OPERATIONS RESEARCH
Vol. 56, No. 6, November–December 2008, pp. 1382–1392
ISSN 0030-364X | EISSN 1526-5463 | 08 | 5606 | 1382

DOI 10.1287/opre.1080.0619

Algorithmic Prediction of Health-Care Costs

Dimitris Bertsimas
Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, dbertsim@mit.edu

Margrét V. Bjarnadóttir
Stanford Graduate School of Business, Stanford, California 94305, margret@stanford.edu

Michael A. Kane
Medical Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, mkane@med.mit.edu

J. Christian Kryder, Rudra Pandey
D2Hawkeye, Waltham, Massachusetts 02453
{ckryder@d2hawkeye.com, rpandey@d2hawkeye.com}

Santosh Vempala
ARC ThinkTank, Georgia Institute of Technology, Atlanta, Georgia 30332, vempala@cc.gatech.edu

Grant Wang
Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, gjw@alum.mit.edu

The rising cost of health care is one of the world's most important problems. Accordingly, predicting such costs with accuracy is a significant first step in addressing this problem. Since the 1980s, there has been research on the predictive modeling of medical costs based on (health insurance) claims data using heuristic rules and regression methods. These methods, however, have not been appropriately validated using populations that the methods have not seen. We utilize modern data-mining methods, specifically classification trees and clustering algorithms, along with claims data from over 800,000 insured individuals over three years, to provide rigorously validated predictions of health-care costs in the third year, based on medical and cost data from the first two years. We quantify the accuracy of our predictions using unseen (out-of-sample) data from over 200,000 members. The key findings are: (a) our data-mining methods provide accurate predictions of medical costs and represent a powerful tool for prediction of health-care costs, (b) the pattern of past cost data is a strong predictor of future costs, and (c) medical information only contributes to accurate prediction of medical costs of high-cost members.

Subject classifications: health care: cost predictions; prediction algorithms; claims data.
Area of review: Special Issue on Operations Research in Health Care.
History: Received January 2007; revisions received August 2007, May 2008, July 2008; accepted July 2008.

1. Introduction

The value of (health insurance) claims data in medical research has often been questioned (Jolins et al. 1993, Dans 1993) because these databases are designed for financial reasons and not for clinical purposes. Nevertheless, claims data has been shown to be useful in many settings and is increasingly used for medical research. Examples include researching differences in the outcomes of adherence to medication (Pladevall 2004), identification of in-hospital complications (Lawthers et al. 2000), length of episodes (Mehta et al. 1999), and medical outcomes (Wennberg et al. 1987). Statistical methods generally used when working with medical data are nicely summarized in Jones (2000), and other publications addressing issues working with health-care cost data include Zhou et al. (1997) and Manning and Mullahy (2001).

The predictive power of claims data became a topic of research in the 1980s (Zhao et al. 2005) and numerous studies have since established the predictive power of administrative data on health-care costs (Ash et al. 2000, Zhao et al. 2001, Farley et al. 2006, Zhao et al. 2005). Van de Ven and Ellis (2000) provides an insightful overview of the developments in risk-based predictive modeling prior to 2000. Cumming et al. (2002) presents a comparison of different predictive models developed in the insurance industry for both risk assessment and population health-care cost prediction. The models compared used both diagnosis and prescription data, and the study further validated the predictive power of claims data. Earlier researchers concentrated on using classical regression models (Zhao et al. 2005, Ash et al. 2000, Zhao et al. 2001, Powers et al. 2005) when predicting total health-care costs, or logistic regression models (LaVange et al. 1986, Roblin et al. 1999) to identify high-risk members. Often these regression models are combined with heuristic classification rules. There has also been significant work in creating comorbidity[1]

scores from administrative data as a method to account for comorbidity differences of comparative populations in medical research (Klabunde et al. 2002), to design fair reimbursement plans (Van de Ven and Ellis 2000, Dunn et al. 2002), and as a basis for predictive modeling of health-care costs (Ash et al. 2000, Farley et al. 2006, Chang and Lai 2005). Numerous studies that predict health-care costs, based on data other than claims data, are available; examples include Fleishman et al. (2006) and Pietz et al. (2004).

In our view, the best way to express the predictability of a method is to perform out-of-sample experiments (that is, use data that the method has not seen) using different performance measures. To the best of our knowledge, the majority of earlier regression studies do not report on the predictability of the method in an out-of-sample experiment, with a few exceptions (Powers et al. 2005, Dove et al. 2003). Traditionally (Cumming et al. 2002), R^2 or adjusted R^2 have been the measures used to evaluate predictive models, but there are some serious drawbacks to their use, which in our opinion makes it unsuitable for a study like the one presented in this paper. The R^2 measure is a relative, not an absolute, measure of fit. It measures the ratio of the improvement of predictability (as measured with the sum of squares of the residuals) of a regression line compared with a constant prediction (see, for example, Bertsimas and Freund 2005). In particular, comparisons based on R^2 can be made when different regression models on the same data set are being compared, but it is not very meaningful to base comparisons with other methods such as the methods we utilize in this paper. Depending on the purpose of the cost prediction (medical intervention, contract pricing, etc.), different error measures may be more appropriate and better suited than R^2. We therefore define new error measures that better describe the prediction accuracy in a variety of ways.

Our objectives in this paper are to utilize modern data-mining methods, specifically, classification trees and clustering algorithms, and claims data from more than 800,000 members over three years to provide predictions of health-care costs in the third year by applying data-mining methods to medical and cost data from the first two years. We quantify the accuracy of our predictions by applying the models to a test sample of more than 200,000 members. The key insights obtained are: (a) our data-mining methods provide accurate predictions of health-care costs and represent a powerful tool for prediction, (b) the patterns of past cost data are strong predictors of future costs, and (c) medical information adds to prediction accuracy when used in the clustering algorithm, whereas with classification trees, cost information alone results in similar error measures.

The rest of this paper is structured as follows: In §2, we describe the data and define the performance measures we consider, and in §3 we present the two principal methods we use: classification trees and clustering algorithms. In §4, we report on the performance of classification trees and clustering, respectively, in forecasting health-care costs; and in §5, we briefly discuss our conclusions and future research directions.

2. The Data and Error Measures

This study uses health-care data generated when hospitals and other health-care providers send claims to third-party payers to receive reimbursement for their services. The study period is from 8/1/2004–7/31/2007, split up into a 24-month observation period from 8/1/2004–7/31/2006 and a 12-month result period from 8/1/2006–7/31/2007. We build our models using information from the observation period to predict outcomes in the result period.

Our data set includes the medical claims data for 838,242 individuals from a commercially insured population, from 2,866 employers and employer groups across the country. The data set includes both medical and pharmaceutical claims, as well as information on the period an individual (and his or her family) was covered by the insurance policy. The data also contain basic demographic information such as age and gender. All members have eligibility starting no later than 8/1/2005 and ending no sooner than 8/1/2006, and all employers had continuous coverage starting no later than 8/1/2005 and ending no sooner than 8/1/2007. This ensures that every employee (and his family) has at least 12 months of data in the observation period and that big populations do not drop out during the result period as a result of change in an employer's insurance carrier. Out of the 838,242 members, 730,918 have eligibility stretching beyond the result period. The difference, just over 108,000 members or 13.8% of the population, drop out during the result period. This is most often due to employee turnover, which is expected to be around 15% per year. A smaller portion, around 3,000 members (based on gender and age distribution of the population), do not have full coverage due to death. Our analysis has shown that including the population with partial coverage in the result period improves the error measures, and therefore in the interest of simplicity we build our models using the population with full coverage in the result period and report these results.

We split the data set, by random assignment, into equally sized parts: a learning sample, a validation sample, and a testing sample. The learning sample is used to build our prediction models, whereas the validation sample is used to evaluate the performance of the various models. The test sample was set aside while building and calibrating the models, and only used at the very end of the experiment to report results of the finalized models. We believe that this methodology appropriately validates our conclusions.

2.1. Aggregation of the Claims Data

The claims include diagnosis, procedure, and drug information. The diagnosis data is coded using the ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) codes, (Centers for Medicare &

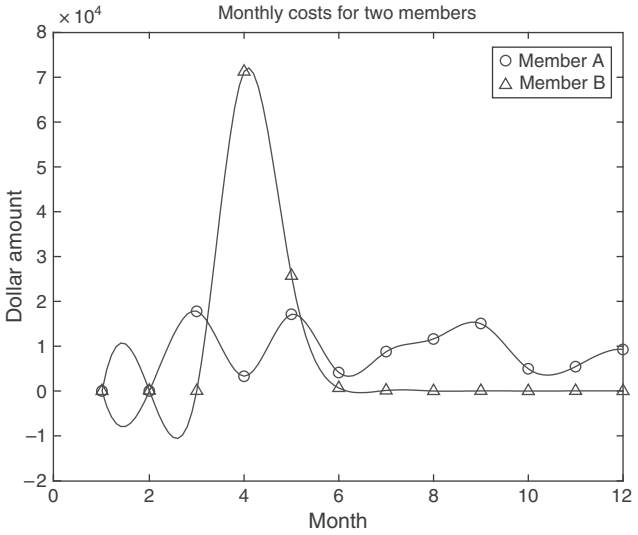
Medicaid Services 2004) the universal codes for medical diagnoses and procedures. The procedures are coded under various coding schemes: ICD9, DRG, Rev Coding, CPT4, and HCPCS—over 22,000 codes altogether. Furthermore, the data include pharmacy claims, that is, it contains information about which, if any, prescription (and some limited over-the-counter) drugs a health plan member is taking, coded in terms of 45,972 drug codes (National Drug Code Directory 2004).

Claims data relies on health-care professionals to encode their diagnoses and procedures in terms of the ICD-9-CM codes. Although coding for medical claims starts with a clinician, it is most often completed and submitted by a separate dedicated billing operator. Because of the inevitable variations in interpretations introduced by these practices, and to reduce the data to a more manageable size, we chose to use coding groups rather than individual codes. We reduced over 13,000 individual diagnoses to 218 diagnosis groups. Medical procedures and drug categories were likewise grouped. Over 22,000 individual procedures are classified into 180 procedure groups, and over 45,000 individual prescription drugs were classified into 336 therapeutic groups. Also included in the analysis are over 700 medically developed quality and risk measures that designate hazardous clinical situations (for example, patients with a pattern of ER care without office visits, diabetics with foot ulcers, etc). We also count the number of diagnoses, procedures, drugs, and risk factors that each member has and include them as additional variables. In summary, the predictive medical variables include: the diagnosis groups, the procedure groups, the drug groups, the risk factors we developed, and their count, for a total of close to 1,500 possible medical variables. We refer the reader to online Appendices A and D for more details. An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

2.2. Cost and Demographic Data

In addition to the medical variables, we utilize 22 cost variables because we believe that cost information gives a global picture of the health of a member. We include age and gender as well. To capture the trajectory of the medical costs (as a proxy of the overall medical condition), we use the monthly costs for the last 12 months in the observation period, the total drug cost and the total medical cost over the entire observation period, as well as the overall cost in the last six months and the last three months of the observation period. Furthermore, to capture the pattern of costs, we developed a new indicator variable that captures whether or not a member's cost pattern exhibits a "spike" pattern, i.e., a sudden increase followed by a sudden decrease in cost. To demonstrate this idea, let us consider Figure 1, which depicts the monthly cost of two members in the last 12 months of the observation period. Although both members have around \$98,000 of paid claims, Member A has constant relatively high medical costs (a typical pattern for a member with a chronic condition), whereas Member B's cost profile has a spike (a typical pattern for a member with an acute condition). The key idea here is that whereas constant high medical costs have a strong tendency to repeat in the future, a cost pattern that exhibits a spike might have a low risk of high future health-care costs: Examples include pregnancy complications, accidents, or acute medical conditions like pneumonia or appendicitis.

Figure 1. 12 months of health-care costs of two members, with overall cost of \$97,500 and \$98,100, respectively.



Notes. A cubic spline curve is fit to the data for easier viewing. The cost profile for Member A has the characteristics of a chronic illness, whereas the characteristics of Member B's profile is acute. The diagnoses behind the most expensive claims of Member A are lymphema and respiratory failure. The reasons behind the highest claims of Member B reflect complications of labor.

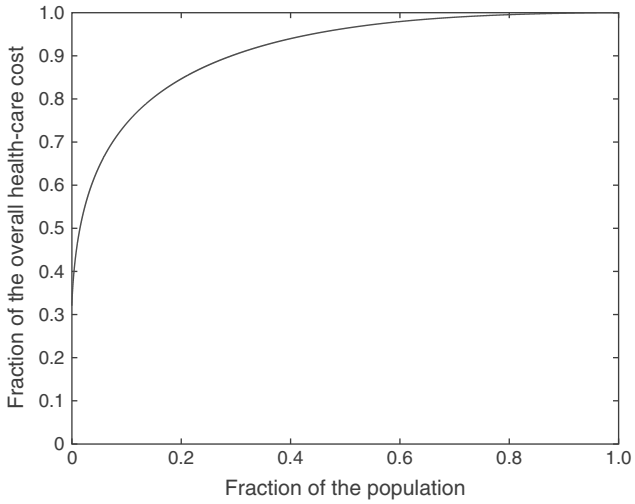
Moreover, we used the following additional four variables: the maximum monthly cost, the number of months with above-average cost, positive trend and negative trend in the last months of the observation period.

Finally, we used gender and age as additional variables. Table 1 summarizes all the variables used in the study, and more details are provided in online Appendix A.

Table 1. Summary of the data elements used.

Variable number	Description
1–218	Diagnosis groups, count of claims with diagnosis codes from each group
219–398	Procedure groups
399–734	Drug groups
735–1,485	Medically defined risk factors
1,486–1,489	Count of members' diagnosis, procedures, drugs, and risk factors
1,490–1,521	Cost variables, including overall medical and pharmacy costs, acute indicator, and monthly costs
1,522–1,523	Gender and age

Figure 2. Cumulative health-care costs of the result period for members in the learning sample.



Notes. On the X-axis is the cumulative percentage of the population and on the Y-axis is the cumulative percentage of the overall health-care costs. For example, we note that 8% of the population (the most expensive members) account for 70% of the overall health-care costs.

2.3. Cost Bucketing

The range of paid amounts for members in the learning sample during the result period is from \$0 up to \$710,000. The population's cumulative cost exhibits known characteristics: 80% of the overall cost of the population originates from only 20% of the most expensive members. Figure 2 shows the cost characteristics of our population. We note that, for our sample, around 8% of the population contributes 70% of the total health-care costs.

To reduce noise in the data and at the same time reduce the effects of extremely expensive members (who can be considered outliers), we partitioned the members' costs into five different bands or cost buckets. We partition in such a way that the sum of all members' costs is approximately the same in each bucket, i.e., the total dollar amount in each bucket is the same (approximately \$117 million per cost bucket). We chose five buckets because it ensures a large enough number of members in the most expensive bucket (we have 1,175 members in the learning sample in bucket five). Table 2 shows the range of each bucket, the percentage, and the number of members of the learning sample that are in each bucket.

Table 2. Cost bucket information.

Bucket	Range	Percentage of the learning sample (%)	Number of members
1	<\$3,200	83.9	204,420
2	\$3,200–\$8,000	9.7	23,606
3	\$8,000–\$18,000	4.2	10,261
4	\$18,000–\$50,000	1.7	4,179
5	>\$50,000	0.5	1,175

Notes. Cost bucket ranges and fraction of the learning sample in each bucket (calculated for the last 12 months of the observation period costs). The sum of members' costs that fall in any one of the buckets is between \$116 and \$119 million.

The knowledge of the predicted bucket of a member is valuable to health care management professionals. Buckets 1 through 5 can be interpreted as representing low, emerging, moderate, high, and very high risk of medical complications. Members predicted to be in buckets 2 and 3 are candidates for wellness programs, members predicted to be in bucket 4 are candidates for disease management programs, whereas those members forecasted to be in the most expensive bucket are candidates for case management programs, the most intense type of patient care program.

2.4. Performance Measures

We measure the performance of our models with three main error measures: the hit ratio, the penalty error, and the absolute prediction error (APE). To be able to compare our results to published studies, we also include R^2 and truncated R^2, and introduce a new similar measure $|R|$. We provide some additional insights into R^2 in §2.4.2 and define the new error measures in §2.4.1.

2.4.1. Definition of Error Measures

The Hit Ratio. We define the hit ratio to be the percentage of the members for whom we forecast the correct cost bucket.

The Penalty Error. The penalty error is motivated by opportunities for medical intervention and is therefore asymmetric. There is a greater penalty for underestimating higher costs, consistent with the greater medical and financial risk in missing these individuals. The penalty of misidentifying an individual as high risk, whose actual costs are low, is smaller than the opposite case, because little harm or cost ensues in this instance. Therefore, the penalties for underestimating a cost bucket are set as twice those for overestimating it. This is motivated by the estimated opportunity loss by doctors. Table 3 shows the penalty table for the five-cost-bucket scheme. We define the penalty error measure to be the average forecast penalty per member of a given sample.

The Absolute Prediction Error. The absolute prediction error is derived from actual health-care costs. We define the absolute prediction error to be the average absolute

Table 3. The penalty table defines the penalty error measure for the five cost buckets.

	Outcome				
Forecast	1	2	3	4	5
1	0	2	4	6	8
2	1	0	2	4	6
3	2	1	0	2	4
4	3	2	1	0	2
5	4	3	2	1	0

Note. A perfect forecast results in an error of zero.

Table 4. Analysis of the sums in the denominator of R^2 and $|R|$.

Bucket	Percentage of the the learning sample	Percentage of overall $\sum((t_i - a)^2)$	Percentage of overall $\sum((t_i - a)^2)$ when truncated	Percentage of overall $\sum(\lvert t_i - m\rvert)$	Percentage of overall $\sum(\lvert t_i - m\rvert)$ when truncated
1	83.9	30.8	36.1	47.0	48.3
2	9.7	12.4	15.9	20.0	20.7
3	4.2	14.0	14.3	14.0	14.2
4	1.7	14.9	16.9	10.9	10.6
5	0.5	27.9	16.8	8.2	6.2

Notes. Contribution to denominator sums of the R^2 and $|R|$ error measures as a function of the cost bucket in the last 12 months of the observation period. (Numbers are based on the testing sample.)

difference between the forecasted (yearly) dollar amount and the realized (yearly) dollar amount. As an example, if we forecast a member's health-care cost to be \$500 in the result period, but in reality the member has an overall health-care cost of \$2,000, then the absolute predicted error for the member is $|\$500 - \$2{,}000| = \$1{,}500$. We define the absolute prediction error (APE) to be the average error over a given sample. APE has been used in recent studies (Cumming et al. 2002, Powers et al. 2005, Dunn et al. 2002) together with the traditional R^2. An advantage of APE is that it does not square the prediction errors, which makes it less sensitive to outliers (members with extreme health-care cost). This is of special concern due to the nature of health-care cost data because there are a few individual members with very unpredictable high costs.

2.4.2. The R^2 Measure

R^2 is defined as

$$R^2 = 1 - \frac{\sum_i (t_i - f_i)^2}{\sum_i (t_i - a)^2},$$

where f_i is the forecasted cost of member i, t_i is the true cost of member i, and a is the average health-care cost in the result period. If we look at the contribution of members in the observation period's cost buckets to the sum in the denominator, it varies greatly, as shown in Table 4. The second column has the fraction of the learning sample in each bucket, and the third column has the contribution to the sum in the denominator. We note that 27.9% of the sum is contributed by the 0.5% of members in bucket 5 in the observation period. R^2 is therefore disproportionately influenced by the members in the most expensive bucket.

R^2 squares each prediction error, which makes it very sensitive to prediction error for members with high health-care costs. A model that does very well for the majority of the population might therefore have low R^2 due to a few extreme unpredictable outliers (for example, members with a sudden onset of a serious condition). In the literature, researchers have dealt with this fact by truncating the health-care cost. We denote the resulting R^2 when claim costs are truncated to \$100,000 by R^2_{100}, and the fourth column of Table 4 shows the contribution to the denominator sum in that case. By truncating these members, the contribution in the denominator sum of bucket 5 reduces to 16%, close to that of buckets 2 through 4.

A natural measure of health-care cost prediction is the absolute value of the prediction error. We therefore define a new R-like measure that has some of the same properties as R^2,

$$|R| = 1 - \frac{\sum |t_i - f_i|}{\sum |t_i - m|},$$

where m is the sample median. We note that $|R| = 0$ if we predict the median of the sample for all members, and $|R| = 1$ if $t_i = f_i$ for all members i. In the same way that R^2 measures the reduction in the residuals squared, $|R|$ measures the reduction in the sum of absolute values of the residuals. In the last two columns of Table 4, we summarize the contributions to the $|R|$ denominator sum for the populations. We note that the contribution is strictly decreasing in the observation period bucket, and is less affected by truncation (noted by $|R_{100}|$). We conclude that $|R|$ is less sensitive to outliers than R^2, and therefore possibly better suited for health-care cost predictions.

3. Methods

3.1. The Baseline Method

To make meaningful comparisons, we define a baseline method against which we compare the results of the prediction models. As our baseline method, we use the health-care cost of the last 12 months of the observation period as the forecast of the overall health-care cost in the result period. Because current health-care cost is a strong indicator of a person's health, this baseline is much stronger than, for example, random assignment. Table 5 shows how

Table 5. The cost bucket distribution of members in the testing sample.

Last 12-month observation period cost bucket	Result period cost bucket (%)				
	1	2	3	4	5
1	75.63	5.54	1.88	0.66	0.20
2	5.03	2.98	1.19	0.39	0.11
3	1.81	1.01	0.91	0.39	0.08
4	0.51	0.38	0.34	0.38	0.11
5	0.10	0.08	0.08	0.10	0.13

Table 6. Baseline performance.

Bucket	Hit ratio (%)	Penalty error	APE ($)
All	80.0	0.431	2,677
1	90.1	0.287	1,279
2	52.3	0.992	4,850
3	41.7	1.358	9,549
4	30.5	1.669	21,759
5	19.3	1.825	75,808

Notes. Performance measures of the baseline method overall and by cost bucket. The cost buckets refer to the cost in the last 12 months of the observation period.

the population falls into the defined cost buckets in the last 12 months of the observation period and the results period. As an example, close to 70% of the population are in bucket 1 in both periods. We further note that for members that fall into cost buckets 1 through 4 in the observation period, the most common bucket in the result period is bucket 1. On the other hand, for members who fall into cost bucket 5 in the observation period, the most common result period bucket is bucket 5. This can be interpreted as follows: Most members who are experiencing moderate cost are, most commonly, getting better, whereas those in the most expensive bucket have a greater tendency to incur high medical costs.

Table 6 summarizes the baseline forecast for all error measures. The baseline prediction model correctly predicts 80.0% of the members, the average penalty error is 0.431, and the absolute prediction error is $2,677. To get a deeper understanding of the baseline method, we examine the effectiveness of the baseline method with respect to the buckets in the observation period. From Table 6 we observe, for example, that for bucket 1 members the hit ratio is 90.1%, the penalty error is 0.287, and the absolute prediction error is $1,279. The fact that most of the members are in bucket 1, have low health-care costs, and continue to have low health-care costs in the result results in a high hit ratio, and low penalty error and average prediction error for the overall population. Note that the performance measures worsen with each increasing cost bucket.

3.2. Data-Mining Methods: Classification Trees

Classification trees (Breiman et al. 1984) have been applied in many fields such as finance, speech recognition, and medicine. As an example, in medicine they have been applied to develop classification criteria for medical conditions such as osteoarthritis of the hip (Altman et al. 1991), the Churg-Strauss syndrome (Masi et al. 1990), and head and neck cancer (Wadsworth et al. 2004). Classification trees recursively partition the member population into smaller groups that are more and more uniform in terms of their known result period cost. This partition can be represented as a tree. This graphical representation makes classification trees easily interpretable, and therefore models that build on them can be medically verified.

Table 7. Classification tree example.

- If a member does not have CAD, predict bucket 1.
- If a member has CAD but does not have diabetes, predict bucket 3.
- If a member has CAD and diabetes, predict bucket 5.

Notes. An example of a classification tree, built on data that has only information about three diagnoses—CAD, diabetes, and acute pharyngitis—from the observation period and the cost bucket of the result period. We note that acute pharyngitis does not appear in the tree, which makes intuitive sense because we do not expect acute pharyngitis to affect the following year's health-care costs.

As an example, consider the simplified case of a data set having information on only three diagnoses in the observation period—coronary artery disease (CAD), diabetes, and acute pharyngitis—as well as the cost bucket of the result period. The classification tree built on this data might result in the classifier depicted in Table 7. The classifier can be used to predict the result period's health-care cost for any unseen member. Assuming we have a new member for whom we want to predict a cost bucket, we first look at whether or not he/she has been diagnosed with CAD. If not, we predict the member to be in cost bucket 1 next period. If the member has been diagnosed with CAD, we examine whether he/she has been diagnosed with diabetes. If he/she has, we predict the member to be in cost bucket 5, and in cost bucket 3 otherwise. We refer the interested reader to online Appendix B for details.

Running the classification tree algorithm on the full data set results in more complicated classifiers than the one depicted in Table 7. Tables 8 and 9 describe characteristics of subgroups predicted to be in buckets 5 and 4 by these more complicated trees. These scenarios demonstrate how the trees use both cost and medical information, along with age, to identify the risky members of the population.

3.3. Data-Mining Methods: Clustering

Clustering algorithms organize objects so that similar objects are together in a cluster and dissimilar objects belong to different clusters. Our prediction clustering method centers around the algorithm behind EigenCluster, a search-and-cluster engine developed in Kannan et al. (2004). The clustering algorithm, when applied to data, automatically detects patterns in the data and clusters together members who are similar. We adapted the original clustering algorithm for the purpose of health-care cost prediction. We first cluster members together using only their monthly cost data, giving the later months of the observation period more weight than the first months (see online Appendix C). The resulting clustering places members within a particular cluster who all have similar cost characteristics. Then, for each cost-similar cluster, we run the algorithm on their medical data to create clusters whose members have both similar cost characteristics as well as medical conditions. We then assign a forecast for a particular cluster based on the known result period's costs of the learning sample. To

Table 8. Predicted cost bucket 5 members.

Examples of members predicted to be in cost bucket 5 in the result period
• Members with overall costs between \$12,300 and \$16,000 in the last 12 months of the observation period and who have acute cost profiles. The members take no more than 14 different therapeutic drug classes during that period, and have not had a heart blockage followed by dose(s) of amiodarone hcl. They have more than 15 individual diagnoses and at least one of the following conditions: (a) have been in the ICU because of congestive heart failure, (b) have chronic obstructive pulmonary disease with more than one prescription for Macrolides or floxins, (c) have renal failure with more than one hospitalization in the observation period, or (d) have both coronary artery disease and depression.
• Members with more than \$24,500 in costs in the observation period, an acute cost profile, and a diagnosis of secondary malignancy (cancer).
• Members in cost bucket 2, with nonacute cost profile, and costs between \$2,700 and \$6,100 in the last 6 months of the observation period, and with either (a) coronary artery disease and hypertension receiving antihypertensive drugs or (b) has peripheral vascular disease and is not on medication for it.
• Members in cost bucket 2, taking between 15 and 34 different therapeutic drug classes during the observation period, with nonacute cost profile, and costs between \$1,200 and \$4,000 in the last 6 months of the observation period, and who have a Hepatitis C related hospitalization during the observation period.
• Members in cost buckets 2 and 3 with nonacute cost profiles, less than \$2,400 in pharmacy costs and on fewer than 13 therapeutic drug classes, but who have received Zyban (prescription medication designed to help smokers quit) after a seizure.

Note. Examples of members that the classification tree algorithm predicts to be in bucket 5.

illustrate, let us give an example (details on the algorithm can be found in online Appendix C). We start with a cluster found by the algorithm using cost characteristics only. The cost profiles of the members are shown in Figure 3. We note that all members have relatively low cost until the last six months of the observation period, but a greater cost in the last months of the period.

The key observation is that when using cost information only we are not able to distinguish between the members in the cluster. The algorithm uses medical information to identify subgroups within the cost cluster and partitions the members into two subclusters. Table 10 shows some of the medical characteristics with the greatest difference in prevalence between the two groups.

The first cluster consists of members that have pathology, cytopathology, infusions, and other indicators of cancer indicating a potentially serious health problem that is likely to lead to higher health-care costs in the future. The second cluster, on the other hand, consists predominantly of members who are in physical therapy and have had orthopedic surgery and have other musculoskeletal characteristics. We can expect that these members will be getting better, and thus will have lower health-care costs in the following year.

4. Results

4.1. Performance of the Data-Mining Methods

We ran the classification tree algorithm using the learning sample, and calibrated the algorithm using the validation sample. We built three distinct classification trees, one for each of the three performance measures. Once we found the right tree for each error measure, we used it to classify the testing sample. We report those results. We ran the clustering algorithm in a similar manner. The resulting clusters contain groups of members with similar cost characteristics and often similar medical characteristics. For each cluster, we assign a prediction based on the learning and validation samples and apply it to the testing sample. We report on the performance of the algorithms on the aggregate level first, and then by bucket.

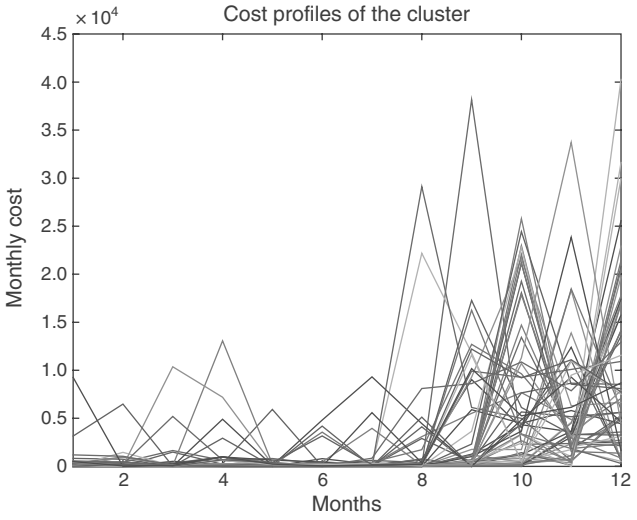
Table 11 shows the performance measures. The trees predict the right bucket for over 84% of the population, the average penalty is 0.385, and the absolute prediction

Table 9. Predicted cost bucket 4 members.

Examples of members predicted to be in cost bucket 4 in the result period
• Members in cost buckets 2 through 5 that have taken more than 34 therapeutic drug classes during the observation period.
• Members in cost bucket 1 that have inpatient days (have been in a hospital) in the last three months with around \$1,300 in health-care costs in the last three months.
• Women in cost bucket 1 that have between \$1,300 and \$1,500 in cost in the last six months of the observation period, that do not have renal failure, but have taken Arava (disease-modifying antirheumatic drug) within 180 days prior to delivery and who do not have prescribed prenatal vitamins during pregnancy.
• Members in cost bucket 1, who have more than \$1,700 in health-care costs in the last six months of the observation period, that have nonacute cost profiles, and who have hypertension but no lab test in the observation period.
• Members with more than \$24,500 in health-care costs in the observation period but less than \$3,200 in pharmacy costs and on fewer than 14 different therapeutic drug classes during the observation period, with nonchronic cost profile, do not have a diagnosis of secondary malignancy, but have more than nine office visits in the last three months of the observation period.

Note. Examples of members that the classification tree algorithm predicts to be in bucket 4.

Figure 3. The monthly costs of the last 12 months of the observation period for all members of a cost-similar cluster.



error is \$2,243. There is a considerable improvement in all the performance measures over the baseline methodology. Most notably, the improvement in absolute prediction error is over 16%. The reduction in the penalty error is 10.5% and there is close to 5% improvement in the hit ratio. For the clustering algorithm, there is again considerable improvement in all the performance measures compared to the baseline method. The results are comparable with the classification tree algorithm, with the clustering algorithm having an edge in the absolute prediction error.

We now take a more detailed view on the accuracy of the algorithms and break down the performance by the observation period's cost bucket. For both algorithms, the improvements are most significant for the top buckets. For the classification tree algorithm, we note that the hit ratio almost doubles, the decrease in the penalty error is 23%, and the decrease in the absolute prediction error is over 50% for the most expensive bucket. The clustering algorithm, similarly, more than doubles the hit ratio, decreases the penalty error by more than 35%, and decreases the average absolute prediction error by over 58% for the most expensive bucket. We note that the classification tree algorithm does a bit better on the lowest-cost buckets for the hit ratio and penalty error, but the clustering algorithm works better on the higher-cost buckets.

Table 10. Distinguishing features of medical clusters.

Frequency in cluster one (%)	Frequency in cluster two (%)	Description
18	**72**	Physical therapy
29	**83**	Durable medical equipment
14	**66**	Orthopedic surgery, exclude endoscopic
4	**48**	Osteoarthritis
39	3	Risk factor: amount paid for injectables greater than \$4,000
71	38	Pathology
32	0	Hematology or oncology infusions
7	**38**	Rehab
21	**52**	Musculoskeletal disorders
25	3	Emetics
25	3	Blood products or transfusions
18	0	Cancer therapies

Notes. Some of the features that differentiate between cost-similar members and separated into two medical subclusters. The first two columns show the percentage of members of each cluster who have a certain diagnosis, have had a procedure, or are taking a drug.

4.2. Prediction Using Cost Information Only

We next investigate the predictability of health-care costs using cost information alone, and compare the prediction to the results when the algorithms use both cost and medical information. We note in Table 12 that for the lower buckets the results are just as good, and in some cases slightly better. The classification trees have better error measures for the lower-cost buckets, but the clustering algorithm does better for the two most expensive buckets. In general, the classification trees do not benefit from adding the medical variables.

Given that an important objective of cost prediction is medical intervention through patient contact, the models with interpretable medical details are preferred. In other cases, the simpler models that achieve good results using only 22 cost variables, as opposed to almost 1,500 medical variables, may be preferred.

4.3. Comparison with Other Studies

We start by noting that comparisons across studies that use different data sets are not fully valid because the average prediction error is highly dependent on the data set. Therefore, as an indication only, we compare our average absolute prediction error to the error reported by two other studies. Cumming et al. (2002) reports an average absolute prediction error of 93% of the actual mean, and Powers et al. (2005) reports an error of 98% of the actual mean. The error for the clustering algorithm is 78.8% of the mean of our testing sample and the classification trees 89.4%, lower than in the other two studies.

Traditionally, prediction software has aimed to minimize R^2. Cumming et al. (2002) reports R^2_{100} from 0.140 to 0.198 (with claims truncated at \$100,000) and R^2 from 0.099 to 0.154 (without truncation). The trees have $R^2 = 0.162$ and $R^2_{100} = 0.204$, and the clustering algorithm has $R^2 = 0.180$ and $R^2_{100} = 0.219$, as can be seen in the top row of Table 13. In the top row of Table 14, we provide $|R|$ and $|R_{100}|$ for both our measures as well as the baseline method.

Finally we note that summarizing the goodness of cost prediction to one number, whether it is R^2 or $|R|$ can be misleading, and important information is lost. To illustrate

Table 11. The resulting performance measures.

	Hit ratio (%)			Penalty error			APE ($)		
Bucket	Trees	Cluster	Baseline	Trees	Cluster	Baseline	Trees	Cluster	Baseline
All	84.6	84.3	80.0	0.386	0.374	0.431	2,243	1,977	2,677
1	90.2	89.9	90.1	0.275	0.259	0.287	1,398	1,152	1,279
2	60.2	58.7	52.3	0.864	0.884	0.992	4,158	4,051	4,850
3	51.9	52.7	41.7	1.038	1.071	1.358	6,598	6,585	9,549
4	43.3	44.4	30.5	1.241	1.177	1.669	12,665	11,116	21,759
5	36.9	42.7	19.3	1.405	1.170	1.825	36,541	31,613	75,808

Notes. The top line shows the measures for the whole population, followed by the measures broken down by the observation's last 12 months cost buckets, for the classification tree algorithm, the clustering algorithm, and the baseline methodology.

this point we have included in Tables 13 and 14 the relative reduction in the error sum for each of the cost buckets. As an example, if $\sum_i (t_i - a)^2 = 100$ for the members in cost bucket 1, and $\sum_i (t_i - f_i)^2 = 95$ for the same members, the relative reduction is $(95-100)/100 = 0.05$, or 5%. We note that for buckets 1–4, the baseline improves over predicting the sample average, but for the most expensive bucket, bucket 5, the baseline does worse. For the most expensive members, repeating the current cost is not a strong prediction rule, and due to the weight that those members carry in the R^2 measure (due to very large residuals), this results in negative R^2.

Our algorithms reduce the relative error for all cost buckets, and the reduction increases with higher-cost buckets, ranging from 5% to 49% for the R^2 and R^2_{100} measures and from 10% to 32% for the $|R|$ and $|R_{100}|$ measures. This shows that our prediction models improve predictions for members in all buckets, and most significantly for the most expensive members.

4.4. Summary of Results

In summary, we observe that both algorithms improve predictions over the baseline method for all performance measures and the improvement is more significant for more costly members (higher buckets). In terms of overall performance measures (overall hit ratio and absolute prediction error), the methods are comparable. The clustering method results in better predictions for current high-cost bucket members and consistently better absolute prediction error, whereas the classification tree algorithm has an edge on lower-cost members when we look at the hit ratio and the penalty error. We believe that the reason that the clustering algorithm is stronger in predicting high-cost members is the hierarchical way cost and medical information are used. Recall that the clustering algorithm first uses cost information and then uses medical information in situations where medical information can further discriminate between members belonging in different cost buckets. Referring back to our clustering sample, we note that all members of a cost-similar cluster have similar cost trajectories of rising costs in the last months of the observation period. Using medical information, the clustering algorithm is able to distinguish between two main groups of patients: higher-risk cancer patients with predicted cost bucket 4, and patients with musculoskeletal and orthopedic characteristics with predicted cost bucket 1. When medical information is not dense—that is for members in the lower buckets—using cost information only results in similar error measures. Furthermore, from our comparison with previous studies we find evidence that our algorithms do well in comparison to current prediction methods, and an analysis of the R^2 and $|R|$ measures showed improved predictions for all cost buckets.

5. Conclusions and Future Research

The algorithms we developed based on modern data-mining methods provide quantifiable predictions of medical costs and represent a powerful tool for the prediction of health-care costs. We also argue that R^2, which has traditionally been used to report prediction accuracy, has some

Table 12. The resulting performance measures using cost information only.

	Hit ratio (%)			Penalty error			APE ($)		
Bucket	Trees	Cluster	Baseline	Trees	Cluster	Baseline	Trees	Cluster	Baseline
All	84.6	84.2	80.0	0.389	0.399	0.431	2,214	2,116	2,677
1	90.1	90.1	90.1	0.279	0.282	0.287	1,395	1,269	1,279
2	60.3	57.5	52.3	0.873	0.920	0.992	4,033	4,146	4,850
3	52.3	49.9	41.7	1.025	1.093	1.358	6,462	6,580	9,549
4	42.7	41.7	30.5	1.256	1.272	1.669	12,310	12,412	21,759
5	35.2	40.5	19.3	1.367	1.220	1.825	35,875	33,907	75,808

Notes. The top line shows the measures for the whole population, followed by the measures broken down by the observation's last 12 months cost buckets, for the classification tree algorithm, clustering algorithm, and the baseline methodology.

Table 13. R^2 results—The R^2 and R^2_{100} for the two algorithms and the baseline.

	Baseline		Trees		Clustering	
Bucket	R^2	R^2_{100}	R^2	R^2_{100}	R^2	R^2_{100}
All	−0.102	−0.050	0.162	0.204	0.180	0.220
1 (%)	−3.3	−5.3	−5.3	−8.3	−5.0	−7.9
2 (%)	−5.6	−8.9	−6.3	−10.9	−5.7	−8.6
3 (%)	−8.7	−13.6	−12.8	−23.3	−12.7	−22.5
4 (%)	−5.7	1.3	−22.6	−34.1	−24.4	−36.5
5 (%)	50.0	60.1	−31.0	−39.4	−37.0	−49.8

Note. Rows 1 through 5 show the relative reduction in the denominator sum for each cost bucket.

Table 14. $|R|$ results—The $|R|$ and $|R_{100}|$ for the two algorithms and the baseline.

	Baseline		Trees		Clustering	
Bucket	$\|R\|$	$\|R_{100}\|$	$\|R\|$	$\|R_{100}\|$	$\|R\|$	$\|R_{100}\|$
All	−0.037	−0.013	0.171	0.182	0.182	0.194
1	−11.5%	−11.9%	−10.4%	−10.8%	−12.7	−13.1
2	−8.5%	−8.8%	−23.9%	−24.9%	−21.7	−22.4
3	10.1%	10.6%	−25.0%	−26.2%	−24.1	−25.3
4	32.5%	35.4%	−23.4%	−25.4%	−24.2	−26.3
5	71.0%	58.2%	−16.6%	−23.4%	−23.7	−33.0

Note. Rows 1 through 5 show the relative reduction in the denominator sum for each cost bucket.

limitations, and the use of more descriptive error measures, specially designed for the application at hand, might give better insight into the prediction accuracy. Despite the relative abundance of clinical information included in our data sets, we found that for all but the highest-cost patients, primary cost information was the most accurate predictor of true costs. It is clear that cost is an efficient surrogate for medical information, except in cases where the very dense medical data are available. The algorithms can be used for cost predictions for individuals and groups and as a basis for patient intervention in health-care management. Future research that builds on these algorithms could be used for financial reimbursement or insurance-pricing purposes, but such an effort requires greater integration with health-care economics and system design.

6. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Endnote

1. Comorbidity is defined as coexisting medical conditions.

Acknowledgments

The authors thank the reviewers for their insightful and helpful comments and Bijay Ghimire and the engineering team at D2Hawkeye for their extensive data support.

References

Altman, R., G. Alarcón, D. Appelrouth, D. Bloch, D. Borenstein, K. Brandt, C. Brown, et al. 1991. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheumatism* **34** 505–514.

Ash, A., R. P. Ellis, G. Pope, J. Ayanian, D. Bates, H. Burstin, L. Iezzoni, E. McKay, W. Yu. 2000. Using diagnoses to describe populations and predict costs—Capitated payment system uses demographic risk adjusted to calculate payments. *Health Care Financial Rev.* **21** 7–28.

Bertsimas, D., R. Freund. 2005. *Data, Models and Decisions*, Dynamic Ideas, Belmont, MA.

Breiman, L., J. Friedman, R. Olshen, J. C. Stone. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Centers for Medicare & Medicaid Services: Diagnosis and Procedure Codes and Their Abbreviated Titles. 2004. Version 22, effective October 1. Retrieved August 28, 2006, http://www.cdc.gov/nchs/datawh.htm#International%20Classification.

Chang, R. E., C. L. Lai. 2005. Use of diagnosis-based risk adjustment models to predict individual health care expenditure under the national health insurance system in Taiwan. *J. Formosan Med. Assoc.* **104** 883–890.

Cheng, D., R. Kannan, S. Vempala, G. Wang. 2008. A divide-and-merge methodology for clustering. *Proc. ACM Sympos. Principles of Database Systems*. Forthcoming.

Classification Rule with Unbiased Interaction Selection and Estimation. Binaries retrieved July 20, 2007, http://www.stat.wisc.edu/loh/cruise.html.

Cumming, R., D. Knutson, B. Cameron, B. Derrick. 2002. A comparative analysis of claims-based methods of health risk assessment for commercial populations. Retrieved August 28, 2006, http://www.soa.org/ccm/cms-service/stream/asset?asset_id=9215098&g11n.

Dans, P. 1993. Looking for answers in all the wrong places. *Ann. Internal Med.* **119** 855–857.

Dove, H., I. Duncan, A. Robb. 2003. A prediction model for targeting low-cost, high-risk members of managed care organizations. *Amer. J. Managed Care* **9** 381–389.

Dunn, D. L., A. Rosenblatt, D. A. Taira, E. Latimer, J. Bertko, T. Stoiber, P. Braun, S. Busch. 2002. A comparative analysis of methods of health risk assessment. Society of Actuaries Monograph M-HB96-1. Retrieved August 28, 2006, http://www.soa.org/ccm/content/research-publications/library-publications/monographs/health-benefits-monographs/.

Farley, J. F., C. R. Harley, J. W. Devine. 2006. A comparison of comorbidity measurements to predict healthcare expenditures. *Amer. J. Managed Care* **12** 110–117.

Fleishman, J. A., J. W. Cohen, W. G. Manning, M. Kosinksi. 2006. Using the SF-12 health status measure to predict medical expenditures. *Med. Care* **44** I-54–I-63.

Jolins, J., M. Ancukiewicz, E. DeLong, D. Pryor, L. Muhlbaier, D. Mark. 1993. Discordance of databases designed for claims payment versus clinical information systems: Implications for outcomes research. *Ann. Internal Med.* **119** 844–850.

Jones, A. M. 2000. Health econometrics. A. J. Culyer, J. P. Newhouse, eds. *Handbook in Health Economics*. Elsevier, Amsterdam, 265–344.

Kannan, R., S. Vempala, A. Vetta. 2004. On clusterings: Good, bad and spectral. *J. ACM* **51** 497–515.

Kim, H., W.-Y. Loh. 2001. Classification trees with unbiased multiway splits. *J. Amer. Statist. Assoc.* **96** 589–604.

Kim, H., W.-Y. Loh. 2003. Classification trees with bivariate linear discriminant node models. *J. Comput. Graphical Statist.* **12** 512–530.

Klabunde, C. N., J. L. Warren, J. M. Legler. 2002. Assessing comorbidity using claims data—An overview. *Med. Care* **40** 26–35.

LaVange, L. M., V. G. Iannacchione, S. A. Garfinkel. 1986. An application of logistic regression methods to survey data: Predicting high cost users of medical care. *Proc. Survey Research Methods Section*, American Statistical Association. Retrieved August 28, 2006, http://www.amstat.org/sections/SRMS/proceedings/papers/1986_049.pdf.

Lawthers, A. G., E. P. McCarthy, R. B. Davis, L. E. Peterson, R. H. Palmers, L. I. Iezzoni. 2000. Identification of in-hospital complications from claims data. Is it valid? *Med. Care* **38** 785–793.

Loh, W.-Y., Y.-S. Shih. 1997. Split selection methods for classification trees. *Statistica. Sinica* **7** 815–840.

Manning, W. G., J. Mullahy. 2001. Estimating log models: To transform or not to transform? *J. Health Econom.* **20** 461–494.

Masi, A. T., G. G. Hunder, J. T. Lie, B. A. Michel, D. A. Bloch, W. P. Arend, L. H. Calabrese, S. M. Edworthy, A. S. Fauci, R. Y. Leavitt 1990. The American College of Rheumatology 1990 criteria for the classification of Churg-Strauss Syndrome (allergic granulomatosis and angiitis). *Arthritis Rheum* **33** 1094–1100.

Mehta, S., H. Glick, S. Suzuki, K. Schulman. 1999. Determining an episode of care using claims data: Diabetic foot ulcer. *Diabetes Care* **22** 1110–1115.

National Drug Code Directory. 2004. List of FDA approved drugs as of December 31, 2004. Retrieved August 28, 2006, http://www.fda.gov/cder/ndc/database/default.htm.

Pietz, K., C. M. Ashton, M. McDonell, N. P. Wray. 2004. Predicting healthcare costs in a population of Veterans Affairs beneficiaries using diagnosis-based risk adjustment and self-reported health status. *Med. Care* **42** 1027–1035.

Pladevall, M. 2004. Clinical outcomes and adherence to medications measured by claims data in patients with diabetes. *Diabetes Care* **27** 2800–2805.

Powers, C. A., C. M. Meyer, M. C. Roebuck, B. Vaziri. 2005. Predictive modeling of total healthcare costs using pharmacy claims data: A comparison of alternative econometric cost modeling techniques. *Med. Care* **43** 1065–1072.

Roblin, D. W., P. I. Juhn, B. J. Preston, R. D. Penna, S. P. Feitelberg, A. Khoury, J. C. Scott. 1999. A low-cost approach to prospective identification of impending high cost outcomes. *Med. Care* **37** 1155–1163.

Van de Ven, W. P. M. M., R. P. Ellis. 2000. Risk adjustment in competitive health plan markets. A. J. Culyer, J. P. Newhouse, eds. *Handbook in Health Economics*. Elsevier, Amsterdam, 756–845.

Wadsworth, J. T., K. D. Somers, L. H. Cazares, G. Malik, B.-L. Adam, B. C. Stack Jr., G. L. Wright Jr., O. J. Semmes. 2004. Serum protein profiles to identify head and neck cancer. *Clinical Cancer Res.* **10** 1625–1632.

Wennberg, J., N. Roos, L. Sola, A. Schori, R. Jaffe. 1987. Use of claims data systems to evaluate health care outcomes. Mortality and reoperation following prostatectomy. *J. American Med. Assoc.* **257** 933–936.

Zhao, Y., A. S. Ash, R. P. Ellis, J. Z. Ayanian, G. C. Pope, B. Bowen, L. Weyuker. 2005. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Med. Care* **43** 34–43.

Zhao, Y., R. P. Ellis, A. S. Ash, D. Calabrese, J. Z. Ayanian, J. P. Slaughter, L. Weyuker, B. Bowen. 2001. Measuring population health risks using inpatient diagnoses and outpatient pharmacy data. *Health Serv. Res.* **36** 180–193.

Zhou, Z. H., C. A. Melfi, S. L. Hui. 1997. Methods for comparison of cost data. *Ann. Internal Med.* **127** 752–756.