Data Analytics to Dissect Healthcare Delivery Patterns

Luke Gallione
University of Colorado Denver
1250 14th St.
Denver, Colorado 80202
Gallione.Luke@gmail.com

Jack Schryver
Oak Ridge National Laboratory
1 Bethel Valley Rd
Oak Ridge, TN 37831
schryverjc@ornl.gov

Arjun Shankar
Oak Ridge National Laboratory
1 Bethel Valley Rd
Oak Ridge, TN 37831
shankarm@ornl.gov

ABSTRACT

We systematically dissect and explain variation in Medicare costs by exploring the correlation structure in a recently released set of Medicare data. A portion of regional variation in Medicare cost occurs due to geographic differences in payment rates due to standard-of-living adjustments for individual services. However, we are concerned with explaining the variance that remains after standardized cost adjustment. By relating the demographic, service utilization, and prevention quality indicator data to the standardized per capita Medicare costs (SPCC), we highlight key service usage patterns which explain the variance in cost. By taking this information into account, it should be possible to develop policy-driving recommendations that encourage more cost-efficient, high-quality medical care. We present a method to identify minimal multiple linear regression models which account for the most variance in Medicare cost. We then focus our attention on a model trained on data from 2007 which explains 95% of this variance with only five predictor variables. Finally, we discuss the results of this method across multiple years of data where we observe a shift of the predictors of SPCC from quantity of services to quality of services. Our contribution is a principled approach to deconstruct a vital real-world data set that can better inform policy decisions about regional differences in Medicare costs and benefits.

1. INTRODUCTION

The Center for Medicare and Medicaid Services (CMS) has released data which contains aggregated demographic, Medicare spending, Medicare utilization, and prevention quality indicator data consisting of over three hundred and fifty variables for its three hundred and dix different Hospital Referral Regions (HRRs) for four years starting in 2007 [8]. Our research explores the correlation structure underlying this data in an attempt to explain variation in Medicare costs, rather than the variables which overtly contribute the most to cost. The focus on variance allows us to highlight inconsistencies in spending and quality of healthcare delivery. By understanding these inconsistencies and their sources, we hope to inform policy decisions which make Medicare more efficient. To our knowledge, no other prior work has taken a comprehensive and systematic look at the several hundred variable dataset to reduce it to a few relevant indicators.

The next section of the paper (Section 2) presents relevant background information including the definition of an HRR and the various measures of cost under consideration. We also use a simple heatmap of the correlation matrix to discuss the ways in which indicators relate (and may vary) with each other. Section 3 describes the three step variable selection process for a multiple linear regression model. Section 4 discusses three standard linear regression diagnostics on this model. Section 5 is an analysis of the predictive performance of the model in time. Section 6 provides general conclusions and thoughts for future work.

2. BACKGROUND

Thomas Bodenheimer's excellent overview of the rising cost of healthcare in the United States essentially concludes that an aging population has only a small influence on cost growth. Instead, he finds that the spread of innovative technologies and a health system in which providers dominate the market explain the rise in expenditures. These factors are not unrelated as he suggests: "when payers cured prices and quantities of medical services in the early 1990s, hospitals consolidated into systems that could command higher prices and fewer restrictions on quantities of services. Because most facilities for new technologies were located at hospitals, hospital market power enabled these technologies to proliferate [1, 2, 3, 4]." These technologies promise improved quality of care if used appropriately. Therefore we should be hesitant to restrict their usage outright. Fisher and Wennberg et al. conclude that the average baseline health status of Medicare beneficiaries "was similar across regions of differing spending levels, but patients in higher spending regions received approximately 60% more care[6, 7]." The increased utilization was explained primarily by three factors. First, more frequent physician visits, especially in the inpatient setting. Second, more frequent tests and minor (but not major) procedures. Third, increased use of specialists and hospitals. Interestingly, they also found no significant correlation between spending and access to care and between spending and quality of care. Thorpe and Howard conclude that virtually all of the growth in Medicare beneficiaries' health care spending is associated with patients who are under medical management for five or more conditions [9]. Our

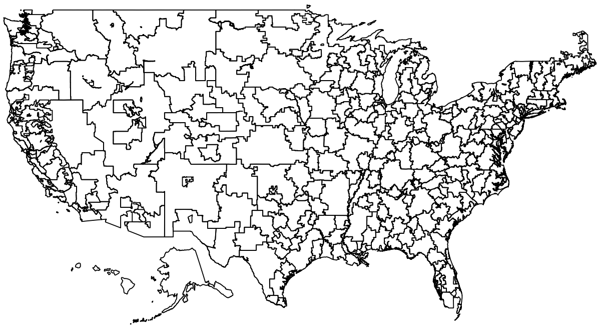

Figure 1: The 306 Hospital Referral Regions. Note that some locations are assigned to the Unknown HRR. This accounts for the holes in the map.

work responds to these above analyses using our encompassing approach that starts with all of the indicator data.

Hospital Referral Region Data

The HRR is a region with locality of medical referrals. HRRs have been defined by determining where patients were referred for major cardiovascular surgery and neurosurgery consultations and procedures. The regions are generally larger than a county, but smaller than a state, and are shown in Figure 1. We analyze a dataset of aggregate data containing the average values at the HRR level. This data set is created from a larger data set containing data at the individual level. This aggregate data does not contain any Personally Identifiable Information (PII) and has been made available to the public by CMS.

The data types are demographic, Medicare spending, Medicare utilization, and prevention quality indicator data.

As an initial exploration into and description of this data set, consider the heatmap in Figure 2, which shows the relative level of correlation between 98 variables in 2007. The data are left in the order given by CMS: demographic, Medicare spending, Medicare service utilization, and prevention quality indicator data. The color red corresponds to complete positive correlation as can be seen by inspecting the main diagonal, while a color of white corresponds to complete negative correlation as can be seen by inspecting the variables "Percent.Female" and "Percent.Male". Certain variables have a strong positive correlation with their neighbor. This is no coincidence. Many variables encode similar information. For example, the variables "Percent of Beneficiaries using Post Acute Care Home Health services," "Post Acute Care Home Health Covered Stays," and "Post Acute Care Home Health Covered Days" can be seen to correlate strongly with each other by observing a three by three red square on the diagonal of Figure 2. The other two by two and three by three red squares on the diagonal indicate other variables which contain similar information to each other.

The Medicare spending in a given HRR is specified by the value of the "Actual Cost" variable, but the CMS data set contains other measures of spending, including "Standardized Cost," "Standardized Risk Adjusted Cost" and per capita versions of all three. The per capita values are the most illu-

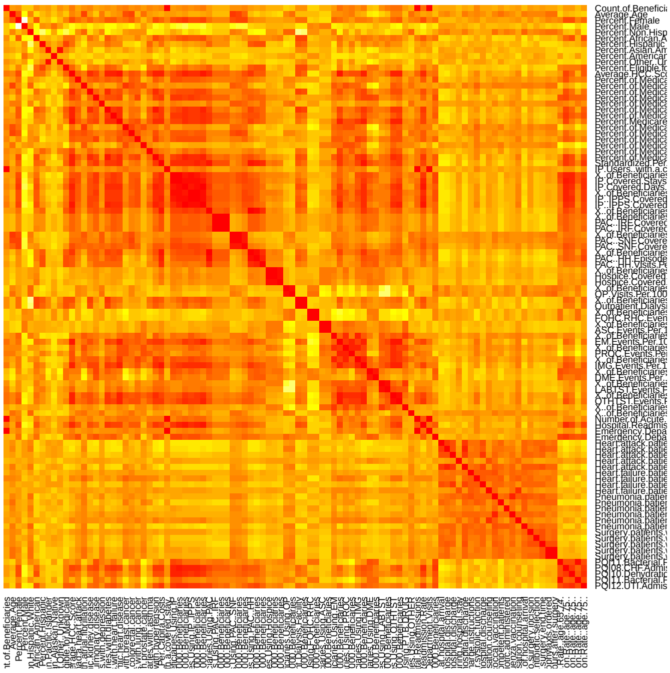

Figure 2: A heatmap showing the relative level of correlation between 98 variables in 2007. The data are left in the order given by CMS: demographic, Medicare spending, Medicare service utilization, and prevention quality indicator data. Red, yellow, and white correspond to correlation coefficient values of 1, 0 and -1, respectively.

minating since they do not depend on the number of Medicare beneficiaries in a given HRR. The standardization is performed in order to remove geographic differences in payment rates for individual services as a source of variation. This allows the utilization of services and their costs to be compared. For these reasons this investigation attempts to optimize a multiple linear regression on the "Standardized Per Capita Costs" (SPCC) variable.

3. MULTIPLE LINEAR REGRESSION VARIABLE SELECTION

The variable identification and parsimonious selection process involves three distinct phases: a pre-selection phase eliminates variables from the data set based on irrelevance, redundancy, and missing-data-driven simplifications, the second phase further winnows the list of variables by use of a recursive algorithm, and the third and final phase is an exhaustive comparison of all linear models producible from the remaining variables. We focus our attention on the results of this method applied to the data from 2007, and investigate the applicability of the method and the insights it produces on the 2008-2010 data sets.

3.1 Pre-Selection

The initial data set consists of 359 variables for 306 different HRRs; however, many of these variables are redundant or irrelevant to our investigation. To begin, only cost variables which contained standardized costs were included since we are trying to predict SPCC solely from demographic, service utilization, and quality indicator variables. This left a remainder of 164 variables. We eliminated variables with count data, since we are concerned with per capita usage rather than totals. We also eliminated Average HCC Score Expressed as a Ratio to the National Average as it is a constant multiple of another variable, and is thus redundant in the context of linear regression. This brought the variable count down to 128. A final step eliminated all variables with missing values in any of the four years. This step is necessary, but may introduce bias. The final step succeeded in bringing the variable count down to 98. (Indeed these are the same 98 variables in Figure 2. The only Medicare spending variable in the heatmap is SPCC.)

3.2 Backwards Algorithm

After pre-selection a simple multivariate regression was performed using SPCC as the desired output and all 98 variables as predictors. Then, the full model's Akaike information criterion (AIC) was calculated. The AIC is a measure of the relative quality of a statistical model for a given set of data. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. The AIC is defined by $AIC = 2k - ln(L)$ where k is the number of parameters in the model and L is the maximized value of the likelihood function for the model, which is proportional to the goodness of fit. Thus, the AIC deals with the trade-off between the goodness of fit of the model and the complexity of the model.

After the full model's AIC was calculated, the predictor that contributes the least to the AIC was removed. A new model was fit to the remaining variables and the process was repeated until removing a variable no longer lowers the AIC.

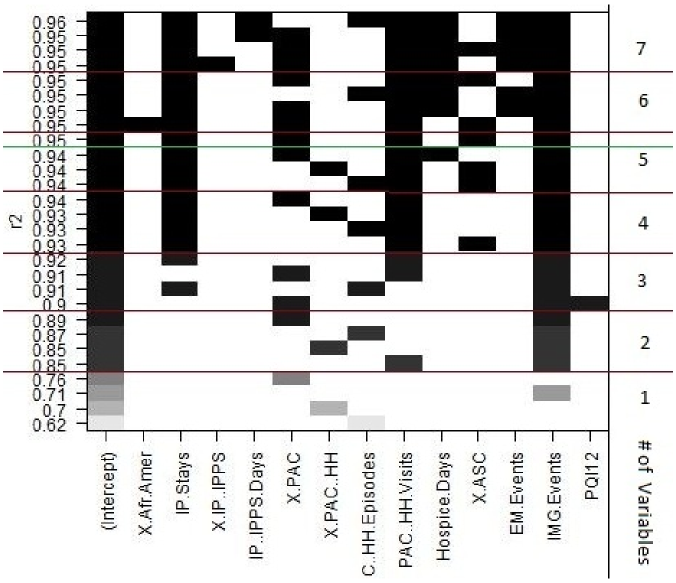

Figure 3: Variables for the top four models for models with up to 7 predictors. The R^2 value for each model is shown on the y-axis, while a black rectangle indicates that a particular x-axis variable is included in the model as a predictor. It is impressive that so few variables can account for nearly all of the variance in cost.

This "backwards" algorithm is in contrast to a "forwards" algorithm which would build the model up from an empty model. When this "backwards" algorithm was used on our 98 variables, the minimal AIC model was found when only 44 variables remained. While an exploration of these 44 variables would be fruitful, the number of variables is unwieldy.

3.3 Exhaustive Comparison

The final step of variable selection compared the models whose predictors correspond to every subset of our remaining variables up to a maximum subset size of 7 predictors. The predictors for the top four models by R^2 value for each variable subset size are shown in Figure 3.

Note that the variables "IP.Stays" (In-Patient Covered Stays per 1000 Beneficiaries), "PAC..HH.Visits" (Post Accute Care Home Health Visits per 1000 Beneficiaries), and "IMG.Events" (Imaging Events per 1000 Beneficiaries) appear in the top 4 models for all models with at least 4 predictors. This is a strong indication that these variables consistently explain different aspects of the variance in SPCC. In Section 5 we shall focus on the variables that appear in the top models, rather than focusing on a single model. We believe that this approach offers deeper insight into our data's correlation structure while also avoiding some of the pitfalls of over-fitting.

Despite this belief, and in order to keep our results tangible, in Section 4 we have chosen to focus our attention on the optimum model with five predictors. The number five was chosen since the value of the coefficient of determination does not increase when the number of variables is increased from five to six; producing a distinct "elbow" on an elbow

Predictors	Coeff.	Stand. C.	t value	P(≥ $\|t\|$)
Intercept	784.3	-1.364e-16	0	1
IP Stays	7.243	.2917	14.065	2e-16
% PAC	9867	.2669	10.524	2e-16
PACHH Visits	.08773	.2886	14.263	2e-16
% ASC	3062	.1093	7.477	8.44e-13
IMG Events	.6860	.3207	15.072	2e-16

Table 1: Standardized coefficients and t-test information for five predictor variables for a multiple linear regression model trained on data from 2007. The variables are: "IP.Stays" (In-Patient Covered Stays per 1000 Beneficiaries), "% PAC" (Percentage of Beneficiaries using Post Acute Care per 1000 Beneficiaries), "PAC..HH.Visits" (Post Accute Care Home Health Visits per 1000 Beneficiaries), "% ASC" (Percentage of Beneficiaries using Ambulatory Surgery Centers per 1000 Beneficiaries), and "IMG.Events" (Imaging Events per 1000 Beneficiaries)

plot of number of variables vs. R^2 value. The five predictors involved in the model are "In-Patient Covered Stays per 1000 Beneficiaries," "Percent of Beneficiaries using Post Acute Care," "Post Acute Care Home Health Visits per 1000 Beneficiaries," "Percent of Beneficiaries using Ambulatory Surgical Centers," and "Imaging Events per 1000 Beneficiaries." Note that even though demographic and prevention quality indicator variables were included in the initial data, our method primarily selects service utilization variables. The predictors are respectively abbreviated as "IP Stays," "% PAC," "PACHH Visits," "% ASC," and "IMG Events" for the readability of Figures 3 and 4.

4. MULTIPLE LINEAR REGRESSION DIAGNOSTICS

4.1 Standardized Coefficients

The actual values of the regression coefficients are not very informative. The standardized coefficients are the estimates resulting from an analysis carried out on independent variables that have been standardized so that their variances are unity. Therefore, the standardized coefficients represent how many standard deviations SPCC will change for every standard deviation increase in the predictor variable. This allows the relative amount of variance in SPCC explained by each predictor to be compared, regardless of the units of the predictors. One must be careful to realize that an increase by one standard deviation has no reason to be "equivalent" to a similar change in another predictor, but conversely, that this method at least allows a baseline comparison to be made. A quick glance at Table 1, shows that four of the predictors, with standardized coefficients of about 0.3, explain roughly the same amount of variance. In contrast, when the value of the fifth predictor, "% ASC," is increased by its standard deviation, with all other variables held fixed at their average, the value of SPCC will only change by about 0.1 standard deviations. It is interesting to note that of the five included predictors in this model, "Percent of Beneficiaries using Ambulatory Surgical Centers" was the least justifiable according to the variable selection process. In general though, these differences are negligible and it is sufficient to say that the five variables explain the same order of magnitude of variance in SPCC.

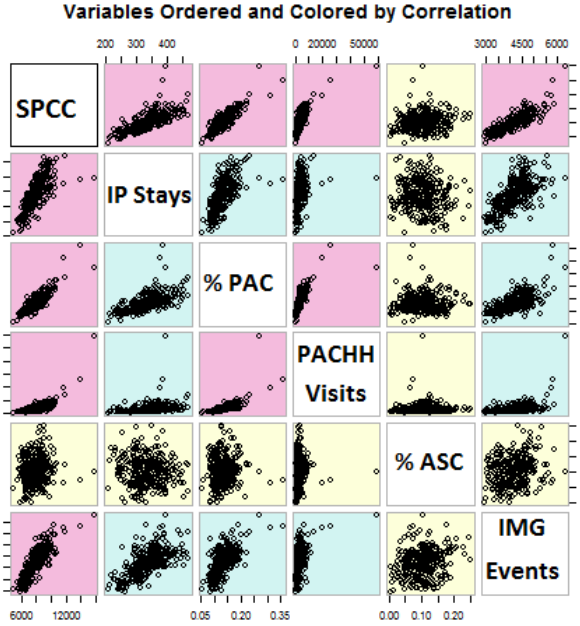

Figure 4: Scatterplot matrix of SPCC and the five predictor variables from 2007. The plotted variables in the pink, blue, and yellow panels have corresponding correlation coefficient values greater than .7, between .3 and .7, and less than .3, respectively. The correlation matrix is symmetric.

4.2 T-Tests

The AIC does not provide a test of a model in the sense of testing a null hypothesis; i.e., AIC can tell nothing about the quality of the model in an absolute sense. If all the candidate models fit poorly, AIC will not give any warning of that. Therefore we consider the standard student's T-test under the null hypothesis that the value of the coefficient is 0, that is, that the predictor and SPCC are not linearly related.

The second to last column of Table 1 contains the t-values for each predictor under this null hypothesis. The final column displays the probability that the measurement of all potential t-values (which follow a t-distribution with 300 degrees of freedom and mean 0) is greater than the observed t-value given that the null hypothesis is true. This measure is commonly referred to as a p-value. The estimated coefficients for four of the predictors are equally improbable under the null hypothesis. The probability value associated with "% ASC" is only slightly greater. These differences are negligible and since all five p-values are small it is sufficient to reject the null hypothesis that the value of the regression coefficient for each of the five predictors is 0. Thus the correlation between the five predictors and SPCC is statistically significant.

4.3 Scatterplot Analysis

The scatterplot matrix in Figure 3 compares the values of SPCC and the five predictor variables from 2007. Four of the variables have a strong linear correlation with cost. The fifth variable, "% ASC" has almost no correlation with SPCC. Recall that this model has an R^2 value (a measure of variance explained by the model) of .95. It is impressive that five variables can account for nearly all of the variance in cost.

5. MODEL VALIDATION AND TEMPORAL ANALYSIS

We now turn our attention to the results of this method across multiple years of data. First, the data was purged of all cost and total count information. As before, we also eliminated all variables with missing values in any of the four years. Then we applied the backwards algorithm to the remaining 98 variables. Between 40 and 50 variables remained, depending on year. Finally, the exhaustive comparison step was applied. We once again decided to look at models with 7 or fewer predictors and only consider the four models within each predictor subset size with the highest R^2 value.

As mentioned above, it is somewhat misleading to focus on only the "top" model. Instead, it is wiser to analyze the top models as a group. Rather than displaying the graphic in Figure 3 for the remaining three years, we have summarized the information in Table 2. Note that the 2007 column matches the totals found in Figure 3.

A few significant trends emerges from this table. First, there is a shift from "Post Acute Care Home Health Visits" to "Post Acute Care Home Health Episodes." Second, there is a shift from "In-Patient Covered Stays" to "percent of Beneficiaries using In Patient services." Since there can be multiple visits within an episode, and since a beneficiary who uses IP services can have multiple stays, both of these trends may suggest an evolution of the predictors from quantity of services to quality of services. To a lesser extent this trend is present in the "Acute Inpatient Prospective Payment System" (IPPS) variables as well. The prevalence of IP services as top predictors for SPCC supports the findings of Fisher and Wennberg et. all who claim that utilization in higher spending regions is partially "explained by more frequent physician visits, especially in the inpatient setting [6, 7].

However, we also notice an increase in the predictive power of Evaluation and Management Events, Skilled Nursing Facility Covered Days, and Hospice Covered Days. The studies of Bodenheimer and of Fisher and Wennberg et. all claimed that hospital market power and increased utilization of hospitals and specialists were primary cost drivers [1, 2, 3, 4, 6, 7]. Our results do not support their findings since skilled nursing facilities and hospices are by definition not hospitals, but rather post-accute care facilities.

The number of Imaging Events is consistently a top predictor of SPCC for all four years of data. This corroborates the findings of Bodenheimer who claims that the spread of technology is a primary cost driver [1, 2, 3, 4]. This also corroborates the findings of Fisher and Wennberg et. all who claim that utilization in higher spending regions is partially explained by tests and technology events, and who further found that the "quality of care in higher spending regions was no better on most measures and was worse for several preventive care measures [6, 7]." We have observed that Prevention Quality Indicator data is not a top predictor of SPCC, which supports their claim.

Average SPCC	\$8,216	\$8,635	\$9,044	\$9,236
% growth of SPCC	N.A.	1.051	1.047	1.021

Table 3: Average Standardized Per Capita Costs for 2007-2010. National SPCC as proportion of previous year's SPCC. While the result cannot be statistically significant with a sample size of only 3, note that 2010 has the least increase in SPCC.

Table 3 shows Average Standardized Per Capita Costs for 2007-2010 and National SPCC as percentage of previous year's SPCC. While the result cannot be statistically significant with a sample size of only 3, note that growth rates appear to decline, and that 2010 has the least increase in SPCC.

6. DISCUSSION AND CONCLUSION

Medicare provides legitimate services to millions of Americans, and any efforts to reduce cost should always be coupled with an effort to increase the quality of care. If one's motivation is to blindly decrease overall Medicare spending without concern for the quality of care, then one is concerned with the variables which contribute the most to cost; the largest slices of pie on a pie chart. But if the motivation is to increase the efficiency of Medicare by removing what are deemed to be spending inconsistencies, then we are more concerned with the variation in cost than the values of cost. Recall that even though demographic and prevention quality indicator variables were included in the initial data, our method selects service utilization variables as the primary explanation of variance in SPCC.

Our analysis is built on multiple linear regression modeling, which always comes with the caveat that correlation does not imply causation. Nevertheless, we believe that it is important to understand the correlation structure of Medicare service utilization as it relates to standardized per capita costs (SPCC). For future work, we are investigating a structural equation model for SPCC. Another caveat is that our linear modeling neglects non-linear relationships. We would also like to find a better way to deal with variables with missing values. While not SPCC related, Cooper et. all found steep, curvilinear relationships between lower income and both increased hospital utilization and increasing percentages of individuals reporting disabilities [5]. Therefore we suspect that there may be important non-linear relationships to SPCC in our dataset.

Our results support previous findings that the spread of technology is a primary cost driver in the U.S. While in patient treatment is still an important cost driver, we have also observed a shift in predictors of SPCC towards evaluation and management events and post acute care facilities in the form of skilled nursing facilities and hospice treatment.

Perhaps the most interesting result is the shift of the predictors from quantity of services to quality of services. This

Predictors of Standardized Per Capita Costs	2007	2008	2009	2010
Percent.African.American	1	1	0	0
Average.HCC.Score	0	0	1	1
Percent.Medicare.Beneficiaries.with.ischemic.heart.disease	0	0	1	1
Percent.of.Beneficiaries.Using.IP	0	2	0	12
IP.Covered.Stays.Per.1000.Beneficiaries	18	17	19	9
IP.Covered.Days.Per.1000.Beneficiaries	0	0	1	0
Percent.of.Beneficiaries.Using.IP..IPPS	1	0	0	9
IP..IPPS.Covered.Stays.Per.1000.Beneficiaries	0	2	7	0
IP..IPPS.Covered.Days.Per.1000.Beneficiaries	2	2	0	0
Percent.of.Beneficiaries.Using.PAC	13	15	0	0
PAC..SNF.Covered.Days.Per.1000.Beneficiaries	0	0	0	8
Percent.of.Beneficiaries.Using.PAC..HH	4	4	0	4
PAC..HH.Episodes.Per.1000.Beneficiaries	6	5	15	20
PAC..HH.Visits.Per.1000.Beneficiaries	19	19	22	1
Percent.of.Beneficiaries.Using.Hospice	0	0	4	0
Hospice.Covered.Stays.Per.1000.Beneficiaries	0	0	7	0
Hospice.Covered.Days.Per.1000.Beneficiaries	8	10	4	14
Percent.of.Beneficiaries.Using.ASC	7	2	0	3
ASC.Events.Per.1000.Beneficiaries	0	0	1	0
EM.Events.Per.1000.Beneficiaries	6	6	9	13
IMG.Events.Per.1000.Beneficiaries	25	25	20	17
PQI12.UTI.Admission.Rate..age.75..	1	1	0	0

Table 2: Total number of models each variable is present in for the top four models for variable subset sizes 1 through 7 for four years of data. Observe that there are only three demographic variables and only one prevention quality indicator variable in this table, while the rest are service utilization variables. Note that the 2007 column matches the totals found in Figure 3. IP:In Patient. IPPS: Accute In Patient Prospective Payment System. PAC: Post Accute Care. SNF: Skilled Nursing Facility. HH: Home Health. ASC: Ambulatory Surgery Center. EM: Evaluation and Management. IMG: Imaging. UTI: Urinary Tract Infections.

coupled with the results in Table 3 suggest that the industry's attempt to switch from fee-for-service to pay-for-performance may be working. If this trend continues, then we expect to see a further decline in SPCC as a percentage of the previous year's SPCC.

7. REFERENCES

[1] T. Bodenheimer. High and rising health care costs. part 1: seeking an explanation. *Annals of Internal Medicine*, 142(10), 2005.

[2] T. Bodenheimer. High and rising health care costs. part 2: technologic innovation. *Annals of Internal Medicine*, 142(11):932–937, 2005.

[3] T. Bodenheimer. High and rising health care costs. part 3: the role of health care providers. *Annals of Internal Medicine*, 143(1):26–31, 2005.

[4] T. Bodenheimer and A. Fernandez. High and rising health care costs. part 4: can costs be controlled while preserving quality? *Annals of Internal Medicine*, 142(10):847–854, 2005.

[5] R. Cooper, M. Cooper, E. L. McGinley, X. Fan, and J. T. Rosenthal. Poverty, wealth, and health care utilization: A geographic assessment. *Journal of Urban Health*, 89(5):828–847, 2012.

[6] E. Fisher, D. E. Wennberg, T. A. Stukel, D. J. Gottlieb, F. L. Lucas, and . L. Pinder. The implications of regional variations in medicare spending. part 1: The content, quality, and accessibility of care. *Annals of Internal Medicine*, 138(4):273–287, 2003.

[7] E. Fisher, D. E. Wennberg, T. A. Stukel, D. J. Gottlieb, F. L. Lucas, and . L. Pinder. The implications of regional variations in medicare spending. part 2: Health outcomes and satisfaction with care. *Annals of Internal Medicine*, 138(4):288–298, 2003.

[8] C. for Medicare & Medicaid Services. Research, statistics, data & systems, June 2014.

[9] K. Thorpe and D. Howard. The rise in spending among medicare beneficiaries: the role of chronic disease prevalence and changes in treatment intensity. *Health Affairs*, 25(5):w378–w388, 2006.