body systems and had a V-code associated with having a physical in the VA, so as to exclude patients who relied primarily on non-VA providers. All analytical files were created from the universe of all VA users, with the exception of the low-risk sample, which was drawn from the 2 million random sample.

For each sample, we computed the DxCG and V21 risk scores using fiscal year 2010 clinical data. The DxCG *Risk Solutions* software generates 394 Hierarchical Condition Categories (HCCs) in the process of calculating three risk scores: (1) DxCG Medicare prospective (excluding Rx), (2) DxCG Medicaid prospective (including Rx), and (3) DxCG Medicare concurrent (excluding Rx). The V21 model creates 189 HCCs and uses a subset of these HCCs to calculate a risk score. The default risk score is based on weights for people living in the community; different weights are available for those living in an

institution or new enrollees. We used the community risk score, unless a person spent more than 90 days in a nursing home, at which point we used the institutional weights.

A challenge in comparing risk models is the distinction between prospective and concurrent risk. Prospective risk models use current year diagnoses to estimate next year's costs, while concurrent risk models use current year diagnoses to estimate current costs. Prospective risk models are thought to place more weight on chronic conditions that influence expenditures over multiple time periods, while concurrent models place more weight on acute care needs. Because they use utilization and outcome data from the same time period, concurrent risk models usually yield higher fit statistics (e.g., *R*-squared) than prospective models. DxCG software produces three risk scores (two prospective and one concurrent), while the V21 provides only a prospective risk score, although it can be modified to create a concurrent risk score. Therefore, we compared all risk models in their ability to measure concurrent and prospective costs.

To compare risk scores, we calculated each person's total costs, which included inpatient, outpatient, pharmacy, and purchased care (Fee Basis) in FY2010. We excluded care provided by and paid for by non-VA insurers (e.g., Medicare). Total costs were based on DSS data, which is an activity-based cost accounting system embedded at local VA medical centers. Although researchers have used these cost data and found them to be accurate and precise (Chapko et al. 2008; Wagner et al. 2011), there remains the potential for unusually high costs (due to mismatched units and unit costs) or negative costs (due to accounting reconciliations). Veterans whose DSS pharmacy costs exceeded \$50,000 were excluded from these analyses. In addition, negative inpatient, outpatient, and pharmacy costs were replaced with zeros. We used HERC Average Cost data, which are encounter-level cost estimates based on non-VA relative value units, in a sensitivity analysis. The results were highly robust to the source of cost data (HERC vs. DSS), so we present results using the DSS data.

For aim 1, we established a base set of covariates including the risk score, age, age-squared, and gender. For aim 2, we included age, age-squared, gender, and a more extensive set of covariates, including race, being married, and 46 psychiatric condition categories, developed by Rosen and colleagues (Montez-Rath et al. 2006; Sloan et al. 2006), to account for behavioral health needs for which existing risk systems struggle (Ettner et al. 2000). We included a variable identifying a lack of health insurance, veteran's priority-level status (Department of Veterans Affairs 2013), and an indicator if the veteran was in a

designated registry (e.g., Agent Orange). Finally, we included 25 drug class categories based on an alphanumeric list of 580 drug types maintained by the VA Pharmacy Benefits Management Service (see Table A1). Each drug class represented whether the patient had had a related prescription in the year (i.e., frequency, quantity, or dose was not recorded). We tried three alternative methods for including pharmacy information: CMS' Rx-HCC model, Medicaid Rx, and veteran's VA pharmacy costs in the prior year. None of these methods worked as well as the 25 VA drug class, so we report the results with the VA drug class information here.

Analyses

The first objective was to compare DxCG and V21 risk scores using clinical and cost data from FY2010. We used regression models to understand how these concurrent risk scores fit the data while controlling for age and gender. We compared ordinary least squares (OLS), log-OLS, square root OLS, and generalized linear model (GLM) with gamma distribution and log link, and GLM with gamma distribution and square root link. Log-OLS and square root OLS refer to models where the natural log or square root is taken of the dependent variable and OLS is performed on the transformed variable. To compare model fit, we examined *R*-squared, root mean squared error (RMSE), and mean absolute error (MAE). In addition, Hosmer–Lemeshow goodness-of-fit tests were run by sorting the predicted values from low to high, and then comparing the mean observed and expected values within each decile. Poorly specified models have extreme differences across the deciles. We examined alternative *R*-squared measures for the GLM models, but in the end chose not to include them in the table as they can create confusion when compared to the traditional *R*-squared. Model fit statistics were calculated when the data were retransformed with the appropriate smearing estimator (Duan et al. 1983; Jones 2010). In separate analyses, we compared the prospective risk models, in which we used FY2010 clinical data to predict FY2011 costs.

The second objective was to compare the DxCG and V21 risk models, after recalibrating the V21 to the VA with some additional control variables. For this aim, we reran the analytical models outlined in aim 1 with the extensive set of control variables, described above. Recalibrated risk scores were calculated by using the regression model to predict the person's costs, which we then divided by the average predicted costs. Recalibration was done by

splitting the samples into two random groups, one for the estimation and the other for validation.

Risk scores can span the range of positive numbers. Unfortunately, different regression specifications can lead to prediction violations, such as negative estimates. The square root OLS model removes the possibility of negative predictions because the predictions are squared and corrected with a smearing estimator (Jones 2010). Squaring values can create rank order problems (i.e., when squared, -2 and 1 change rank order). We examined the raw square root OLS predictions and confirmed that we did not have any negative predicted values.

RESULTS

Sample Characteristics

Table 1 shows the descriptive statistics for the six analytical samples. In the general sample, the average age was 62 years and 94 percent were male. In terms of age and gender, the multimorbid and high cost samples were similar to the general sample, whereas the low-risk sample had a younger mean age and had a higher proportion of women.

In the random sample, the average annual cost was \$8,819 (SD: \$24,976), with a maximum of \$1.66 million (see Table 1). The distribution of costs varied across the samples. The MH-SUD and multimorbid sample had 2–3 times higher average costs than the general sample. The high cost subsample, which represented the top 5 percent of high cost patients, had an average cost of \$76,920. The low-risk sample had an average cost of \$2,435, although it was based on a smaller sample ($n = 78,032$ persons), in part because of the operational definition and limitations in defining low risk using administrative data without relying directly on utilization or costs.

Aim 1: Comparison of Risk Scores

For all of the samples, the average V21 score yielded a higher average score than the average DxCG prospective risk score excluding pharmacy (Table 2, columns 1 and 2). However, the V21 score was much lower than the average DxCG prospective risk score based on Medicaid data with pharmacy (Table 2). Note that all of these risk scores were being compared in concurrent models.

Table 1: Descriptive Statistics

	General Sample	Older Sample	High-Cost Sample	MH-SUD Sample	Multimorbid Sample	Low-Risk Sample
N	1,995,620	644,524	261,487	830,832	817,951	78,032
Mean age (SD)	62.0 (15.9)	81.4 (4.6)	62.5 (13.4)	56.9 (15.2)	62.2 (13.8)	48.2 (17.4)
Male	94%	98%	95%	91%	94%	86%
Total costs*						
Mean	8,819	8,067	76,920	15,067	21,345	2,435
Median	2,563	1,908	52,954	5,637	9,337	1,093
SD	24,976	25,624	76,697	33,560	40,603	5,203
Maximum	1,660,240	1,597,986	2,979,525	2,476,373	2,979,525	275,166
Percent of costs from pharmacy	18.9	25.3	5.3	17.4	15.3	6.5

Notes: *Total Costs include inpatient, outpatient, pharmacy, and Fee Basis care. Veterans whose DSS Rx costs exceed \$50,000 were excluded from these analyses. Negative DSS VA in/out costs and DSS Rx costs were also replaced with zeros.

Table 2: Average Risk Scores by Analytical Sample

	<i>CMS V21 Concurrent Mean (SD)</i>	<i>DxCG Medicare</i>		<i>DxCG Medicaid</i>
		<i>Prospective without Rx Mean (SD)</i>	<i>Concurrent without Rx Mean (SD)</i>	<i>Prospective with Rx Mean (SD)</i>
Random sample of users	0.756 (0.730)	0.661 (0.698)	0.497 (0.879)	1.756 (2.126)
Older users	1.065 (0.750)	0.921 (0.656)	0.504 (0.874)	2.020 (2.267)
High cost users	2.234 (1.580)	2.077 (1.628)	2.684 (2.243)	7.228 (4.323)
MH-SUD users	0.893 (0.850)	0.802 (0.802)	0.770 (1.092)	2.487 (2.708)
Multimorbid users	1.160 (1.024)	1.044 (1.002)	1.004 (1.343)	3.146 (3.027)
Low-risk sample	0.295 (0.234)	0.236 (0.219)	0.152 (0.244)	0.708 (0.708)

Notes: Risk scores were computed using VA clinical data from 2010.

The V21 score was most highly correlated with the DxCG Medicare prospective score without pharmacy (results not shown). The correlations ranged from .83 in the general sample to .75 in the high-cost and low-risk sample. The similarities between the V21 and DxCG Medicare prospective score was striking when computing the difference score; more than 99 percent of the cases in the general sample had a difference score between $-.1$ and $.1$. Correlations between the V21 and DxCG Medicaid prospective score with pharmacy were lowest for the low-risk sample ($r = .33$). The correlations were higher for the general sample (.64), older sample ($r = .65$), high-cost sample (.57), MH-SUD sample ($r = .65$), and multimorbid sample (.63).

Four goodness-of-fit statistics showed that no single statistical model was universally superior, especially when the data were viewed across the six analytical samples. Table 3 shows the fit statistics for the square root OLS model across the six analytical samples. The fit statistics for the general sample and the complete Hosmer–Lemeshow deciles for the OLS and square root OLS models are available in additional tables online.

The large volume of results makes it easy to lose sight of the primary comparison of interest: the performance of the alternative risk models. The CMS V21 and DxCG prospective model yielded very similar fit statistics. In estimating concurrent costs, neither fared as well as the concurrent DxCG model nor the DxCG prospective model based on Medicaid with pharmacy data. As shown in Table 4, R -squared estimates were between 8 and 20 percentage points higher for the DxCG prospective model with pharmacy than the CMS V21 model. All of the models struggled to fit the data from the low-

Table 3: Concurrent Recalibrated Model Fit Statistics from Ordinary Least Squares (OLS) Regression with Square Root Transformation*

	<i>CMS V21</i>		<i>DxCG Medicare</i>		<i>DxCG Medicaid</i>
	<i>Concurrent</i>	<i>Concurrent with VA Drug Class Indicators</i>	<i>Prospective without Rx</i>	<i>Concurrent without Rx</i>	<i>Prospective with Rx</i>
<i>R²</i>					
General	0.5793	0.6924	0.5819	0.6274	0.6351
Older	0.5728	0.6772	0.5677	0.6233	0.6397
MH-SUD	0.5820	0.6810	0.5896	0.6268	0.6509
High cost	0.3559	0.4281	0.3544	0.4244	0.4241
Multimorbid	0.5350	0.6331	0.5326	0.5957	0.5943
Low risk	0.2922	0.4573	0.3113	0.3508	0.3778
<i>HL F-stat</i>					
General	9,206	7,499	20,251	20,687	11,726
Older	3,803	2,182	6,683	7,074	3,418
MH-SUD	2,883	2,941	6,699	6,068	3,288
High cost	197	298	214	167	159
Multimorbid	1,671	1,972	4,282	5,297	1,992
Low risk	53	91	112	127	113
<i>RMSE</i>					
General	16,560	15,531	17,053	16,590	15,540
Older	15,972	14,894	16,415	16,108	14,652
MH-SUD	22,277	20,922	22,713	21,920	20,667
High cost	57,306	54,756	57,422	54,254	54,473
Multimorbid	27,709	26,105	28,234	26,761	26,069
Low risk	4,393	4,023	4,418	4,269	4,164
<i>MAE</i>					
General	5,977	5,084	5,965	5,590	5,541
Older	5,535	4,765	5,582	5,138	5,040
MH-SUD	9,224	8,026	9,142	8,668	8,400
High cost	33,186	31,216	33,167	31,127	31,463
Multimorbid	12,414	10,920	12,466	11,492	11,533
Low risk	1,892	1,643	1,862	1,815	1,772

Notes: *R²* represents *R*-squared; HL *F*-stat is the Hosmer–Lemeshow goodness-of-fit *F* statistic; RMSE is root mean squared error; MAE is mean absolute error.

*All analyses are based on risk scores computed with 2010 clinical data in relation to 2010 costs.

risk sample, but again the DxCG prospective model with pharmacy accounted for more variance.

The results described above use the risk scores computed with 2010 data to model 2010 costs. In these concurrent comparisons, *R*-squared estimates ranged from .4287 to .5682. As expected, model fit, as measured by *R*-squared, decreased when the 2010 risk scores were used to predict 2011 costs.

Table 4: Concurrent Model Fit Statistics from Ordinary Least Squares (OLS) Regression with Square Root Transformation*

		<i>DxCG Medicare</i>		<i>DxCG Medicaid</i>
	<i>CMS V21</i>	<i>Prospective</i>	<i>Concurrent</i>	<i>Prospective</i>
	<i>Concurrent</i>	<i>without Rx</i>	<i>without Rx</i>	<i>with Rx</i>
<i>R</i> ²				
General	0.4287	0.4308	0.5122	0.5682
Older	0.4108	0.3876	0.4802	0.5907
MH-SUD	0.3985	0.4191	0.4876	0.5738
High cost	0.1920	0.1999	0.2650	0.3779
Multimorbid	0.3910	0.3906	0.4790	0.5377
Low risk	0.1646	0.1966	0.2694	0.2701
<i>HL F</i> -stat				
General	9,516	18,274	33,496	9,454
Older	3,759	5,111	9,318	2,463
MH-SUD	3,812	5,017	8,848	3,906
High cost	454	130	56	158
Multimorbid	2,519	4,071	6,212	1,760
Low risk	34	90	156	127
<i>RMSE</i>				
General	20,576	21,829	22,060	17,884
Older	22,018	23,377	23,761	18,464
MH-SUD	27,942	29,215	28,865	23,895
High cost	70,312	70,003	67,206	62,716
Multimorbid	34,035	35,043	33,708	29,888
Low risk	4,945	5,045	4,782	4,605
<i>MAE</i>				
General	7,415	7,423	6,783	6,398
Older	7,320	7,552	6,812	6,077
MH-SUD	11,843	11,607	10,774	9,942
High cost	41,640	41,266	39,120	36,720
Multimorbid	15,225	15,236	13,868	13,234
Low risk	2,087	2,035	1,937	1,941

Notes: R^2 represents R -squared; HL F -stat is the Hosmer–Lemeshow goodness-of-fit F statistic; RMSE is root mean squared error; MAE is mean absolute error.

*All analyses are based on risk scores computed with 2010 clinical data in relation to 2010 costs.

When used in a prospective fashion, the R -squared for the risk scores ranges from .2097 to .2487 (see supplemental tables).

Aim 2: Recalibrating the Model for VA

As was found in aim 1, different regression specifications yielded different fit statistics. The log-transformed and GLM log-link models provided good fits

across the Hosmer–Lemeshow deciles, with the exception of the top decile, where the fit was very poor. This was evident in very large RMSE and MAE fit statistics. Convergence problems were also more common with the GLM models. In comparison, the square root OLS model yielded the highest *R*-squared values, good fits across the Hosmer–Lemeshow deciles, and respectable RMSE and MAE statistics.

Recalibrating the risk models with additional control variables resulted in higher *R*-squared values for all risk scores. In the general sample, the recalibrated V21 model yielded an *R*-squared of .58. This fit was still lower than the DxCG Medicaid model with pharmacy (*R*-squared of .64). Adding the 26 drug class variables to the V21 model provided *R*-squared values that were similar to or higher than the DxCG Medicaid with pharmacy model. The addition of pharmacy information was most dramatic with the low-risk sample. The Hosmer–Lemeshow deciles for the square root OLS model are available in an additional table online.

Recalibration was also influential in improving *R*-squared when the 2010 risk scores were being used to predict 2011 cost data. As shown in Table 5, model fit improved for all models, but model fit for these prospective models was still substantially lower than the *R*-squared for the concurrent models shown in Table 3.

Table 5: Prospective *R*-Squared Fit Statistics*

		<i>DxCG Medicare</i>		<i>DxCG Medicaid</i>
	<i>CMS V21</i>	<i>Prospective without Rx</i>	<i>Concurrent without Rx</i>	<i>Prospective with Rx</i>
2011 costs				
Basic model				
General	0.2117	0.2228	0.2097	0.2487
Older	0.1743	0.1771	0.1732	0.2538
MH-SUD	0.1819	0.1970	0.1873	0.2408
High cost	0.0401	0.0676	0.0403	0.0949
Multimorbid	0.1527	0.1690	0.1520	0.1903
Low risk	0.0600	0.0684	0.0673	0.0470
Recalibrated with additional variables				
General	0.3485	0.3558	0.3485	0.3391
Older	0.3428	0.3480	0.3440	0.3405
MH-SUD	0.3190	0.3252	0.3182	0.3148
High cost	0.1423	0.1631	0.1434	0.1399
Multimorbid	0.2640	0.2757	0.2631	0.2544
Low risk	0.1321	0.1326	0.1319	0.1272

Notes: *R*² is based on square root OLS regression models.
*All analyses are based on risk scores computed with 2010 clinical data in relation to 2011 costs.

In the recalibrated models, the average risk score in the general sample is 1 (by construction). Average risk increased considerably for the high-cost group (average 5.58), and averaged .038 for the low-risk group. Interestingly, the older population had a slightly lower average risk score (.91). We expect this is due to an increased reliance on non-VA providers when veterans become Medicare eligible and this is scheduled for future examination. The split-sample analysis confirmed that model fit was very similar in the estimation dataset as compared to the validation dataset (results not shown). The models also confirmed that the results were robust when we used HERC costs instead of DSS costs.

DISCUSSION

Many organizations face a difficult choice in selecting risk adjustment techniques for cost data. Proprietary techniques like DxCG have established reputations and a long history of use, but they have not usually been the subject of direct comparisons with free alternatives. Our analyses show that the choice depends, in part, on the anticipated use and whether the health care system can recalibrate the model. If users have a strong preference for a concurrent risk model (e.g., risk adjustment in a sample with most acute care needs) or cannot recalibrate the model, then the V21 software is unlikely to be as satisfying as some of the DxCG risk scores. The DxCG concurrent and prospective with RX risk scores provided stronger fit statistics than the V21 model when used as a concurrent risk score.

One advantage of the DxCG system over the V21 is its ability to compute a risk score using pharmacy data. In the analyses with our samples, the pharmacy information added 8 to 20 percentage points to the *R*-squared estimates when compared to the base models without pharmacy, and the pharmacy data were particularly valuable in the low-risk group. The advantage of the DxCG software diminished greatly when we recalibrated the V21 risk score to include pharmacy information; in a few cases, the recalibrated V21 score yielded fit statistics that surpassed the DxCG scores.

We examined alternative methods for including pharmacy information into the risk score: CMS Rx-HCC model, Medicaid Rx (Gilmer et al. 2001), Rx risk (Fishman et al. 2003), and pharmacy costs in the prior year. Interestingly, we found that the extensive margin (any use of pharmacy in the drug classes) was more useful than the intensive margin (pharmacy spending in the prior year). We also found that while the Rx-HCC and Medicaid Rx model

yielded substantial improvements in fit over the V21 model alone, none of these methods worked as well as the VA drug class. Because the choice of risk adjustment techniques depends on the availability of pharmacy data, we separately report the data with the pharmacy variables so that readers can interpret the results accordingly. The VA drug class is maintained by the VA PBM, and the national drug file has both the National Drug Code (NDC) and the VA drug class (VA Pharmacy Benefits Management Service). Therefore, researchers outside of VA can also use this information to link NDC codes to the VA drug class. To expedite this process, we provide a dataset linking VA drug class and NDC for 2014 data (Wagner et al. 2015).

Health services researchers can easily download the V21 model and compute risk scores for their data. At a minimum, researchers would need age, gender, and ICD-9 diagnostic information. Whether V21 risk scores are sufficient depends in part on the analyst's sample. Given the models were developed on a wide distribution of people, they tend to yield poor fits on either end of the distribution (low-risk individuals or high-cost users). Thus, researchers conducting work in the areas of health promotion and prevention may need to consider other risk adjustment models. It was in the low-risk subpopulation where pharmacy data greatly improved the model fit, as many people may use little health care in part because they are adhering to medication regimens.

Robustness and Sensitivity

The DxCG and V21 risk scores were tested across six samples, using a range of regression models and a number of fit statistics. The results did not identify a preferred statistical approach, but they do provide some reassurances on the model specification. Like prior research (Montez-Rath et al. 2006), the square root transformation generally performed quite well in its ability to fit the data, as evidence with the Hosmer–Lemeshow deciles and *R*-squared. Readers should be cautious about using these results to conclude that one risk model was always preferred to another risk model; the results do not support that conclusion. The OLS and square root models can also create prediction problems (negative or changes in rank order). Although we did not encounter rank order problems with the square root models, analysts who use this approach would want to check for such possibilities.

Despite all of the analyses, there are some limitations with this research that are worthy of discussion. There were many alternative risk adjustment systems that we could have chosen for this comparison. For

example, Charlson and Chronic Illness and Disability Payment System (CDPS) are in the public domain and CRGs and ACGs are common proprietary systems. We limited our comparison to DxCG and V21 because they are commonly used for utilization and cost outcomes, and because the history and potential for use in the VA. However, we believe the methods we present can guide future comparisons of other risk adjustment systems. Second, these risk models were all compared using cost data as the dependent variable. Risk adjustment techniques designed for one outcome (e.g., costs) may not perform well on other outcomes (e.g., mortality) (Quan et al. 2005; Wang et al. 2012a,b). Finally, these models focused on care provided by or paid by VA. Most older veterans are eligible for Medicare and Medicare use is common (Petersen et al. 2010; Liu et al. 2011). It is possible to include Medicare data in risk scores, although those data are subject to lags in availability. Ongoing work is examining how these risk scores change when Medicare data are included.

Gaming and Risk Selection

One concern with risk adjustment models is gaming to make the patient's risk appear to be as high as possible (Schone and Brown 2013). In the current study, we included pharmacy data, which proved to increase the fit of the regression models substantially. Pharmacy data were based on any use of a medication in one of 25 VA drug classes. While VA drug class outperformed other models for pharmacy use, caution is warranted because measuring pharmacy use in this manner could lead to gaming. In VA, the issue of gaming is less acute because the VA uses a separate tiered capitation model, known as Veterans Equitable Resource Allocation, to distribute funding without using pharmacy data.

Another concern with risk adjustment is risk selection (c.f., Brown et al. 2014). In VA, the issue of risk selection is less immediately apparent because the models developed in this paper were designed for performance monitoring and are distinct from VA's financial allocation model. One area of potential concern for risk selection is among veterans who are dually eligible for VA and other insurance, namely Medicare. Because the VERA is calculated only on VA data, there is an implicit incentive for VA medical centers to provide enough services that "qualify" patients for the highest possible reimbursement level, but there is no incentive to provide all of the veterans' care. An alternative model that estimates patient risk based on complete clinical information

(VA and non-VA claims) and then allocates VA funds based on a prorated algorithm would change the incentives associated with dual coverage. More research would be needed to identify a suitable algorithm for prorating the funds, but it could potentially reduce the notable distortions caused by risk selection and dual use (Trivedi et al. 2012).

Implementation

The VA has computed the V21 and the recalibrated models to estimate patient risk, and to date risk scores for 2006–2015 have been created on its data warehouses. The recalibrated risk scores have been named Nosos, which means “chronic condition” in Greek. The square root OLS model has consistently provided excellent fit and across the 9 years of data, and we have not encountered any rank order problems when retransforming the data. Further data documentation for the implementation of V21 and Nosos in VA is available online (Wagner et al. 2015).

CONCLUSION

We find high concordance between the DxCG prospective risk score based on Medicare data and the V21 model provided by CMS. The DxCG system provides alternative risk scores that are not provided by V21 (a prospective Medicaid score with pharmacy and a concurrent Medicare score). Others may place more value on these other risk scores, and we found that the DxCG Medicaid score with pharmacy frequently produced the best fit statistics in concurrent models. The V21 model can be recalibrated, and with the inclusion of pharmacy data, the recalibrated V21 model yielded risk scores that provided similar fit statistics to the DxCG software.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: Funding for this study was provided by the VA Office for Operational Analytics and Reporting. We thank Mel Ingber (RTI), Paul Fishman (Group Health Cooperative), the associate editor, and two anonymous reviewers for suggestions. We appreciate the comments and insights from James Campbell, Yu-Fang Li, Mei-Ling Shen, Bruce Kinoshian, Amy Rosen, and Maria Rath Montez.

Disclosures: This work was funded through the Operational Value Initiative from the VA Office of Analytics and Business Intelligence/Office of Informatics and Analytics and VA Operational Analytics and Reporting.

Disclaimers: The views, opinions, and content of this publication are those of the authors and do not necessarily reflect the views, opinions, or policies of the Department of Veterans Affairs. The authors had full control of the study design, methods used, outcome parameters and results, analysis of data, and production of the written report.

REFERENCES

- Agency for Healthcare Quality and Research. 2013. "Chronic Condition Indicator (CCI) for ICD-9-CM" [accessed on April 20, 2013]. Available at <http://www.hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp>
- Bohmer, R. M. 2011. "The Four Habits of High-Value Health Care Organizations." *New England Journal of Medicine* 365 (22): 2045–7.
- Brown, J., M. Duggan, I. Kuziemko, and W. Woolston. 2014. "How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program." *American Economic Review* 104 (10): 3335–64.
- Chapko, M. K., C. F. Liu, M. Perkins, Y. F. Li, J. C. Fortney, and M. L. Maciejewski. 2008. "Equivalence of Two Healthcare Costing Methods: Bottom-Up and Top-Down." *Health Economics* 18 (10): 1188–201.
- Department of Veterans Affairs. 2013. "Health Benefits: Priority Group Tables" [accessed on April 16, 2013]. Available at http://www.va.gov/healthbenefits/resources/priority_groups.asp
- Duan, N., W. G. Manning, C. N. Morris, and J. P. Newhouse. 1983. "A Comparison of Alternative Models for the Demand for Medical Care." *Journal of Business and Economic Statistics* 1 (2): 115–26.
- Ettner, S. L., R. G. Frank, T. Mark, and M. W. Smith. 2000. "Risk Adjustment of Capitation Payments to Behavioral Health Care Carve-Outs: How Well Do Existing Methodologies Account for Psychiatric Disability?" *Health Care Manag Sci* 3 (2): 159–69.
- Fishman, P. A., M. J. Goodman, M. C. Hornbrook, R. T. Meenan, D. J. Bachman, and M. C. O. K. Rosetti. 2003. "Risk Adjustment Using Automated Ambulatory Pharmacy Data: The RxRisk Model." *Medical Care* 41 (1): 84–99.
- Gao, J., E. Moran, P. L. Almenoff, M. L. Render, J. Campbell, and A. K. Jha. 2011. "Variations in Efficiency and the Relationship to Quality of Care in the Veterans Health System." *Health Affairs (Millwood)* 30 (4): 655–63.
- Gilmer, T., R. Kronick, P. Fishman, and T. G. Ganiats. 2001. "The Medicaid Rx Model: Pharmacy-Based Risk Adjustment for Public Programs." *Medical Care* 39 (11): 1188–202.

- Jones, A. 2010. "Models for Health Care." In *The Oxford Handbook of Economic Forecasting*, edited by D. Hendry, and M. Clements, pp. 625–54. Oxford, England: Oxford University Press.
- Liu, C. F., W. G. Manning, J. F. Burgess Jr, P. L. Hebert, C. L. Bryson, J. Fortney, M. Perkins, N. D. Sharp, and M. L. Maciejewski. 2011. "Reliance on Veterans Affairs Outpatient Care by Medicare-Eligible Veterans." *Medical Care* 49 (10): 911–7.
- Montez-Rath, M., C. L. Christiansen, S. L. Ettner, S. Loveland, and A. K. Rosen. 2006. "Performance of Statistical Models to Predict Mental Health and Substance Abuse Cost." *BMC Medical Research Methodology* 6: 53.
- Petersen, L. A., M. M. Byrne, C. N. Daw, J. Hasche, B. Reis, and K. Pietz. 2010. "Relationship between Clinical Conditions and Use of Veterans Affairs Health Care among Medicare-Enrolled Veterans." *Health Services Research* 45 (3): 762–91.
- Phibbs, C. S., A. Bhandari, W. Yu, and P. G. Barnett. 2003. "Estimating the Costs of VA Ambulatory Care." *Medical Care Research and Review* 60 (3 Suppl): 54S–73S.
- Pope, G. C., J. Kautter, R. P. Ellis, A. S. Ash, J. Z. Ayanian, L. I. Lezzoni, M. J. Ingber, J. M. Levy, and J. Robst. 2004. "Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model." *Health Care Financ Rev* 25 (4): 119–41.
- Pope, G. C., J. Kautter, M. J. Ingber, S. Freeman, R. Sekar, and C. Newhart. 2011. *Evaluation of the CMS-HCC Risk Adjustment Model*. Prepared by RTI for Centers for Medicare & Medicaid Services, Medicare Plan Payment Group, Division of Risk Adjustment and Payment Policy.
- Quan, H., V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J. C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali. 2005. "Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data." *Medical Care* 43 (11): 1130–9.
- Schone, E., and R. Brown. 2013. "Risk Adjustment: What is the Current State of the Art, and How Can It Be Improved?" In *Mathematica Policy Research: Research Synthesis Report No. 25*. Minneapolis, MN: Robert Wood Johnson Foundation.
- Sloan, K. L., M. E. Montez-Rath, A. Spiro 3rd, C. L. Christiansen, S. Loveland, P. Shokeen, L. Herz, S. Eisen, J. N. Breckenridge, and A. K. Rosen. 2006. "Development and Validation of a Psychiatric Case-Mix System." *Medical Care* 44 (6): 568–80.
- Smith, M., R. Saunders, L. Stuckhardt, and J. M. McGinnis. 2012. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Washington, DC: The National Academies Press, Institute of Medicine.
- Trivedi, A. N., R. C. Grebla, L. Jiang, J. Yoon, V. Mor, and K. W. Kizer. 2012. "Duplicate Federal Payments for Dual Enrollees in Medicare Advantage Plans and the Veterans Affairs Health Care System." *Journal of the American Medical Association* 308 (1): 67–72.
- VA Pharmacy Benefits Management Service. n.d. [accessed on March 27, 2015]. Available at <http://www.pbm.va.gov/nationalformulary.asp>

- Wagner, T. H., S. Chen, and P. G. Barnett. 2003. "Using Average Cost Methods to Estimate Encounter-Level Costs for Medical-Surgical Stays in the VA." *Medical Care Research and Review* 60 (3 Suppl): 15S–36S.
- Wagner, T. H., G. Sethi, W. Holman, K. Lee, F. G. Bakaeen, A. Upadhyay, E. McFalls, H. G. Tobler, R. F. Kelly, M. D. Crittenden, H. Thai, and S. Goldman. 2011. "Costs and Quality of Life Associated with Radial Artery and Saphenous Vein Cardiac Bypass Surgery: Results from a Veterans Affairs Multisite Trial." *American Journal of Surgery* 202 (5): 532–5.
- Wagner, T. H., E. H. Cowgill, J. Cashy, and M.-L. Shen. 2015. "Risk Adjustment: Guide to the V21 and Nosos Risk Score Programs" [accessed on March 27, 2015]. Available at <http://www.herc.research.va.gov/include/page.asp?id=technical-report-risk-adjustment>
- Wang, L., B. Porter, C. Maynard, C. Bryson, H. Sun, E. Lowy, M. McDonell, K. Frisbee, C. Nielson, and S. D. Fihn. 2012a. "Predicting Risk of Hospitalization or Death among Patients with Heart Failure in the Veterans Health Administration." *American Journal of Cardiology* 110 (9): 1342–9.
- Wang, L., B. Porter, C. Maynard, G. Evans, C. Bryson, H. Sun, I. Gupta, E. Lowy, M. McDonell, K. Frisbee, C. Nielson, F. Kirkland, and S. D. Fihn. 2012b. "Predicting Risk of Hospitalization or Death among Patients Receiving Primary Care in the Veterans Health Administration." *Medical Care* 51 (4): 368–73.
- Yu, W., T. H. Wagner, S. Chen, and P. G. Barnett. 2003. "Average Cost of VA Rehabilitation, Mental Health, and Long-Term Hospital Stays." *Medical Care Research and Review* 60 (3 Suppl): 40S–53S.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Table A1. VA Drug Groups (First Two Digits of VA Drug Class).

Table A2. Model Fit Statistics for the General Random Sample.

Table A3. Hosmer–Lemeshow Deciles for Basic OLS Regressed on Total Cost.

Table A4. Hosmer–Lemeshow Deciles for Basic OLS with Square Root Transformation Regressed on Total Cost.

Table A5. Hosmer–Lemeshow Deciles for Recalibrated OLS with Square Root Transformation Regressed on Total Cost.

Table A6. Recalibrated Risk Scores in VA Samples from 2010.