Data-Driven Approach to Health Care: Applications Using Claims Data

by

Margrét Vilborg Bjarnadóttir

B.S., Mechanical and Industrial Engineering
University of Iceland, 2001

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2008

Author .....
Sloan School of Management
August 14, 2008

Certified by .....
Dimitris J. Bertsimas
Boeing Professor of Operations Research
Thesis Supervisor

Accepted by .....
Cynthia Barnhart
Professor
Co-Director, Operations Research Center

Data-Driven Approach to Health Care: Applications Using Claims Data

by

Margrét Vilborg Bjarnadóttir

Submitted to the Sloan School of Management
on August 14, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Large population health insurance claims databases together with operations research and data mining methods have the potential of significantly impacting health care management. In this thesis we research how claims data can be utilized in three important areas of health care and medicine and apply our methods to a real claims database containing information of over twi million health plan members. First, we develop forecasting models for health care costs that outperform previous results. Secondly, through examples we demonstrate how large-scale databases and advanced clustering algorithms can lead to discovery of medical knowledge. Lastly, we build a mathematical framework for a real-time drug surveillance system, and demonstrate with real data that side effects can be discovered faster than with the current post-marketing surveillance system.

Thesis Supervisor: Dimitris J. Bertsimas
Title: Boeing Professor of Operations Research

Acknowledgments

First and foremost I would like to thank my thesis advisor Professor Dimitris Bertsimas for all his guidance, support and friendship during my time at the ORC. I am truly grateful for the opportunity to work with and learn from such an inspiring mentor. I especially thank him for encouraging me to embark on the next phase of my academic journey.

I would like to thank Dr. Michael Kane with whom we have collaborated over the past few years. His positive attitude towards applying new science to medicine is an inspiration to the further development of data-driven approaches in health care.

I am also very grateful to other members of my thesis committee - Professor Arnold Barnett and Gabriel Bitran. Professor Barnett, is an inspiration in the classroom and I thank him for his help during my academic job search. I thank Professor Bitran for the valuable suggestions and positive comments during my thesis writing process. I would also like to thank Professors Georgia Perakis for welcoming me to MIT and always keeping her door open. In addition, I would also like to thank several professors who have been part of my excellent MIT experience: Richard Larson, Amedeo Odoni, Roy Welsch, Jim Orlin and Retsef Levi. I would also like to thank the ORC staff: Andrew, Laura, Paulette and Veronica.

This thesis would not exist without the data and support from of D2hawkey. I would like to thank the founders Chris Kryder and Rudra Pandey for their help in making this project happen. Most importantly I would like to thank the wonderful research and engineering team at D2Hawkeye, who answered frantic panic data and technical support requests during odd hours: Bijay, Sanjay, Anil and the team, thank you.

The ORC was a great place to call home on campus. I would like to thank Mike W, who despite my best efforts still sticks with his debatable political beliefs and to

Amr for the advise and friendship. To Katy for solving the first LP proofs together and her positive attitude through the ups and downs. I would like to thank my office mate David for the collaboration, Tim for showing us how TA is really played and Theo for always making me look punctual. Hamed, Pavithra, Ruben, Ilan, Kostas, Juliane, Mike Jr, Melanie, Yann, Dan, Guillaume, Lincoln, Jose, Alex, Carol, Doug and the rest of the gang for the great fun.

I would also like to thank my non-ORC friends for making my life in Boston so enjoyable. Trying to list everyone would ensure that I would forget someone important, so instead I would like to thank the Icelandic crowd at MIT for the happy times at the Muddy, the Icelandic crew around town for making things interesting, and the Icelandic Boston Sowing club, for the white wine sipping and women talk. To the Fulbrighters and S&P-ers, thank you for the trips, brunches and all the other get-togethers. To all the friends from around the globe that visited me during my stay in Boston and to the many friends at home, that kept in touch over the past six years and were only a phone call away.

To my family for their constant support. My mum for showing herself around Boston when I was too busy, my dad for making sure that every crappy apartment I rented was fire and thief save, Sindri for afternoons with House and beer and Geir for allowing me to exercise my shopping skills on his budget while spoiling Hekla. And to my second family down in Atlanta for their support and for providing a warm home in the US.

And to Nelson, for all the good times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis explores applications of operations research methods and data mining to health care insurance claims data. This research was possible through collaboration with D2Hawkeye, a medical data mining company in Waltham, Massachusetts. A collection of members' claims has the benefit of giving a "bird's eye view" of a patient's health care, providing the opportunity to recognize patterns in a member's records that is not visible from a single specialist's view. Large population claims databases therefore provide a wealth of research opportunities, and together with advanced data mining models, have the potential of significantly impacting health care management in the US. In this thesis, we research how claims data can be utilized for cost prediction, medical knowledge discovery and drug surveilance.

Below, we give an introduction to claims data and the contents of the subsequent chapters.

1.1 Health Insurance Claims Data

Health insurance claims data, or simply "claims data", includes two types of claims, medical and pharmaceuitical, as well as information about the members, such as age, gender and his/her geographical location. Medical claims data get generated when hospitals and other health care providers send claims to third-party payers to receive

reimbursement for their services. Each time a member visits a doctor or a hospital, a claim line gets generated with the reason for the visit, the diagnois and the procedure performed. If multiple services are performed, a single visit might result in multiple claim lines. The data does not include results of tests or procedures, although sometimes the results can be ineffered from the subsequent treatment. Pharmaceutical claims data get generated when a member fills a prescription and includes, for example, information about the drug, the prescribing doctor, and the number of days of supply.

The value of claims data in medical research has often been questioned [36, 23] 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 being used for medical research. Examples include researching differences in the outcomes of adherence to medication [53], in length of episodes [50], and of medical outcomes [61] as well as identification of in-hospital complications [44]. Statistical methods generally used when working with medical data are nicely summarized in [37], and other work addressing issues working with health care cost data include [65, 47].

Claims data relies on health care professionals to encode their diagnoses and procedures in terms of the ICD-9-CM codes. There are numerous studies that investigate the reliability of claims data, which compare the information in the claims data to actual medical records; we refer the reader to [22] for a nice summary of those studies. In short, the sensitivity[1] of a diagnosis varies from 50% to over 90%, depending on the diagnosis. Procedures in the claims data have a high correlation with the medical record, and prescription claims have been found to be a more reliable record for drugs actually dispensed than the medical record. In summary, claims data has limitations to its accuracy, but the availability and the size of the data make it an attractive

[1]The Sensitivity of a diagnosis in a claims data is the probability of the diagnosis appearing in the claims data among all patients with disease. A sensitivity of 100% for a diagnosis means that for all members with the disease, the appropriate diagnosis appears in the claims data.

option when conducting research in medicine and health care.

1.2 Cost Prediction and Discovery of Medical Knowledge

The rising cost of health care is one of the world's most important problems. Accordingly, predicting such costs accurately is a significant first step in addressing this problem. In Chapter 2, we focuse on health care cost prediction. Earlier researchers concentrated on using classical regression models or logistic regression models often combined with heuristic classification rules [21], and traditionally, researchers have reported the accuracy of their models using in-sample R^2. In our view, the best way to express the predictability of a method is to perform out-of-sample experiments using different performance measures. We therefore introduce new error measures and report our results out-of-sample, that is, on data that was not used in developing the models. We introduce the concept of a "cost bucket" — a predefined range of cost. We first forecast the cost bucket and then translate the prediction into dollar terms. The "bucketing" helps reduce the noise in the data and the effects of outliers. We also introduce a baseline method of "repeating costs," that we use to compare our results with. We apply data-mining algorithms, in particular clustering and classification trees, to cost prediction and outperform previously published results.

Through our work on cost prediction, we identify opportunities for medical discovery, using an modified version of the clustering algorithm Eigencluster [3]. The algorithm can take a global view of the data and identify new patterns and may therefore reveal unexpected associations among diagnoses, procedures and drugs. We identify a recently suspected link between osteoporosis and depression. In addijton, we identify nonsteroidal anti-inflammatory agents as a risk factor for cardiac patients. Further analysis of the data and the existing medical literature confirmed our discoveries.

1.3 Drug Surveillance

After the withdrawal of rofecoxib (commonly known as Vioxx) from the pharmaceutical market in 2004, post-FDA-approval drug safety and surveillance has come under serious scrutiny. In a 2006 study by the Institute of Medicine, it was pointed out that efforts to monitor the risk-benefit tradeoff of medications decreases after FDA approval, and that this issue needs to be addressed [55]. Currently the Center for Drug Evaluation and Research, a part of the Food and Drug Administration (FDA), handles the post-FDA-approval drug surveillance, which is conducted using the Adverse Event Reporting System (AERS). The AERS is a voluntary system where patients and health care professionals can submit reports of adverse events. Although the system has often proved useful in identifying serious side effects of drugs, it has been insufficient in identifying potential safety signals, as not all events (some even suggest very few) get reported and events that can be indicators of increased risk might not be considered important by individual patients or health care professionals.

Claims data holds great potential for real-time drug surveillance, due to its fast availability and size, which is crucial when trying to detect rare events. In Chapter 3, we develop a framework for drug surveillance and address two of its fundament issues: a) how do we choose comparison groups? and b) how do we compare the two groups? We test serveral methods on real data from D2Hawkeye and report on the results.

1.4 Contributions

There are several contributions of this thesis. In our work on cost prediction we indentify the past cost trajectory to be a powerful predictor of future cost. This observation could refocus the research effort in the area away from detailed disease based modeling. We also raise questions about the limitations and validity of current measures of predictive accuracy and propose alternatives, guided by the use of the cost predictions. Finally, we introduced new methods that have not been applied

in this context before, and showed that they outperform previous published results. We discuss how medical knowledge is often obtained from small studies. As a result, large-scale claim databases has the potential to add to that knowledge, and we provide two examples to support that argument. Lastly, in Chapter 3 we perform, to our knowledge, the first full drug-surveillance experiment, that tests across the whole spectrum of possible side effects. Our findings discourage the use of a comparison population as a direct comparison, the current method of choice. Our work shows that a successful drug-surveillance system can be built, based on claims data analysis and could become one of FDA's standard tools for post-marketing surveillance.

Chapter 2

Prediction of Health Care Costs and Algorithmic Discovery of Medical Knowledge

2.1 Introduction

The predictive power of claims data became a topic of research in the 1980s [63] and numerous studies since have established the predictive power of administrative data on health care costs, [10, 64, 27, 63]. Van de Ven *et al.* [59] provides an insightful overview of the developments in risk based predictive modeling prior to 2000. Cumming *et al.* [21] presents a comparison analysis 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 have concentrated on using classical regression models [63, 10, 64, 54] when predicting total health care costs or logistic regression models, [43, 57] 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 ad-

[1]Comorbidity is defined as coexisting medical conditions.

ministrative data, as a method to account for comorbidity differences of comparative populations in medical research [41], to design fair reimbursement plans [59, 25] and as a basis for predictive modeling of health care costs [10, 27, 18]. Numerous studies that predict health care cost, based on data other than claims data are available, examples include [29, 52].

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 [54, 24]. Traditionally [21] R^2 (or adjusted R^2) have been the measures used to evaluate predictive models but there are some serious drawbacks to its use, which in our opinion makes it unsuitable for a study like the one presented in this chapter. 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, [12]). 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 chapter. 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 chapter 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, c) medical information adds to prediction accuracy when used in the clustering algorithm, while with classification trees, cost information alone results in similar error measures.

The rest of the chapter is structured as follows: In Section 2.2, we describe the data and define the performance measures we consider and in Section 2.3, we present the two principal methods we use: classification trees and clustering algorithms. In Section 2.4, we report on the performance of classification trees and clustering respectively in forecasting health care costs, and in Section 2.6, we briefly discuss our conclusions and future research directions.

2.2 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 long 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 2866 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/her family) was covered by the insurance policy. The data also contains basic geographic information such as age and gender. All members have eligibility starting no later then 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

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 employers 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, expected 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, while 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.

2.2.1 Aggregation of the Claims Data

The claims include diagnosis, procedure and drug information. The diagnosis data is coded using the ICD-9-CM [1] (International Classification of Diseases, Ninth Revision, Clinical Modification) codes, 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 includes 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 [5].

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 which 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 diagnosis, procedures, drugs and risk factors that each member has and include it as additional variables. In summary, the predictive medical variables include: the diagnosis groups, the procedure groups, the drug groups, the risk factors that we have developed, and their count, for a total of close to 1500 possible medical variables. We refer the reader to Appendix 2.A for more details.

2.2.2 Cost and Demographic Data

In addition to the medical variables, we utilize 22 cost variables, since we believe that cost information gives a global picture of the health of a member and include age and gender as well. In order to capture the trajectory of the medical costs (as a proxy of the overall medical condition) we use the monthly costs for the last twelve 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 6 months and the last 3 months of the observation period. Furthermore, in order to capture the pattern of costs, we developed a new indicator variable that captures whether or not the 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 2-1 that depicts the monthly cost of two members in the last twelve months of the ob-

servation period. While 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), while Member B has a spike in the cost profile (a typical pattern for a member with an "acute" condition). The key idea here is that while 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, for example in the case of pregnancy complications, accidents, or acute medical conditions like pneumonia or appendicitis.

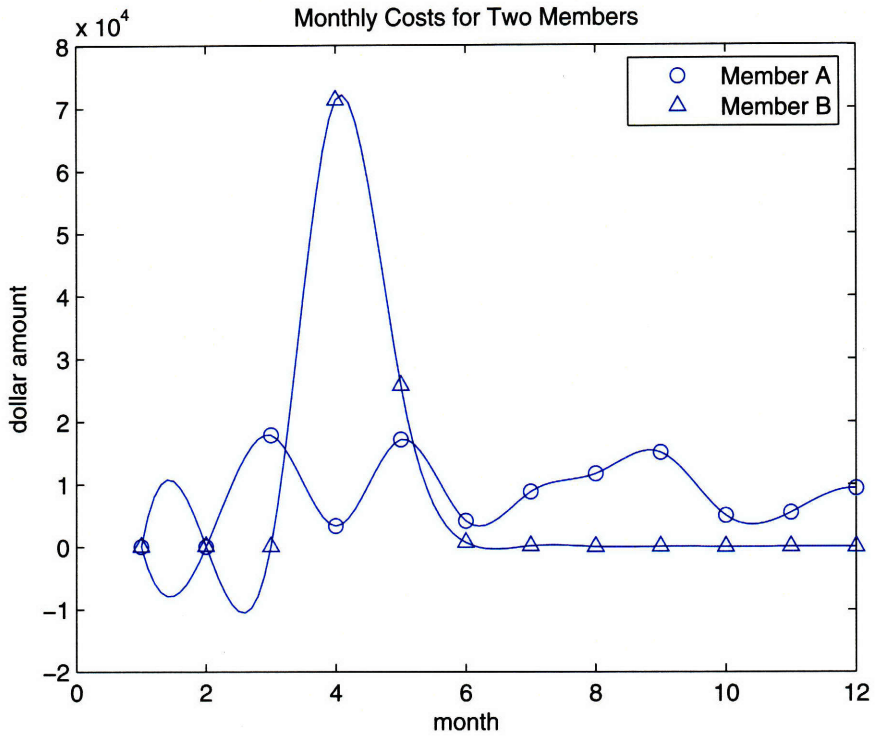

Figure 2-1: 12 months health care costs of two members, with overall cost of $97,500 and $98,100 respectively. A cubic spline curve is fit to the data for easier viewing. The cost profile for Member A has the characteristics of a chronic illness while, the characteristics of member B's profile is "acute". 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 and negative trend in the last months of the observation period.

Finally, we used gender and age as additional variables. Table 2.1 summarizes all the variables used in the study and more details are provided in Appendix2.A.

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 - 1485	Medically defined risk factors
1486 - 1489	Count of members diagnosis, procedures, drugs and risk factors
1490 - 1780	Cost variables, including overall medical and pharmacy costs, acute indicator and monthly costs
1522-1523	Gender and age

Table 2.1: Summary of the Data Elements Used.

2.2.3 Cost Bucketing

The range of paid amounts for members in the learning sample during the result period is from no cost 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. In Figure 2-2, that 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.

In order 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

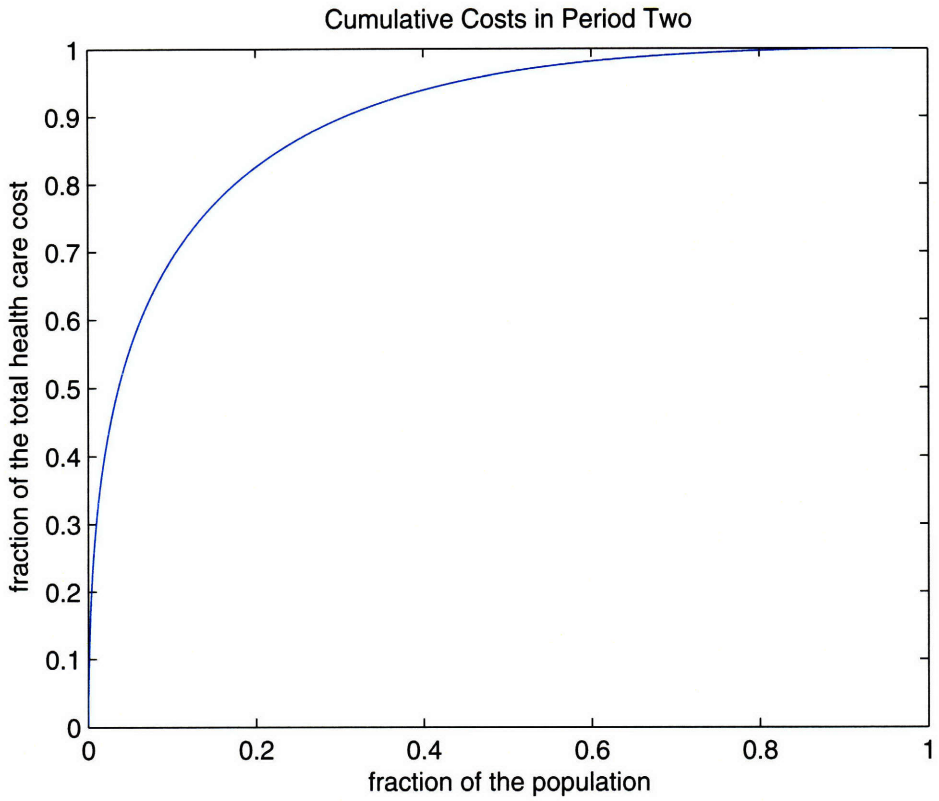

Figure 2-2: Cumulative health care costs of the result period for members in the learning sample. On the x-axis is the cumulative percentage of the population and on the y-axis is the cumulative percentage of the overall health care costs. For example we note that 8% of the population (the most expensive members) account for 70% of the overall health care costs.

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 top bucket (we have 1175 members in the learning sample in bucket five). Table 2.2 shows the range of each bucket, the percentage and the number of members of the learning sample that are in each bucket.

The knowledge of the predicted bucket of a member is valuable to health care management professionals. The buckets from one through five can be interpreted as representing low, emerging, moderate, high and very high risk of medical complications.

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

Table 2.2: Cost Bucket Information. Cost bucket ranges and fraction of the learning sample in each bucket (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.

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, while those members forecasted to be in the top bucket are candidates for case management programs, the most intense patient care program.

2.2.4 Performance Measures

We will measure the performance of our models with three main error measures: the hit ratio, a 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 Section 2.2.4 and define the new error measures in Section 2.2.4.

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 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, as little harm or cost ensues in this instance. Therefore, the penalties for underestimating a cost bucket are set as twice those for overestimating it, the estimated opportunity loss by doctors. Table 2.3 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.

		Outcome				
		1	2	3	4	5
Forecast	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

Table 2.3: The Penalty Table defines the penalty error measure for the five cost buckets. A perfect forecast results in an error of zero.

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 difference between the forecasted (yearly) dollar 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 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 [21, 54, 25] 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, as there are a few individual members with very unpredictable high costs.

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 2.4. The second column has the fraction of the learning sample in each bucket, and in the third column 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 the top bucket in the observation period. R^2 is therefore disproportionably influenced by the members in the top 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 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 claims are truncated to \$100,000 by R^2_{100}, and the fourth column of Table 2.4 shows the contribution to the denominator sum in that case. By truncating the these members the contribution in the denominator sum of the top bucket reduces to 16%, close to that of bucket 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 as R^2 measures the reduction in the residuals squared, $|R|$ measures the reduction

Bucket	% of the Learning Sample	% of Overall $\sum((t_i - a)^2)$	% of Overall $\sum((t_i - a)^2)$ Truncated	% of Overall $\sum(\|t_i - m\|)$	% of Overall $\sum(\|t_i - m\|)$ 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%

Table 2.4: Analysis of Denominator Sums of R^2 and $|R|$. Contribution to denominator sums of R^2 and the $|R|$ error measures as a function of the bucketed cost in the last 12 months of the observation period (numbers are based on the testing sample).

in the sum of absolute values of the residuals. In the last two columns of Table 2.4, 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.

2.3 Methods

2.3.1 The Baseline Method

In order 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 twelve months of the observation period as the forecast of the overall health care cost in the result period. Since current health care cost is a strong indicator of a person's health, this baseline is much stronger than, for example, random assignment. Table 2.5 shows how 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 seventy percent of the population are in bucket one 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

one. 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 most members who are experiencing moderate cost are, most commonly, getting better, while those in the most expensive bucket have a greater tendency to incur high medical costs.

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 2.5: The Cost Bucket Distribution of Members in the Testing Sample.

Table 2.6 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. In order 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 2.6 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 period significantly affects the baseline error measures. Note that the performance measures worsen with each increasing cost bucket.

2.3.2 Data Mining Methods: Classification Trees

Classification trees [15] 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

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

Table 2.6: 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.

[9], the Churg-Strauss syndrome [48] and head and neck cancer [60]. 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, and this graphical representation makes classification trees easily interpretable and therefore models that build on them can be medically verified.

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 2.7. 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 one 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 five, and in cost bucket three otherwise. We refer the interested reader to Appendix 2.B for details.

Running the classification tree algorithm on the full data set results in more complicated classifiers than the one depicted in Table 2.7. Tables 2.8 and 2.9 describe characteristics of subgroups predicted to be in bucket 5 and 4 by these more complicated trees. These scenarios demonstrate how the trees use both cost and medical

- 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.

Table 2.7: An example of a classification tree, built on data that has only information about three diagnosis, 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 as we do not expect acute pharyngitis to affect the following year's health care costs.

information along with age to identify the risky members of the population.

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 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 diagnosis 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 non-acute cost profile and between $2,700 and $6,100 in costs 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 non-acute cost profile and between $1,200 and $4,000 paid in the last 6 months of the observation period and finally have a Hepatitis C related hospitalization during the observation period. • Members in cost buckets 2 and 3 with non-acute cost profiles, less $2,400 in pharmacy costs and on fewer than 13 therapeutic drug classes, but have received Zyban (prescription medication designed to help smokers quit) after a seizure.

Table 2.8: Examples of members that the classification tree algorithm predicts to be in bucket 5.

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 one that have inpatient days (have been in a hospital) in the last three months with around $1,300 dollars in health care costs in the last 3 months. • Women in cost bucket one that have between $1,300 and $1,500 in cost in the last 6 months of the observation period, that do not have Renal failure, but have taken Arava (disease-modifying anti-rheumatic drug) within 180 days prior to delivery and do not have prescribed prenatal vitamins during pregnancy. • Members in cost bucket one, who have more than $1,700 in health care costs in the last 6 months of the observation period, that have non-acute cost profiles and have hypertension but no lab test in the observation period. • Members with more than $24,500 in healthcare 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 non-chronic cost profile, do not have a diagnosis of secondary malignancy, but have more than nine office visits in the last 3 months of the observation period.

Table 2.9: Examples of members that the classification tree algorithm predicts to be in bucket 4.

2.3.3 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 [38]. 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 Appendix 2.C.) The result 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 both have 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 illustrate let us give an example (details on the algorithm can be found in Appendix 2.C). We start with a cluster, found by the algorithm using cost characteristics only. The cost profiles of the members are shown in Figure 2-3. 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 sub clusters. Table 2.10 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

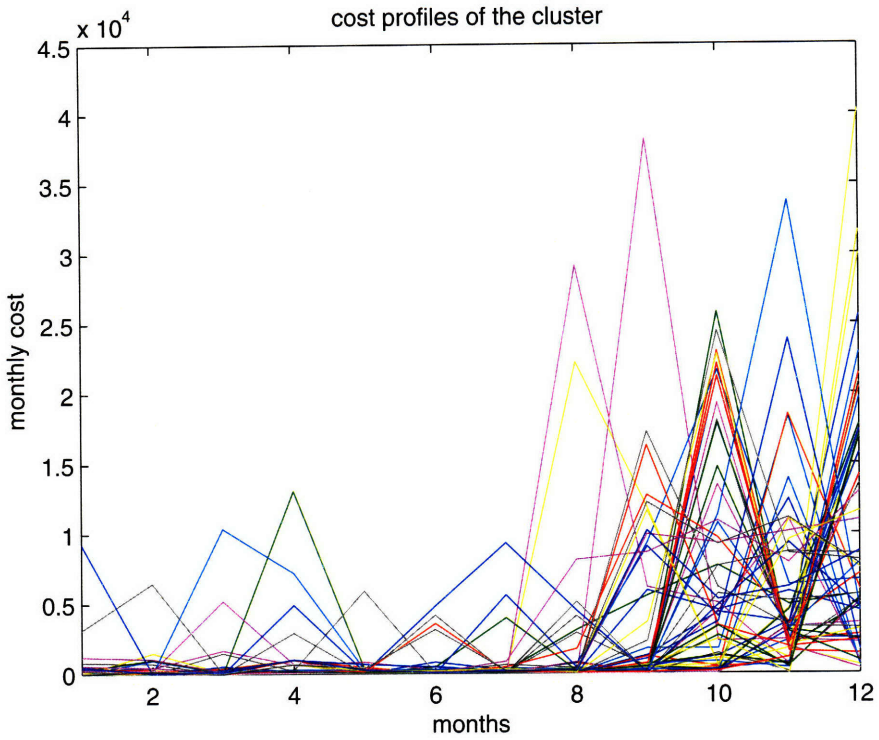

Figure 2-3: The monthly costs of the last 12 months of the observation period for all members of a cost similar cluster.

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 have lower health care costs in the following year.

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

Table 2.10: Some of the features that distinguish between cost similar members and separates them into two medical sub clusters. 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.

2.4 Results

2.4.1 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 had found the right tree for each error measure we used it to classify the testing sample and we report those results. Similarly we ran the clustering algorithm. 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.

	Hit Ratio (%)			Penalty Error			APE ($)		
Bucket	Trees	Cluster	BL	Trees	Cluster	BL	Trees	Cluster	BL
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

Table 2.11: The Resulting Performance Measures. The top line shows the measures for the whole population, followed by the measures broken down by the observation's last twelve months cost buckets, for the classification tree algorithm, clustering algorithm and the base line (BL) methodology.

Table 2.11 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 error is $2,243. There is a considerable improvement in all the performance measures compared to the baseline methodology, particularly in the absolute prediction error where the improvement 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. In comparison with the classification tree algorithm, the results are comparable, 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 top bucket. The clustering algorithm similarly more than doubles the hit ratio by, decreases the penalty error by more than 35% and the decrease in average absolute prediction error is over 58% for the top 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.

2.4.2 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 2.12 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 low 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, models with interpretable medical details are preferred. In other cases, simpler models, that achieve good results using only 22 cost variables, compared to close to 1500 medical variables, may be preferred.

	Hit Ratio (%)			Penalty Error			APE($)		
Bucket	Trees	Cluster	BL	Trees	Cluster	BL	Trees	Cluster	BL
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

Table 2.12: The Resulting performance measures using cost information only. The top line shows the measures for the whole population, followed by the measures broken down by the observation's last twelve months cost buckets, for the classification tree algorithm, clustering algorithm and the base line (BL) methodology.

2.4.3 Comparison with Other Studies

We start by noting that, comparisons across studies that use different data sets, are not fully valid as 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. [21] reports an average absolute prediction error of 93% of the actual mean, and [54] 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, lower than in the two other studies.

Traditionally prediction software have aimed to minimize R^2. Cumming *et al.* [21] 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 2.13. In the top row of Table 2.14 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 this point we have included in Tables 2.13 and 2.14 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 one, 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, 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, most significantly for the most expensive members.

	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%

Table 2.13: The R^2, and R^2_{100} for the two algorithms and the baseline. Rows 1 through 5 show the relative reduction in the denominator sum for each cost bucket.

	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

Table 2.14: The $|R|$, and $|R_{100}|$ for the two algorithms and the baseline. Rows 1 through 5 show the relative reduction in the denominator sum for each cost bucket.

2.4.4 Summary of Results

In summary, we observe that the 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, while 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 is 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, identifying one as higher risk cancer patients, predicting cost bucket 4 while predicting cost bucket 1 for the patients with musculoskeletal and orthopedic characteristics. Where 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 analysis of the R^2 measure and $|R|$ showed improved predictions for all cost buckets.

2.5 Algorithmic Discovery of Medical Knowledge

New medical information is often obtained through small controled studies tht focus on few detailed factors rather than the big picture. Large-scale datasets coupled with advanced algorithm have to potential to discover information, that is only visible with large populations. Through our work on cost prediction, we have identified

opportunities for medical discovery, using an adopted version of Eigencluster. We applied Eigencluster to selected subgroups of the populations and analyzed the results. The algorithm can take a global view of the data and identify new patterns and therefore reveal unexpected associations among diagnoses, procedures and drugs. In this Chapter, we demonstrate through two examples how medical insights can be extracted from the data.

2.5.1 Association of Estrogens with Antidepressants

The clustering algorithm has the ability to take a global view of the data and potentially identify new patterns in the data. In this example, we ran the algorithm on all women in our sample between 45 and 65 years of age. When interpreting the resulting clustering a strong association between estrogens and antidepressants was observed. Information on a cluster consisting of 26,651 members that demonstrates this relationship is shown in Table 2.15.

Prevalence	Description
50%	Estrogens & comb.
32%	Antidepressants
28%	Antihyperlipidemic drugs
27%	Hypertension
22%	Ace inhibitors & comb.
20%	Beta blockers & comb.
20%	Thyroid agents / hormones
16%	Gastrointestinal drugs
14%	Calcium channel blockers & comb.
11%	Cardiac drugs

Table 2.15: Antidepressants and estrogen cluster, the top ten most distinguishing features of the cluster. We note that both antidepressants and estrogens are among the distinguishing features. This pattern of estrogen coupled with antidepressants (and in some cases depression) was repeated throughout the clustering and led us to our analysis of the relationship between estrogens and antidepressants.

After noting this relationship, we analyzed our dataset and found that the probability of a member being on an antidepressant goes up by 166% [2] if we know the member is

[2]In our data we are unable to verify that all members have complete pharmacy claims data. For comparison, we excluded all members whom we could not verify having complete pharmacy claims,

taking estrogens, compared to members not taking estrogens. The difference in the number taking antidepressants was assessed by a z test and we found the difference to be statistically significant at the $p < 0.001$ level. [3] Our observation is consistent with recent reports of relationships between osteoporosis and depression [20, 56, 8].

2.5.2 Association of Nonsteroidal Anti-Inflammatory Agents and Increasing Costs

Table 2.16 shows information on two additional clusters that both contain cardiac patients from a general study that included numerous clusters. The members of the first cluster have significantly higher health care costs (bucket 4) in the result period, while members of the second cluster have significantly lower health care costs (bucket 2). Comparing the characteristics of each cluster, we observed that they are very similar except for the fifth factor of the first cluster, the presence of nonsteroidal anti-inflammatory agents[4].

This observation, together with recent reports of increased cardiac risk in patients taking Cox 2 inhibitors, led us to analyze the costs and cardiac outcomes of members taking Cox 2 drugs. We compared overall costs and the prevalence of CAD between patients of the study sample who had taken only Cox 1 drugs and patients who had taken only Cox 2 drugs, as presented in Table 2.17. Differences in Coronary Artery Disease prevalence were significant at the p<0.001 level. Since Cox 1 and Cox 2 are prescribed for the same condition, we expect the differences to be explained by the drugs class.

---

in which case the increase is 33%. In the authors opinion the true increase is somewhere between the two.

[3]We used a z test for differences in proportions. The p value was <0.001, both for the whole sample, and excluding all members that could not be verified having complete pharmacy information.

[4]There are two additional factors that are also not the same in the two clusters, Opiates and GI drugs. Those factors appeared frequently in clusters in a nonspecific fashion - their use was common to many conditions. On the other hand, except for musculoskeletal clusters, the association between anti-inflammatory drugs and medical conditions not related to the musculoskeletal system was limited to this association with cardiac events.

[5]The cost has been adjusted for the cost of the Cox 1 and Cox 2 prescriptions.

Cluster 1	
Prevalence	Description
41%	Antihyperlipidemic drugs
38%	Ace inhibitors & comb.
34%	Calcium channel blockers&comb.
33%	Hypertension
25%	Nonsteroidal /Anti-Inflam. agent
25%	Opiate agonists
24%	Beta blockers & comb.
13%	Cardiac drugs
13%	Hypotensive agents
13%	Antidiabetic agents, misc.
Cluster 2	
Prevalence	Description
44%	Antihyperlipidemic drugs
37%	Hypertension
33%	Ace inhibitors & comb.
31%	Beta blockers & comb.
30%	Chest pain
27%	Gastrointestinal drugs
27%	CAD
23%	Calcium channel blockers&comb.
22%	Diabetes mellitus
15%	Cardiac drugs

Table 2.16: Top ten most distinguishing features of two clusters of cardiac patients. Cluster 1 contains members that have significantly higher health care costs in the result period compared to the observation period while Cluster 2's members have significantly lower health care costs in the observation period. A distinguishing feature of a cluster is a diagnosis/procedure or drug that has more prevalence among members of the cluster than in the general population.

	Cox 1	Cox 2	p-value
Number of Members	30,277	11,046	
Average Cost[5]	\$4,954	\$8,306	<0.001 [6]
CAD Prevalence	0.05	0.09	<0.001 [7]

Table 2.17: Comparison of costs and cardiac outcomes in the observation period for members taking Cox 1 and Cox 2.

[6]The difference in the average cost was assessed by a t test and a z test as described in [65]. Both were significant at the $p<0.001$ level.

2.6 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 prediction of health care costs. We also argue that R^2, that has traditionally been used to report prediction accuracy has some 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 most dense medical data are available. The algorithms can be used for cost predictions for individuals and groups and as a base for patient intervention in health care management. Future research that builds on these algorithms could be used for financial reimbursement/insurance pricing purposes, but such an effort requires greater integration with health care economics, and system design.

The clustering algorithm discovered autonomously that cardiac patients taking non-steroidal anti-inflammatory agents have higher costs in the following year. Moreover identifies a suspected link between osteoporosis and depression The algorithm may reveal unexpected associations among diagnoses, procedures and drugs and may identify potential safety issues with drugs in common use. In general, the algorithm reveals associations, but further studies would be needed to establish causality. Effort should also be put on automating the disocvery process, which is a current undertaking of the author.

[7]The difference in the number with CAD coding was assessed by a z test.

2.A Appendix - Detailed List of Variables and Examples of Coding Groups

Table 2.1 on page 27 summarizes the variables used in our study, what follows is a more detailed list.

1. Variables 1 thought 218 are the number of claims for a member, that have coding belonging to different diagnosis groups. A diagnosis groups is a collection of ICD-9 codes that have been put together to form a group. As an example, Table 2.18 in Appendix 2.D shows all ICD-9 codes that fall into the diabetic group.

2. Variables 219 through 398 are the number of claims for a member, that have coding belonging to different procedure groups. These groups were created in the same fashion as the diagnosis groups. Table 2.19 in Appendix 2.D contains an example of all ICD-9 codes that fall into our MRI-scan group.

3. Variables 399 through 734 are the number of claims for a member, that have coding belonging to different drug groups. The grouping is bases on the NDC codes, and drugs from the same drug class are grouped together. Table 2.20 in Appendix 2.D contains an example of all drugs in the Insulin group.

4. Variables 7350 through 1485 are indicators of additional specified risk factors. We can divide the risk factors into 4 main categories:

   (a) Interaction between illnesses, an example is: diabetes and obesity.

   (b) Interaction between diagnosis an age, an example is: CAD and age above 65.

   (c) Noncompliance to treatment an example is: a pattern of ER care without office visits.

   (d) Illness severity, an example is: diabetes with foot ulcers.

5. Variables 1486 through 1489 are counts of the number of different diagnosis, procedures, drugs and risk factor a member has during the observation period.

6. Variables 1490 through 1511 are cost variables. Those are:

   (a) Monthly costs of the last twelve months of the observation period (12 variables).

   (b) Overall pharmacy costs (1 variable).

   (c) Overall medical costs (1 variable).

   (d) Overall cost (the sum of medical and pharmacy costs) (1 variable).

   (e) Overall cost in the last 6 months of the observation period (1 variable).

   (f) Overall cost in the last 3 months of the observation period (1 variable).

   (g) Positive and negative trends, found by fitting a line through the last monthly costs of the observation period (2 variables).

   (h) Acute indicator, a indicator variable found by comparing the highest month with the average monthly cost. If these are significantly different, the indicator takes on the value 1. (1 variable)

   (i) Number of months above average. This variable is an indicator of the shape of the cost profile. If the cost is relatively constant over the period, this variable takes on a value around six, which is an indicator for a chronic cost profile (1 variable).

   (j) The cost of the highest month in the observation period.

7. Variable 1512 is an indicator variable for the gender being female.

8. Variable 1513 is the age of the member at the beginning of the observation period.

2.B Appendix - Classification Trees

Classification trees recursively partition the independent variable space into a set of subspaces and assign a separate classification rule to each subspace. This partitioning can be represented as a tree. We start with the whole sample space at a root node, and then partition the data set into two subsets according to a splitting rule designed to minimize a node impurity measure that has been defined. This first split is shown in Figure 2-4. The process then continues dividing up the subspaces, until a defined stopping criteria is satisfied.

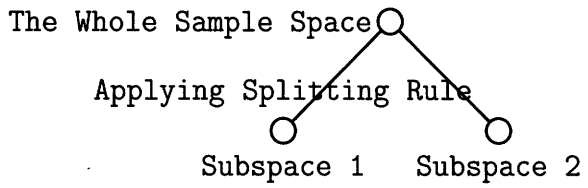

Figure 2-4: The first step in recursive partitioning, creating the first two sub nodes.

We used the software CRUISE [2] to create our classification trees. What follow is a discussion on some of the specifics of the algorithm, we refer the reader to [2] and [40] for further details.

Notation

Let X be a vector of independent variables and Y be a categorical dependent variable that takes k different values. Let t be any node in a classification tree and let p_{it} be the fraction of the observations for which $y = i$ at node t. Let N_t be the collection of observations at node t.

Node impurity measure

The most common node impurity measure, and the one that we use, is the Gini-index, which is defined as

$$1 - \sum_{i=1}^{k} p_{it}^2,$$

which can be rewritten as

$$\sum_{i \neq j} p_{it} \cdot p_{jt}.$$

In the case of uneven misclassification costs, as in the case of the penalty matrix error measure, the Gini-index is adjusted to

$$\sum_{i \neq j} W(i,j) p_{it} \cdot p_{jt}$$

where $W(i,j)$ is the cost of misclassifying as i a class j case. Note that if all the observations belong to a single class, then the Gini-index is zero.

Splitting rule

The split is chosen to give the largest reduction in the defined node impurity measure. The split can either rely on a single variable or multiple variables. To preserve interpretability, we choose to use single variable splits. There are two main categories of univariate splitting methods: exhaustive search and methods that do statistical hypothesis tests at each node to assess the significance of a split.

Loah and Shih [46] show that the key to avoiding selection bias is the separation of variable selection from split point selection. This separation differs from the exhaustive search approach of simultaneously finding the variable to split on and the splitting value. It has also been noted that exhaustive search methods are biased toward categorical variables over numerical variables as well as toward continuous variables over discrete variables, because the continuous variables afford more splits. An algorithm that overcomes this bias is 1D as described in [40], where at each node an analysis of variance (ANOVA) F-statistic is calculated for each variable. The variable with the largest F-statistic is selected, and linear discriminant analysis is applied to it to find the split value. This is the algorithm used in this chapter.

Classification rule at end nodes

At an end node, the class that minimizes a error measure is assigned to the node. For example, for our penalty error we assign class i to an end node t if i minimizes

$$\sum_{j \in N_t} P(i, y_j),$$

where P is the penalty matrix defined in Section 2.2.4 and y_j is the observed class of member j. In the case when the costs for all misclassifications are equal, the assignment rule simplifies to assign the most frequently observed class at an end node as the classification rule. For the average absolute dollar error we assign the median cost of the learning sample at each node.

Pruning and selecting the "best" tree

At some point, the recursive partitioning needs to be stopped. This stopping point can be predefined, limited by the number of levels, minimum number of observations at a node, or when improvement in the node impurity measure is negligible. A methodology introduced by Breiman *et al.* [15], which is used here, is to overgrow the tree and then prune it back using the pruning sample. After overgrowing the tree, the classification rule is applied to the pruning sample and the misclassification cost is calculated at each node and at each parent node. We then cut off the nodes that result in the smallest increase (or the biggest decrease) in the overall misclassification cost. The result is a sequence of trees, each associated with a certain misclassification cost. The tree with the smallest misclassification cost is called 0-SE [15]. The smallest tree within one standard error of the minimum is called 1-SE and is the tree we choose to use.

2.C Appendix - Clustering

This Appendix explains how the Eigencluster algorithm was adapted for medical data mining. For details on the original version of Eigencluster we refer the reader to [19, 38].

2.C.1 Notation and Outline

Given a learning set L and a validation set V of patients for whom we know the result period's cost, we predict the cost for the test set T according to the below procedure. We view each member as a vector of features, which belong in one of two categories – cost and medical. Let L_i, V_i, T_i be the subset of members in the learning, validation and test set respectivly that belong to bucket i in the the observation period. The procedure is as follows, and applies to any measure (the hit ratio, the penalty error, the average absolute prediction error, etc.).

1. Define feature weights
2. Apply feature weights to $L_i \cup V_i \cup T_i$.
3. Use EigenCluster to cluster $L_i \cup V_i \cup T_i$ based only on cost features. Let R_i be the resulting hierarchical clustering tree.
4. Using L_i, and V_i compute the frontier F_i of R_i for which clustering based on medical information is at least as good as clustering based on cost information.
5. For each node C in F_i, let x be the single prediction that optimizes the sum of the measure on $C \cap (L_i \cup V_i)$. Use x as a prediction for each test member in $C \cap T_i$.

Below we briefly discuss each of the outlined steps.

2.C.2 Define Feature Weights

We define two sets of weights: cost weights and medical weights. The cost weights apply to the cost features, whereas the medical weights apply to the medical features. As the last months of the observation period have stronger correlation with the result period, the last months are given more weight than the first. Equal weight is given to each of the medical features.

2.C.3 Applying Weights

For every member u (vector of features) in $L_i \cup V_i \cup T_i$, we apply the weights $\vec{w}$ by setting $u_i \leftarrow \sqrt{w_i} u_i$. Thus, the inner product between members is now the weighted similarity.

2.C.4 Using EigenCluster

The goal in applying EigenCluster is to put members who have similar cost patterns together in a "cluster". The hypothesis is that members with similar cost patterns in the observation period will have similar future cost patterns. In each "cluster", there will be members of the learning, validation and test sets. Thus, we will make a prediction for each of the members of the test set based on the result period behavior of the learning set and validation sets.

Technical details

We apply EigenCluster to the set of members $L_i \cup V_i \cup T_i$, where each member is only described by cost features. The result is a hierarchical clustering tree R_i. Each node is a subset of the members and the root node is the entire set $(L_i \cup V_i \cup T_i)$. Each interior node has two child nodes, whose subsets comprise the subset at the parent node. Each leaf node (a node with no child nodes) represents a subset of size at least 50.

2.C.5 Compute the Frontier

We would like to make predictions that are based on medical information, as well as based on cost information. It appears that cost information can distinguish members with different result period costs at a coarse level, but medical information cannot. On the other hand, medical information can distinguish members with different result period costs at a more fine level, whereas cost patterns cannot. This is the motivation behind the frontier – the "coarsest" level at which medical information can distinguish members.

The frontier consists of nodes in R_i for which we can improve the clustering using medical information, that is the resulting prediction is at least as good as if we had clustered those nodes further using cost information. We describe next how to compute this frontier.

Technical details

We walk up the tree R_i and apply EigenCluster to the member subset at each interior node, but only using medical features. Suppose we are at some interior node, and c_1 and c_2 are the best error measures for our child nodes C_1 and C_2 as determined by the learning sample ($L_i \cap C_1$ and $L_i \cap C_2$) and applied to the validation sample ($V_i \cap C_1$ and $V_i \cap C_2$). We apply EigenCluster to the subset at our current node to obtain a hierarchical clustering tree $\hat{R}$. For every leaf node $\hat{C}$ in $\hat{R}$, we compute the single answer $\hat{x}$ that optimizes the sum of the error measure on $\hat{C} \cap L_i$ and apply it to the validation sample $\hat{C} \cap V_i$. Let $\hat{c}$ be the cost incurred by $\hat{x}$. If $\hat{c}$, summed over all leaf nodes is more optimal than the sum of c_1 and c_2, we designate this interior node to lie on the frontier F_i, replace its subtree with $\hat{R}$, set its cost to be the sum of the $\hat{c}$'s and continue up R_i. After we have walked up the whole tree we have replaced parts of the tree R_i, which was built using cost information only, with number of new subtrees $\hat{R}_i$'s that use medical information and improve prediction.

2.C.6 Prediction

Each leaf node contains a subset of patients. Roughly two thirds are in the learning and validation sets, and the a third are in the test set. The idea is that every member of this node is similar – otherwise they would not be put in the same node. Therefore, it is natural to think that the result period behavior of the patients in the learning and validation sets is similar to the result period behavior of the patients in the test set. This motivates our prediction technique, described below.

Technical details

We now have at our disposal leaf nodes $C_1, \ldots, C_n$. Each C consists of members of the learning, validation and the test sets. We compute the answers x_j that optimizes the sum of the measure on $C \cap L_i \cap V_i$, and use this as a prediction for each member in $C \cap T_i$.

2.D Examples of Group Coding

Table 2.18

ICD-9 Code	Description
250	Diabetes Mellitus
2500	Diabetes Mellitus without complications
2500x	Diabetes Mellitus without complications
2501	Diabetes with Ketoacidosis
2501x	Diabetes with Ketoacidosis
2502	Diabetes with Hyperosmolarity
2502x	Diabetes with Hyperosmolarity
2503	Diabetes with Coma
2503x	Diabetes with Coma
2504	Diabetes with Renal Manifestations
2504x	Diabetes with Renal Manifestations
2505	Diabetes with Ophthalmic Manifestations
2505x	Diabetes with Ophthalmic Manifestations
2506	Diabetes with Neurological Manifestations
2506x	Diabetes with Neurological Manifestations
2507	Diabetes with Peripheral Circulatory Disorders
2507x	Diabetes with Peripheral Circulatory Disorders
2508	Diabetes with Manifestations
2508x	Diabetic Hypoglycemia
2509	Diabetes with Complication

Continued on next page...

Table 2.18 – Continued

ICD-9 Code	Description
2509x	Diabetes with Complication
3572	Polyneuropathy in Diabetes
3620	Diabetic Retinopathy
36201	Background Diabetic Retinopathy
36202	Proliferative Diabetic Retinopathy
36203	Nonproliferative Diabetic Retinopathy
36204	Mild Nonproliferative Diabetic Retinopathy
36205	Moderate Nonproliferative Diabetic Retinopathy
36206	Severe Nonproliferative Diabetic Retinopathy
36207	Diabetic Macular Edema
36641	Diabetic Cataract
6480	Diabetes Mellitus - Complications of Delivery
6480x	Diabetes Mellitus - Complications of Delivery
V4585	Insulin Pump Status
V5391	Fitting/Adjust Insulin Pump

Table 2.18 An example of ICD-9 diagnosis codes in a diabetes diagnosis group ("x" at the end of a code stands for any number.

Table 2.19

Code	Description	Code Origin
0159T	Computer-aided detection, including computer algorithm analysis of MRI	CPT4
0160T	Therapeutic repetitive transcranial magnetic stimulation treatment pla	CPT4
70336	Magnetic Image, Jaw Joint	CPT4
7054x	MRI of Face, Neck and Head	CPT4

Continued on next page...

Table 2.19 – Continued

Code	Description	Code Origin
7055x	MRI of the Brain	CPT4
7155x	MRI Chest	CPT4
7214x	MRI Neck, Lumbar or Chest Spine	CPT4
72150	Magnetic Resonance (proton)	CPT4
72156	MRI (proton) of Chest, Lumbar or Angio Spine W/O&w Dye	CPT4
7219x	MRI Pelvis	CPT4
73218/9	MRI Upper Extremity	CPT4
7322x	MRI Uppr Extremity	CPT4
73718/9	MRI Lower Extremity	CPT4
7372x	MRI Lower Extremity	CPT4
7418x	MRI Abdomen	CPT4
7555x	Heart/Cardiac MRI	CPT4
76093	Magnetic Image, Breast	CPT4
76094	Magnetic Image, Both Breasts	CPT4
76394	MRI for Tissue Ablation	CPT4
76400	Magnetic Image, Bone Marrow	CPT4
76498	MRI Procedure	CPT4
7702x	Magnetic resonance guidance	CPT4
77084	Magnetic resonance (eg, proton) imaging, bone marrow blood supply	CPT4
C8903-8	MRI , Breast	HCPCS
C9723	Dynamic Infrared Blood Perfusion Imaging	HCPCS
Q0070	Magnetic Image, Spine	HCPCS
I8891	MRI of Brain & Brainstem	ICD9
I8892	MRI Chest & Heart	ICD9
I8893	MRI Spinal Canal	ICD9

Continued on next page...

Table 2.19 – Continued

Code	Description	Code Origin
I8894	MRI Musculoskeletal	ICD9
I8895	MRI Pelvis,prostate,bladder	ICD9
I8896	Other Intraoperative Magnetic Resonance Imaging	ICD9
I8897	Magnetic Resonance Image Unspecified	ICD9
I8899	Unspecified MRI	ICD9
R483	MRI	Rev Code
R61x	MRI	Rev Code

Table 2.19 An example of procedure codes in a procedure group. The table displays all codes within the MRI-scan group ("x" at the end of a code stands for any number). In general the codes in a procedure group come from various sources: ICD-9, DRG, Rev Coding, CPT4 and HCPCS.

Table 2.20

NDC Code	NDC Description	Rx10 Description
00003378015	Insulin	Insulins
00069006119	Exubera Chamber	Insulins
00069009741	Exubera Release Unit	Insulins
00002811201	Iletin Ii Pzi Beef	Insulin - Beef
00002821201	Iletin Ii Reg. Beef	Insulin - Beef
00002831201	Iletin Ii Nph Beef	Insulin - Beef
00002841201	Iletin Ii Lente Beef	Insulin - Beef
00003244510	Insulin, Purified Ultralente Beef	Insulin - Beef
00169352215	Insulin Standard Nph	Insulin - Beef
00169352815	Insulin Standard Lente	Insulin - Beef
00169357805	Insulin Standard Semilente	Insulin - Beef
00169357215	Insulin Standard Ultralente	Insulin - Beef

Continued on next page...

Table 2.20 – Continued

NDC Code	NDC Description	Rx10 Description
00002811101	Iletin Ii Pzi Pork	Insulin - Pork
00002821101	Iletin Ii Regular Pork	Insulin - Pork
00002831101	Iletin Ii Nph Pork	Insulin - Pork
00002841101	Iletin Ii Lente Pork	Insulin - Pork
00002850001	Iletin Ii Regular Pork	Insulin - Pork
00003244110	Insulin, Purified Semilente Pork	Insulin - Pork
00169010001	Insulin Purified	Insulin - Pork
00169020001	Insulin Purified	Insulin - Pork
00169030001	Insulin Purified	Insulin - Pork
00169244010	Insulin Purified Regular Pork	Insulin - Pork
00169244210	Insulin Purified Lente Pork	Insulin - Pork
00169244710	Insulin Purified Nph Pork	Insulin - Pork
00169351215	Insulin Standard Regular	Insulin - Pork
54569165200	Iletin Ii Reg. Pork	Insulin - Pork
54569165202	Iletin Ii Reg. Pork	Insulin - Pork
54569281600	Insulin Purified Lente Pork	Insulin - Pork
54569281700	Insulin Purified Regular Pork	Insulin - Pork
54569289100	Iletin Pork Nph	Insulin - Pork
54569289101	Iletin Pork Nph	Insulin - Pork
00002811001	Iletin Pzi	Insulin - Beef & Pork
00002824001	Iletin Regular I	Insulin - Beef & Pork
00002831001	Iletin Nph I	Insulin - Beef & Pork
00002844001	Iletin Lente I	Insulin - Beef & Pork
00002851001	Iletin Semilente	Insulin - Beef & Pork
00002864001	Iletin Ultralente	Insulin - Beef & Pork
54569165101	Iletin Nph I	Insulin - Beef & Pork
54569165102	Iletin Nph I	Insulin - Beef & Pork

Continued on next page...

Table 2.20 – Continued

NDC Code	NDC Description	Rx10 Description
54569295100	Iletin Regular I	Insulin - Beef & Pork
54569295101	Iletin Regular I	Insulin - Beef & Pork
54868142801	Iletin Nph I	Insulin - Beef & Pork
54868208901	Iletin Regular I	Insulin - Beef & Pork
00002751001	Humalog	Insulin - Human
00002751101	Humalog Mix 75/25	Insulin - Human
00002751559	Humalog	Insulin - Human
00002751659	Humalog	Insulin - Human
00002821501	Humulin R	Insulin - Human
00002821601	Humulin Br	Insulin - Human
68115083910	Lantus	Insulin - Human

Table 2.20 An example of drugs in a drug group. The table contains examples of drugs that belong to the Insulin group, as well as their NDC codes.

Chapter 3

Drug Surveillance

The Food and Drug Administration is responsible for the evaluation of new drugs. A large part of their role is to ensure drug safety. After a drug is approved (if the FDA approves it as safe and effective), the drug maker is responsible for reporting any adverse drug events it learns of to the FDA. In addition, the FDA runs a post-FDA-approval drug surveillance system called the Adverse Event Reporting System (AERS), which is a voluntary system in which patients and health care professionals can submit reports of adverse events. Although the system has often proved useful in identifying serious side effects of drugs, it has been insufficient in identifying potential safety signals, especially since events that can be indicators of increased risk of common conditions might not be considered important by individual patients or health care professionals.

A safety concern raised through the AERS system, or through different channels such as the medical literature, results in warnings to consumers or in serious cases a withdrawal of the drug. A black box warning is the strongest warning the FDA can require. The warning appears on the package insert in a black box stating that a drug can potentially have serious or life-threatening adverse effects. Among some of the more widely covered warnings in recent years are the following:

- The FDA has required that black box warnings be placed on all antidepressant

medications stating they may result in increased risk of suicidal tendencies in children and adolescents.

- FDA advisors have recommended that Pfizer be required to place a black box warning on their Non-Steriodal Anti-Inflammatory (NSAID) drug celecoxib (U.S. trade name Celebrex).

- In November 2004, the FDA required a black box warning on the Depo-Provera contraceptive injection, due to the risk of significant loss of bone density with long-term use.

- In October 2006, the FDA added a black box warning to the anticoagulant warfarin due to the risk of bleeding to death.

- In November 2007, the FDA added a black box warning to the diabetes medication Avandia, citing the risk of heart failure or heart attack to patients with underlying heart disease, or those at a high risk for heart attack.

After the withdrawal of rofecoxib (known outside the medical profession under the marketing name of Vioxx) from the pharmaceutical market in 2004, post-FDA-approval drug safety and surveillance has come under serious scrutiny. A 2006 study by the Institute of Medicine pointed out that efforts to monitor risk-benefits tradeoff of medications decrease after FDA approval and that this issue needs to be addressed [55]. Large claims databases, with near-real time[1] information coupled with advanced statistical models have the potential to greatly improve post-approval drug surveillance, by analyzing individual outcomes in a very large population. Events that on individual level might not look significant (and therefore not get reported to the AERS) can be serious risk indicators when aggregated over large populations. Such approaches can not only discover side effects but also potential added benefits.

[1]One can expect about a three month delay for claims processing and data processing.

3.1 Previous Work

To the best of our knowledge, there is not a large literature on real-time drug surveillance. The most recent [16] is the first paper to attempt detection of adverse drug events within a population as a function of time, using claims data. The major drawback of the paper is that the authors do not test across all possible adverse events, but rather just for a single known side effect and one other effect as a control. They successfully show that adverse events can be detected. Moreover, studies have also started to analyze what a post-marketing surveillance system should entail [11]. A couple of studies focus on vaccine surveillance and monitoring of the vaccines known side effects. The most recent study [45] uses maximized sequential probability ratio testing to test for known side effects, the same method is used in the most recent drug surveillance study [16].

Numerous studies use claims data to investigate drug effects. [34], led one of its authors to testify before the US Senate in the rofecoxib case, as the study drew attention to how much and when the FDA knew about the increased cardiac risk for patients taking rofecoxib. That study and others [58, 33] demonstrate that claims data can be an effective means to monitor drugs and their side effects.

3.2 Overview, Terminology and Notation

According to the FDA's Orange book [7], over 2000 different (or different combinations of) approved pharmaceuticals are marketed under over 5500 different names. Many of the drugs are commonly used, while others are rarely found in claims databases. In this thesis we focus our attention to drugs in common use, although much of the methodology is transferable to less common drugs (in some cases more direct methods, such as randomized trials, may be more appropriate). We choose to focus on these drugs as they play to the strength of claims data, its near immediate availability and large size, which allows us to detect changes in occurrences of rare events,

relatively fast.

Drugs are most commonly prescribed for a single reason or diagnosis. Often there is a choice of more than one drug to treat the condition; for example, atorvastatin (marketed under the name Lipitor) and simvastatin (marketed under the name Zocor) are both drugs that treat high cholesterol. These drugs are very similar [66], and there are no clinical reasons why a doctor would prescribe one over the other[2]. In this case, we can compare populations on one drug to populations taking another. If a drug has quickly become the treatment of choice, a comparison group might not be available, as similarly sick members are not being treated without the drug. With access to older data, we can build a comparison group from previously treated patients and compare the outcomes of similar patients who were treated before the new drug became popular. Care needs to taken to make sure that changes in coding behavior do not affect the outcome, but the same methodology applies. In this chapter we propose different methods to adjust for differences in populations and suggest a methodology to compare the two populations once selected.

We will call the members taking the drug of interest a *treatment group*, and the group of members that we compare them with the *control group*. The *exposure* of a member to a drug, is the time a member is taking the drug. An adverse event can be an onset of a disease such as asthma or a single event, for instance stroke. We define adverse events in terms of the members' claims data, for example by counting the number of claims associated with a particular adverse event. We define the *post-toxicity time period* as the time after a member stops taking a treatment drug until its toxicity no longer affects him or her, which varies from drug to drug.

We consider one-month time intervals, as that is the most common rate of claims

[2]The main clinical advantage of atorvastatin over simvastatin is that it is not metabolized by certain liver enzymes, and thus its blood concentration is not increased when combined with grapefruit juice which inhibits these enzymes. Simvastatin patients should therefore avoid drinking large amounts of grapefruit juice for this reason.

data updates. We define p_i as the probability that an adverse event i happens in any particular month, for any member.

3.3 Selecting Comparison Populations

When we compare effects of a drug, we need to select a comparison population to serve as a baseline. Ideally this population is very similar to the treatment population in terms of age, gender, disease burden and other important factors for medical outcomes. Carefully choosing the comparison group is vital to an efficient drug surveillance system, as a wrong baseline will result in a drift away from the expected number of events, resulting in multiple false alarms. When there is an alternative drug given for the same condition as the treatment drug in question, such as in the case of simvastatin and atorvastatin, the two populations are very similar and can possibly be used without further adjustments. In other cases adjustments are needed. Below we discuss methods for making the adjustments.

3.3.1 General Methods

When a similar drug to the one under study is on the market, but the groups do not fully overlap (for example, in the case of rofecoxib and naproxen; naproxen is more common in children than rofecoxib), simple adjustment methods can be used to select the comparison population. These include

- Randomly select members from the comparison population so that the age and gender (and even cost-bucket) distribution is similar.

- Estimate the baseline rates for each of the sub-populations and combine them in an estimate reflecting the age and gender (and cost) distribution of the treatment population.

Section 3.7 reports on how these methods work with real data. Previously, researchers have used comorbidity-scores and propensity-scores to adjust for differences in the two

populations. We found the simple selection methods mentioned above to be more successful, in our experiments.

3.3.2 Maximal Pairing

Perhaps the most accurate way to select a comparison group is to select, for each member in the treatment group, a member (or members) in the control group with very similar pre-treatment medical history and conditions. To do so, we can solve a maximum (multiple) pairing optimization problem, in which constraints are placed on age, gender, important conditions and risk factors as well as overall medical similarity.

We propose the following formulation to measure medical similarity. Let $x_i(t)$ be a vector of medical features (diagnosis, procedures and drugs) for member i, at time t. In particular let $x_{ij}(t)$ be the number of days with claims that include code (or code group) j, for some pre-specified time prior to starting on a drug. Let w be a weight vector, that weights the features to account for different importance of different codes. For example, we place a greater weight on cancer conditions than on a common cold, or a routine doctor's visit. We define a rescaled vector $\bar{x}_i(t)$ for member i as:

$$\bar{x}_i(t) = \frac{w' x_i(t)}{\sqrt{|w' x_i(t)|}},$$

and the medical similarity of any two members as the inner product of their rescaled vectors. This way, the weighted medical similarity of a member with himself is one and the weighted similarity between any two members i and k is between zero and one.

We can now write a standard optimization problem, that will maximize the number of control-treatment member pairs in the study. Let y_{ik} be an indicator variable that equals one if members i and k being paired up for the study, and is zero otherwise.

The standard maximum pairing problem is then written as,

$$\begin{aligned}
\text{maximize} \quad & \sum_i \sum_k y_{ik} & \\
\text{subject to} \quad & \sum_k y_{ik} \leq 1 & \text{for all } i \\
& \sum_i y_{ik} \leq 1 & \text{for all } k \\
& y_{ik} \in \{0,1\} & \text{for all } i \text{ and } k.
\end{aligned}$$

To reduce the size of the problem (recall that our treatment and control populations might be on the order of 100,000, and therefore the number of possible pairs of the order of $100,000^2$) we include only those y_{ik} in the input that satisfy the medical similarity constraints, as well as age, gender and risk factor requirements. The formulation can easily be extended to allow for multiple control members to be paired with a treatment member.

One of the major drawbacks of the method, is that it cuts down on the number of members that can be used for the study significantly, especially if the population is on the smaller side (as the probability of a similar member being found in the control group is smaller). We can relax the formulation by matching members only based on known risk factors for a given condition for example age, gender, hypertension, cholesterol drugs and diabetes in the case of cardiac events (this would result in solving multiple optimization problems, one for each condition under surveillance).

Considerations when a Similar Drug does not Exist

If a similar drug exists, the medical feature vector for a member on the comparison drug has a natural "anchor" date, the day he or she starts the comparison treatment. The medical feature vector can then be built from the data over the pre-specified period prior to starting treatment. When no comparison drug exists, we need to define

a fixed date, to use as "anchor" date for the comparison population, to avoid biases in the selection process. This date can be some time (τ) in the past, which has the added benefit that we already know the post-treatment outcomes for the comparison population. Knowing the outcomes for the comparison populations results in a more stable baseline. In contrast, selecting τ as the present, we would be following two random processes, the outcomes of the treatment members on one hand and the control members on the other. This would cause more fluctuation in the estimated baseline rates.

3.3.3 Population Maximization

The idea is to select the largest comparison group possible that is similar (pre-treatment) to the treatment group, with the goal of finding stable baseline estimates. Below we assume a similar drug exists and refer the reader to the end of Section 3.3.2 for considerations when building the data vectors in the absence of a similar drug.

Let x_i be a decision variable, indicating whether or not member i of the control group is included in the study. Let a_{ij} be the value for condition j for member i. In each month we would solve the following optimization problem

$$\begin{aligned} \text{maximize} & \sum_i x_i \\ \text{subject to} & \sum_i a_{ij}x_i - (\alpha_j + \delta_j)\sum_i x_i \leq 0 \qquad \text{for all } j \\ & -\sum_i a_{ij}x_i + (\alpha_j - \delta_j)\sum_i x_i \geq 0 \qquad \text{for all } j \\ & x_i \in \{0,1\}, \end{aligned} \tag{3.1}$$

where α_j is the average diagnosis value for the treatment group for control j and δ_j is the allowed perturbation around it. The controls can include cost, age, gender, disease burden, average lenght of pre-treatment history, average time on drug, etc. From the chosen comparison group, which is similar to the treatment group prior to

drug use, we then estimate the baselines for different outcomes.

Beyond Averages

One of the pitfalls of this formulation is that it takes into account only the average, which means the distribution of a control can be quite different between the two groups. (An extreme example would be one population with very young and very old members, and the other with only middle-aged members, both groups could have the same average age.) We can account for these differences by constraining controls over subgroups. For example, we can require the same fraction of the population to be under the age of b. Let f_b be the fraction of members under b in the treatment group. Then by adding the constraints

$$\sum_{i|age_i<b} x_i - (f_b + \delta_b) \sum_i x_i \leq 0 \text{ and}$$
$$- \sum_{i|age_i<b} x_i + (f_b - \delta_b) \sum_i x_i \leq 0$$

we make sure that the fraction of members under b in the comparison group is within δ_b of the treatment group. In general, we can constraint the average of a subpopulation (S_k) to be between $\alpha_{jk} \pm \delta_{jk}$ by adding

$$\sum_{i \in S_k} a_{ij} x_i - (\alpha_{jk} + \delta_{kj}) \sum_{i \in S_k} x_i \leq 0 \text{ and}$$
$$- \sum_{i \in S_k} a_{ij} x_i + (\alpha_{jk} - \delta_{kj}) \sum_{i \in S_k} x_i \leq 0,$$

to the formulation. This formulation therefore has flexibility in terms of the restrictions it can place on the comparison population.

Accounting for Better General Health

When one population is generally healthier than the other, as we will see is the case of rofecoxib and naproxen, the optimal solution to (3.1) will tend to have the lower bounds of the diagnosis condition constraints tight. Therefore, the comparison population will be on average healthier across most if not all code groups compared to the treatment population. We can address this problem by introducing new constraints that ensure that only fraction of the conditions in question can be lower for the comparison compared to the treatment population. We suggest two ways to achieve this objective. First, we can sum up over all the conditions, the differences between the target disease burden (of the treatment group) and the actual disease burden of the comparison group and set the sum to approximately zero. Mathematically, we can write this relaxation as

$$\sum_j \sum_i \bar{a}_{ij} x_i - \left(\sum_j \bar{\alpha}_j + \delta_{tot} \right) \sum_i x_i \leq 0 \text{ and}$$
$$-\sum_j \sum_i \bar{a}_{ij} x_i + \left(\sum_j \bar{\alpha}_j - \delta_{tot} \right) \sum_i x_i \leq 0$$

where all data has been re-scaled to the same mean and standard deviation, $\bar{\alpha}_j$ is the rescaled average value for diagnosis j for the treatment population, $\bar{a}_{ij}$ is the rescaled data value of diagnosis j for the comparison population member i, and finally δ_{tot} is the flexibility we allow in the overall health.

Another approach to even the disease burden is to add additional variables to the formulation to limit the number of conditions that have higher (or lower) average values for the comparison population compared to the treatment group averages. To

achieve this, we introduce new variables and rewrite (3.1) as

$$
\begin{aligned}
\text{maximize} \sum_i x_i & \\
\text{subject to} \sum_i a_{ij} x_i - (\alpha_j + \delta_j - \delta_j y_{jL}) \sum_i x_i \leq 0 & \qquad \text{for all } j \\
-\sum_i a_{ij} x_i + (\alpha_j - \delta_j + \delta_j y_{jU}) \sum_i x_i \leq 0 & \qquad \text{for all } j \\
\sum_j y_{jL} \geq k_L & \\
\sum_j y_{jU} \geq k_U & \\
y_{jL}, y_{jU} \in \{0, 1\}, &
\end{aligned}
$$

where k_L is the number of conditions we want to ensure that the comparison population has at least the same prevalence (or higher) when compared to the treatment population and k_U is the number of conditions we want to ensure that the comparison population has at most (or lower) prevalence of the disease than the treatment group.

3.3.4 Not Adjusting the Population

So far we have discussed methods for selecting a comparison group, so that pre-treatment, the populations have the same characteristics. Another approach is not to adjust the comparison group at all. Instead of applying often complicated selection methods, we simply observe the two populations independently and analyze the changes. The intuition is that if the drug has no effect, the underlying rate of adverse events should not change, and the before and after rates should be similar. Monitoring the comparison group at the same time has the added benefit of accounting for the effect of the underlying disease as well as possible coding changes. Section 3.5 discusses how to monitor drugs without a direct comparison baseline.

3.4 Mathematical Modeling of Surveillance System with a Comparison Group Baseline

We use the comparison group to estimate the probability p_i of an adverse event i occurring for a member in any given month. This estimated probability serves as a baseline for the treatment population. If we view the population as a homogenous population, then the expected number of events in a particular month t for a treatment population of size n is $n \cdot p_i$. Assuming that events in different months are independent, then the sum of events i over τ months is a binomial random variable: $\text{Bin}(\sum_{t=1}^{\tau} n(t), p_i)$, where $n(t)$ is the number of members in the treatment population in month t.

We can reject the null hypothesis that the probability of an event equals p_i if the number of observed events falls far enough from $N_\tau = \sum_{t=1}^{\tau} n(t) \cdot p_i$, the expected number of events. In particular, for some significant level (1-α), we reject the null hypothesis if N_τ falls outside the (smallest possible) interval $[k_1, k_2]$ defined by

$$\max_{k_1} P(k \leq k_1) \leq \alpha/2,$$
$$\max_{k_2} P(k \geq k_2) \geq \alpha/2,$$

where k is $\text{Bin}(\sum_{t=1}^{\tau} n(t), p_i)$.

When n is sufficiently large and

$$p_i \cdot n \approx n \cdot p_i \cdot (1 - p_i)$$

we can approximate the binomial distribution with a Poisson distribution and build confidence intervals in the same way.

3.4.1 Poisson Approximation for Non-Homogeneous Groups

Very commonly, adverse events for example heart attacks vary greatly with both gender and age (as well as underlying health status). Therefore, the probability p_i of an event i occurring is not uniform over the group. To overcome this problem, we divide the population into subgroups. In particular if j is a subgroup of the population with n_j members and

$$p_{ij} \cdot n_j \approx n_j \cdot p_{ij} \cdot (1 - p_{ij})$$

holds for all subgroups $j = 1, \ldots, J$, then the number of events up to time t has an approximate Poisson distribution, and the expected number of events up to and including τ is equal to

$$\sum_{t=1}^{\tau} \sum_{j=1}^{J} p_{ij} \cdot n_j(t).$$

At any point in time we can therefore draw a (1-α)% confidence interval around the expected mean and reject the null hypothesis of the rates of adverse events being the same if the observed number of events is far enough from our expected number.

3.4.2 Controlling for False Positives

So far we have not taken into account that we are looking at multiple possible adverse events at the same time and therefore putting ourselves at risk for false positives. If we have defined M possible adverse events, at any point in time we expect $M \cdot \alpha$ to be outside the confidence intervals (prior to observing any data), in the absence of any true effect of the treatment. We also need to keep in mind that conditions are not independent. For example, a member who gets a lung infection is more likely to get asthma, making controlling for false positives even more important.

When some $N_i(t) > U_i(t)$, where $U_i(t)$ is the upper $(1 - \alpha/2)$ % confidence interval, there is the possibility that we have observed this increased rate by chance alone. Below we list how this issue can be addressed.

1. We can make α very small, or choose α such that we optimize the tradeoff

between the number of false positives and the detection of side effects.

2. We can start a new process that includes only new members (members starting on the drug) to avoid any biases and monitor this new process. This approach has the serious drawback of cutting down the population and could result in a very long additional ramp-up time while building up a new population.

3. We can continue to monitor the same process but define a new target Δ and target time T (based on the Poisson distribution). We accept the diagnosis as a side effect, only if the number of additional events reaches Δ on or before time T.

In our case studies we chose the third option, as our data did not allow us the luxury of restarting the process, and the third option has a lower false positive rate than the first option, when the time to discovery of a true signal is kept constant.

Medical coding is constantly changing, and medicine is often complex. Coding patterns change when reimbursement policies change. This change can affect the baseline estimates (especially those that are estimated from historical data). A two-step system allows for flexibility to investigate potential explanations behind the increase at the time of first alert, both systematic changes in coding behavior and undiscovered underlying explanatory variables for the conditions: even though we have selected a "similar" comparison group, for rare events there might be a underlying reason that is missed in the process.

3.5 Mathematical Modeling of Surveillance System without a Comparison Group Baseline

Section 3.4 discussed how to compare a treatment population to a comparison population, when the comparison population is selected to be similar to the treatment group. A different approach is to not adjust the comparison population prior to monitoring the drugs, rather accept that they may be dissimilar and follow the changes

for the two groups independently. We assume that the comparison group is large, and therefore the observed rates of the comparison population can be considered known and constant. We therefore want to monitor the change in the adverse event rate and to detect if a) it significantly increases and b) the relative change significantly differs from the change in the comparison group.

Let $p_{cb}(i)$ and $p_{ca}(i)$ be the before- and after-treatment probabilities for adverse event i for the comparison group (assumed non-random), and let $p_{tb}(i)$ and $p_{ta}(i)$ be defined similarly for the treatment group. We want to raise a safety concern, if we observe that $p_{tb}(i) < p_{ta}(i)$ and that the relative change[3] is greater for the treatment group than the control group, that is

$$\frac{p_{ca}(i) - p_{cb}(i)}{p_{cb}(i)} \leq \frac{p_{ta}(i) - p_{tb}(i)}{p_{tb}(i)}.$$

Equivalently,

$$\begin{aligned} \frac{p_{ta}(i) - p_{tb}(i)}{p_{tb}(i)} \geq \frac{p_{ca}(i) - p_{cb}(i)}{p_{cb}(i)} &\Leftrightarrow \\ \frac{p_{ta}(i)}{p_{tb}(i)} - 1 \geq \frac{p_{ca}(i)}{p_{cb}(i)} - 1 &\Leftrightarrow \\ \frac{p_{ta}(i)}{p_{tb}(i)} \geq \frac{p_{ca}(i)}{p_{cb}(i)}. & \end{aligned}$$

The resulting hypothesis test of the relative change for the treatment population being greater than that of the comparison population results in a chi-square test with one degree of freedom [14, 28]. As before we compare our outcomes to an upper $(1 - \alpha/2)\%$ confidence interval. We note that the confidence interval depends on $p_{tb}(i)$ and $p_{ta}(i)$, which are both a function of the size of the treatment population and the prevalence of the event.

[3]We chose to use relative change rather than absolute change, as the starting probabilities $(p_{cb}(i), p_{tb}(i))$ might be quite different and make a direct comparison difficult.

3.6 Practical Considerations

3.6.1 Time on Drug and Toxicity Period

The literature [16] suggests that the minimum time that a member needs to take the drug and the definition of the post-toxicity period are very important. Our case studies, however, show that the minimum time a member takes a drug is in many cases not important, and the trade-off of having a larger population by having the minimum time shorter (even down to a single prescription) makes including everyone preferable. We illustrate this point in our rofecoxib case study in Section 3.7. Intuitively, one expects the toxicity of a drug to go down as time passes after a member stops taking the drug. The exception to this phenomenon occurs when a permanent damage has been done. Due to limitations on the size of the data set, we were unable to research the effects of varying the post-toxicity period.

3.6.2 Definitions of Events

We define events using the occurrences of specific ICD-9 codes in the claims data. An event can be defined in a number of different ways.

- *Claim-line event*: Each claim line (a data entry) with the particular diagnosis code of interest is an event. This definition can result in multiple counts per days, if a diagnosis is associated with multiple procedures or financial transactions. This count can be useful as it is an indicator of the severity of the event, but on the other hand it introduces a lot of variability into the parameter estimates.

- *Claims-day event*: Each day that a diagnostic code appears in the data defines an event.

- *Claims-period event*: This event is defined as a period of time within which all occurrences of the code count as a single event. Claims-period event is parameterized by the number of consecutive days d without the diagnosis code

allowed, without ending the event. As an example, if $d = 5$ then a single event can have a 5 day long break in coding, if there are 6 or more days before a code reappears in the claims data, then the reappearance counts as a new event. It is natural to allow for some break in coding, as in many clinical cases there may be follow-up appointments, prescriptions to be filed or long-term treatment. If $d = 0$, the claims-period event definition becomes the claims-day event. If $d = T$, where T is the study horizon, the claims-period event becomes a count of first occurrences of conditions.

It is important to note that for some adverse effects, such as heart attacks, the way an event is defined has very little effect on drug surveillance. On the other hand, for other diseases it can be misleading to define events in certain ways. This difference depends on the nature of the events in question. Some events are a "one time thing," such as complications of labor; others may take a long time to resolve (resulting in multiple claims over a long period); finally, some events can be the start of a long and irreversible condition, such as the onset of Alzheimer's disease. The definition of events may therefore not be the same for every diagnosis.

3.6.3 Grouping of Codes

The ICD-9 codes are organized in a hierarchical structure, by the organ systems. One can view the structure as a tree, and every level of the tree represents an additional digit of the ICD-9 code, and the descriptions of the conditions get correspondingly more detailed. There is a significant variability in ICD-9 coding, as some health care professionals may code only to the third digit, while others to the fourth or fifth. At the same time, the further down the tree, the less common the conditions get, and therefore the harder it is to observe a significant shift in risk. Lastly, sometimes neighboring codes (that have the same parent) are very similar and can be joined into one for the purpose of drug surveillance. All these considerations should be taken into account when a drug surveillance system is implemented, and medically trained professionals who are knowledgeable about the codes are needed to construct

a good code grouping at the right coding level.

3.6.4 Stability of the Estimates

When surveiling a new drug, it takes a while to observe enough events to get a stable estimates. This is particularly true when we follow the relative change. We therefore need to wait until we have observed enough events to be able to raise a flag with some confidence. We define $\Delta(e_\tau(t))$ as the maximum fluctuation allowed in a estimator over the past τ months at time t, before a flag can be raised. In our numerical experiments, we have found that defining $\Delta(e_1(t)), \Delta(e_2(t)), \Delta(e_3(t))$, the change over the past one, two and three months to work well (using a single $\Delta(e_\tau(t))$ allows for unwanted conditions to slip through).

3.7 Case Study: Rofecoxib and Naproxen

Rofecoxib is a nonsteroidal anti-inflammatory drug (NSAID)[4] developed by Merck & Co. to treat osteoarthritis, acute pain conditions, and dysmenorrhoea. Rofecoxib was approved FDA on May 20, 1999 and was subsequently marketed under the brand names Vioxx, Ceoxx and Ceeoxx.

In November 2000, the *New England Journal of Medicine* published the VIGOR (Vioxx GI Outcomes Research) study [13]. The goal of the study was to assess whether rofecoxib is associated with a lower incidence of upper gastrointestinal events when compared to naproxen. The study showed reduced gastrointestinal events for the rofecoxib population, but also reported a four-fold increase in cardiac events. This increase, however, was deemed not statistically significant and attributed to a positive effect of naproxen. The results of the VIGOR study were submitted to the

[4]Nonsteroidal anti-inflammatory drugs (NSAIDs) come in two classes, non-selective COX-1 and selective COX-2. COX-1 mediates the synthesis of prostaglandins responsible for protection of the stomach lining, while COX-2 mediates the synthesis of prostaglandins responsible for pain and inflammation. The COX-2 therefore delivered the same pain relief, but with reduced risk of peptic ulcers.

FDA in February 2001, which led to the introduction of warnings on rofecoxib labeling concerning the increased risk of cardiovascular events (heart attack and stroke) in April 2002.

In October 2001, another study [42] was published that compared rofexocib to a placebo. This study found an elevated death rate among rofecoxib patients, although the deaths were not generally heart-related. However, the study did not find any elevated cardiovascular risk due to rofecoxib. Before 2004, Merck cited this study as evidence of rofecoxib's safety, contrary to VIGOR. In the following months, more studies and articles appeared [17, 51], and researchers debated the cardiac side effects of rofecoxib. Around the same time, another study [58] demonstrated the protective cardiac effects of naproxen compared to other NSAIDs.

In 2001, Merck started the APPROVe study, a three-year trial with the primary aim of evaluating the efficacy of rofecoxib for the prophylaxis of colorectal polyps. The APPROVe study was terminated in 2004, when the preliminary data from the study showed an increased relative risk of adverse cardiovascular events (including heart attack and stroke) beginning after 18 months of rofecoxib therapy. At the same time, information about an FDA study [34] that supported previous findings of increased risk of heart attack due to rofecoxib came out. The study estimated that rofecoxib caused between 88,000 and 139,000 heart attacks, 30 to 40 percent of which were probably fatal, in the five years the drug was on the market.

On September 30, 2004, Merck voluntarily withdrew rofecoxib from the market because of concerns about the increased risk of heart attack. Rofecoxib was one of the most widely used drugs ever to be withdrawn from the market. In the year before withdrawal, Merck had sales revenue of $2.5 billion from rofecoxib.

The goal of this case study is to assess whether an active drug surveillance system could have led to faster withdrawal of rofecoxib from the market.

3.7.1 The Data Set

The data set used in this study contains claims from close to 2.4 million members between 5/1/1999 and 12/1/2005. Most members are active during only a fraction of this period, and the data contains a total of 33,215,756 member months. We have a total of 24,044 members who were prescribed rofecoxib and 71,100 members who took naproxen. The average time in the data set prior to treatment is 12.9 months for the naproxen members and 9.7 months for the rofecoxib members.

As in the previous chapter, the data is acquired through D2Hawkeye. D2Hawkeye is a growing company founded in 2001, and therefore very little data is available from the early years of rofecoxib. The first prescription for rofecoxcib in our data set is in early July 2000. Figure 3-1 shows the number of members starting on rofecoxib each month, the number of members stopping each month, the overall number of members taking rofecoxib each month, and the cumulative number of members who have at some point taken the drug. The sharp increase in number of members starting on rofecoxib about 30 months into the study is explained by new clients being acquired by D2Hawkeye.

Days on drug

Table 3.1 show the number of days a member is on the drug (assuming that he or she takes all her dispensed medication). As we can see, close to half of the rofecoxib members take rofecoxib for 30 days or less. Therefore in this case study, we cannot be too stringent about the minimum time a member takes the drug. Section 3.7.4 shows how varying the minimum requirement affects the results of the study.

Given the late adoption of rofecoxib in this data set and the short period of time members stay on rofecoxib, we choose to set the toxicity period to the length of a member's stay in the database. Restricting the toxicity period to be a fixed time

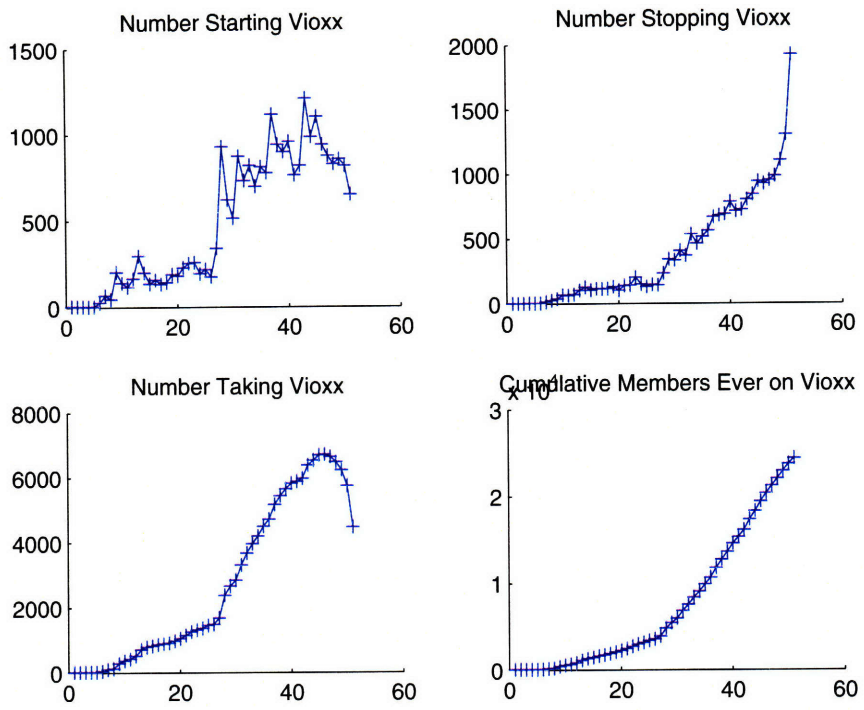

Figure 3-1: The top-left figure shows the number of members starting rofecoxib each month; the top-right figure the number of members who stop taking rofecoxib. The bottom-left figure shows the number of active members at each point in time and finally the bottom-right picture shows the cumulative number of members who have ever taken rofecoxib in our data set. The x-axis represents the number of months from the start of the study (July 2000).

period would cut down our data set too fast, and as a result we would lose the ability to detect rare adverse events.

Age and Gender Distribution

When we compare the rofecoxib population and the naproxen population, we find that the average age differs by 9 years: it is close to 40 years for the naproxen population, and 49 years for those taking rofecoxib. Naproxen is more prominent among women and children, as Table 3.2 shows. We define the following age groups that we use throughout the study: 1) 0-19 years old, 2) 20-30 years old, 3) 30-34 years old, 4) 35-39 years old, 5) 40-44 years old, 6) 45-49 years old, 7) 50-54 years old 8) 55-59

Days on Drug	Rofecoxib	Naproxen
0-10	2,957	19,989
11-30	8,767	35,083
31-60	2,974	8,501
61-90	1,987	2,862
91-120	1,092	1,444
121-150	757	716
151-180	954	648
181-210	491	361
221-240	416	298
241-270	610	227
271-300	290	206
301-330	226	145
331-360	382	199
more	2,055	961

Table 3.1: Length of naproxen and rofecoxib therapy.

years old, 9) 60-64 years old, and 10) over 64 years of age.

	Naproxen		Rofecoxib	
Age Group	Male	Female	Male	Female
1	3.0%	4.9%	0.9%	1.1%
2	5.1%	9.2%	2.3%	3.3%
3	3.9%	6.3%	2.4%	2.9%
4	4.6%	6.9%	3.4%	4.5%
5	5.2%	7.9%	4.9%	6.4%
6	5.2%	8.0%	6.1%	8.7%
7	4.6%	6.5%	6.3%	9.6%
8	3.4%	4.6%	6.0%	9.0%
9	2.0%	2.5%	4.4%	6.5%
10	2.7%	3.5%	4.7%	6.6%

Table 3.2: The age-gender distribution of rofecoxib and naproxen members.

Pre-treatment Costs and Diagnosis

Since naproxen and rofecoxib were approved for treatment of the same conditions, one would expect the populations prior to starting the treatments to be similar. The data, however, tell a different story. Table 3.3 shows the average monthly health care cost prior to starting treatment, by age group. We note that the rofecoxib

members uniformly have higher pre-treatment costs. The 300% cost difference in the lowest age group is in part explained by the fact that the average age in age group 1 for naproxen is lower than for rofecoxib. Table 3.4 shows the average number of claim-line events per month for seven randomly chosen diseases. We note that for 6 out of 7 diseases the prevalence is higher for the rofecoxib population. In fact, "thyroid problems" is one of the few diagnoses that is virtually equal between the two populations. These differences are only in small part explained by the difference in age-gender distribution, as verified by a more detailed analysis.

Age Group	Naproxen	Rofecoxib	Difference
1	$132	$527	299.1%
2	$195	$281	44.2%
3	$224	$365	62.5%
4	$219	$358	63.9%
5	$224	$376	67.8%
6	$243	$319	31.2%
7	$279	$397	42.1%
8	$312	$403	29.1%
9	$348	$477	37.1%
10	$225	$414	83.8%

Table 3.3: Comparison of pre-treatment costs. The second and third columns show the average cost per month before starting treatment of naproxen and rofecoxib, respectively.

Disease	Naproxen	Rofecoxib	Difference
Anemia	0.0058	0.0083	44%
Benign Neoplasms	0.0103	0.0119	16%
Endocrine	0.0450	0.0548	22%
Malignant Neoplasms in Bone	0.0056	0.0139	150%
Nutritional Deficiencies	0.0005	0.0010	104%
Other Blood Diseases	0.0015	0.0043	177%
Thyroid	0.0103	0.0102	0%

Table 3.4: Comparison of pre-treatment diagnosis prevalence. The second and third columns show the average number of claim-line events per month before starting treatment of naproxen and rofecoxib, respectively.

3.7.2 Event Definitions and Code Grouping

As previously mentioned, the ICD-9 codes are structured in a hierarchical way. Due to coding inaccuracies at the lower levels, we chose to focus on three-digit codes (including all subsequent four- and five-digit codes). We furthermore merge codes that have the same parent code and represent the same condition. As an example, we group together all tuberculosis codes, which all fall under the parent category of infectious and parasitic diseases. Table 3.5 shows the eight different level-3 codes for tuberculosis. Details of the group coding appear in Appendix 3.A. Tables 3.6 and 3.7 summarize the codes used for identifying the two known side effects of rofecoxib at the time of withdrawal, heart attacks and stroke.

ICD-9 Code	Description
010	Primary tuberculosis infection
011	Pulmonary tuberculosis
012	Other respiratory tuberculosis
013	Tuberculosis of meninges and central nervous system
014	Tuberculosis of intestines, peritoneum, and mesenteric glands
015	Tuberculosis of bones and joints
016	Tuberculosis of genitourinary system
017	Tuberculosis of other organs
018	Miliary tuberculosis

Table 3.5: Level 3 ICD-9 codes grouped together in the tuberculosis group, a part of infections and parasitic diseases.

ICD-9 Code	Description
410	Acute myocardial infarction
411	Other acute and subacute forms of ischemic heart disease
4110	Postmyocardial infarction syndrome
4118	Other
41189	Other
4130	Angina decubitus

Table 3.6: ICD-9 coding used to identify heart attacks.

ICD-9 Code	Description
433	Occlusion and stenosis of precerebral arteries including basilar artery, carotid artery, and vertebral artery, etc.
434	Occlusion of cerebral arteries including cerebral thrombosis and cerebral embolism and unspecified cerebral artery occlusion.
435	Transient cerebral ischemia

Table 3.7: ICD-9 coding used to identify strokes.

3.7.3 Using Optimization to Select the Comparison Groups

In our research we implemented all of the strategies discussed in Section 3.3. Our observations on using optimization methods appear below, and Sections 3.7.4 and 3.7.5 reports the results of other methods.

Using Maximum Pairing

Pairing each rofecoxib member with a naproxen member based on their medical histories prior to starting the treatment results in data samples that are very similar. We implemented this approach but found it to be unsuccessful due to how fast it cuts down on the data. To be able to compute members' similarity, we need at least 6 months of data prior to starting treatment, for each member. This criterion excludes 27,727 naproxen and 11,892 rofecoxib members from the study; we lose almost half of our rofecoxib population. The minimum number of days on a drug and other inclusion requirements leave very few members in the sample. Table 3.8 shows the number of *potential* members to be paired up in each month for two years, from May 2001 to May 2003, when we required at least 90 days of taking the drug and at least 180 days in the sample before and after the start of treatment. In the optimization, age and gender were ignored, and minimum similarity was set to the average similarity of members. This combination of rather stringent inclusion criteria but loose optimization criteria results in zero observed cardiac events over the two year period. We therefore conclude that although matching based on past history would take care of all discrepancies in the pre-treatment disease burden, due to the limited number of members that have sufficient prior history in our data set, this method was unsuc-

cessful. This method may potentially work better with larger data sets that include more members over longer periods.

Month	Rofecoxib Members	Naproxen Members
May 2001	47	41
June 2001	27	28
July 2001	26	22
August 2001	31	22
September 2001	36	23
October 2001	21	27
November 2001	42	24
December 2001	44	31
January 2002	52	35
February 2002	44	36
March 2002	45	36
April 2002	40	32
May 2002	34	29
June 2002	33	42
July 2002	36	38
August 2002	83	41
September 2002	69	40
October 2002	53	39
November 2002	109	58
December 2002	114	45
January 2003	114	53
February 2003	116	57
March 2003	156	99
April 2003	116	86
May 2003	171	78

Table 3.8: The number of potential members for pairing.

Using Population Maximization

As opposed to trying to match every single rofecoxib member with a similar naproxen member, population maximization tries to adjust for the pre-treatment disease burden over the whole population at once. When this approach is implemented, selecting the right δ_j's, the slack on the control constraints takes some care. When a diagnosis is common, we can be more stringent than when the diagnosis is rare. Similarly, our

results suggest that including very sick members (in the pre-treatment period) who have a lot of diagnoses constrains the optimal solution of the optimization problem. Removing the pre-treatment "top spenders" improves the optimal solution (the sample size). Our optimization method produced results similar to those from simpler selection methods and we therefore omit them.

3.7.4 Results from Methods with a Baseline Rate

The study period is from July 2000 (the date of the first rofecoxib prescription in our data set) through March 2005 (6 months after rofecoxib was withdrawn from the market). Our experiment tests across all 328 coding groups. We compare the results from finding a baseline by adjusting the comparison population by age, gender and cost bucket to the results from following the relative changes in coding in both populations independently. We furthermore investigate different requirements for the minimum number of days taking the drugs and different definitions of events.

We analyze the time until the known side effects; cardiac events and stroke are flagged through the drug surveillance. We also pay special attention to renal coding since a meta-study from 2006 [62] found an increase in renal risk as a result of taking rofecoxib.

We first analyze the results of adjusting the comparison population for age and gender. The results are functions of both how we limit our population (minimum number of days taking the drug) and how we define an event (claims-line, claims-day, claims-event). Moreover, whether we raise an alarm is a function of how we set our confidence intervals. Finally, to avoid false positives due to unstable risk estimators, we set a minimum number of observed events in the post-treatment period that need to be observed, before an flag can be raised.

Table 3.9 compares the results for different minimum days and definitions of events

Settings (MinDays, Event)	Cardiac	Stroke	Renal Failure	Secondary Renal Coding	Number of False Positives
(0,line)	-	43	47	-	121
(0,event(30))	-	54	-	-	88
(0,event(1500))	-	51	-	-	73
(30,line)	-	41	36	-	188
(30,event(30))	-	55	-	-	26
(30,event(1500))	53	40	36	-	176

Table 3.9: The result of running age- and gender-adjusted population selection. If a known side effect was detected, the number represents the month of detection. The primary and secondary upper bounds were set to 97.5% and 95% respectively, and we required a minimum of 20 observed events in the post-treatment-period.

for the age-gender adjusted methodology. We note that we succeed in detecting stroke across different settings, but the same cannot be said for the cardiac events, which we detect only once. These succesful detections are outweighed by a very high false alarm rate, which makes this approach unsuccessful. It is therefore clear that adjusting only for age and gender does too little to adjust for prior differences between the two populations. Figure 3-2 shows how the numbers of cardiac events, stroke, and renal failure evolve over time compared to the confidence interval around the expected number of events. In the same figure we display the event data for the diagnosis of thyroid problems, which was signaled as a side effect, but is a false positive.

A member's health care cost is a good proxy for his or her overall health condition. We therefore split our population into cost buckets similar to those introduced in Section 2.2.3, adjusted for increases in health care cost, (as this is an older dataset). We adjust our estimates based on cost buckets and gender (we exclude age-groups in the adjustment as it would segment the data into too many subgroups, resulting in unstable baseline estimates). Table 3.10 shows the results for the same parameter settings as before for the age-gender adjusted methodology in Table 3.9. We note that we still have too many false positives, but all settings correctly identify stroke

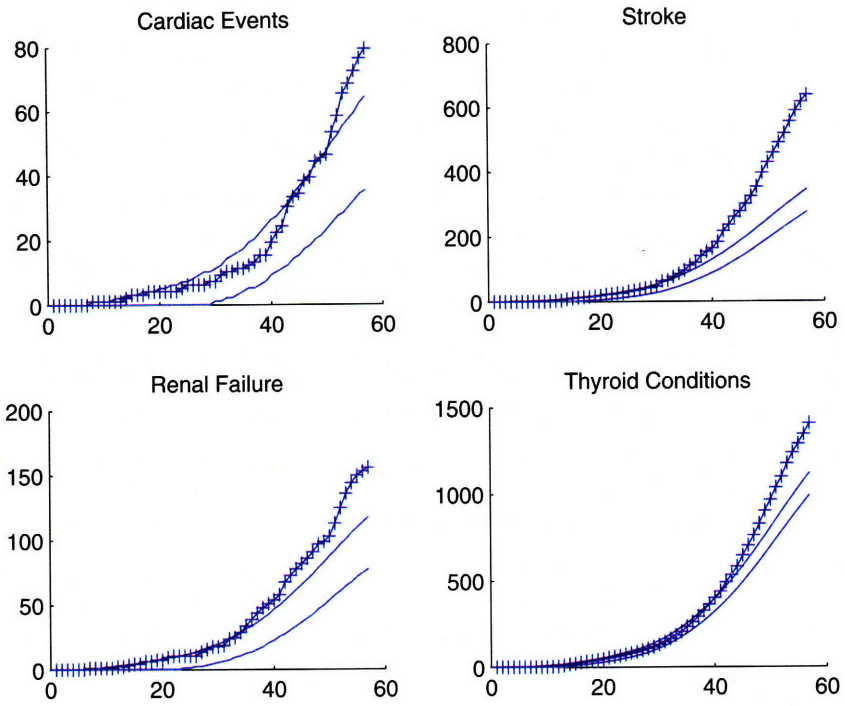

Figure 3-2: The number of observed events compared to a 95% confidence intervals constructed from the naproxen population estimates, for 4 conditions: 3 known side effects as well as thyroid problems. The x-axis is the number of months from the start of the study.

and renal failure as a side effects, and they do so faster than the age-gender adjusted method. The number of false positives is similar or higher. The two adjustment methods therefore performed similarly.

3.7.5 Results from Methods without Baseline Rate

We now look at the results from running the drug surveillance without a predefined baseline. In order to raise an alarm, the estimate of the relative risk needs to have stabilized. As we see from Figure 3-3, this stabilization does not occur until around week 50 for many of the conditions. One of the main reasons is that for the first 3 years of the study, there are very few members in the data, and it is not until 2004

Settings (MinDays , Event)	Cardiac	Stroke	Renal Failure	Secondary Renal Coding	Number of False Positives
(0,linect)	-	35	37	-	142
(0,event(30))	-	36	40	-	121
(0,event(1500))	46	36	41	-	109
(30,linect)	56	28	32	-	179
(30,event(30))	-	40	31	-	40
(30,event(1500))	-	31	40	-	171

Table 3.10: The result of running cost bucket- and gender-adjusted population selection. If a known side effect was detected, the number represents the month of detection. The primary and secondary upper bounds were set to 97.5% and 95% respectively, and we required a minimum of 20 observed events in the post-treatment-period.

that we have a significant number of members taking rofecoxib. We therefore first test for stability of our estimates, and only if a condition passes the stability test do we analyze if the relative risk is significantly larger for the rofecoxib members than for the naproxen members. If a condition raises an alarm, it enters a six-month false-positive-period, designed to catch false positives due to temporary stabilization of the estimates.

Analyzing the Whole Population

Running the analysis resulted in 21 flags being raised, two of which are cardiac events and stroke. From Table 3.11 we note that at the time of alarm, we estimate the relative risk of strokes to be 4.38 and those of heart attack to be 4.04. We did not detect increased risk of renal failure. Out of the 19 other alarms, 8 of the conditions are related to the circulatory system and might therefore be related to a wider cardiac effect of rofecoxib than previously acknowledged.

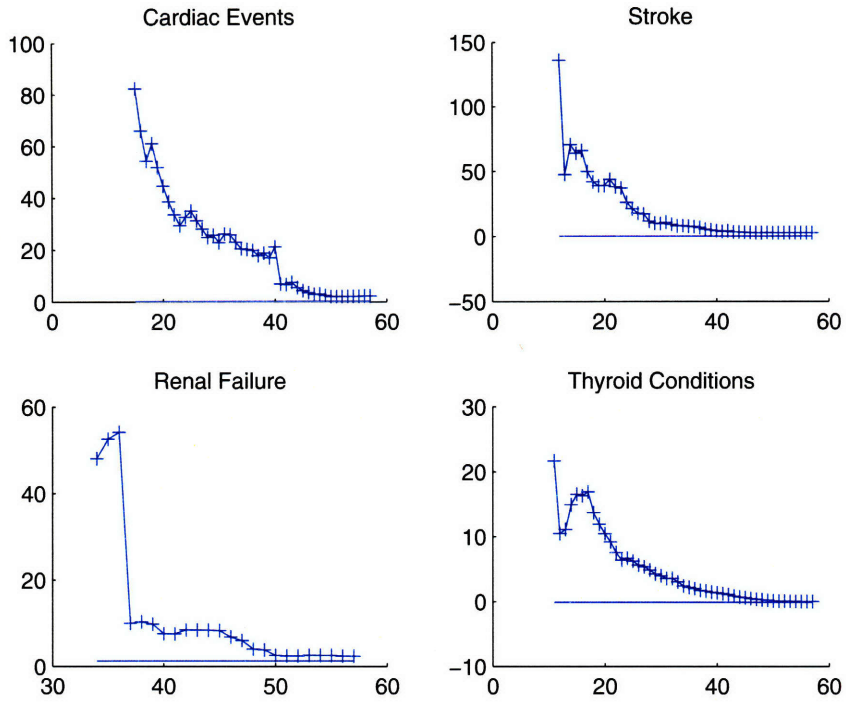

Figure 3-3: The estimates for the relative risk for four conditions over time. The horizontal line is the estimated relative change for naproxen.

Condition	Time at Detection	Relative Risk at Detection
Cardiac Events	55 (29)	4.002
Stroke	41 (15)	4.384
Renal Failure	Not detected	-

Table 3.11: Relative risk for known side effects at the time of detection. The number in parentheses is the number of months since September 2002.

Analyzing the Population by Cost Bucket

Table 3.12 shows the estimated risk for the known side-effects and thyroid problems, by cost buckets. We observe that the expected elevated risks of the known side effects are visible in the lower buckets, but they disappear in upper buckets. Sick members have a lot of pre-treatment coding, resulting in increased variability in the estimates. This observation motivates a further study into how the lower cost buckets (the healthier population) can be used for drug surveillance, rather than the popu-

lation as a whole. Figure 3-4 shows the estimates of the relative risk, for bucket 1 members for cardiac events, stroke, renal failure and thyroid problems as a function of time. We note that the estimates again do not stabilize until around week 50.

Minimum Days on Drug	Cost Bucket	Cardiac Events	Renal Failure	Stroke	Thyroid Problems
0	1	1.487	1.381	2.138	-0.110
	2	0.022	4.434	0.179	-0.407
	3	-0.006	0.129	-0.215	-0.456
	4	-0.844	-0.533	-0.524	-0.434
	5	-0.324	-0.831	-0.493	-0.480
30	1	2.829	5.982	2.592	-0.113
	2	0.181	3.724	0.229	-0.412
	3	-0.193	0.110	-0.242	-0.485
	4	-0.910	-0.369	-0.519	-0.492
	5	0.008	-0.856	-0.535	-0.395

Table 3.12: Relative risk for known side effects and thyroid problems. For example, we observe an almost six-fold risk in renal failure for bucket 1 members (this in large part due to very few events in the pre-treatment history) if we require members to have been prescribed a minimum of 30 days of rofecoxib.

Running the drug surveillance using bucket 1 members only results in a reduced number of flags being raised significantly. Table 3.13 summarizes the information about the flags raised. Seven conditions were flagged, including stroke and cardiac events, the known side effects of rofecoxib. Two other conditions related to the circulatory system were flagged, raising questions about wider cardiac effects of rofecoxib (as we saw previously when we ran the analysis on all buckets). Another condition that was flagged is hypertensive chronic kidney disease, a sign of deteriorating kidney health, which is connected to the known effects of rofecoxib on the kidneys. Two other conditions get flagged, one of which, "other conditions originating in the perinatal period," we catch as a false positive in the six-month-long follow-up period, leaving "congenital anomalies of urinary system" as a false positive. This condition has very few events in both the pre-treatment and post-treatment periods with a pre-treatment

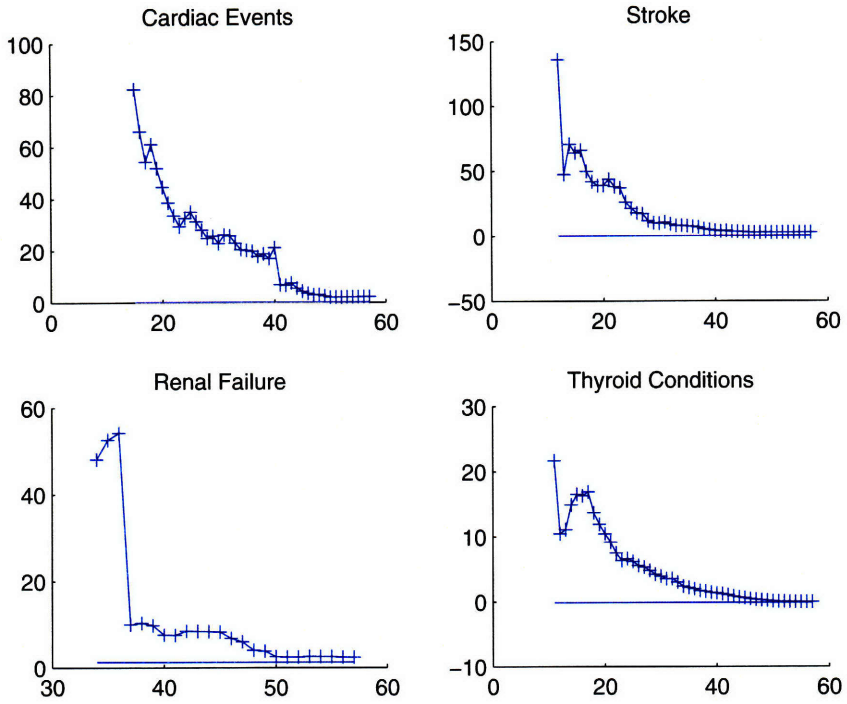

Figure 3-4: The figure shows, for bucket 1 members, how the estimate for the relative risk for rofecoxib members changes, for four different conditions: Cardiac events, stroke, renal failure and thyroid problems. The horizontal line in each subfigure is the relative change for the naproxen population.

p=0.0000391. Furthermore the corresponding relative risk for naproxen members is 1.96, indicating that there might be a systematic increase in relative risk and that our estimate has not gone down enough when the study ends (due to very few observed events). Using only the bucket one population reduces the number of false positives, while detecting both cardiac events and stroke.

Renal Failure

As previously discussed, renal failure became a known side effect of rofecoxib after the drug was withdrawn. When running the analysis on bucket 1 only, we did not flag renal failure (however, we did observe kidney disease). When the surveillance is

Condition	Time until Discovery (t)	Relative Risk at t	Relative Risk at $min(t+6, 57)$	Change between t and $t+6$
Cardiac Events	55(29)	2.74	2.83	3.12%
Stroke	51 (23)	2.72	2.59	-4.8%
Hypertensive chronic kidney disease	45 (19)	10.3	12.0	16.5%
Angina pectoris , excluding 413.0	56 (30)	2.24	2.24	0%
Portal vein thrombosis, venous embolism and throbosis	52 (26)	2.41	2.54	5%
Congenital anomalies of urinary system	52 (26)	3.80	3.73	-2%
Other conditions originating in the perinatal period	49 (23)	2.13	1.92	-10%

Table 3.13: The side effects identified by the relative risk methodology, using population from bucket 1 only. The numbers in parenthesis are the number of months from September 2002.

run on buckets 2 through 5, only one condition is flagged, renal failure in bucket 2. One of the reasons we see renal failure in bucket 2 and not in bucket 1 is that healthy kidneys do not go into renal failure right away (unlike a heart attack which may have no prior indications and therefore no prior costs). Kidneys progressively get worse until they reach the stage of renal failure. Due to the biologics of rofecoxib, healthy kidneys (as are most kidneys in members in bucket 1) are not affected by rofecoxib as much as troubled kidneys, such as those that might be found in members in the upper buckets. Table 3.14 shows the time, relative risk and change during the six-month false-positive control period for the renal failure condition in bucket 2.

This observation raises the point that looking at bucket 1 may not be enough when looking for adverse events that cannot (or rarely) originate from bucket 1. Those events are few, and we leave it to implementation with medical expertise to define the appropriate surveillance buckets for those events.

Condition	Time until Discovery (t)	Relative Risk at t	Relative Risk at $t+6$	Change between t and $t+6$
Acute renal failure and unspecified renal failure	51 (25)	3.59	3.94	10%

Table 3.14: The side effects identified by the relative risk methodology, using population from buckets 2 through 5. The number in parenthesis is the number of months from September 2002.

3.7.6 Conclusions

In this section we have shown that we can effectively way follow the changes in relative risk and discover false positives sooner than with the current post surveillance system. We have also shown that the key to a successful system, is to follow bucket one members only, except for few conditions that cannot arise among bucket one members. This approach reduces the variance in the system and leads to more stable estimates and many fewer false positives. We have also found that the most appropriate definition of events is to count only first occurrences of conditions (claims-period events, with large d), rather than define events as claims-lines or claim-day, as this approach also reduces the variability in the system.

From our analysis, we have learned that it takes between 2 and 3 years, after we get a significant amount of data, to detect cardiac events, stroke and kidney problems. It is therefore clear that many adverse events could have been prevented if active real-time drug surveillance had been in place during the rofecoxib years.

3.8 Case Study: Atorvastatin vs. Simvastatin

Atorvastatin, marketed under the name Lipitor, is now one of the largest selling drugs in the world, with US sales in 2006 exceeding $12.9 billion. Marketed by Pfizer, the drug belongs to the class of pharmaceuticals called statins. It is used to control

elevated cholesterol levels and, as a result, lowers the risk of cardiovascular disease. Another statin is simvastatin, better known under the marketing name Zocor, which in medical studies has been shown to have efficacy and toxicity profile similar to atorvastatin [35, 66].

3.8.1 The Data

In our dataset we have over 57,000 members who have taken atorvastatin (thereof over 56,000 in bucket 1) and close to 23,000 members (close to 22,000 in bucket 1) who have taken simvastatin. Table 3.15 summarizes some of the key parameters for the statin members. For example, we note that average age is 55 and 56 years for atorvastatin and simavastatin respectively. The study period was from September 1999 through August 2005.

Drug	Cost Bucket	Count	Avg. Age	Avg. Days on Drug	Avg. Pre-treatment Months	Avg. Post-treatment Months	Avg. Monthly Pre-treatment Costs
Atorvastatin	1	56,035	55	1196	10	28	229
Simvastatin	1	21,935	56	832	11	31	266
Atorvastatin	all	57,578	55	1203	10	28	567
Simvastatin	all	22,812	56	868	11	31	891

Table 3.15: Data summary for atorvastatin (Lipitor) and simvastatin (Zocor).

3.8.2 Results

We applied the same methodology and parameter settings as with rofecoxib and naproxen. Table 3.16 shows the flags that were raised, and the changes in the six-month follow-up period. We note that all the flags raised get caught as false positives in the follow-up period.

Drug	Condition	Time until Discovery (t)	Relative risk at t	Relative risk at $min(t+6, 57)$	Change in 6 months
Zocor	Malignant neoplasm of bone, connective tissue, skin, and breast	45	6.59	3.99	-39.5%
Lipitor	Symptoms involving nervous and musculoskeletal systems	48	4.27	1.71	-60.0%
Zocor	Other psychoses	48	13.62	6.26	-54.1%
Lipitor	Cataract	49	3.08	1.80	-41.5%
Lipitor	Disorders of the autonomic nervous system	50	3.74	1.96	-47.5%
Lipitor	Other disorders of soft tissues	51	2.52	1.04	-58.9%
Zocor	Hyperkinetic syndrome	52	7.80	3.17	-59.4%
Zocor	Varicose veins of lower extremities and other sites	52	2.99	2.28	-23.8%
Zocor	Fitting and adjustment of prosthetic device, implant or other device	60	3.50	1.99	-43.0%
Zocor	Disorders resulting from impaired renal function	68	2.40	2.02	-15.9%

Table 3.16: The side effects identified by the relative risk methodology, using population from bucket 1 only and the settings from the rofecoxib study.

Looking at the estimates for the relative risk, we notice less variance than with the estimates using the rofecoxib data. One reason is that the average post-treatment period is longer and the population is taking the drug for longer periods. We therefore adjust the parameters accordingly and rerun the analysis. As a result of the new parameter settings, one flag was raised, for other psychoses. As Figure 3-5 shows, a flag is raised when the relative risk is stable for a couple of months, 4 years into the study. The risk goes down in the following months, and therefore the condition is labeled as a false positive in the six-month follow-up period.

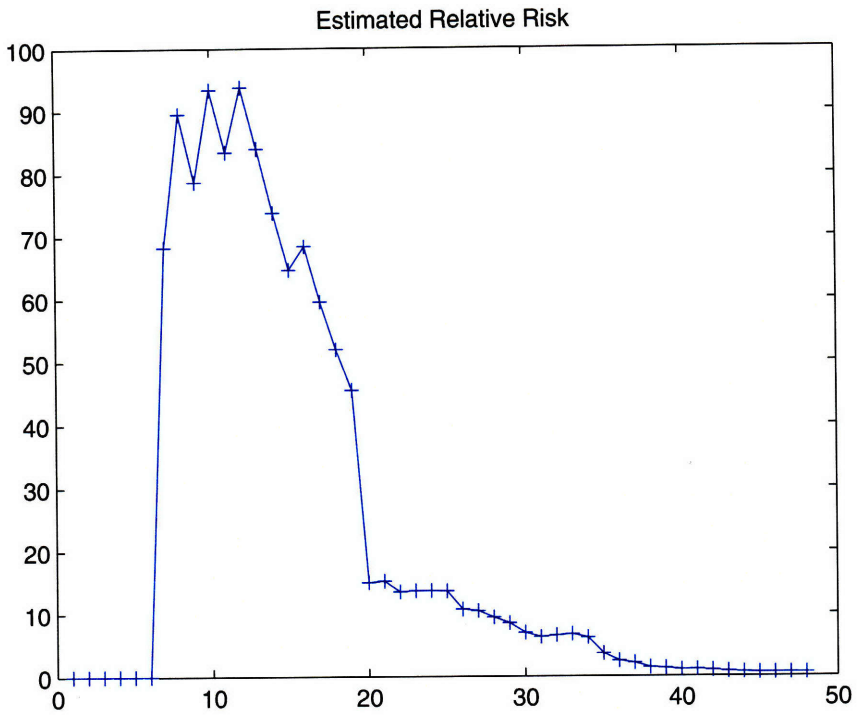

Figure 3-5: The estimate of the relative risk for "other psychoses" during the study period. The x-axis represents the number of months from September 2000.

3.8.3 Conclusions

The results from this study indicate that atorvastatin and simvastatin have similar side effects, as suggested in the medical literature. Furthermore, the study indicates that the methodology developed in previous sections can work across different drugs. An interesting future study would be to compare simvastatin and atorvastatin to a sample from the general population, to better understand the drugs' toxicity profiles.

3.9 Case Study: Sildenafil vs. Tadalafil

Sildenafil, known under the marketing names Viagra, Revatio and others, is a drug used to treat male erectile dysfunction (impotence) and pulmonary arterial hypertension. Viagra was launched in 1998 and had the fastest initial sales growth of any pharmaceutical product following its launch [39]. In 2000, it had 92 percent of the

global market for prescribed erectile dysfunction pills. One of Viagra's main competitors today is tadalafil, better known under the marketing name of Cialis. Tadalafil was approved by the FDA in November 2003. One advantage tadalafil has over sildenafil (and vardenafil, known under the marketing name Levitra) is its 17.5 hour half-life, compared to 4 hours for sildenafil, earning Cialis the nickname "the weekend pill."

Sildenafil has some rare but serious side effects, including the following: priapism, severe hypotension, myocardial infarction, ventricular arrhythmias, stroke and increased intraocular pressure, especially in men with heart conditions. More common, but less serious side effects include diarrhea, dizziness, dyspepsia, facial flushing, headache, nasal congestion, rash, sneezing, palpitations, photophobia, upset stomach and urinary tract infection [6, 49, 32]. These side effects reflect the ability of sildenafil to cause blood vessels to widen. Tadalafil has many side effects in common with sildenafil [26], as both drugs belong to the same drug class and therefore work in similar ways. The most common side effects of tadalafil are headache, indigestion, back pain, muscle aches, flushing, and stuffy or runny nose.

In May of 2005, the U.S. Food and Drug Administration issued a warning that sildenafil (as well as tadalafil and vardenafil) could lead to vision impairment [30], and in October 2007, the FDA required labeling for sildenafil (and other similar drugs including tadalafil) to warn users of the potential risk of sudden hearing loss [31].

In this case study, we apply the relative risk methodology and compare sildenafil to tadalafil. As the drugs share many side effects, the goal is to evaluate if there is a difference between the two drugs, as opposed to flagging all side effects. An interesting study would be to compare those men taking sildenafil to the general population, but that is outside the scope of this chapter.

3.9.1 The Data

The study uses the same database as in the previous two case studies. The data includes 11,401 sildenafil members and 1,931 tadalafil members after excluding women (277 members) and members under 20 (46 additional members). Most of the members are in bucket one, as Table 3.17 shows. The first prescription for tadalafil is in December 2003, and the last is in October 2005. Due to the late introduction of Cialis to our dataset, the average number of months after starting treatment is much lower for tadalafil than for sildenafil. To overcome this difference (longer post-treatment period allows for a longer time for events to happen), we adjust the post-treatment histories for the sildenafil members. For same reasons, the pre-treatment period is longer for the tadafil members and we adjust the pre-treatment data in the same way. We furthermore compare the surveillance results using adjusted histories with the outcome when we do not adjust the length of pre- and post-histories. We note that in other ways the populations are quite similar: the average age of the two populations is almost the same (52 and 53 years) and the average pre-treatment monthly cost is close ($223 and $290 for bucket one members).

Drug	Cost Bucket	Count	Avg. Age	Avg. Days on Drug	Avg. Pre-treatment Months	Avg. Post-treatment Months	Avg. Monthly Pre-treatment Costs
Tadalafil	1	1,882	52	80	18	12	290
Sildenafil	1	11,158	53	126	11	23	223
Tadalafil	all	1,931	52	81	18	12	475
Sildenafil	all	11,401	53	128	11	23	406

Table 3.17: Data summary for sildenafil (Viagra) and tadalafil (Cialis).

3.9.2 Results

Tadalafil has a short timeframe in our datasets, and a small sample size. We therefore do not expect to be able to detect many differences as not all estimators of adverse events will have stabilized in 2 years with only 2,000 members. We ran the experiment both with and without requirements for the minimum number of days that the members took the drugs, and with and without adjustments to pre- and post-treatment histories. Table 3.18 shows the flags raised as a result of our analysis. No flags were raised when the analysis was run with no requirement on minimum number of days on the drugs (both with and without adjusting the histories). When we required 30 days as the minimum number of days taking the drugs, one flag was raised (both with and without history adjustment), but it had very few events behind it in the post-treatment period (nine events), and was caught by the six-months false-positive surveillance period. Figure 3-6 shows the relative risk estimate during the study period; we note that it temporarily levels off around month 15 but then continues to fall.

Min Days on Drug	History Adjustment	Condition	Sildenafil Rel. Risk	Tadalafil Rel. Change	Month
30	No	Special screening for endocrine, nutritional, metabolic, and immunity disorders	-30.5%	235.3%	17
30	Yes	Special screening for endocrine, nutritional, metabolic, and immunity disorders	-30.7%	235.3%	17

Table 3.18: Flags raised for sildenafil (Viagra) and tadalafil(Cialis). No flags were raised for when no requirement was put on the minimum numbers of days a member needed to take the drug in order to be included in the analysis. Month refers to the number of months after the first tadalafil prescription.

In conclusion, due to limitations in our data (very few events) we are unable to detect

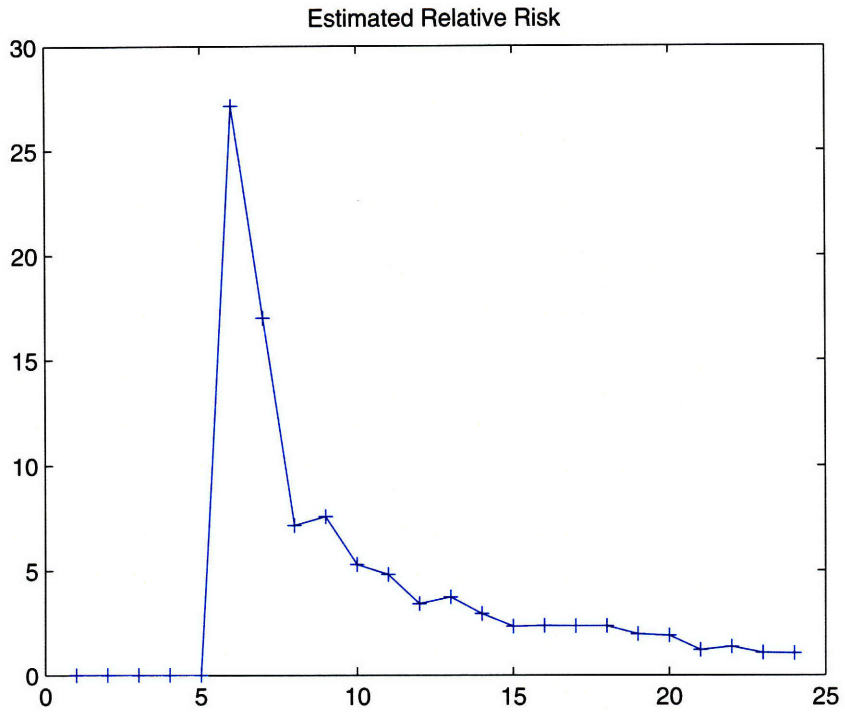

Figure 3-6: The estimate of the relative risk for "special screening for endocrine, nutritional, metabolic, and immunity disorders". The x-axis represents the number of months from the start of the study (December 2003).

any differences in side effects between tadalafil and sildenafil, if there are any. The condition that was flagged had less than ten observed events in the post-treatment period and was labeled as a false positive in the six-month follow-up period. This finding indicates that our methodology does not raise unwanted alarms when dealing with smaller populations.

3.10 Conclusions

In this chapter we have laid the foundations for a surveillance system for drugs in common use. We ran three case studies that show promising results and encourage further study.

The rofecoxib case study highlighted some of the difficulties with setting up a robust system. We found that it is not enough to select similar drugs: prior differences need to be accounted for. We introduced pre-treatment and post-treatment relative change as the measure to achieve that goal. We also introduced several mathematical programming formulations of the population selection problem and believe, although they did not prove to be completely successful in our case study (in part due to a small sample size) that they have potential to aid in population selection when drug surveillance is run against a comparative baseline.

The two other case studies showed that the methodology developed did not raise any unexpected flags or have a high rate of false positives (none in our two studies) when applied to broader range of drugs classes.

This study shows that drug surveillance using claims data could become one of FDA's standard tools for post-marketing surveillance. The methodology introduced here should be further developed on drugs with known side effects, which will highlight more challenges and be a valuable learning process. Some of the refinements that will further fine-tune the methods are a) to use medical knowledge to create a better grouping of ICD-9 codes; b) to implement definitions of events as functions of the underlying conditions, with acute events treated differently than chronic conditions; and c) to analyze the optimal setting of the upper confidence intervals, to strike a balance between the expected time until discovery of a true side effect and the probability of a false positive.

In conclusion, in this chapter we have emphasized finding side effects. With the same methodology we can potentially find unexpected benefits of drugs currently on the market.

3.A Appendix - Grouping of Codes Used in Case Studies

Table 3.19 shows the ICD-9 groupings used in the study. The group descriptions are based on [4].

Table 3.19

Description	ICD-9 Code
Intestinal infectious diseases	001.x - 009.x
Tuberculosis	010.x - 018.x
Zoonotic bacterial diseases	020.x - 027.x
Other bacterial diseases	030.x - 041.x
Human immunodeficiency virus infection	042.x
Poliomyelitis and other non-arthropod-borne viral diseases of central nervous system	045.x - 049.x
Viral diseases accompanied by exanthem	050.x - 057.x
Other human herpesviruses	058.x
Arthropod, borne viral diseases	060.x - 066.x
Other diseases due to viruses and chlamydiae	070.x - 079.x
Rickettsioses and other arthropod-borne diseases	080.x - 088.x
Syphilis and other venereal diseases	090.x - 099.x
Other spirochetal diseases	100.x - 104.x
Mycoses	110.x - 118.x
Helminthiases	120.x - 129.x
Other infectious and parasitic diseases	130.x - 136.x
Late effects of infectious and parasitic diseases	137.x - 139.x
Malignant neoplasm of lip, oral cavity, and pharynx	140.x - 149.x
Malignant neoplasm of digestive organs and peritoneum	150.x - 159.x
Malignant neoplasm of respiratory and intrathoracic organs	160.x - 165.x
Malignant neoplasm of bone, connective tissue, skin, and breast	170.x - 176.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Malignant neoplasm of genitourinary organs	179.x - 189.x
Malignant neoplasm of other and unspecified sites	190.x - 199.x
Malignant neoplasm of lymphatic and hematopoietic tissue	200.x - 208.x
Benign neoplasms	210.x - 229.x
Carcinoma in situ	230.x - 234.x
Neoplasm of uncertain behavior	235.x - 238.x
Neoplasm of unspecified nature	239.x
Disorders of thyroid gland	240.x - 246.x
Diseases of other endocrine glands	250.x - 259.x
Nutritional deficiencies	260.x - 269.x
Other metabolic and immunity disorders	270.x - 279.x
Anemias	280.x - 287.x
Diseases of white blood cells	288.x
Other diseases of blood and blood-forming organs	289.x
Dementias	290.x
Alcohol and drug induced mental disorders	291.x
Drug-induced mental disorders	292.x
Transient mental disorders due to conditions classified elsewhere	293.x
Persistent mental disorders due to conditions classified elsewhere	294.x
Schizophrenic disorders	295.x
Episodic mood disorders	296.x
Other psychoses	298.x
Pervasive developmental disorders	299.x
Anxiety, reactions to stress or adjustment reaction	300.x, 308.x, 309.x
Personality disorders	301.x
Sexual and gender identity disorders	302.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Alcohol dependence, drug dependence or nondependent abuse of drugs	303.x, 304.x, 305.x
Physiological malfunction arising from mental factors, special symptoms or syndromes	306.x, 307.x
Specific nonpsychotic mental disorders due to brain damage	310.x
Depressive disorder, not elsewhere classified	311.x
Disturbance of conduct, not elsewhere classified	312.x
Disturbance of emotions specific to childhood and adolescence	313.x
Hyperkinetic syndrome of childhood	314.x
Specific delays in development	315.x
Psychic factors associated with diseases classified elsewhere	316.x
Mental retardation	317.x - 319.x
Meningitis	320.x, 321.x, 322.x
Encephalitis, myelitis, and encephalomyelitis	323.x
Intracranial and intraspinal abscess	324.x
Phlebitis and thrombophlebitis of intracranial venous sinuses	325.x
Late effects of intracranial abscess or pyogenic infection	326.x
Organic sleep disorders	327.x
Cerebral degenerations	330.x, 331.x
Parkinson's disease	332.x
Other extrapyramidal disease and abnormal movement disorders	333.x
Spinocerebellar disease	334.x
Anterior horn cell disease	335.x
Other diseases of spinal cord	336.x
Disorders of the autonomic nervous system	337.x
Pain, unspecified by location	338.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Multiple sclerosis	340.x
Other demyelinating diseases of central nervous system	341.x
Hemiplegia and hemiparesis	342.x
Other paralytic syndromes	344.x
Epilepsy and recurrent seizures	345.x
Migraine	346.x
Cataplexy and narcolepsy	347.x
Other conditions of brain	348.x
Other and unspecified disorders of the nervous system	349.x
Trigeminal nerve disorders	350.x
Facial nerve disorders	351.x
Disorders of other cranial nerves	352.x
Nerve root and plexus disorders	353.x
Mononeuritis of upper limb, lower lip and mononeuritis multiplex	354.x, 355.x
Hereditary and idiopathic peripheral neuropathy	356.x
Inflammatory and toxic neuropathy	357.x
Myoneural disorders	358.x
Muscular dystrophies and other myopathies	359.x
Disorders of the globe, retinal (exluding 361.x) iris, ciliary body, refraction, accommodation, conjunctiva, lacrimal, cornia orbit, optic nerve and visual pathways.	360.x, 362.x, 364.x, 367.x, 371.x, 372.x, 375.x, 376.x, 377.x, 379.x
Retinal detachments and defects	361.x
Chorioretinal inflammations, scars, and other disorders of choroid	363.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Glaucoma	365.x
Cataract	366.x
Visual disturbances	368.x
Blindness and low vision	369.x
Keratitis	370.x
Inflammation of eyelids and other disorders of eyelids	373.x, 374.x
Strabismus and other disorders of binocular eye movements	378.x
Disorders of external ear	380.x
Nonsuppurative otitis media and eustachian tube disorders	381.x
Suppurative and unspecified otitis media	382.x
Mastoiditis and related conditions	383.x
Vertiginous syndromes and other disorders of vestibular system	386.x
Otosclerosis	387.x
Other disorders of ear(including tympanic membrane, middle ear and mastroid	388.x, 385.x 386.x
Hearing loss	389.x
Acute rheumatic fever	390.x - 392.x
Chronic rheumatic heart disease	393.x - 398.x
Essential or secondary hypertension	401.x, 405.x
Hypertensive heart disease	402.x
Hypertensive chronic kidney disease	403.x
Hypertensive heart and chronic kidney disease	404.x
Acute myocardial infarction, and other acute and subacute forms of ischemic heart disease	410.x, 411.x
Old myocardial infarction	412.x
Angina pectoris , excluding 413.0	413.x
Other forms of chronic ischemic heart disease	414.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Diseases of pulmonary circulation	415.x - 417.x
Other forms of heart disease	420.x - 429.x
Cerebrovascular disease	430.x - 438.x
Diseases of arteries, arterioles, and capillaries	440.x - 449.x
Phlebitis and thrombophlebitis	451.x
Portal vein thrombosis, venous embolism and throbosis	452.x, 453.x
Varicose veins of lower extremities and other sites	454.x, 456.x
Hemorrhoids	455.x
Noninfectious disorders of lymphatic channels	457.x
Hypotension	458.x
Other disorders of circulatory system	459.x
Acute respiratory infections	460.x - 466.x
Other diseases of the upper respiratory tract	470.x - 478.x
Pneumonia and influenza	480.x - 488.x
Chronic obstructive pulmonary disease and allied conditions	490.x - 496.x
Pneumoconioses and other lung diseases due to external agents	500.x - 508.x
Other diseases of respiratory system	510.x - 519.x
Disorders of tooth development and eruption	520.x
Diseases of hard tissues of teeth ,	521.x
Diseases of pulp and periapical tissues	522.x
Gingival and periodontal diseases	523.x
Dentofacial anomalies, including malocclusion	524.x
Other diseases and conditions of the teeth and supporting structures	525.x
Diseases of the jaws	526.x
Diseases of the salivary glands	527.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Diseases of the oral soft tissues, excluding lesions specific for gingiva and tongue	528.x
Diseases and other conditions of the tongue	529.x
Diseases of esophagus	530.x
Gastritis and duodenitis	535.x
Gastric, duodenal, petic, gastrojejunal ulcers and gastric mucositis	531.x - 534.x, 538.x
Disorders of function of stomach, and other disorders of stomach and duodenum	536.x, 537.x
Appendicitis	540.x - 543.x
Hernia of abdominal cavity	550.x - 553.x
Noninfectious enteritis and colitis	555.x - 558.x
Condition of the liver	570.x - 573.x
Cholelithiasis and other disorders of gallbladder and biliary tract	574.x - 576.x
Diseases of pancreas	577.x
Gastrointestinal hemorrhage	578.x
Intestinal malabsorption	579.x
Acute glomerulonephritis, nephrotic syndrome, chronic glomerulonephritis, nephritis and nephropathy, not specified as acute or chronic	580.x - 583.x
Acute renal failure and unspecified renal failure	584.x, 586.x
Chronic kidney disease (ckd)	585.x
Renal sclerosis, unspecified	587.x
Disorders resulting from impaired renal function	588.x
Small kidney of unknown cause	589.x
Infections of kidney	590.x
Hydronephrosis	591.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Calculus of kidney, ureter or lower urinary tract	592.x, 594.x
Other disorders of kidney and ureter	593.x
Cystitis and other disorders of bladder	595.x, 596.x
Urethritis, not sexually transmitted, and urethral syndrome, urethral stricture and other disorders of urethra and urinary tract	597.x - 599.x
Hyperplasia of prostate	600.x
Inflammatory diseases of prostate	601.x
Other disorders of prostate	602.x
Hydrocele	603.x
Orchitis and epididymitis	604.x
Redundant prepuce and phimosis	605.x
Infertility, male	606.x
Disorders of penis	607.x
Other disorders of male genital organs	608.x
Benign mammary dysplasias	610.x
Other disorders of breast	611.x
Inflammatory disease of female pelvic organs	614.x - 616.x
Endometriosis	617.x
Genital prolapse	618.x
Fistula involving female genital tract	619.x
Noninflammatory disorders of ovary, fallopian tube, broad ligament, cervix, vagina, vulva or perineum	620.x, 622.x - 624.x
Disorders of uterus, not elsewhere classified	621.x
Pain and other symptoms associated with female genital organs	625.x
Disorders of menstruation and other abnormal bleeding from female genital tract	626.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Menopausal and postmenopausal disorders	627.x
Infertility, female	628.x
Other disorders of female genital organs	629.x
Ectopic and molar pregnancy	630.x - 633.x
Other pregnancy with abortive outcome	634.x - 639.x
Complications mainly related to pregnancy	640.x - 649.x
Normal delivery, and other indications for care in pregnancy, labor, and delivery	650.x - 659.x
Complications occurring mainly in the course of labor and delivery	660.x - 669.x
Complications of the puerperium	670.x - 677.x
Carbuncle and furuncle	680.x
Cellulitis and abscess	681.x, 682.x
Acute lymphadenitis	683.x
Impetigo	684.x
Pilonidal cyst	685.x
Other local infections of skin and subcutaneous tissue	686.x
Erythematosquamous dermatosis	690.x
Atopic dermatitis and related conditions, contact dermatitis and other eczema	691.x ,692.x
Dermatitis due to substances taken internally	693.x
Bullous dermatoses	694.x
Erythematous conditions	695.x
Psoriasis and similar disorders	696.x
Lichen	697.x
Pruritus and related conditions	698.x
Corns and callosities	700.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Other hypertrophic and atrophic conditions of skin	701.x
Other dermatoses	702.x
Diseases of nail	703.x
Diseases of hair and hair follicles	704.x
Disorders of sweat glands	705.x
Diseases of sebaceous glands	706.x
Chronic ulcer of skin	707.x
Urticaria	708.x
Other disorders of skin and subcutaneous tissue	709.x
Diffuse diseases of connective tissue	710.x
Arthropathy associated with infections	711.x
Crystal arthropathies	712.x
Arthropathy associated with other disorders classified elsewhere	713.x
Rheumatoid arthritis and other inflammatory polyarthropathies	714.x
Osteoarthrosis and allied disorders	715.x
Other and unspecified arthropathies	716.x
Internal derangement and other unsprecified disorders of joins	717.x, 718.x,
(includes knees)	719.x
Ankylosing spondylitis and other inflammatory spondylopathies	720.x
Spondylosis and allied disorders	721.x
Intervertebral disc disorders	722.x
Other disorders of cervical region and back	723.x, 724.x
Polymyalgia rheumatica	725.x
Peripheral enthesopathies and allied syndromes	726.x
Other disorders of synovium, tendon, and bursa	727.x
Disorders of muscle, ligament, and fascia	728.x
Other disorders of soft tissues	729.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Osteomyelitis, periostitis, and other infections involving bone	730.x
Osteitis deformans and osteopathies associated with other disorders classified elsewhere	731.x
Osteochondropathies	732.x
Other disorders of bone and cartilage	733.x
Flat foot	734.x
Acquired deformities of limbs (excluding curvature of spine)	735.x, 736.x, 738.x
Curvature of spine	737.x
Nonallopathic lesions, not elsewhere classified	739.x
Anencephalus and similar anomalies	740.x
Spina bifida	741.x
Other congenital anomalies of nervous system	742.x
Congenital anomalies of eye, ear, face, and neck	743.x, 744.x
Bulbus cordis anomalies and anomalies of cardiac septal closure	745.x
Other congenital anomalies of heart and the circulatory system	746.x, 747.x
Congenital anomalies of respiratory system	748.x
Cleft palate and cleft lip	749.x
Other congenital anomalies of upper alimentary tract and digestive system	750.x
Congenital anomalies of genital organs	752.x
Congenital anomalies of urinary system	753.x
Certain congenital musculoskeletal deformities	754.x
Other congenital anomalies of limbs and musculoskeletal anomalies	755.x, 756.x
Congenital anomalies of the integument	757.x
Chromosomal anomalies	758.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Other and unspecified congenital anomalies	759.x
Maternal causes of perinatal morbidity and mortality	760.x - 763.x
Other conditions originating in the perinatal period	764.x - 779.x
General symptoms	780.x
Symptoms involving nervous and musculoskeletal systems	781.x
Symptoms involving skin and other integumentary tissue	782.x
Symptoms concerning nutrition, metabolism, and development	783.x
Symptoms involving head and neck	784.x
Symptoms involving cardiovascular system	785.x
Symptoms involving respiratory system and other chest symptoms	786.x
Symptoms involving digestive system	787.x
Symptoms involving urinary system	788.x
Other symptoms involving abdomen and pelvis	789.x
Nonspecific findings on examination of blood	790.x
Nonspecific findings on examination of urine	791.x
Nonspecific abnormal findings in other body substances	792.x
Nonspecific abnormal findings on radiological and other examination of body structure	793.x
Nonspecific abnormal results of function studies	794.x
Other and nonspecific abnormal cytological, histological, immunological and dna test findings	795.x
Other nonspecific abnormal findings	796.x
Ill-defined and unknown causes of morbidity and mortality	797.x - 799.x
Persons with potential health hazards related to communicable diseases	v01.x - v06.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Persons with need for isolation, other potential health hazards and prophylactic measures	v07.x - v09.x
Persons with potential health hazards related to personal and family history	v10.x - v19.x
Persons encountering health services in circumstances related to reproduction and development	v20.x - v29.x
Liveborn infants according to type of birth	v30.x - v39.x
Persons with a condition influencing their health status	v40.x - v49.x
Elective surgery for purposes other than remedying health states	v50.x
Aftercare involving the use of plastic surgery	v51.x
Fitting and adjustment of prosthetic device, implant or other device	v52.x, v53.x
Other orthopedic aftercare	v54.x
Attention to artificial openings	v55.x
Encounter for dialysis and dialysis catheter care	v56.x
Care involving use of rehabilitation procedures	v57.x
Encounter for other and unspecified procedures and aftercare	v58.x
Donors	v59.x
Persons encountering health services in other circumstances	v60.x - v69.x
General medical examination	v70.x
Observation and evaluation for suspected conditions not found	v71.x
Special investigations and examinations	v72.x
Special screening examination for viral and chlamydial diseases	v73.x
Special screening examination for bacterial and spirochetal diseases	v74.x
Special screening examination for other infectious diseases	v75.x
Special screening for malignant neoplasms	v76.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Special screening for endocrine, nutritional, metabolic, and immunity disorders	v77.x
Special screening for disorders of blood and blood-forming organs	v78.x
Special screening for mental disorders and developmental handicaps	v79.x
Special screening for neurological, eye, and ear diseases	v80.x
Special screening for cardiovascular, respiratory, and genitourinary diseases	v81.x
Special screening for other conditions	v82.x
Genetics	v83.x - v84.x
Body mass index	v85.x
Estrogen receptor status	v86.x
A transport accident	e800.x - e848.x
Place of occurrence	e849.x
Accidental poisoning	e850.x - e858.x, e860.x - e869 .x
Misadventures to patients during surgical and medical care, and abnormal reaction to treatment	e870.x - e876.x, e878.x - e879.x
Accidental falls	e880.x - e888.x
Other accidents	e890.x - e928.x
Late effects of accidental injury	e929.x
Drugs, medicinal and biological substances causing adverse effects in therapeutic use	e930.x - e949.x
Suicide and self - inflicted injury	e950.x - e959.x
Homicide and injury purposely inflicted by other persons	e960.x - e969.x
Legal intervention	e970.x - e978.x

Continued on next page...

Table 3.19 – Continued

Description	ICD-9 Code
Injury undetermined whether accidentally or purposely inflicted	e980.x - e989.x
Injury resulting from operations of war, or terrorism	e979.x, e990.x - e999.x

Table 3.19 Grouping of ICD-9 Codes, "x" at the end of a code stands for any number.

Bibliography

[1] Centers for medicare & medicaid services: Diagnosis and procedure codes and their abbreviated titles. version 22, effective october 1, 2004. (Accessed August 28, 2006, at http://www.cdc.gov/nchs/datawh.htm#International%20Classification.).

[2] Classification rule with unbiased interaction selection and estimation (cruise). (Binaries accessed July 20th, 2007, at http://www.stat.wisc.edu/ loh/cruise.html).

[3] Eigencluster version 1.1. Accessed August 28, 2006, at http://eigencluster.csail.mit.edu/.

[4] Icd-9 codes online. (Accessed July 1, 2008, at http://icd9cm.chrisendres.com/).

[5] National drug code directory (a list of fda approved drugs), as of december 31, 2004. (Accessed August 28, 2006, at http://www.fda.gov/cder/ndc/database/default.htm.).

[6] Pfizer inc. information about viagra. (Accessed July 28th, 2008, at http://www.viagra.com).

[7] *Approved Drug Products with Therapeutic Equivalence Evaluations.* U.S. Department of Health and Human Services, 2008.

[8] S. Alesci, P. E. Martinez, S. Kelkar, I. Ilias, D. S. Ronsaville, S. J. Listwak, A. R. Ayala, J. Licinio, H. K. Gold, M. A. Kling, G. P. Chrousos, and P. W. Gold. Major depression is associated with significant diurnal elevations in plasma

interleukin-6 levels, a shift of its circadian rhythm, and loss of physiological complexity in its secretion: clinical implications. *J Clin Endocrinol Metab*, 90(5):2522–2530, May 2005.

[9] R. Altman, G. Alarcn, D. Appelrouth, D. Bloch, D. Borenstein, K. Brandt, C. Brown, T. D. Cooke, W. Daniel, and D. Feldman. The american college of rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum*, 34(5):505–514, May 1991.

[10] A. S. Ash, R. P. Ellis, G. C. Pope, J. Z. Ayanian, D. W. Bates, H. Burstin, L. I. Iezzoni, E. MacKay, and W. Yu. Using diagnoses to describe populations and predict costs. *Health Care Financ Rev*, 21(3):7–28, 2000.

[11] J. A. Berlin, S. C. Glasser, and S. S. Ellenberg. Adverse event detection in drug development: Recommendations and obligations beyond phase 3. *Am J Public Health*, Jun 2008.

[12] D. Bertsimas and R. Freund. *Data, Models and Decisions: the Fundamendals of Management Science.* Dynamic Ideas, Belmont, Massachusetts, 2005.

[13] C. Bombardier, L. Laine, A. Reicin, D. Shapiro, R. Burgos-Vargas, B. Davis, R. Day, M. B. Ferraz, C. J. Hawkey, M. C. Hochberg, T. K. Kvien, T. J. Schnitzer, and V. I. G. O. R. S. Group. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. vigor study group. *N Engl J Med*, 343(21):1520–8, 2 p following 1528, Nov 2000.

[14] P. L. Bonate. *Analysis of Pretest-Posttest Designs.* Chapman & Hall/CRC, 2000.

[15] L. Breiman, J. Friedman, R. Olshen, and J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, California, 1984.

[16] J. S. Brown, M. Kulldorff, K. A. Chan, R. L. Davis, D. Graham, P. T. Pettus, S. E. Andrade, M. A. Raebel, L. Herrinton, D. Roblin, D. Boudreau, D. Smith, J. H. Gurwitz, M. J. Gunter, and R. Platt. Early detection of adverse drug

events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiol Drug Saf*, 16(12):1275–1284, Dec 2007.

[17] T. G. Burnakis. Cardiovascular events and cox-2 inhibitors. *JAMA*, 286(22):2808; author reply 2811–2808; author reply 2812, Dec 2001.

[18] R.-E. Chang and C.-L. Lai. Use of diagnosis-based risk adjustment models to predict individual health care expenditure under the national health insurance system in taiwan. *J Formos Med Assoc*, 104(12):883–890, Dec 2005.

[19] D. Cheng, R. Kannan, S. Vempala, and G. Wang. A divide-and-merge methodology for clustering. *Transactions on Database Systems*, 31:1499–1525, 2006.

[20] G. Cizza, P. Ravn, G. P. Chrousos, and P. W. Gold. Depression: a major, unrecognized risk factor for osteoporosis? *Trends Endocrinol Metab*, 12(5):198–203, Jul 2001.

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

[22] D. Czerwinski. *Quality of care and drug surveillance: A data-driven perspective.* PhD thesis, Massachusetts Institute of Technology, 2008.

[23] P. E. Dans. Looking for answers in all the wrong places. *Ann Intern Med*, 119(8):855–857, Oct 1993.

[24] H. G. Dove, I. Duncan, and A. Robb. A prediction model for targeting low-cost, high-risk members of managed care organizations. *Am J Manag Care*, 9(5):381–389, May 2003.

[25] D. Dunn, A. Rosenblatt, D. Taira, and et al. A comparative analysis of methods of health risk assessment. Technical report, Society of Actuaries Monograph M-HB96-1, 2002. (Accessed August

28, 2006, at http://www.soa.org/ccm/content/research-publications/library-publications/monographs/health-benefits-monographs/.).

[26] I. Epocrates. Epocrates online: Cialis tadalafil adverse reactions. Accessed July 28th 2008 at: https://online.epocrates.com/noFrame/.

[27] J. F. Farley, C. R. Harley, and J. W. Devine. A comparison of comorbidity measurements to predict healthcare expenditures. *Am J Manag Care*, 12(2):110–119, Feb 2006.

[28] C. P. Farrington and G. Manning. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9:1447–1454, 1990.

[29] J. A. Fleishman, J. W. Cohen, W. G. Manning, and M. Kosinski. Using the sf-12 health status measure to improve predictions of medical expenditures. *Med Care*, 44(5 Suppl):I54–I63, May 2006.

[30] Food and D. Administration. Alert for healthcare professionals sildenafil citrate (marketed as viagra). (Accessed July 28th, 2008, at http://www.fda.gov/cder/drug/InfoSheets/HCP/sildenafilHCP.htm).

[31] Food and D. Administration. Fda announces revisions to labels for cialis, levitra and viagra. (Accessed on July 28th, 2008, at http://www.fda.gov/bbs/topics/NEWS/2007/NEW01730.html).

[32] Food and D. Administration. Lab-0221-2.4 viagra (sildenafil citrate) tablets. (Accessed July 28th, 2008, at http://www.fda.gov/cder/foi/label/2005/020895s021lbl.pdf).

[33] F. D. Gianfrancesco, A. L. Grogg, R. A. Mahmoud, R. hua Wang, and H. A. Nasrallah. Differential effects of risperidone, olanzapine, clozapine, and conventional antipsychotics on type 2 diabetes: findings from a large health plan database. *J Clin Psychiatry*, 63(10):920–930, Oct 2002.

[34] D. J. Graham, D. Campen, R. Hui, M. Spence, C. Cheetham, G. Levy, S. Shoor, and W. A. Ray. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet*, 365(9458):475–481, 2005.

[35] W. Insull, J. K. Ghali, D. R. Hassman, J. W. Y. As, S. K. Gandhi, E. Miller, and S. O. L. A. R. S. Group. Achieving low-density lipoprotein cholesterol goals in high-risk patients in managed care: comparison of rosuvastatin, atorvastatin, and simvastatin in the solar trial. *Mayo Clin Proc*, 82(5):543–550, May 2007.

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

[37] A. Jones. *Handbook in Health Economics*, chapter Health Econometrics, pages 265–344. North Holland, 2000.

[38] R. Kannan, S. Vempala, and A. Vetta. On clusterings: good, bad and spectral. *Journal of the ACM*, 51:497–515, 2004.

[39] A. Keith. The economics of viagra. *Health Aff (Millwood)*, 19(2):147–157, 2000.

[40] H. Kim and W.-Y. Loh. Classification trees with unbiased multiway splits. *Journal of the American Statistics Association*, 96:589–604, 2001.

[41] C. N. Klabunde, J. L. Warren, and J. M. Legler. Assessing comorbidity using claims data: an overview. *Med Care*, 40(8 Suppl):IV–26–35, Aug 2002.

[42] M. A. Konstam, M. R. Weir, A. Reicin, D. Shapiro, R. S. Sperling, E. Barr, and B. J. Gertz. Cardiovascular thrombotic events in controlled, clinical trials of rofecoxib. *Circulation*, 104(19):2280–2288, Nov 2001.

[43] L. LaVange, V. Iannacchione, and S. Garfinkel. An application of logistic regression methods to survey data: Predicting high cost users of medical care. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1986. (Accessed August 28, 2006, at http://www.amstat.org/sections/SRMS/proceedings/papers/1986_049.pdf.).

[44] A. G. Lawthers, E. P. McCarthy, R. B. Davis, L. E. Peterson, R. H. Palmer, and L. I. Iezzoni. Identification of in-hospital complications from claims data. is it valid? *Med Care*, 38(8):785–795, Aug 2000.

[45] T. A. Lieu, M. Kulldorff, R. L. Davis, E. M. Lewis, E. Weintraub, K. Yih, R. Yin, J. S. Brown, R. Platt, and for the Vaccine Safety Datalink Rapid Cycle Analysis Team. Real-time vaccine safety surveillance for the early detection of adverse events. *Med Care*, 45(10 Supl 2):S89–S95, Oct 2007.

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

[47] W. G. Manning and J. Mullahy. Estimating log models: to transform or not to transform? *J Health Econ*, 20(4):461–494, Jul 2001.

[48] A. T. Masi, G. G. Hunder, J. T. Lie, B. A. Michel, D. A. Bloch, W. P. Arend, L. H. Calabrese, S. M. Edworthy, A. S. Fauci, and R. Y. Leavitt. The american college of rheumatology 1990 criteria for the classification of churg-strauss syndrome (allergic granulomatosis and angiitis). *Arthritis Rheum*, 33(8):1094–1100, Aug 1990.

[49] J. G. McMurray, R. A. Feldman, S. M. Auerbach, H. Deriesthal, N. Wilson, and O. behalf of the Multicenter Study Group. Long-term safety and effectiveness of sildenafil citrate in men with erectile dysfunction. *Ther Clin Risk Manag*, 3(6):975–981, Dec 2007.

[50] S. S. Mehta, S. Suzuki, H. A. Glick, and K. A. Schulman. Determining an episode of care using claims data. diabetic foot ulcer. *Diabetes Care*, 22(7):1110–1115, Jul 1999.

[51] D. Mukherjee, S. E. Nissen, and E. J. Topol. Risk of cardiovascular events associated with selective cox-2 inhibitors. *JAMA*, 286(8):954–959, 2001.

[52] K. Pietz, C. M. Ashton, M. McDonell, and N. P. Wray. Predicting healthcare costs in a population of veterans affairs beneficiaries using diagnosis-based risk adjustment and self-reported health status. *Med Care*, 42(10):1027–1035, Oct 2004.

[53] M. Pladevall, L. K. Williams, L. A. Potts, G. Divine, H. Xi, and J. E. Lafata. Clinical outcomes and adherence to medications measured by claims data in patients with diabetes. *Diabetes Care*, 27(12):2800–2805, Dec 2004.

[54] C. A. Powers, C. M. Meyer, M. C. Roebuck, and B. Vaziri. Predictive modeling of total healthcare costs using pharmacy claims data: a comparison of alternative econometric cost modeling techniques. *Med Care*, 43(11):1065–1072, Nov 2005.

[55] B. M. Psaty and S. P. Burke. Protecting the health of the public–institute of medicine recommendations on drug safety. *N Engl J Med*, 355(17):1753–1755, Oct 2006.

[56] J. Robbins, C. Hirsch, R. Whitmer, J. Cauley, and T. Harris. The association of bone mineral density and depression in an older population. *J Am Geriatr Soc*, 49(6):732–736, Jun 2001.

[57] D. W. Roblin, P. I. Juhn, B. J. Preston, R. D. Penna, S. P. Feitelberg, A. Khoury, and J. C. Scott. A low-cost approach to prospective identification of impending high cost outcomes. *Med Care*, 37(11):1155–1163, Nov 1999.

[58] D. H. Solomon, R. J. Glynn, R. Levin, and J. Avorn. Nonsteroidal anti-inflammatory drug use and acute myocardial infarction. *Arch Intern Med*, 162(10):1099–1104, May 2002.

[59] M. Van de Ven, P. Wynand, and R. Ellis. *Handbook in Health Economics*, chapter look at me look at me Risk Adjustment in Competitive Health Plan Markets, pages 756–845. North Holland, 2000.

[60] J. T. Wadsworth, K. D. Somers, L. H. Cazares, G. Malik, B.-L. Adam, B. C. Stack, G. L. Wright, and O. J. Semmes. Serum protein profiles to identify head and neck cancer. *Clin Cancer Res*, 10(5):1625–1632, Mar 2004.

[61] J. E. Wennberg, N. Roos, L. Sola, A. Schori, and R. Jaffe. Use of claims data systems to evaluate health care outcomes. mortality and reoperation following prostatectomy. *JAMA*, 257(7):933–936, Feb 1987.

[62] J. Zhang, E. L. Ding, and Y. Song. Adverse effects of cyclooxygenase 2 inhibitors on renal and arrhythmia events: meta-analysis of randomized trials. *JAMA*, 296(13):1619–1632, Oct 2006.

[63] Y. Zhao, A. S. Ash, R. P. Ellis, J. Z. Ayanian, G. C. Pope, B. Bowen, and L. Weyuker. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Med Care*, 43(1):34–43, Jan 2005.

[64] Y. Zhao, R. P. Ellis, A. S. Ash, D. Calabrese, J. Z. Ayanian, J. P. Slaughter, L. Weyuker, and B. Bowen. Measuring population health risks using inpatient diagnoses and outpatient pharmacy data. *Health Serv Res*, 36(6 Pt 2):180–193, Dec 2001.

[65] X. H. Zhou, C. A. Melfi, and S. L. Hui. Methods for comparison of cost data. *Ann Intern Med*, 127(8 Pt 2):752–756, Oct 1997.

[66] Z. Zhou, E. Rahme, and L. Pilote. Are statins created equal? evidence from randomized trials of pravastatin, simvastatin, and atorvastatin for cardiovascular disease prevention. *Am Heart J*, 151(2):273–281, Feb 2006.