Predicting High-Cost Privately Insured Patients Based on Self-Reported Health and Utilization Data

Peter J. Cunningham, PhD

The shift to value-based payment strategies and healthcare providers assuming greater financial risk for the care of patients has generated considerable interest in risk-adjustment methods that can identify patients who will incur high healthcare costs.[1,2] Such methods often rely on medical claims and electronic health records (EHRs) to predict future health needs and healthcare costs of patients based on their prior diagnoses, utilization, and costs.[3]

About 13% of privately insured individuals change health plans in a given year and about 9% change their usual source of care,[4] and this appears to have increased since implementation of the Affordable Care Act (ACA).[5] However, claims or medical records data are usually not available to private insurance plans or practitioners when they accept and treat new patients.

Due to the lack of data on health history and prior use, health needs assessments (HNAs) or health risk appraisals (HRAs) are often administered by plans to identify patients' care needs, although it is largely unknown how useful these tools are in predicting which patients are likely to incur high healthcare costs and therefore may require more intensive care management.[6,7] Among other potential limitations, HNAs and HRAs rely entirely on patient self-reporting of health conditions and medical care use, which may be subject to reporting error.

Research based on the nationally representative Medical Expenditure Panel Survey and the Medicare population shows that common measures of self-reported health status—including the single-item self-rated general health measure, the SF-12 measure, and self-reported chronic conditions—generally performed as well as risk scores derived from claims or medical records.[8-11] The results of 2 studies of Medicaid enrollees showed that self-reported information on health status, health behavior, prior utilization, and other measures were also good predictors of future use of high healthcare costs.[12,13]

Whether these results are consistent for individuals with private insurance is unknown. In addition, most studies have not directly compared the predictive power of self-reported health measures

ABSTRACT

OBJECTIVES: To examine how well self-reported data on health, health behaviors, and healthcare utilization by a sample of privately insured patients predict whether they will incur high healthcare costs the following year.

STUDY DESIGN: A 2012 mail survey of autoworkers from Chrysler, Ford, and General Motors, with 3983 survey respondents linked to their health insurance claims data for 2012 and 2013.

METHODS: High healthcare costs are defined as being in the 75th percentile or higher of healthcare expenditures. Models that include combinations of claims-based measures of expenditures and morbidity and self-reported measures of health, health behaviors, and healthcare utilization are compared.

RESULTS: Claims-based measures of healthcare costs and comorbidity for 2012 were strong predictors of whether a patient would incur high healthcare costs in 2013 (C statistic = 0.78). Self-reported measures of chronic conditions, health status, health behaviors, and hospital use are also good predictors of high healthcare costs. However, even the most comprehensive model that included self-reported measures was not as accurate in predicting high healthcare costs (C statistic = 0.73).

CONCLUSIONS: Efficient targeting of high-cost patients is crucial to the success of innovative care delivery models that attempt to lower costs and improve quality of care through more intensive care management of patients. The results of this study show that in the absence of claims data on prior use and expenditures, patient-reported measures of health status and prior healthcare use are reasonable predictors of future healthcare costs for a privately insured population.

Am J Manag Care. 2017;23(7):e215-e222

TAKEAWAY POINTS

Identifying patients who are likely to incur high healthcare costs is a crucial goal of innovative care delivery models. Insurance claims or electronic health records are often used to identify high-cost patients, but sometimes they are unavailable. The results of this study show that:

▶ Self-reported health and healthcare utilization based on survey data can be useful in predicting whether privately insured patients will incur high healthcare costs when claims-based expenditure and health information from prior years is not available.

▶ Among the survey-based measures, questions on inpatient stays and emergency department visits in the previous year were the strongest predictors of incurring high healthcare costs in the following year.

with claims-based measures of previous expenditures for the same population. The objective of this paper was to determine how well patient-reported data on health conditions, overall health status, health behaviors, and experiences with the healthcare system predict high health costs in the future. To our knowledge, this is the first study that assesses the ability of a comprehensive set of self-reported measures of health, health behaviors, and healthcare utilization to predict high healthcare costs for a privately insured population, similar to studies that have been conducted for the Medicaid population.[12,13] Also, because the data include both prior-year claims data on health expenditures and self-reported measures for the same time period, the study compares self-reported measures and claims-based measures of expenditures in predicting high healthcare costs.

METHODS

Data

The data for this study were based on the 2012 Autoworker Health Care Survey, a survey of active and retired hourly wage workers from Chrysler, Ford, and General Motors. The survey was sponsored by the National Institute for Health Care Reform, a nonprofit, nonpartisan organization established by the International Union, United Autoworkers; Chrysler Group, LLC; Ford Motor Company; and General Motors. The total survey sample included 8624 hourly wage workers, retirees younger than 65 years (ie, not eligible for Medicare), and their spouses. The sample was randomly selected, with some oversampling of active workers so that the proportion of active and retired workers in the sample was about evenly split.

The survey questionnaire was administered by mail. A consent form was included that asked respondents for permission to obtain insurance claims data for themselves and their spouses. Survey respondents were asked to sign and return the consent form to the survey firm if they agreed to allow their insurance claims data to be linked to their survey data. There were 3983 survey respondents who provided consent and had claims data that we were able to link to their survey responses. The **eAppendix Figure** (eAppendices available at **ajmc.com**) describes the process for determining the final sample for this study.

The survey response rate was 64%. Among survey responders, claims data were available for 46%, for a combined response rate of 29.4% for the sample with linked claims data. Survey weights used in this analysis adjusted for survey nonresponse and differences between individuals who provided signed consent to access their claims data and those who did not provide consent. With these statistical adjustments, the full survey sample and the linked survey/claims sample were very similar with respect to age, gender, race/ethnicity, education, income, health status, and prevalence of chronic conditions, as is shown in the **eAppendix Table**.

Dependent Variable

The main dependent variable is a binary indicator of whether the individual had high healthcare costs in 2013 (the year following the survey), defined as being in the 75th percentile or higher with respect to total expenditures for 2013 (about $7000 or higher). Total expenditures include combined costs for hospital care (inpatient, emergency department [ED], outpatient), prescription drugs, office-based physicians, tests, procedures, medical equipment, and other services covered by the health plan.

Independent Variables

Two variables from sample individuals' 2012 claims data were constructed for this analysis, as was a categorical measure reflecting total healthcare expenditures for 2012. Categories include being in the 90th percentile of spending or higher ($16,660 and above), between the 75th and 90th percentiles ($6610 and $16,660), between the 50th and 75th percentiles ($2431 and $6610), and below the 50th percentile ($2431 and under). Diagnoses based on *International Classification of Diseases, Ninth Revision, Clinical Modification* codes were used to construct a Charlson comorbidity index.[14] This is a commonly used measure of health status based on 22 conditions that has been used in prior research to predict mortality and future healthcare costs.

Measures obtained from the survey questionnaire are used to identify all other independent variables in the analysis (see **Table 1** for more detailed definitions of these variables). These include demographic and socioeconomic characteristics (age, gender, race/ethnicity, educational attainment, family income). The survey also asked about a selected number of chronic conditions that respondents reported had been diagnosed by a physician, including hypertension, heart disease, congestive heart failure, diabetes, chronic obstructive pulmonary disease (COPD), arthritis, depression, high cholesterol, and cancer. Also included were distinct measures of both self-rated physical and mental health, as well as work and activity limitations, needing help with personal

care, and requiring special equipment due to health problems.

The survey also asked about health behaviors or lifestyle factors that may be related to healthcare costs, including the amount of physical activity respondents engage in on a weekly basis, whether they are a current smoker, and height and weight (to identify individuals who are obese). Self-reported measures of healthcare utilization are also included, such as having a usual source of care, number of hospital stays, and ED visits.

Analysis

The analysis assessed: 1) how well self-reported information on health, health behaviors, and healthcare use predict being a high-cost patient in 2013 compared with claims-based measures of health expenditures and comorbidity and 2) which self-reported measures appear to be the most important in predicting whether a patient will be considered high-cost in 2013. Logistic regression models for the likelihood of being a high-cost patient in 2013 were estimated using combinations of the independent variables described above. One model included only information on 2012 expenditures and comorbidity (based on claims data), as well as age, gender, and race/ethnicity. These results were compared with a series of logistic regressions that included only survey variables. These regressions sequentially added groups of survey variables to assess their relative contribution to predicting high-cost patients. A final model included all of the 2012 claims-based expenditures and survey measures described above.

For each model, we computed a C statistic, which is a common measure used to quantify how well a model predicts an outcome better than chance.[15] Values range from 0.5 to 1.0, with a value of 0.5 indicating that the model predicts high-cost patients no better than chance and a value of 1.0 indicating that the model perfectly predicts high-cost patients. Models that have a C statistic of 0.7 or higher are considered to be good predictors of the outcome measure.

A pseudo R^2 reflects the proportion of explained variation in the dependent variable

TABLE 1. Characteristics of Autoworkers in 2012 and Percent Who Were High-Cost Patients in 2013

Characteristic	Percent of Autoworkers (SE)	Percent With High Costs in 2013 (SE)
All individuals (n = 3983)	100.0	25.0
Healthcare costs in 2012		
≥90th percentile	10.0 (0.5)	54.9 (2.5)
75th-90th percentiles	15.0 (0.6)	46.8 (2.1)
50th-75th percentiles	25.0 (0.7)	22.9 (1.4)
<50th percentile	50.0 (0.8)	9.2 (0.7)
Age, years		
<40	6.8 (0.4)	16.0 (2.3)
40-49	16.2 (0.5)	22.6 (1.7)
50-55	23.8 (0.7)	22.6 (1.4)
56-64 (ref)	53.2 (0.7)	23.9 (0.9)
Gender		
Male	51.3 (0.8)	22.1 (0.9)
Female (ref)	48.7 (0.8)	23.6 (1.0)
Race/ethnicity		
White, non-Hispanic (ref)	79.5 (0.7)	23.3 (0.7)
Black, non-Hispanic	13.8 (0.6)	19.9 (1.8)
Mixed	2.4 (0.2)	23.3 (4.4)
Other	4.3 (0.3)	22.7 (3.3)
Education		
Less than HS graduate (ref)	5.0 (0.4)	30.5 (3.3)
HS graduate	36.3 (0.8)	21.8 (1.1)
Vocational or technical school	12.2 (0.5)	25.7 (2.0)
Some college	31.5 (0.7)	21.9 (1.2)
College graduate	15.0 (0.6)	22.3 (1.7)
Family income		
<$35,000 (ref)	17.4 (0.6)	23.9 (1.7)
$35,000-$49,000	23.8 (0.6)	20.9 (1.3)
$50,000-$74,999	27.2 (0.7)	25.0 (1.3)
$75,000-$99,000	18.5 (0.6)	22.6 (1.5)
>$100,000	13.2 (0.5)	20.9 (1.7)
Self-reported chronic conditions		
Hypertension	43.6 (0.8)	27.6 (1.1)
Heart disease	10.1 (0.5)	34.0 (2.4)
Congestive heart failure	3.9 (0.3)	33.3 (3.9)
Diabetes	19.2 (0.6)	36.0 (1.8)
COPD	14.2 (0.6)	36.9 (2.1)
Arthritis	36.9 (0.8)	30.6 (1.2)
Depression	23.9 (0.7)	32.5 (1.5)
High cholesterol	48.0 (0.8)	27.0 (1.0)
Cancer	7.7 (0.4)	31.7 (2.7)

(continued)

Characteristic	Percent of Autoworkers (SE)	Percent With High Costs in 2013 (SE)
Number of chronic conditions		
None (ref)	17.5 (0.6)	11.2 (1.2)
1	23.0 (0.7)	15.2 (1.2)
2	23.0 (0.7)	20.9 (1.4)
≥3	36.5 (0.8)	34.4 (1.3)
Self-reported physical health		
Excellent	10.5 (0.5)	12.6 (1.6)
Very good	35.3 (0.8)	15.4 (1.0)
Good	40.1 (0.8)	26.7 (1.1)
Fair or poor	14.1 (0.5)	38.1 (3.1)
Self-reported mental health		
Excellent	31.8 (0.7)	18.5 (1.1)
Very good	39.6 (0.8)	20.8 (1.0)
Good	22.3 (0.7)	28.6 (1.5)
Fair or poor	6.3 (0.4)	40.3 (4.2)
Disabilities		
Work limitations due to health	15.6 (0.6)	38.9 (2.0)
Other activity limitations due to health	13.0 (0.5)	39.3 (2.2)
Needs help or supervision with activities	1.6 (0.2)	46.4 (6.4)
Requires special equipment	4.7 (0.3)	38.1 (3.6)
Amount of physical activity per week		
None	20.6 (0.6)	28.1 (1.6)
1-2 days	31.7 (0.7)	23.8 (1.2)
3-4 days	28.4 (0.7)	18.9 (1.2)
≥5 days	19.2 (0.6)	21.2 (1.5)
Current smoker		
No	76.8 (0.7)	22.0 (0.8)
Yes	23.2 (0.7)	25.7 (1.5)
Body mass index		
Normal/underweight	22.4 (0.7)	18.4 (1.3)
Overweight	35.7 (0.8)	20.7 (1.1)
Obese	42.0 (0.8)	26.7 (1.1)
Has a usual source of care		
No	8.7 (0.5)	15.3 (2.0)
Yes	91.3 (0.5)	23.6 (0.7)
Self-reported hospital stays (mid-2011 through mid-2012)		
None	87.5 (0.5)	20.7 (0.7)
1	8.4 (0.4)	36.0 (2.7)
≥2	4.1 (0.3)	41.7 (3.9)
Self-reported ED visits (mid-2011 through mid-2012)		
0	76.6 (0.7)	20.0 (0.7)
1	16.0 (0.6)	29.4 (1.8)
≥2	7.4 (0.4)	39.0 (2.9)

based on a logistic regression model.[16] Values can range from 0 to 1.0, with 0 indicating that the model explains none of the variation in the dependent variable and 1.0 indicating that the model explains all of the variation in the dependent variable.

A third measure is the discrimination slope, which, for each model, reflects the difference in the average predicted values for individuals with high healthcare costs in 2013 and the average predicted values for those who did not have such costs. For the best performing models, sensitivity, specificity, positive predictive values (PPVs), and negative predictive values (NPVs) were computed.

RESULTS

Characteristics of the Sample and Percent With High Healthcare Costs in 2013

Table 1 shows the weighted characteristics of the sample and the percentage with high healthcare costs in 2013. Autoworkers with high healthcare costs in 2012 also tended to have high costs in 2013. Among those in the 90th percentile of costs for 2012, 54.5% were in the 75th percentile or higher of spending for 2013. Among those who were below the 50th percentile of spending for 2012, only 9.2% had high healthcare costs for 2013.

More than half of autoworkers were between the ages of 56 and 64, with an additional 23.8% between the ages of 50 and 55. Most autoworkers were white or African American, not college graduates, and had annual family incomes less than $100,000. The percentages of high-cost patients were lower for younger workers, African Americans, and those with at least some college.

Autoworkers had high prevalence of self-reported chronic conditions, especially hypertension (43.6%), arthritis (36.9%), depression (23.9%), and high cholesterol (48%). One-third had 3 or more chronic conditions during the survey year. The percentage identified as having high healthcare costs in 2013 was especially high for autoworkers with 3 or more chronic conditions.

TABLE 2. Results of Multivariate Logistic Regression Models for Predicting High-Cost Patients for 2013

Models	C Statistic	Pseudo R^2	Discrimination Slope
1. Includes 2012 healthcare expenditures and comorbidity index from claims data; age, gender, race/ethnicity from self-report	0.78 (0.009)[a]	17.8	0.209
Models based on 2012 survey data only[b]			
2. Includes demographic and socioeconomic measures	0.57 (0.011)[c]	1.1	0.014
3. Adds self-reported chronic conditions	0.70 (0.010)[c,a]	8.4	0.098
4. Adds self-reported health status and disabilities, and health behavior	0.72 (0.01)[c,a]	10.4	0.122
5. Adds self-reported hospital inpatient and ED use in 2012	0.73 (0.01)[c,a]	11.2	0.131
6. Includes all survey and claims-based variables used in models 1-6	0.80 (0.009)[a]	20.0	0.241

ED indicates emergency department.
[a]Difference with model 2 is statistically significant at .01 level.
[b]Variables are added sequentially in models 2-6.
[c]Difference with model 1 is statistically significant at .05 level.
Source: 2012 Autoworker Health Care Survey, linked to health insurance claims for 2012 and 2013.

Despite a high prevalence of chronic conditions, most autoworkers perceived their physical and mental health as excellent, very good, or good. About 1 in 6 (15.6%) reported that they were limited in their ability to work due to health, and 13% reported other health-related activity limitations. High healthcare costs in 2013 were strongly associated with these measures of health status.

About one-fifth of autoworkers reported little or no physical activity in a typical week, 23.2% were smokers, and 42% were classified as obese. The percentage with high costs in 2013 was higher among those who had no physical activity (compared with individuals who were physically active) and were smokers (compared with nonsmokers), and obese persons (compared with the nonobese workers).

Most autoworkers (91.3%) had a usual source of care, 12.5% reported an inpatient stay in 2012, and 7.4% reported 2 or more visits to the ED in 2012. Autoworkers with self-reported inpatient and ED use were much more likely to have high healthcare costs in 2013 compared with those with no self-reported hospital use.

Models Predicting High-Cost Patients

Table 2 compares the performances of different models predicting high-cost patients for 2013. Model 1 includes 2012 expenditures, the comorbidity index, age, gender, and race/ethnicity. This model had a C statistic of 0.78, a pseudo R^2 of 17.8%, and a discrimination slope of .209. Models 2 to 5 include only patient-reported information from the surveys. Model 2 includes patient-reported demographics, education, and income; model 3 adds self-reported chronic conditions; model 4 adds health status and health behavior measures; and model 5 adds self-reported inpatient and ED use. The results for the C statistics, R^2, and discrimination slope are consistent in that they show: 1) adding self-reported health, health behaviors, and utilization (models 3-5) substantially improves predictions of high healthcare costs compared with the model that includes only demographics and socioeconomic status, and

TABLE 3. Detailed Statistics on Accuracy of Predicting High-Cost Patients for Selected Models

	Risk Threshold	
Models From Table 2	**50th percentile**	**75th percentile**
Model 1: Based on claims data for 2012 expenditures, comorbidity, and demographics from survey data		
Sensitivity	42.6	3.4
Specificity	91.3	99.7
Positive predictive value	62.0	79.5
Negative predictive value	82.7	75.6
Model 5: Survey data only (demographics, SES, self-reported chronic conditions, health status, health behaviors, ED visits, and hospital stays)		
Sensitivity	20.9	2.9
Specificity	95.9	99.7
Positive predictive value	62.8	75.7
Negative predictive value	78.6	75.6
Model 6: Includes all claims and survey-based variables		
Sensitivity	41.6	11.1
Specificity	92.7	99.3
Positive predictive value	65.4	84.2
Negative predictive value	82.8	77.2

ED indicates emergency department; SES, socioeconomic status.

2) models based on survey measures have high predictive power (model 5, C statistics = 0.73), but not quite as high as the model that includes claims-based measures and demographic characteristics (model 1, C statistic = 0.78).

Table 3 shows measures of sensitivity, specificity, PPV, and NPV for the 3 best-performing models (models 1, 5, and 6), computed at both the 50th and 75th percent risk thresholds. The most noteworthy finding from these results is that measures of sensitivity

TABLE 4. Probability of Being a High-Cost Patient in 2013 Based on Patient Self-Reports of Health, Health Behaviors, and Healthcare Utilization

Characteristic	Marginal Probabilities (SE)
Age, years	
<40	19.8 (2.7); (P <.001)
40-49	24.0 (1.3)
50-55	23.8 (1.4)
56-64 (ref)	26.4 (1.0)
Gender	
Male	25.6 (1.0)
Female (ref)	24.1 (1.0)
Race/ethnicity	
White, non-Hispanic (ref)	25.3 (0.8)
Black, non-Hispanic	22.1 (1.9)
Mixed	20.9 (4.0)
Other	27.1 (3.5)
Education	
Less than HS graduate (ref)	26.5 (3.2)
HS graduate	23.1 (1.1)
Vocational or technical school	27.1 (2.0)
Some college	25.2 (1.3)
College graduate	26.2 (1.9)
Family income	
<$35,000 (ref)	22.3 (1.8)
$35,000-$49,000	21.6 (1.4)
$50,000-$74,999	27.1 (2.0)
$75,000-$99,000	25.2 (1.3)
>$100,000	26.2 (1.9)
Self-reported chronic conditions	
Hypertension	25.3 (1.3)
Heart disease	28.8 (2.5)
Congestive heart failure	26.4 (0.5)
Diabetes	33.3 (2.1); P <.001
COPD	32.0 (2.3); P <.001
Arthritis	27.7 (1.4); P <.01
Depression	28.5 (1.8); P <.01
High cholesterol	25.8 (1.2)
Cancer	33.1 (3.1); P <.001
Number of chronic conditions	
None (ref)	23.9 (3.4)
1	22.3 (2.2)
2	24.3 (1.5)
≥3	26.7 (2.2)

(continued)

(ie, the percentage of individuals who had high costs in 2013 who were accurately predicted by the models) are low, relative to similar studies of the Medicaid population.[12,13] In fact, few high-cost cases for 2013 were predicted accurately based on the 75th percentile threshold. Models that include claims data (models 1 and 6) perform better on sensitivity compared with the model that includes only survey data (model 5). The model that includes both claims-based and survey variables performed the best on sensitivity at the 75th percentile risk threshold.

Importance of Individual Self-Reported Health Measures in Predicting High-Cost Patients

Table 4 shows the marginal probabilities computed from the model with survey-only variables (model 5). The probability of being a high-cost patient in 2013 was significantly higher among older (relative to younger) individuals, as well as among those with diabetes, COPD, arthritis, depression, and cancer; among those in good, fair, or poor self-reported health (compared with excellent or very good); and among those with work limitations due to health. Those with self-reported hospital stays and ED visits in 2012 were more likely to have high costs in 2013 compared with those with no hospital use.

DISCUSSION

The results of this analysis show that self-reported information on health, health behaviors, and healthcare use commonly obtained through HNAs or HRAs is a reasonably good predictor of future healthcare costs for a privately insured population in the absence of claims or EHRs. Although the models with only self-reported measures do not perform quite as well as models that include claims-based information on spending and morbidity, the results are similar to studies that examined the usefulness of self-reported measures in predicting high-cost patients among Medicaid beneficiaries.[12,13]

Although error in patient-reported data is a longstanding concern, one advantage is that it is less susceptible than claims or EHR data to "upcoding," or the tendency by some plans and providers to aggressively code patient diagnoses to make patients appear sicker in order to maximize payment. A recent study found that risk-adjustment scores based on claims data were significantly higher for enrollees in Medicare Advantage health plans—which are compensated by the federal government and partially based on risk scores—than they would be if the enrollees were in fee-for-service plans.[17] Similar risk-adjustment methods are used in the ACA's federal and state marketplaces: higher rates are paid to plans with sicker enrollees, funded in part through lower rates paid to plans with healthier enrollees. Self-reported health information from patients or plan enrollees is less susceptible to such bias.

TABLE 4. Probability of Being a High-Cost Patient in 2013 Based on Patient Self-Reports of Health, Health Behaviors, and Healthcare Utilization *(continued)*

Characteristic	Marginal Probabilities (SE)
Self-reported physical health	
Excellent or very good (ref)	20.9 (1.2)
Good	27.4 (1.2); *P* <.001
Fair or poor	28.7 (2.2); *P* <.001
Self-reported mental health	
Excellent or very good	24.4 (0.9)
Good	25.7 (1.5)
Fair or poor	26.5 (2.9)
Disabilities	
Work limitations due to health	29.7 (2.6); *P* <.01
Other activity limitations due to health	27.1 (2.6)
Needs help or supervision with activities	22.4 (5.6)
Requires special equipment	27.9 (3.8)
Amount of physical activity per week	
None	25.7 (1.5)
1-2 days	25.2 (1.2)
3-4 days	23.8 (1.4)
≥5 (ref)	24.9 (1.6)
Current smoker	
No (ref)	24.5 (0.8)
Yes	26.0 (1.5)
Body mass index	
Normal/underweight (ref)	24.0 (1.6)
Overweight	24.6 (1.2)
Obese	25.5 (1.1)
Has a usual source of care	
No (ref)	23.3 (2.6)
Yes	25.0 (0.7)
Self-reported hospital stays	
None (ref)	24.0 (0.8)
1	29.8 (2.7); *P* <.05
≥2	32.8 (4.2); *P* <.05
Self-reported ED visits	
0 (ref)	23.5 (0.8)
1	28.1 (1.9); *P* <.05
≥2	31.1 (3.1); *P* <.01

COPD indicates chronic obstructive pulmonary disorder; ED, emergency department; HS, high school; ref, reference; SE, standard error.
[a]Marginal probabilities computed from logistic regression analysis.
Source: 2012 Autoworker Health Care Survey linked to health insurance claims for 2012 and 2013.

Nevertheless, there are no perfect predictors of which patients will be high-cost in the future. Only about half of the study sample who had high healthcare costs in 2012 also had high healthcare costs in 2013, based on the definition used in this study. This may reflect greater variability among the sample for the study compared with other studies, either due to the relatively small sample size (n = 3983) or because the 25% of the study sample with high healthcare costs was more heterogeneous than a similarly defined group for the Medicaid population. Regardless, the best measures in this study used to predict high-cost patients still have a relatively high rate of error.

The success of innovative care delivery models that focus on care management for high-cost patients depends, in part, on whether the additional resources needed for more intensive care management results in greater cost savings in the long run by preventing unnecessary or avoidable utilization. The key to this success is the efficient targeting of patients who will incur high healthcare costs unless diverted into care management programs. If such targeting includes a large number of patients who will not incur high costs even without the intervention, the effectiveness of care management practices in reducing healthcare costs may be greatly diminished. The high scores for specificity in this study (ie, the proportion of non–high-cost cases in 2013 that were accurately identified as such) suggest the models estimated in this study would be relatively successful in preventing costly case management or other specialized services to patients who would not benefit from them. On the other hand, the relatively low sensitivity scores suggest that a large percentage of patients who would potentially benefit from these services may not be selected to receive them and therefore would be at higher risk for incurring higher costs.

Limitations

There are several limitations to this analysis that should be noted. First, the sample is limited to US autoworkers and therefore may not be generalizable to other privately insured populations. Predicting high costs for the autoworker population, which tends to be older and have a high prevalence of chronic conditions, may be quite different than for a younger and healthier population. In addition, the small sample size, compared with those in other studies that examined self-reported measures, may lead to lower precision in the predictive ability of self-reported measures than if a larger sample had been available. Also, the results may differ for specific conditions, which is important because many disease management programs are designed to improve quality of care and lower costs for specific diseases (eg, diabetes) rather than high-risk patients in general.

CONCLUSIONS

Information from health needs assessments or health risk appraisals are increasingly used for a variety of purposes to

improve delivery of care, but little is known as to how effective they could be in targeting privately insured patients who are likely to incur high healthcare costs. The results from this study indicate that self-reported information on health conditions, health status, and healthcare use can be useful in predicting high healthcare costs when prior year claims or medical records are not available. ∎

REFERENCES

1. Cucciare MA, O'Donohue W. Predicting future healthcare costs: how well does risk-adjustment work? *J Health Organ Manag*. 2006;20(2-3):150-162.
2. Ash AS, Zhao Y, Ellis RP, Schlein Kramer M. Finding future high-cost cases: comparing prior cost versus diagnosis-based methods. *Health Serv Res*. 2001;36(6, pt 2):194-206.
3. Levine SH, Adams J, Attaway K, et al. Predicting the financial risks of seriously ill patients. California Health-Care Foundation website. http://www.chcf.org/~/media/MEDIA%20LIBRARY%20Files/PDF/PDF%20P/PDF%20 PredictiveModelingRiskStratification.pdf. Published December 2011. Accessed March 4, 2016.
4. Cunningham PJ. Few Americans switch employer health plans for better quality, lower costs. National Institute for Health Care Reform website. http://nihcr.org/analysis/improving-care-delivery/prevention-improving-health/health-plan-switching/. Published January 2013. Accessed July 25, 2016.
5. Sung I. How is health reform impacting insurance switching patterns? The Health Care Blog website. http://thehealthcareblog.com/blog/2015/07/17/how-is-health-reform-impacting-insurance-switching-patterns/. Published July 17, 2015. Accessed July 23, 2016.
6. Lafata JE, Shay LA, Brown R, Street RL. Office-based tools and primary care visit communication, length, and preventive service delivery. *Health Serv Res*. 2016;51(2):728-745. doi: 10.1111/1475-6773.12348.
7. Leininger L, Avery K. The capacity of self-reported health measures to predict high-need Medicaid enrollees. State Health Access Data Assistance Center website. http://www.shadac.org/sites/default/files/Old_files/shadac/publications/Brief_PredictingHighNeedsEnrollees_Leininger_1.pdf. Published February 2015. Accessed March 4, 2016.
8. Perrin NA, Stiefel M, Mosen DM, Bauck A, Shuster E, Dirks EM. Self-reported health and functional status information improves prediction of inpatient admissions and costs. *Am J Manag Care*. 2011;17(12):e472-e478.
9. Fleishman JA, Cohen JW, Manning WG, Kosinski M. Using the SF-12 health status measure to improve predictions of medical expenditures. *Med Care*. 2006;44(suppl 5):I54-I63.
10. Fleishman JA, Cohen JW. Using information on clinical conditions to predict high-cost patients. *Health Serv Res*. 2010;45(2):532-552. doi: 10.1111/j.1475-6773.2009.01080.x.
11. DeSalvo KB, Jones TM, Peabody J, et al. Health care expenditure prediction with a single item, self-rated health measure. *Med Care*. 2009;47(4):440-447. doi: 10.1097/MLR.0b013e318190b716.
12. Leininger LJ, Friedsam D, Voskuil K, DeLiere T. Predicting high-need cases among new Medicaid enrollees. *Am J Manag Care*. 2014;20(9):e399-e407.
13. Wherry LR, Burns ME, Leininger LJ. Using self-reported health measures to predict high-need cases among Medicaid-eligible adults. *Health Serv Res*. 2014;49(suppl 2):2147-2172. doi: 10.1111/1475-6773.12222.
14. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373-383.
15. Pencina MJ, D'Agostino RB Sr. Evaluating discrimination of risk prediction models: the C statistic. *JAMA*. 2015;314(10):1063-1064. doi: 10.1001/jama.2015.11082.
16. Cox DR, Snell EJ. *The Analysis of Binary Data*. 2nd ed. London: Chapman and Hall; 1989.
17. Geruso M, Layton T. Upcoding: evidence from Medicare on squishy risk adjustment. National Bureau of Economic Research website. http://www.nber.org/papers/w21222.pdf. Published May 2015. Accessed March 21, 2016.

Full text and PDF at **www.ajmc.com**

eAppendix

eAppendix Table. Characteristics of Autoworkers in 2012 and Percent Who Were High-Cost Patients in 2013

	Full sample who completed survey (n = 8624)	Sample linked to claims data (n = 3983)
All persons	100.0	100.0
Age		
<40	7.6 (0.3)	6.8 (0.4)
40-49	16.3 (0.3)	16.2 (0.5)
50-55	22.8 (0.5)	23.8 (0.7)
56-64 (ref)	53.3 (0.5)	53.2. (0.7)
Gender		
Male	51.0 (0.6)	51.3 (0.8)
Female (ref)	49.0 (0.6)	48.7 (0.8)
Race/ethnicity		
White/non-Hispanic (ref)	76.9 (0.5)	79.5 (0.7)
Black, non-Hispanic	16.5 (0.4)	13.8 (0.6)
Mixed	2.5 (0.2)	2.4 (0.2)
Other	4.1 (0.2)	4.3 (0.3)
Education		
Less than HS graduate (ref)	5.1 (0.3)	5.0 (0.4)
HS graduate	36.2 (0.5)	36.3 (0.8)
Vocational or technical school	11.7 (0.4)	12.2 (0.5)
Some college	31.4 (0.5)	31.5 (0.7)
College graduate	15.5 (0.4)	15.0 (0.6)
Family income		
<$35,000 (ref)	19.6 (0.4)	17.4 (0.6)
$35,000 - $49,000	22.5 (0.5)	23.8 (0.6)
$50,000 - $74,999	26.8 (0.5)	27.2 (0.7)
$75,000 - $99,000	17.8 (0.4)	18.5 (0.6)
> $100,000	13.3 (0.3)	13.2 (0.5)
Number of chronic conditions		
None (ref)	17.8 (0.4)	17.5 (0.6)
1	21.9 (0.5)	23.0 (0.7)
2	22.4 (0.5)	23.0 (0.7)
3 or more	37.8 (0.5)	36.5 (0.8)
Self-reported physical health		
Excellent	9.6 (0.3)	10.5 (0.5)

	Full sample who completed survey (n = 8624)	Sample linked to claims data (n = 3983)
Very good	34.3 (0.5)	35.3 (0.8)
Good	39.0 (0.5)	40.1 (0.8)
Fair or poor	17.1 (0.4)	14.1 (0.5)
Body mass index		
Normal/underweight	22.2 (0.5)	22.4 (0.7)
Overweight	36.4 (0.5)	35.7 (0.8)
Obese	41.4 (0.6)	42.0 (0.8)

HS indicates high school.

Source: 2012 Autoworker Health Care Survey, linked to health insurance claims for 2012 and 2013.

eAppendix Figure. Identifying the Study Sample

```
                    ┌─────────────────────────┐
                    │  13,500 autoworkers and │
                    │    spouses randomly     │
                    │   selected for survey   │
                    │         sample          │
                    └────────────┬────────────┘
                                 │
                                 ▼
                    ┌─────────────────────────┐
              ┌─────│    8624 completed and   │──────────┐
              │     │    returned survey      │          │
              │     │     questionnaires      │          │
              │     └─────────────────────────┘          │
              ▼                                           ▼
  ┌───────────────────────┐              ┌───────────────────────┐
  │  4018 did not provide │              │  4606 provided signed │
  │  signed consent to    │              │  consent to link to   │
  │  link claims and were │              │  claims data          │
  │  excluded from the    │              │                       │
  │  study                │              └───────────┬───────────┘
  └───────────────────────┘                          │
                                                      ▼
                                          ┌───────────────────────┐
                                          │  3983 linked to claims │
                                          │  data and comprise the │
                                          │  final sample for this │
                                          │  study                 │
                                          └───────────────────────┘
```