Quality of Care and Drug Surveillance: A Data-Driven Perspective

by

David Czerwinski

B.S., Stanford University (1998)

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

June 2008

Author .....
Sloan School of Management
May 15, 2008

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

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

Quality of Care and Drug Surveillance:

A Data-Driven Perspective

by

David Czerwinski

Submitted to the Sloan School of Management
on May 15, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

In this thesis, we describe the use of medical insurance claims data in three important areas of medicine. First, we develop expert-trained statistical models of quality of care based on variables derived from insurance claims. Such models can be used to identify patients who are receiving poor care so that interventions can be arranged to improve their care. Second, we develop an algorithm that utilizes claims data to perform post-marketing surveillance of drugs to detect previously unknown side effects. The algorithm performed strongly in several realistic simulation tests, detecting side effects a large fraction of the time while controlling the false detection rate. Lastly, we use insurance claims data to improve our understanding of the costs of care for patients who suffer from depression and a chronic disease.

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

Acknowledgments

Neither this thesis nor my survival at MIT would have been possible if not for the help of many, many others whom it is very gratifying to thank in print. I would like to thank my advisor, Dimitris Bertsimas, for inviting me to explore insurance claims data, for teaching me many of his tricks for learning from numbers, and for his guidance, encouragement, and counsel throughout my research. Thanks to Arnie Barnett for his kindness that began even before my arrival on campus, for introducing me to academic research during my first two years at MIT, and for the opportunity to soak up his teaching skills as his TA. Thanks to Georgia Perakis for her help and invaluable advice about all matters academic and for being a role model in the MBA classroom. Thanks to Retsef Levi for a valuable TA experience and for his help during my job search. Thanks to the ORC staff – Paulette, Laura, and Andrew – for making things run smoothly. And thanks to Anna Piccolo, Christine Liberty, Shiba Nemat-Nasser, and Conor Murphy – it was always nice to see a friendly face in E53.

This research would not have been possible without D2 Hawkeye and the farsighted leadership of Chris Kryder and Rudra Pandey. I also want to thank Bijay Ghimire and Anil Shrestha of D2 for their invaluable assistance. Thanks to my collaborators, Dr. Michael Kane of MIT Medical and Dr. Charles Welch of Massachusetts General Hospital, for their enthusiasm, their tolerance of my ignorance of health care, and all they taught me to diminish that ignorance. Thanks to Dr. Thorvadur Love of Brigham and Women's Hospital for reading through cases for our study of quality of care. Thanks to the gang at the ORC – my office mate Margret, my classmates Nelson, Hamed, Pavithra, Pamela, Carine, Stephen, and Mohamed, the wise ORC alums who have since graduated and the young 'uns too numerous to mention, but you know who you are.

Of course, I wasn't at the ORC 24 hours a day! A thank you too big to fit on the printed page to my wife Maria for so many things – for getting my butt in gear in the first place to return to grad school, for suffering through more sub-zero weather than any California girl deserves to, for putting up with many nights with me at the office, and even more nights with the three of us crammed into our shoebox-sized apartment. But mostly for her love, support, good spirits, and encouragement. Kiely, for a good game of soccer, the wake-up calls, and the purrs. My mom and dad, Laura, Martina, Jeanie, Dale, Tressie, Ellie, Katie, and Jon for all of their love and support. My grandparents, especially my Grandpa Larry. The young ladies of McCormick Hall for being so well-behaved. Kathy and Charles and the McCormick house team.

And of course, for providing sustenance, Bartley's Burgers, Pizzeria Regina, The Uppercrust, Emma's, UBurger, Rebecca's Cafe, and the food trucks. A special thank you to Anna's Taqueria for being a constant reminder that this is not California. The student DJs at WERS. RAND for an enjoyable summer in Santa Monica.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis concerns three areas of health care: quality of care, drug surveillance, and cost of care. The research makes use of a large database of health insurance claims which cover a two-year period for 600,000 patients. The data capture all interactions with the health care system that are covered by insurance including outpatient visits, hospital stays, and prescriptions.

The thesis is organized as follows. In Chapter 1 we discuss the development of statistical models to measure quality of care in diabetes patients. In Chapter 2 we develop algorithms for performing large scale post-marketing drug surveillance. Chapter 3 reports our findings on the cost of care of patients who suffer from depression and a chronic disease.

1.1 Insurance claims data

The research was conducted using medical claims data compiled by D2 Hawkeye, Inc., a medical analytics company based in Waltham, Massachusetts. D2 Hawkeye provides computerized medical claims analyses for insurers, third-party administrators, risk-bearing medical groups, and self-insured employers. All insurance-based health care utilization by study subjects was reflected in this database.

Medical claims are generated when a patient visits a doctor or has a hospital stay. Hospital stays generally generate many claims while doctors visit will typically

generate one to five claims. A medical claim records the diagnosis, the procedure performed, the date, the doctor, the patient (identified by a unique ID number), and several monetary values for billing purposes. Monetary values include the amount that the doctor charged, the amount that the insurance company allows for the service provided, the amount the insurer paid, and the amount the patient paid. The amount that the insurer paid is the most reliably recorded, and it is the value that we use in our research. The diagnosis must be handled carefully – if the patient is tested for a particular disease, that disease will be recorded as the diagnosis on the insurance claim, regardless of the test's outcome. For a routine doctors visit, if the patient has two or more diseases only one disease might be recorded on the claim.

Pharmaceutical claims are generated when a patient fills a prescription. The claim records the drug that was dispensed, the doctor who prescribed it, the patient, the date the prescription was dispensed, the number of days of supply that were dispensed, and the monetary values for billing purposes.

Aside from the claims themselves the database also contains other administrative information. The only other table relevant for our work is the eligibility table. For each person, it records when their insurance coverage began (and if they are no longer insured, when it ended), their birthdate, gender, ZIP code, and the type of member that they are (an employee, a spouse, or a dependent of the employee). A person may have multiple eligibility records if they left the insurance plan and later rejoined. Their identification number would remain the same for all records in the eligibility table and for all of their claims.

The database contains approximately 2.1 million people, 70 million medical claims, and 29 million pharmaceutical claims. The 10 most frequent diagnoses, procedures, and prescribed drugs are shown in Tables 1.1, 1.2, and 1.3.

Common diagnoses include high blood pressure (hypertension and benign essential hypertension), high cholesterol (hyperlipidemia), and chest pain. Well-child care is a diagnosis recorded when a child visits the doctor for a checkup. Diabetes, lumbago (lower back pain), and acute pharyngitis (a sore throat) are also frequent diagnoses.

The most frequent procedures are office visits and different types of lab work and

Diagnosis	Claims
Well Child Care	1,818,465
Routine General Medical Examination	1,090,132
Hypertension	895,062
Chest Pain	866,573
Diabetes Mellitus without complications	859,408
Hyperlipidemia	808,467
Gynecological Examination	775,985
Lumbago	691,741
Benign Essential Hypertension	686,958
Acute Pharyngitis	650,261

Table 1.1: The 10 most frequent diagnoses in the database.

Procedure	Claims
Office/Outpatient Visit	8,449,886
Laboratory - Clinical	1,530,925
Venipuncture, Routine	1,387,652
Lab / Chemistry	1,233,029
Blood, hemogram& platelet count	945,782
Emergency Room	835,572
Lipid Profile	745,112
Psychiatry Visit	705,292
Physical Therapy	701,301
Chiropractic Manipulation	692,368

Table 1.2: The 10 most frequent procedures in the database.

blood draws. Emergency room visits are frequent, as are psychiatry visits, physical therapy, and visits to chiropractors.

The most frequently prescribed drugs include antibiotics (amoxicillin and azithromycin) and blood pressure medications (lisinopril, metoprolol, atenolol, and amlodipine besylate). Levothyroxine is frequently prescribed for hypothyroidism, hydrocodone for pain relief, atorvastatin for high cholesterol, and albuterol for asthma.

Table 1.4 shows two months of claims data for a typical diabetic in the database. On September 16, 2003 the patient had an office visit and some tests were performed to monitor their diabetes. Later that month they filled prescriptions for glucose test strips (One Touch Ultra), insulin (Novolog), and a drug to treat neurological side

Drug	Claims
Amoxicillin	816,377
Levothyroxine	738,064
Hydrocodone	728,628
Atorvastatin	580,112
Lisinopril	557,330
Albuterol	448,860
Azithromycin	442,960
Metoprolol	406,891
Atenolol	390,484
Amlodipine Besylate	336,035

Table 1.3: The 10 most frequently prescribed drugs in the database.

Date	Provider	Diagnosis	Procedure/Rx
2003-09-16	C LANCASTER MD	Diabetes	Urinalysis, By Dip Stick or
2003-09-16	C LANCASTER MD	Diabetes	Glucose Blood Test
2003-09-16	C LANCASTER MD	Diabetes	Hemoglobin, Glycated
2003-09-16	C LANCASTER MD	Diabetes	Office/Outpatient Visit, Est
2003-09-17	C LANCASTER MD	Pharmacy	ONE TOUCH ULTRA
2003-09-22	C LANCASTER MD	Pharmacy	NOVOLOG
2003-09-26	SINK, E DVM	Pharmacy	NEURONTIN
2003-11-26	R BEDGOOD MD	Influenza	Comprehensive Metabolic
2003-11-26	R BEDGOOD MD	Influenza	Blood, Occult, By Peroxid
2003-11-26	R BEDGOOD MD	Influenza	Flu Vaccine, Whole, Im

Table 1.4: Two months of claims data for a typical diabetic in the database. Not all fields are shown. Pharmaceutical claims are designated with "Pharmacy" in the Diagnosis column.

effects of diabetes (Neurontin). In November they received a flu vaccine and had a blood test.

Because the primary purpose of claims data is for billing, not analysis, its suitability for analysis has been studied previously. Most studies of the accuracy of diagnostic coding in claims data have used medical records as a benchmark.Two common measures of accuracy in the medical literature are sensitivity and specificity. As it pertains to claims data, sensitivity measures the fraction of time claims data accurately record something that happened (e.g., the doctor made a diagnosis or performed a procedure). For example, if the claims data for 100 diabetic patients were examined and

it was found that a diagnosis of diabetes was only recorded for 75 of the patients, the sensitivity of the claims data would be 75%. Specificity measures the fraction of time claims data accurately reflect that something didn't happen. For example, if the claims data for 100 patients who were hospitalized but didn't have x-rays taken were examined and it was found that 5 patients' claims data actually included charges for x-rays, the specificity of the claims data would be 95%.

The sensitivity of coding for individual diagnoses in claims data varies widely, but reported aggregate sensitivity in US diagnostic data is more consistent, from a low of under 50% [49] to a high of 78% [69]. Higher sensitivity is reported for diagnoses that are acute or symptomatic, and lower for diagnoses that are chronic or asymptomatic [82, 69, 113, 46]. The sensitivity of coding for comorbidities is lower than for primary diagnoses [48], and decreases as the number of comorbidities increase [73]. The reported sensitivity of individual diagnostic codes is more consistent. For instance, diabetes was coded with a reported sensitivity of 81-83% in three separate studies of claims data, using medical records as a benchmark [49, 35, 76].

By contrast, recent studies of sensitivity of procedure codes show fairly close agreement between claims data and the medical record for those procedure codes which are applicable to both, with reported correlation rates of 94%-97% [33, 64, 29]. However, minor procedures performed as part of routine care are more likely to be absent from both types of data [77].

Prescription claims data correlate closely with medical record data [101], and give a more accurate record of drugs actually dispensed [19]. In addition, combining prescription codes and diagnostic codes for a particular disease significantly improves both the sensitivity and specificity of diagnostic coding [15].

1.2 Quality of care

The quality of care patients receive in the US varies considerably. If markers of poor care can be identified in a timely, automated fashion, then interventions (such as case management) could be arranged to improve the care of individual patients. Our ob-

jective was to develop a statistical model of quality of care for diabetes patients based on insurance claims data. We used an expert-trained logistic regression model. The model was developed on a set of 101 diabetes patients whose quality of care was rated by a physician. An out-of-sample validation was performed on an additional set of 30 patients. A second physician also reviewed the set of 30 patients so that inter-rater reliability could be assessed. The data set consisted of medical and pharmaceutical claims over the period 2003-2005. The patients were diabetic, ages 35-55, with annual health-care costs between $10,000 and $20,000. The main outcome measure was the out-of-sample classification accuracy of the logistic regression model. Patients were classified as receiving either good care or poor care. Two models performed particularly well. The best model achieved an out-of-sample accuracy of 80%, compared with a baseline of 63%. We conclude that expert-trained statistical models based on insurance claims data can identify patients receiving poor care accurately enough to be of use in practice. Such models could be used to select and prioritize patients for interventions to improve care.

1.3 Drug surveillance

At the point when a drug is released on the US market, only a relatively small number of people have been exposed to it during clinical trials. When a larger number of people begin using the drug, unexpected side effects (positive or negative) may be discovered and we use insurance claims data to detect them in near "real-time." Claims data is promising for this use because it is available electronically for a large number of people and is updated frequently. A key challenge in drug surveillance is avoiding the loss in statistical power associated with testing a large number of hypotheses. We present several promising methods for addressing this challenge. The methods were tested using simulation and performed strongly, detecting side effects a high percent of the time while maintaining a low false detection rate.

1.4 Depression and cost of care

Over the past 25 years, a growing body of evidence has established an association between depression and high utilization of general medical services. We used claims data to study the association between depression and health-care costs in 11 chronic diseases. We found that depressed patients have higher costs than not depressed patients across all 11 diseases. In most of the diseases, the cost increase occurs mainly in outpatient services and pharmaceuticals. Depressed patients also have a higher mean number of comorbidities than not depressed patients, though this doesn't account for the total cost increase. The prevalence of depression is higher in each of the 11 comorbid diseases than in the total research cohort, and the prevalence of each chronic comorbid disease is higher in the depressed cohort than in the total research cohort. There is a linear association between annual cost of care and prevalence of depression in 10 of 11 chronic comorbid diseases. Depression is associated with an increased risk for subsequent onset of all 11 comorbid diseases. All 11 comorbid diseases are associated with an increased risk for subsequent onset of depression.

1.5 Contributions

The contributions of this thesis are several. In our work on quality of care, we have shown how a decidedly non-quantitative concept, quality, can be modeled statistically with the involvement of a subject matter expert. The findings show that quality of care can be measured accurately enough using claims data to improve the way that case management and other methods of health care intervention are targeted. Our work on drug surveillance introduces a new approach to multiple hypothesis testing. The approach combines techniques from both statistics and operations research. It can reduce the time needed to discover harmful side effects, potentially saving many lives. Finally, our work on depression and cost of care sheds light on the interplay between depression and chronic diseases. Leveraging the size of our claims database, we were able to study a wider spectrum of diseases than had been previously stud-

ied. We were also able to provide a detailed examination of the sources of increased costs. Such an understanding will hopefully lead in the future to better treatment for patients suffering from depression and a chronic disease.

Chapter 2

Measuring Quality in Diabetes Care: An Expert-based Statistical Approach

It has been demonstrated in recent years that many patients in the United States do not receive high-quality health care [89, 21]. In this chapter, we address the problem of identifying, in an automated fashion, diabetes patients who may be receiving poor care so that interventions can be arranged to improve their care. We measure the quality of care with an expert-trained statistical model using variables derived from medical insurance claims data.

We focus on patients with diabetes for several reasons. First, it is a widespread, costly disease. Over 20 million Americans are diabetic – about 7% of the US population – and the annual cost of diabetes is estimated at $132 billion. One in every 10 health care dollars goes towards treating Diabetes [17, 3]. Second, there are well established guidelines for its treatment, and third, limiting the study to one disease minimizes variations in care from patient to patient that aren't related to quality.

We are interested in identifying individual patients with poor care in "real-time" so that interventions can be arranged to improve their care. This is somewhat distinct from previous studies which have measured quality in order to assess the US healthcare system [89, 21, 10, 23, 2, 88, 6, 61, 42], to provide rankings of doctors and hospitals

[104, 45, 37], and to rate providers for "pay-for-performance" type reimbursement [74, 80].

To measure quality, we use statistical models trained on a set of patients whose care was assessed by a physician. The variables in the models are derived from the patients' insurance claims data. We use claims data because in practice they are the only electronically available, timely source of information about the care a patient has received. With them, the care received by a large population of patients can be monitored on an ongoing basis. Other methods of measuring quality, such as reviewing paper medical records, may be more thorough but they do not scale because of the manual labor involved.

That is not to say that claims data are without drawbacks. They lack clinical details such as symptoms, test results, and severity of disease. They reflect little about the patient's quality of life. Though overall the coding of diagnoses and procedures in claims data are accurate, they can somtimes be vague [51]. When there are multiple diagnoses during a single visit some may not be captured. Minor non-monetized procedures – such as counseling a patient to stop smoking – are usually not recorded.

Our statistical models measure the quality of the process of care [13, 12]. In trying to improve the care for a particular patient, structural aspects of care are less relevant since they are fixed over the short term. Ideally, we would measure outcomes of care but in general outcomes are difficult to infer from claims data since lab results, symptoms, etc. are not captured [111]. For example, an insurance claim may record that a diabetes patient had a glycated hemoglobin test, but it will not record the results of the test. Whether the glucose level is improving or not cannot be determined. Furthermore, because we have data only over a two-year period, long-term outcomes cannot be measured. Though we didn't instruct the physician to look specifically at process of care, that is what *de facto* was available to him.

Of course, process measures of quality of care for diabetes exist in the form of the guidelines of the American Diabetes Association [7] and others. These guidelines have been developed based on the best available evidence and, where conclusive evidence is still lacking, consensus of expert opinion. However, there are many aspects of a

patient's care which are beyond the purview of the guidelines. Measuring quality of care for patients with multiple diseases can be problematic [57]. Guidelines in general focus on the optimal treatment of a single condition, where in reality an individual may have several coexistent disorders, and treatment demands for one disease may conflict with recommendations for others. Some have argued that there are not enough hours in the day for a physician to provide all of the care that is specified by each of the guidelines when a patient has multiple diseases. Recent work shows that this may not be the case, however [44]. Still, when a patient has multiple diseases the guidelines cannot be taken literally in cases where the treatment for one disease conflicts with the treatment for the other. Finally, intangible aspects of care might be difficult to capture in written guidelines.

By having a physician review the claims data, we obtain a holistic view of the patient's care. We are able to take into account not just the care for their diabetes but for comorbidities and routine preventive care. Nevertheless, the guidelines are relevant and below we discuss how compliance with the guidelines correlated with the physician's assessment of care.

Though we have built a statistical model to identify poor quality care, we do not claim that the model *defines* poor care. Identifying poor care is not equivalent to defining it. For example, consider the statistical models used by credit card companies to identify fraudulent patterns of transactions. The use of a credit card in rapid succession at gas stations may be a red-flag that the card has been stolen. But that is not to say that it is wrong for a person to use their own credit card in rapid succession at gas stations. It is simply a fact that such behavior is correlated with fraud. In the same way, if our statistical models incorporate the use of narcotics as a flag for poor care this does not mean that all uses of narcotics are inappropriate. It simply means that there is a correlation between the use of narcotics and poor care. We feel that an advantage of our approach is that it doesn't rely on an explicit definition of quality.

2.1 Methods

From a large claims database, we randomly selected 101 diabetes patients aged 35-55 with costs over the two-year study period (September 1, 2003 to August 31, 2005) between $10,000 and $20,000. The lower bound on the cost was to ensure that each patient had enough claims data so that the reviewer could make an assessment of the care they received. The upper bound was to ensure that the claims record was not so long that it became impractical to review. To identify patients with diabetes we required that over the two-year period they had either two outpatient diagnoses of diabetes or one inpatient diagnosis of diabetes.

The claims data consists of all insurance-based healthcare utilization for the patients in the study. Claims for medical services record the date of service, provider, diagnoses, procedures performed, and the amount paid. Claims for prescription drugs record the date the prescription was filled, the prescribing physician, the drug, the number of days of supply, and the amount paid.

We attempted to oversample patients who might have received poor care so as to ensure their representation in the sample. Of course, without a measure of the quality of care at the outset, we couldn't do this exactly. As an approximation, we scored the patients based on the presence of hemoglobin HbA1c tests, lipid profiles, and eye exams in their claims data [110]. We then drew a stratified random sample by score, oversampling the lower scores (i.e., people with zero or one of the above procedures performed).

Dr. Michael Kane, a physician at MIT Medical, reviewed the claims record for each of the 101 patients and scored the quality of care they received. He rated the care on a three-point scale: poor, average, or good. He also rated his confidence in his assessment on a two-point scale: confident or not confident. In addition, he wrote a summary description of each patient and the care they received and noted aspects of it that influenced his rating. This information was used later in developing variables for the statistical models.

Dr. Kane reviewed and rated 30 additional patients (not used to develop the

models) in order to validate the models. A second physician, Dr. Thorvadur Love of Brigham and Women's Hospital in Boston, MA, also rated the patients in the validation set independently of Dr. Kane. Having a second physician rate the patients allowed us to assess the extent to which the models reflected beliefs about quality specific to Dr. Kane. The contrast between the backgrounds and experience of the two doctors is marked. Dr. Kane was trained in the United States, while Dr. Love was trained abroad. Dr. Kane has over 30 years of experience, whereas Dr. Love recently completed his residency.

2.1.1 Model Development and Evaluation

We modeled the data using logistic regression. We evaluated other modeling approaches as well, including classification trees, random forests, support vector machines, the Lasso, and several *ad hoc* optimization-based methods that we developed ourselves. Logistic regression outperformed the other methods. Perhaps with a larger data set the other approaches would have been more advantageous.

The dependent variable in our models is quality. Since we are mainly concerned with identifying poor quality care, we grouped the average and good care patients together into a single group which we will refer to as the good care group. A value of 1 for the quality variable indicates good quality, and a value of 0 indicates poor quality.

We used independent variables that could be calculated from the patients' claims data. Most of these variables capture general aspects of care but some are specifically inspired by the physician's comments. However, we avoided defining variables that would only apply to one or two patients in the sample since we wouldn't be able to make any statistically meaningful statements about such variables. For example, one patient was judged to have received poor care because she was treated over a long period with an antibiotic for a urinary tract infection but without regular gynecological exams. (When she finally did have a gynecological exam uterine cancer was discovered.) As this situation arose with only one patient, we would not be able to statistically assess the value of a quality indicator such as "on antibiotic for a urinary

tract infection without gynecological exams."

The variables fall into categories related to diabetes treatment, patient demographics, healthcare utilization, markers of good care, markers of poor care, providers, claims, and prescriptions. The full list of variables with their definitions can be found in the online appendix.

We also tried incorporating the information about the physician's confidence as well as disaggregating the good care group out into average care and good care. Neither of these improved the models' ability to accurately identify patients whose care was poor.

We calculated the accuracy of a model as the percent of patients that it classified correctly (i.e., that matched the physician's classification). The assessment of the models was based on both their in-sample and out-of-sample accuracy. Because we had a limited number of observations, rather than splitting our 101 observations into a training and a validation set we used bootstrap resampling to estimate out of sample accuracy. The resampling procedure operates as follows. We draw 101 samples, with replacement, from the data set. In general, some patients will be sampled more than once and others not at all. We fit the model to the 101 sampled observations and then used this model to classify the patients who were not sampled. We repeat this process a large number of times (500 in practice) and estimate the out-of-sample accuracy of the model across the 500 bootstrap trials, adjusting for bias as discussed in [41]. After all model selection and fitting was complete and we arrived at a final set of ten models, we performed a true out of sample test on 30 additional cases.

We began by studying the individual relationships between each variable and quality. We used logistic regression to classify the patients, using a separate model for each variable. We next moved on to logistic regression models with three predictor variables. We performed an exhaustive search of the model space by fitting each possible three-variable model to the 101 observations and calculating their accuracy, then further assessing the 50 models with the highest accuracy by calculating their out-of-sample accuracy using bootstrap validation.

Because the majority of patients received good care, the simplest predictive model

would be to blindly classify each patient's care as good; 78% of patients would be accurately classified. This serves as a useful baseline against which to assess our models. Another natural baseline model is one using only variables based on the diabetes treatment guidelines. A logistic regression model based on only these variables also had an accuracy of 78%. Comparing the performance of our models to such a model would reveal whether looking at aspects of care beyond the guidelines is of value in matching the physician with respect to assessing quality.

2.2 Results

2.2.1 Data Summary

Table 2.1 shows a summary of the scores for the 101 patients. 78% received average or better care. The physician had high confidence in 76% of his assessments. Most of the cases of low confidence occurred in the "average care" group ($\chi^2 = 10.6$, $p = 0.005$).

	Low Confidence	High Confidence
Low Quality	5	17
Average Quality	16	25
High Quality	3	35

Table 2.1: Summary of the physician's quality ratings and his confidence in them.

Here is an example descriptive paragraph for a patient who received good care:

> 45 year old type 2 diabetic on metformin and glyburide. Also took lexapro and ambien regularly, and crestor. He carried a diagnosis of sarcoid for first part of the analysis period and was treated with prednisone for a while-that's appropriate for sarcoid, even in a diabetic. He was also appropriately covered with fosamax initially. He seems to have changed PCP's in mid cycle. Had first labs 2/04. Had only one ER visit for a forehead laceration. He had a stress test 5/16/05 for chest pain, followed by a catheterization on 5/27/05. Apparently nothing worrisome was found. Despite sarcoid diagnosis he had no chest X-rays or pulmonary function

tests. Had one podiatry visit in June '05. No home testing, no eye exams. Overall, given pulmonary and mental health comorbidities, care looks good with high confidence.

To determine how much variability there was in diabetes care among the patients in the sample we assessed the compliance of the patients' care with three measures from the diabetes guidelines (glycated hemoglobin tests, lipid profiles, and eye exams). Recall that this isn't a random sample, so inferences can't be drawn about the care received by the entire patient population. 36% of the patients in the sample had evidence of at least one eye exam. Since eye exams may be covered by a separate insurance plan, this number should be treated as a lower bound. 54% had evidence of a glycated hemoglobin test and 54% had evidence of a lipid profile. However, when laboratory work is done in a hospital the claims often don't describe the exact work performed. 59 patients in the data set had instances of such lab work. If we are generous and assume that when lab work was done it was the *correct* lab work (according to the guidelines), then the compliance would be 91% for hemoglobin tests and 92% for lipid profiles. Alternately, if we limit ourselves to the 36 patients all of whose lab work was precisely recorded, 75% of them had hemoglobin tests and 78% had lipid profiles. The correlation between the performance of hemoglobin tests and the performance of lipid profiles was 0.46

	Poor	Average	Good
Percent receiving eye exams	25	34	43
Percent receiving hemoglobin tests	40	53	62
Percent receiving lipid profiles	50	55	54

Table 2.2: Percent of patients receiving eye exams, glycated hemoglobin tests, and lipid profiles by quality rating. For example, of the patients whose care was rated poor, 26% of them received eye exams whereas 50% of the patients whose care was rated good received eye exams.

Table 2.2 shows the compliance with each measure for the poor, average, and good care groups. For eye exams and hemoglobin tests compliance tends to increase as the physician's rating of quality increases. The differences are not statistically significant

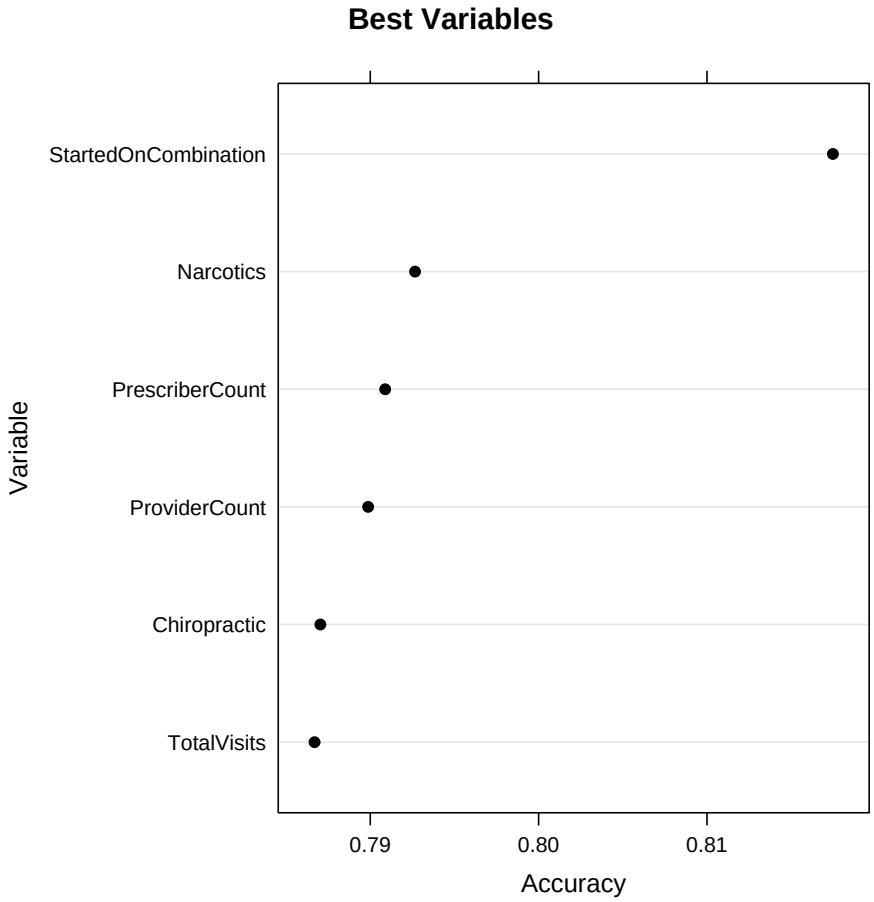


Figure 2-1: Estimated out-of-sample accuracy of classification models built on a single variable.

however (p= 0.95).

2.2.2 Single-Variable Models

The six variables shown in Figure 2-1 all classified the patients more accurately than the baseline. StartedOnCombination was the most accurate predictor, followed by Narcotics, PrescriberCount, ProviderCount, Chiropractic, and TotalVisits. StartedOnCombination is a 0/1 variable which indicates that the patients' drug therapy for diabetes started with a combination of drugs, rather than a single drug. The variable Narcotics measures the number of narcotics prescriptions that the patient filled over the two-year period. PrescriberCount is the number of different doctors who prescribed drugs for the patient during the study period, and ProviderCount is

the number of different providers that the patient saw. (Providers can include individual physicians, clinics, and hospitals.) The variable Chiropractic is the number of chiropractic visits that the patient had, and TotalVisits is the number of all visits, inpatient and outpatient, that the patient had.

Table 2.3 shows the optimal classification rules based on each of the variables. Note that in each case larger values of the variable result in a classification of poor care. For example, based on the Chiropractic variable a patient with more than 57 chiropractic visits over the two-year period is classified as receiving poor care. While it seems reasonable that the excessive use of narcotics is positively correlated with poor care, a positive correlation between PrescriberCount, ProviderCount, Chiropractic, and TotalVisits and poor quality is less intuitive. These four variables are all, in a sense, measures of the *quantity* of care. So, there seems to be an inverse relationship between the quantity of some aspects of care and quality.

Variable	Classify as Poor Quality when...
StartedOnCombination	True
Narcotics	> 22
PrescriberCount	> 19
ProviderCount	> 33
Chiropractic	> 57
TotalVisits	> 47

Table 2.3: Classification rules based on a single variable.

2.2.3 Three-Variable Models

Figure 2-2 shows the accuracy of the 10 best three-variable models. The best model has an accuracy of nearly 85% and as the figure shows there are a number of models with roughly comparable accuracy above 83%.

There are several variables that are included in more than one of these models. StartedOnCombination is a part of seven of the top 10 models. LongestOfficeGap is a part of four of the models and ChronicDrugsBeginning is a part of three. Mammogram and the related variable binary.mammogram are also a part of three of the models.

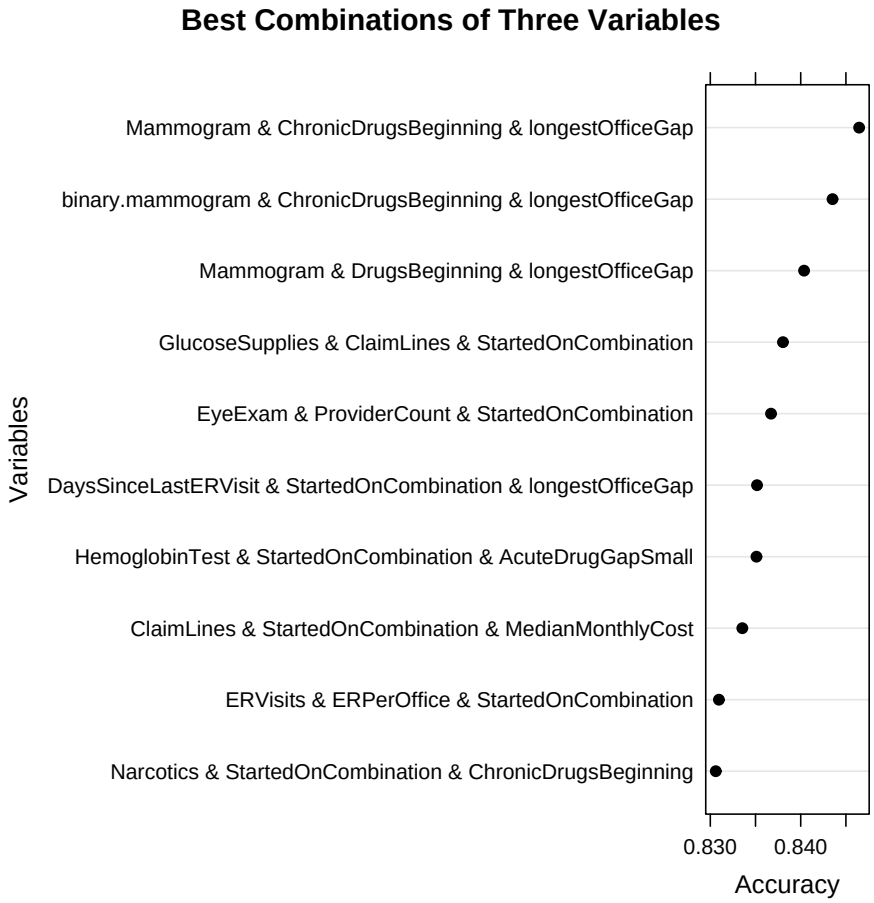


Figure 2-2: Estimated out-of-sample accuracy of classification models with three variables.

We also see variables directly related to diabetes care among the top models: GlucoseSupplies, EyeExam, and HemoglobinTest.

As an illustrative example, the form of the first model from Figure 2-2, involving the variables Mammogram, ChronicDrugsBeginning, and LongestOfficeGap is:

$$\begin{aligned}\text{logit}(\text{binary.quality}) = &-0.79 \\ &+ 0.57 \times \text{Mammogram} \\ &+ 0.2 \times \text{ChronicDrugsBeginning} \\ &+ 0.01 \times \text{longestOfficeGap}\end{aligned}$$

Table 2.4 shows the classification accuracy of this model. In-sample accuracy is 87.1% while the estimated out-of-sample accuracy is 84.6% as shown in Figure 2-2. 45% of the cases of poor care were identified correctly and only one patient who received good care was misclassified.

	Predicted	
Actual	Poor	Good
Poor	10	12
Good	1	78

Table 2.4: The actual classification of the patients' care based on the physician's review compared with the predicted classification of the model.

Table 2.5 shows the in-sample accuracy, sensitivity, and specificity for each of the models. We see the general pattern of very high specificity and lower sensitivity. Very few patients who received good care were misclassified by any of the models. However only from 36% to 55% of the patients who received poor care were correctly identified by the models. In practice, where there may be more cases of poor care then there are resources available to intervene, this level of detection may be quite sufficient.

Actually, the model allows us to trade off sensitivity and specificity. Table 2.5 were obtained by setting the cutoff value between poor care and good care to maximize overall accuracy. Figure 2-3 shows the range of sensitivity and specificity that can be

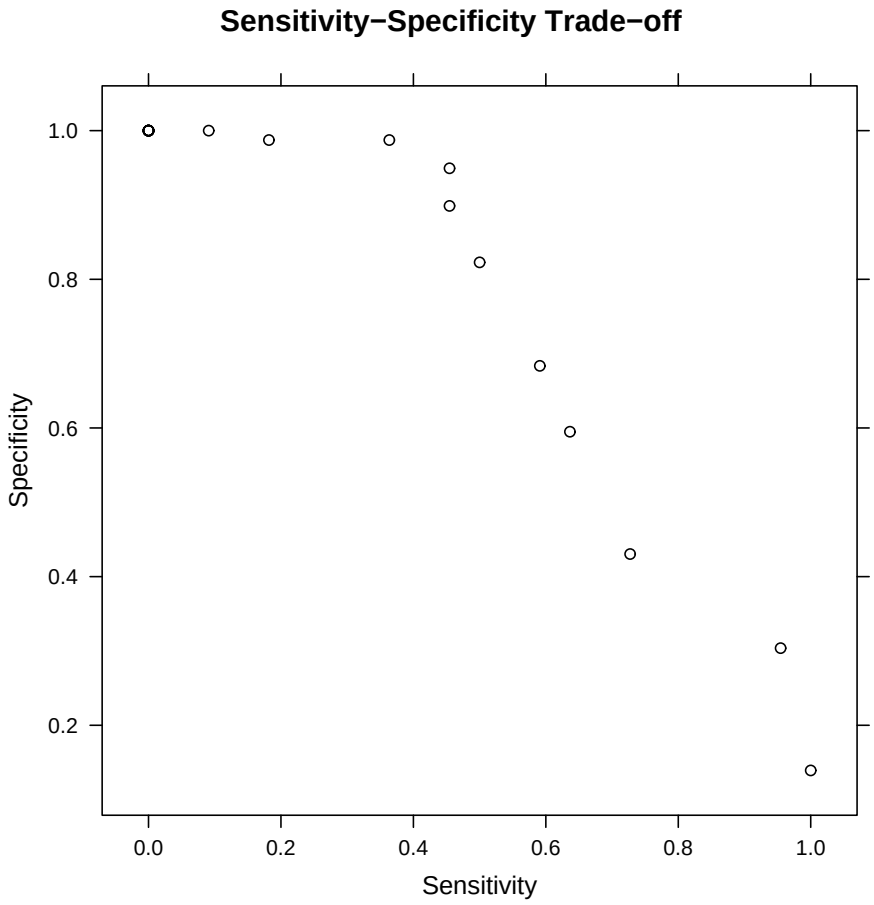


Figure 2-3: Trade-offs between sensitivity and specificity obtainable using the first logistic regression model.

obtained using the first logistic regression model in the table. A higher sensitivity will result in a higher percentage of poor care cases being correctly identified. In practice, this may be more important than the overall accuracy.

	Accuracy	Sensitivity	Specificity
Model 1	0.87	0.45	0.99
Model 2	0.87	0.50	0.97
Model 3	0.87	0.45	0.99
Model 4	0.86	0.45	0.97
Model 5	0.86	0.55	0.95
Model 6	0.86	0.41	0.99
Model 7	0.86	0.55	0.95
Model 8	0.86	0.41	0.99
Model 9	0.85	0.36	0.99
Model 10	0.85	0.41	0.97

Table 2.5: The accuracy, sensitivity, and specificity of the 10 best three-variable models.

Table 2.6 shows the coefficients of the top 10 models. For the purposes of this table the predictor variables were standardized by subtracting their mean and dividing by their standard deviation so that the coefficients could be compared to each other in a meaningful way. (Indicator variables, such as StartedOnCombination, were not standardized.) The table gives a sense of the direction and the strength of the relationship between each of the variables and quality of care. Variables with positive coefficients in the table are positively correlated with good care. Across the models, StartedOnCombination is the single largest determinant of quality of care. There is a second tier of variables consisting of ChronicDrugsBeginning, LongestOfficeGap, EyeExam, ProviderCount, HemoglobinTest, AcuteDrugGapSmall, ERVisits, and Narcotics. A third tier of variables has a smaller effect on the quality score: Mammogram, binary.mammogram, DrugsBeginning, GlucoseSupplies, ClaimLines, DaysSinceLastERVisit, MedianMonthlyCost, and ERPerOffice. Of the variables directly related to diabetes care, EyeExam and HemoglobinTest are in the second tier and GlucoseSupplies is in the third.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Mammogram	0.51		0.55							
ChronicDrugsBeginning	0.69	0.71								0.87
longestOfficeGap	0.79	0.75	0.67			0.42				
binary.mammogram		0.41								
DrugsBeginning			0.45							
GlucoseSupplies				0.48						
ClaimLines				−0.50				−0.79		
StartedOnCombination				−3.12	−2.99	−3.19	−4.23	−3.24	−2.95	−2.73
EyeExam					0.94					
ProviderCount					−0.74					
DaysSinceLastERVisit						0.59				
HemoglobinTest							0.82			
AcuteDrugGapSmall							−0.81			
MedianMonthlyCost								0.39		
ERVisits									−0.69	
ERPerOffice									0.46	
Narcotics										−0.86

Table 2.6: The coefficients of the 10 best three-variable logistic regression models. Variables have been normalized so that the coefficients can be compared directly.

2.2.4 Out Of Sample Validation

	Dr. Love		
Dr. Kane	Poor	Average	Good
Poor	4	4	3
Average	6	5	2
Good	1	2	3

Table 2.7: A comparison of the two physicians' ratings.

Table 2.7 shows the level of agreement between the two doctors' ratings on the 30 out-of-sample cases. The doctors were in complete agreement on 12 of the 30 patients. On an additional 14 their ratings differed by a single level (e.g. a case rated as good by one doctor was rated average by the other). In only four cases was there a complete divergence of ratings, with one doctor rating the care as good and the other doctor rating it as poor. In three of these four cases, Dr. Kane was the one who rated the care poor and in all three cases Dr. Kane's comments indicated that he felt the patient was on an inappropriate combination of drugs. Dr. Love did not make comments about drug combinations for any of the 30 patients.

Here are sample paragraphs written by the two physicians, chosen for one of the patients for which both doctors rated the care as good with high confidence. We begin with Dr. Love's assessment:

> This patient was closely monitored for blood glucose, had ophthalmology follow up, multiple urinalysis, on an ACE ARB. Treated with multiple oral agents. Seen for foot problems. Might have benefited from a platelet inhibitor, but otherwise high quality care.

And Dr. Kane's assessment:

> She is a Type 2 diabetic on several oral agents. She also had regular prescriptions for an ACE inhibitor and for diabetes testing supplies. Had regular prescriptions for nortriptyline (antidepressant) and lorazepam with no formal mental health care, but I saw no sign of excess care or other issues that would indicate active mental health problems. She had eye, gyn and podiatry care and a mammogram, and regular visits with her PCP. She had an ER visit for abdominal pain in July '04 with a prompt follow up visit afterward with her PCP. There were no hospitalizations. Care was orderly and looks to be good care with high confidence.

Neither doctor tended to be harsher in their overall ratings than the other. Dr. Kane rated 11 patients' care as poor, 13 as average, and six as good. Dr. Love rated 11 patients' care as poor, 11 as average, and eight as good.

The doctors also provided confidence scores for their ratings. Dr. Kane rated his confidence as high for 25 of the patients, Dr. Love for 21 of the patients. Since Dr. Kane has more experience reviewing claims data, this difference is not surprising. The doctors were jointly confident in their ratings of 18 of the patients. Their quality ratings were in complete agreement on nine of these 18 patients, a rate marginally better than for the whole set of 30 patients. They still reached opposite conclusions in three cases.

Of the top 10 three variable models identified in Figure 2-2, two of them performed very well out of sample when compared with Dr. Kane. The model based on the variables StartedOnCombination, HemoglobinTest, and AcuteDrugGapSmall had an accuracy of 80%. (Because the patient mix was different for the sample of 30, the relevant baseline statistic for comparison is not 78% but 63%. Direct comparison

of the accuracy rates to those in Table 2.5 aren't meaningful, though comparisons of sensitivity and specificity are.) The model has an out-of-sample sensitivity of 54% and specificity of 95% (See Table 2.8). The model based on StartedOnCombination, ChronicDrugsBeginning, and Narcotics has an accuracy of 83%, with a sensitivity of 54% and specificity of 100%. This latter model is notable for two reasons. First, it is not based on variables specifically related to diabetes so it has the potential to be applied more generally. Second, it was also one of two models that did relatively well when compared with Dr. Love's ratings. It had an accuracy of 67% in matching Dr. Love's ratings with a sensitivity of 36% and specificity of 84%. None of the other models did better in matching his ratings, though the model involving StartedOnCombination, EyeExam, and ProviderCount did equally well. Overall, the models matched Dr. Kane's ratings more closely.

Because these two models performed well both in-sample and out-of-sample, we consider them the two best candidates. The use of these models also seems justified by the fact that their coefficients and variables have reasonable medical interpretations. In the first model, we have already discussed the drawback to starting treatment on a combination of drugs (StartedOnCombination). Glycated hemoglobin tests (HemoglobinTest) are recommended by the diabetes treatment guidelines. The variable AcuteDrugGapSmall, which is negatively correlated with quality of care, indicates the repeated use of an acute drugs. Such repeated use may indicate that the diagnosis or the choice of drug is incorrect. In the second model, the chronic use of narcotics may indicate that the doctor is unable to determine the underlying cause of the pain and there is also the serious hazard that the patient may become addicted to the narcotics. ChronicDrugsBeginning is positively correlated with good care. This variable measures the number of chronic drugs that the patient was on at the beginning of the study period. A larger number of drugs may indicate that the physician has recognized the patient's comorbidities and is taking measures to address them.

	Model	
Dr. Kane	Poor	Good
Poor	6	5
Good	1	18

Table 2.8: Dr. Kane's classification of the 30 out-of-sample cases compared with the classification of the model based on the variables StartedOnCombination, HemoglobinTest, and AcuteDrugGapSmall.

2.3 Conclusions

We have demonstrated that an expert-trained statistical model using insurance claims data can accurately identify patients who are receiving poor care. Furthermore, only a simple model is required to capture a majority of the cases of poor care while maintaining a very low false positive rate. We developed several competitive models and validated them out of sample.

Though we have focused on the use of logistic regression to *classify* patients, in practice the fitted probabilities could be used directly as quality scores. Rather than treating patients in the poor care group as homogeneous, it would make sense for the reviewers or case managers to begin with the patient whose quality score was the lowest. Next, the patient with the second lowest quality score, and so on. In this way, resources are focused first on the patients who may be the likeliest to be receiving poor quality care.

In practice, the model can be improved over time. If a reviewer disagrees with the model's rating for a particular patient, the reviewer can record their own rating. As these ratings accumulate in the database, the model can then be re-fit. This approach could be a cost-effective way to create much larger training data sets with minimal additional overhead.

In several of the models a pattern emerged of an inverse relationship between the quantity of care and quality. There could be many reasons for this. For example, the more interactions a patient has with the healthcare system the more opportunities there are for a mistake or other error to occur. An alternative explanation could be that some of these patients require so much care because the care they are receiving

is poor: that is, the care is not making them better and so they continue to seek more care.

There are several characteristics of the patients in our study that may limit the generality of our model. The patients are all insured, and this is necessary because our method cannot be used without the electronic insurance claims data. The patients are between the ages of 35 and 55. The general approach would apply to patients outside this range, though the specific regression model may change. The patients have relatively high health care costs. It may be difficult to apply this method to patients below a cost threshold, since the density of the claims data may not provide enough information on which to base a judgment of quality. Although high-cost patients are of the most interest to insurers implementing case management, a specific weakness of this methodology is its insensitivity to evaluating the quality of care received by low-cost patients.

Another limitation of the study is that we have used a relatively small sample size. More data will lead to more accurate and more generalizable models. Furthermore, we've only consulted two physicians. Assessments about quality will differ from physician to physician. Ideally, we would include the opinions of several experienced physicians when building the model. Finally, a question of immediate interest is how well the models identified in this study would perform on a general population not limited to diabetics.

Chapter 3

An Optimization Approach to Large Scale Drug Surveillance

3.1 Introduction

At the point when a drug is released on the US market, only a relatively small number of people have been exposed to it during clinical trials. When a larger number of people begin using the drug, unexpected side effects (positive or negative) may be discovered and in this chapter we present an approach to detect such side effects in near "real-time." would need for such surveillance was highlighted by the drug Vioxx which was withdrawn from the market in 2004 after being linked to increased rates of heart attacks and strokes. From 1999 to 2004 it is estimated that Vioxx may have been responsible for tens of thousands of fatal heart attacks [40]. Improved surveillance therefore has the potential to save many lives.

Claims data is ideally suited for systematic drug surveillance because it is available electronically for a large number of people and is updated frequently. The use of claims data for drug surveillance was presented in [40] and [14]. However, they consider the case when the side effect of interest is known *a priori*. A key challenge that we consider is conducting surveillance across all possible side effects while avoiding the loss in statistical power associated with testing a large number of null hypotheses.

To detect previously unknown side effects to a drug we compare the insurance

claims data for patients taking the drug (the treatment group) with the claims data for a suitable control group (i.e., an appropriately chosen set of patients not on the drug, possibly on a comparable drug). From the claims data, we determine the rates at which different diagnoses occur in the two groups and identify discrepancies in the rates that might suggest side effects.

Each possible diagnosis is identified by a three digit code, the ICD-9-CM code (International Classification of Diseases, Clinical Modification, 9th Revision). There are approximately 900 such codes. Some diagnoses occur quite frequently, such as 462 Acute pharyngitis (i.e., a sore throat), while others occur very rarely, such as 032 Diphtheria. (Actually, there is a further level of specificity available by adding a fourth and fifth digit to the code for some diagnoses. However, these "modifier" codes aren't consistently recorded in the claims database and so we restrict ourselves to the three digit codes.)

An interesting aspect of the codes is that they fit in a hierarchical structure from general categories to specific diagnoses. For example, all of the diagnoses related to the circulatory system form a category with codes between 390 and 459. The diagnoses of the circulatory system can be broken down into 9 more specific categories. The codes 390-392 correspond to Acute Rheumatic Fever, for example, while the codes 430-438 correspond to Cerebrovascular Disease. Each of these nine categories can be broken out once more into individual diagnoses, each with its own ICD-9 code. For example, within Acute Rheumatic Fever there exists Rheumatic fever without mention of heart involvement (390), Rheumatic fever with heart involvement (391), and Rheumatic chorea (392).

Thus, the ICD-9 codes form a tree with a root node and three levels beneath it. There are 17 large categories of diagnoses below the root, 110 smaller categories beneath them, and 913 individual diagnoses at the leaves of the tree. We will let $\mathcal{N}$ represent the set of nodes of the tree. $\mathcal{N}$ can be partitioned into subsets $\mathcal{D}$ and $\mathcal{G}$, where $\mathcal{D}$ is the set of nodes corresponding to individual diagnoses (i.e., the leaves of the tree) and $\mathcal{G}$ is the set of nodes corresponding to categories of diagnoses. For the tree of ICD-9 codes, $|\mathcal{N}| = 1041$, $|\mathcal{D}| = 913$, and $|\mathcal{G}| = 128$.

For each diagnosis, we would like to test the null hypothesis that the drug in question has no effect on the rate at which the diagnosis occurs. We thus have a set of approximately 1000 hypothesis tests that we would like to perform using the claims data. Because of the large number of tests, there are challenges involved with trading off the risk of false detections with the need for statistical power (or equivalently, the risk of a missed detection).

Note that the problem of selecting an appropriate control group is not an easy one, but for the purpose of this analysis we assume that one exists. Such a group should satisfy the condition that, in the absence of a treatment effect, each diagnosis would occur in the treatment group at the same rate that it occurs in the control group.

We propose several new approaches to this problem and conducted several simulation studies to compare the performance of these approaches to each other and to the traditional approach to the problem. We begin by considering a single-period setting in Section 3.2 to set the groundwork and introduce several key concepts. We introduce several approaches for this setting and in Section 3.3 briefly compare their performance using simulation. In Section 3.4, we extend one of these approaches to create a dynamic algorithm for the multi-period setting and report on simulation results in Section 3.5. We conclude with a discussion of our results in Section 3.6.

3.2 The Single-Period Setting

We begin with a single-period setting in order to set the foundation and build intuition before addressing the more realistic multi-period setting. In the single-period setting, all of the data become available at the same time and there is a single round of hypothesis testing. The data consist of the number of times each diagnosis and category of diagnosis occurred in the treatment group and the control group. Let $x_i^T, i \in \mathcal{N}$ denote the number of occurrences of i in the treatment group and let $x_i^C, i \in \mathcal{N}$ denote the number of occurrences of i in the control group. Here i can represent either a single diagnosis or a category of diagnosis. From the observed data,

we can construct the 2×2 table:

	Treatment	Control
People having i	x_i^T	x_i^C
People not having i	$N^T - x_i^T$	$N^C - x_i^C$

Based on the number of occurrences of i and the total number of people in the treatment and control groups, N^T and N^C, respectively.

Because the rate of occurrence of some i may be small (for example, where i corresponds to a rare disease), we test each null hypothesis using the Fisher exact test [79] on the 2×2 table rather than the χ^2 test. More formally, for each i we perform a hypothesis test of

$$\begin{aligned} H_0 : \mu_i^T &= \mu_i^C \\ H_0 : \mu_i^T &\neq \mu_i^C \end{aligned}$$

where μ_i^T and μ_i^C are the true rate of occurrence of i in the entire population for people taking the drug and not taking the drug, respectively.

Before proceeding, we review a few relevant statistical concepts. The *p-value* of a statistical test is the probability that the observed data (or more extreme data) would have occurred under the null hypothesis. Accordingly, small p-values work against the null hypothesis. The *significance level* of a statistical test, usually denoted by α, is a value that we set. We will reject the null hypothesis if the p-value is less than α. The significance level can be interpreted as the probability of rejecting the null hypothesis when it is true, known as a Type I error. In the context of drug surveillance, we will refer to a Type I error as a "false detection." The *power* of a statistical test is the probability of rejecting the null hypothesis when it is false. One minus the power is the probability of accepting the null hypothesis when it is false, known as a Type II error. In the context of drug surveillance, we will refer to a Type II error as a "missed detection."

The power of a statistical test depends on several factors. It depends on the sample size – the larger the sample, the more powerful the test. It depends on the true size of the effect, for example whether the drug makes a particular diagnosis twice as likely or 100 times as likely. The larger the effect, the easier it will be to detect, hence the more powerful the test will be. Note that the sample size is constrained by the size of our claims database and the true effect of the drug is naturally beyond our control. Therefore, most importantly for our work, the power of a test depends on the significance level – the larger α is, the more powerful the test will be. We will also make use of the fact that, as a function of α, the power of a test is an increasing concave function. The power of a test is 0 when $\alpha = 0$ and 1 when $\alpha = 1$.

Note that in practice the power of a test is generally not known because the true effect size is not known (if it were, there would be no need for the statistical test). In the single-period setting, then, our use of the power of a test will only be notional. We will make the naive assumption that the drug has the same effect on every i. In the multi-period setting we will make successively more accurate estimates of the effect size as time goes on and use these estimates to approximate the curve that relates the power to α.

For a set of multiple hypothesis tests, the *family-wise* significance level, also referred to as α, controls the probability of at least one of the null hypotheses in the set being falsely rejected. Note that performing each individual test at an α significance level will not lead to a family-wise significance level of α. For example, suppose 100 hypothesis tests are carried out, each at a 5% significance level. Then, the probability that at least one null hypothesis is falsely rejected (assuming all of them are true) is $1 - .95^{100} = .994$, much higher than 5%. Therefore, in order for a family-wise significance level of α to be maintained, each individual test in the set must be carried out at a smaller significance level. A review of approaches to multiple hypothesis testing is provided in [90].

3.2.1 Approach 1: Bonferroni

The classical approach to multiple hypothesis testing is the Bonferroni approach [79]. If k null hypotheses are to be tested at a family-wise significance level of α, then each hypothesis can be tested individually at the significance level α/k. Applying this approach to our diagnosis data, we would conduct 1041 tests and test each one at a significance level of .05/1041 so that the family-wise error is controlled at 5%. (Of course, levels of α other than 5% could be used. Since we are concerned with the comparative performance of different approaches, the selection of α is not crucial since changing it would affect all of the approaches.)

We include the Bonferroni approach in the study as a baseline because it is a standard approach which we hope to improve upon. It will likely suffer from a lack of power: because the significance level for each individual test is so low it is unlikely that any particular null hypothesis will be rejected.

3.2.2 Combining tests

Because combining diagnoses (i.e., testing nodes further up on the tree rather than at the leaves) will be a central feature of several of the approaches, we take a moment to consider it more fully.

Whether combining two (or more – here, we consider only two for simplicity) diagnoses into a single group results in a better or worse test partly depends on the nature of the diagnoses and the effect of the treatment upon them. In general, combining two diagnoses leads to a larger number of events and a more powerful test. For example, if among the patients on a drug there were 4 heart attacks and 6 strokes, we could compare the combined 10 strokes and heart attacks to the combined number of heart attacks and strokes in the control group. However, suppose the treatment only has a true effect on one of the diagnoses. Adding in the number of times that the other diagnosis occurred only adds noise and will not increase the power of the test. If the treatment has a true effect on both of the diagnoses, and the "direction" of the effect is the same (e.g., it increases the frequency of both) then combining

the tests will increase the power. If the treatment has opposite effects on the two diagnoses, making one more frequent and the other less frequent, then combining the tests can lead to neither being detected even though on their own each might have been detected. These are the trade-offs involved in combining diagnoses.

3.2.3 Determining whether a test has power

Another key concept for several of our approaches is whether or not a test has power. As it turns out, some diagnoses are very rare and, given the amount of data on hand, effects of the treatment upon the frequency of the diagnosis are unable to be detected. Whether they can be detected or not can be determined *in advance* of performing any hypothesis tests. Those that cannot be detected should not be tested, and in this way we can reduce the denominator k in the Bonferroni approach.

To determine whether a test of i has power, we count the number of times that i occurred in the entire sample of patients (the treatment and control groups combined), blinding ourselves to whether or not each occurrence came in the treatment group or the control group. For example, if there are 1000 patients in the treatment group and 1000 patients in the control group we would allow ourselves to be privy to the fact that there were (say) eight occurrences of i among all 2000 patients, but we would not "peek" to see how many of the eight occurrences were in the treatment group and how many were in the control group.

Suppose there were x_i occurrences of a particular diagnosis all together and suppose we are conducting the test at an α significance level. The most unlikely situation under the null hypothesis (and hence the situation leading to the smallest p-value) would be if all x_i occurred in a single group, say the treatment group. The corresponding 2×2 table would be:

	Treatment	Control
People having i	x_i	0
People not having i	$N^T - x_i$	N^C

(If the treatment and control groups are the same size, then whether all occurred in the control group or the treatment group would be irrelevant. If one group is larger

than the other group, then the smallest p-value will occur when all of the diagnoses occur in the smaller group.) We assume that this is the case and perform a Fisher exact test on this *hypothetical* situation and obtain a p-value. This p-value represents the *smallest obtainable p-value* under all possible allocations of the x_i occurrences between the two groups. If the p-value is larger than α then we can conclude that the test has no power. That is, given the data at hand we would be unable to reject the null hypothesis of no difference in the rate of occurrence of i between the treatment and control groups. On the other hand, if we obtain a p-value less than α, then the test does have power. Albeit, the power could be small or large and we make no assessment of the magnitude of the power. We only create the dichotomy: tests with power and tests without power.

We can generalize this situation to the case when we are not testing only a single null hypothesis, but k null hypothesese simultaneously. In this case, note that each individual test will be performed at an α/k significance level and so whether the test of a particular i has power depends on how many other tests are being performed. The test of a given i will decrease in power as the number of simultaneous tests increases until at some point it may no longer have power. For example, a given test may have power when only 10 tests are being performed (and a significance level of $\alpha/10$ is used) but not when 20 tests are performed (and a significance level of $\alpha/20$ is used).

For each $i \in \mathcal{N}$ we define s^i to be the maximum number of tests that can be performed simultaneously (inclusive of the test of i) without causing the test of i to have no power. For example, suppose we have computed that the smallest obtainable p-value for a test of a particular i is 0.009. At the 5% significance level, if i were the only node being tested its test would have power since $0.009 < 0.05$. If five nodes were being tested, each test would be performed at a $0.05/5 = 0.01$ significance level and so the test of i would still have power since $.009 < 0.01$. However, if six nodes were being tested, each test would be performed at a $0.05/6 = 0.008$ significance level and so the test of i would not have power, since $0.009 \not< 0.008$. In this example, then, $s^i = 5$. For computational purposes, we limit s^i to be at most $|\mathcal{N}|$ (no more than $|\mathcal{N}|$

j	$s^{i_{(j)}}$
1	20
2	17
3	9
4	9
5	8
6	7
7	6
8	5
9	4
10	4
11	2
12	1
13	1
14	1
15	1
16	0
17	0
18	0
19	0
20	0

Table 3.1: Example of possible values of s^i, in descending order.

can be performed anyways).

3.2.4 Approach 2: A simple algorithm

We use this idea about power to reduce the number of tests performed. The tests that remain will have increased power. Note that this approach will increase the probability of a detection and, even though not all tests are performed, it will not increase the probability of a missed detection.

Using a straightforward algorithm, we find the largest subset of tests to perform such that all of the tests being performed have power. To find such a subset, we compute s^i for each test and order the tests in decreasing order of s^i. We label the ordered tests $i_{(1)}, i_{(2)}, \ldots, i_{(|\mathcal{N}|)}$ and perform all tests such that $j < s^{i_{(j)}}$.

For example, if we have the values shown in Table 3.1 we would perform tests 1

through 6. We would not perform test 7 since it only has power when 6 or fewer tests are performed. Note that this algorithm not only finds the largest subset of tests to perform such that all of the tests being performed have power, but in some sense it also finds the most "powerful" such subset because it chooses the tests with the largest s^i, a quantity which increases with the power of the test.

3.2.5 Approach 3: A mixed integer optimization approach

The previous algorithm discards tests that are not powerful enough. Alternatively, rather than discarding tests we can combine them with their neighbors on the tree, in effect "rolling them up" to their parent node and performing a test on the parent. As with Approach 2, this approach will identify a subset of tests S such that $|S| < s^i \quad \forall i \in S$.

It also seems desirable to enforce the constraint that if a test is performed at a given node, then no tests will be performed at its children, and vice versa. This will reduce redundancy among the tests performed.

Of course, it is possible to find many different subsets of tests that satisfy these requirements. What criteria should be used to determine the optimal subset is not obvious. It seems that a reasonable objective would be to choose the largest subset. That is, to maximize the number of tests performed. This objective isn't without its drawbacks, because the more tests performed the less powerful each test will be. We address this drawback later.

The optimization problem can be formulated as follows. For a given node i, the decision variable p_i will be 1 if the test of node i is to be performed and 0 otherwise.

$$\begin{aligned}
&\text{maximize} && \sum_{i \in \mathcal{N}} p_i && \\
&\text{subject to} && p_i \leq \frac{1 + |\mathcal{N}| - \sum_{i \in \mathcal{N}} p_i}{1 + |\mathcal{N}| - s^i} && \forall i \in \mathcal{N} \\
& && p_k + p_l \leq 1 && \forall k, l \text{ where } k \text{ is a descendant of } l \\
& && p_i \in \{0, 1\} && \forall i \in \mathcal{N}.
\end{aligned}$$

The objective function maximizes the number of tests performed. The first constraint enforces the condition that we may only perform a test if it has power. For a particular i, the value on the right hand side of this constraint will be at least 1 when the number of tests performed is less than or equal to s^i, allowing the test of i to be performed. When the number of tests being performed is greater than s^i, the value on the right hand side will be strictly between zero and one, forcing $p_i = 0$. The expression on the right hand side was carefully constructed to prevent it from ever being negative, which would render the constraint infeasible. The second constraint ensures that we only perform a test at a node if we don't perform a test at any descendant nodes.

The objective function formulated above ignores the fact that tests performed at the leaves of the tree are more specific than tests performed higher up in the tree. That is, they have the potential to identify a particular diagnosis as a side effect rather than a more general category of diagnosis. All else begin equal, given a choice between performing two tests, one of which is further down on the tree, we will generally prefer to perform the one that is further down and hence more specific.

To enforce this preference, we weight the nodes lower down on the tree more heavily in the objective function. We can do this in the following way. If there are k leaves, we can weight the test at each leaf by $1 + \frac{1}{k+1}$. The tests at the next level up are assigned weights $1 + \frac{1}{k+2}$, and so on. These weights are chosen small enough so that a subset of $n + 1$ tests is always preferable to a subset of n tests, regardless of where on the tree the tests are.

Our objective function becomes:

$$\text{maximize} \sum_{i \in \mathcal{N}} w_i p_i$$

where the w_i's are the weights.

Note that this formulation also allows us the flexibility to assign more weight to diagnoses we are particularly interested in. For example, we may want to weight more serious diagnoses more heavily or weight diagnoses by their average annual treatment cost. Based on the clinical trials for a drug or biological aspects of the drug, certain

side effects might be of particular concern and medical experts can set the weights accordingly.

3.2.6 Approach 4: Unequal significance levels

For Approaches 2 and 3, we have used a simple dichotomy: a test either has power or it doesn't. However, even if two tests have power they may differ substantially in their power and it may improve our testing algorithm if we take this into account. To illustrate this idea, we begin with a simple example.

Consider a case in which there are only two diagnoses of concern, but one is rarer than the other. Suppose we have a treatment group of 10,000 patients and a control group of 10,000 patients. In the control group the "common" diagnosis occurs at a rate of 4 in 1000 and the "rare" diagnosis occurs at a rate of 1 in 1000. Finally, suppose that the treatment doubles the rate of occurrence of both diagnoses. That is, in the treatment group the common diagnosis occurs at a rate of 8 in 1000 and the rare diagnosis occurs at a rate of 2 in 1000.

Not knowing that this is the true effect of the treatment, we would like to test the two null hypotheses that the treatment does not affect the rate of the common diagnosis nor the rare diagnosis. We would like to use a family-wise significance level of 5%. This 5% can be "distributed" between the two tests in a variety of ways and how it is distributed affects the power of our two tests. For example, we could test for an effect of the treatment on the common diagnosis at the 5% level and not test for an effect on the rare diagnosis at all. In this case, the test of the common diagnosis would have power of over 90%. The rare diagnosis test would have no power, since it is not conducted. On the other hand, we could test only for an effect on the rare diagnosis and not test for an effect on the common diagnosis. In this case, the rare diagnosis test would have power of approximately 40% (note the power is lower because the diagnosis is less common). The common diagnosis test would have no power, since it is not conducted.

These are the two extreme cases. In general, to maintain the 5% family-wise Type I error rate, we can conduct the two tests at any combination of levels α_c and α_r ("c"

for common, "r" for rare) such that $\alpha_c + \alpha_r = 0.05$. Conducting the tests at the levels $\alpha_c = \alpha_r = 0.025$ would be equivalent to the Bonferroni approach. However, it may be advantageous to deviate from the Bonferroni levels to increase the combined power of the two tests.

There are different ways one could measure, and hence optimize, the overall power of the two tests. One reasonable objective is to minimize the expected number of missed detections. Other objectives are possible as well, such as minimizing the probability of no detection or maximizing the probability of detecting both effects. The latter seems overly aggressive for drug surveillance, a situation in which we would be happy to detect even one of a number of side effects.

The expected number of missed detections is given by $(1 - p_c) + (1 - p_r) = 2 - (p_c + p_r)$ where p_c and p_r are the power of the tests of the common diagnosis and the rare diagnosis, respectively. Figure 3-1 shows how the expected number of missed detections varies as we vary p_c, observing the constraint that $p_c + p_r = 0.05$.

Note that the function is minimized when a slightly higher significance level is used to test the rare diagnosis than is used for the common diagnosis.

This approach can easily be generalized to an arbitrary number of diagnoses (and categories of diagnosis) of varying levels of rarity. Although the power functions are nonlinear, they can readily be approximated by piecewise linear functions. The problem of determining the optimal significance level for each test can be formulated as a mixed integer linear optimization problem. As we noted before, since the true effect of the treatment on the rate of each diagnosis is not known, the power functions themselves are not known. (In the multi-period setting, this isn't a grave problem because we can build successively more accurate estimates of the power functions as we move forward through time.) The approach we take in the single-period setting is to assume that the treatment has the same effect on all diagnoses, say doubling their rate of occurrence. For a very common diagnosis, this assumption may be quite unrealistic. For example, it would in fact be impossible to double the rate of occurrence of a diagnosis that occurs at a rate of 700 in 1000. Therefore, it might make sense to assume a smaller effect size for more common diagnoses. Indeed, if there were

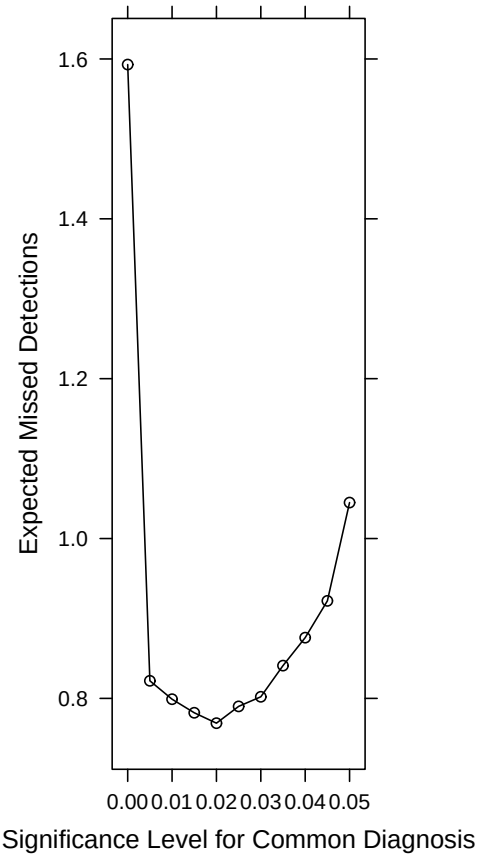


Figure 3-1: Varying the significance level of the two tests yields (maintaining a combined 5% significance level) affects the expected number of missed detections

a large effect on a common diagnosis it would most likely be detected in clinical trials or through the FDA's Adverse Event Reporting System and the techniques discussed here would be unnecessary.

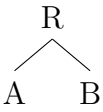
Approach 4 can be formulated as follows. For each node i we approximate the power curve of its test with n linear pieces with slopes m_j^i and intercepts n_j^i for $j = 1, ..., n$. As with Approach 3, we assign a weight w_i to each test based on its height in the tree. The decision variables α_i are the significance levels at which to perform each test. Because our objective is piecewise linear in each α_i we introduce a set of variables r_i, representing the power of each test, in the objective and handle the piecewise linear aspect in the constraints. We use a binary decision variable p_i to indicate whether each test is performed or not.

$$
\begin{aligned}
\text{maximize} \quad & \sum_{i \in \mathcal{N}} w_i r_i & & \\
\text{subject to} \quad & \sum_{i \in \mathcal{N}} \alpha_i = 0.05 & & \\
& \alpha_i \leq p_i & & \forall i \in \mathcal{N} \\
& r_i \leq \alpha_i m_j^i + b_j^i & & \forall i \in \mathcal{N},\ j = 1, \ldots, n \\
& p_k + p_l \leq 1 & & \forall k, l \text{ where } k \text{ is a descendant of } l \\
& p_i \in \{0, 1\}, \alpha_i \geq 0 & & \forall i \in \mathcal{N}.
\end{aligned}
$$

The objective function maximizes the sum of the power of the tests, which is equivalent to minimizing the expected number of missed detections. The first constraint ensures that the significance levels add to 0.05. The second constraint relates the binary variables to the continuous ones so that if a test is not performed its significance level must be zero. The third constraint corresponds to the piecewise linear functions, which give the relationship between the significance level and the power.

3.3 Simulation of the Single-Period Setting

We have performed simulations comparing the performance of these four approaches. Rather than running the simulations on the whole ICD-9 tree we used a very simple tree structure with one root node and two leaves. There are two diagnoses, diagnosis A and diagnosis and B. We also have a category that includes A and B, so the tree has three nodes: the root, R, and two leaves, A and B, as shown below.



There are 1000 people in the treatment group and 1000 people in the control group. In each trial of the simulation, for each person each individual diagnosis either occurs or does not occur according to the appropriate probability (specified below). We constructed 21 different scenarios in which we varied the size and direction of the true effect at each leaf and performed 1000 trials for each scenario. We used each of the four approaches to test for side effects and counted the fraction of times that the null hypothesis at each node was rejected, accepted, or not tested by each approach. Two scenarios are presented in detail below to demonstrate the simulation method and highlight some of the findings, followed by a summary of the whole set. The details of the scenarios can be found in Table B.1.

The algorithms were implemented in R [78], as were the simulations. The optimization models were expressed in ZIMPL [59] and solved using the freely available SCIP solver [1], using SoPlex [112] as the LP solver.

In the first scenario each diagnosis occurred at a rate of 2% in the control group, and the treatment had the effect of doubling the rate of occurrence of each diagnosis. Note that in this scenario the effect size is the same for both diagnoses and the direction is the same.

Since in this case there is a true effect at each node in the tree, we will compare the approaches based on how often they reject the null hypothesis at each node. Table 3.2 shows the results.

	Approach			
Node	1	2	3	4
R	5.9%	9.0%	1.3%	14.5%
A	1.5%	1.8%	2.4%	0%
B	2.0%	2.1%	4.5%	0%

Table 3.2: Detection rates of the four approaches on the three nodes, scenario 1.

At the root node, Approach 4 outperformed the others by a large amount. It was followed by Approach 2 and Approach 1. Approach 3 was a distant fourth.

The leaves were rejected less often than the root. This is because the number of occurrences is smaller at the leaves and so the statistical tests have less power than the combined test at the root. On leaf A, Approach 3 had the highest rejection rate. Approach 2 was second and Approach 1 was close behind. The standings are similar for leaf B, with Approach 3 performing best, and Approaches 1 and 2 having approximately half the rejection rate of Approach 3.

Across all the nodes, Approach 2 strictly dominates Approach 1. Approach 3 dominates all other approaches on the leaves, but does not perform as well on the root. Recall that in Approach 3 the leaves are weighted more heavily, so it often chooses to test one or both leaves rather than testing the root. These results demonstrate the trade-off between performing a test at a parent or a child. Interestingly, Approach 4 almost exclusively tests the root and ignores the leaves.

In the second scenario, diagnosis A occurred at a rate of 4% in the control group and 2% in the treatment group. Diagnosis B occurred at a rate of 2% in the control group and 4% in the treatment group. Note that the treatment has opposite effects on the two diagnoses in this case: it increases the frequency of one and decreases the frequency of the other. This will cause trouble for approaches that favor testing at the root, because the effects will cancel out to some extent. The results are shown in Table 3.3

In this case, Approach 4 was again the best at detecting an effect at the root node. But the 1.5% of the time that it detected an effect at the root is less than the

	Approach			
Node	1	2	3	4
R	0.6%	1.4%	0.3%	1.5%
A	1.2%	1.4%	1.5%	0.2%
B	1.6%	1.7%	3.3%	0.1

Table 3.3: Detection rates of the four approaches on the three nodes, scenario 2.

	Approach			
Rank	1	2	3	4
1	3	3	18	0
2	1	3	3	0
3	7	15	0	0
4	10	0	0	21

Table 3.4: Ranking of approaches, leaf A

percentage of time that some of the other approaches detected effects at the leaves. Approach 3, for example, detected effects at each of the leaves at least 1.5% of the time. As before, Approach 3 dominates the other approaches on the leaves.

Across all 21 scenarios that we simulated, the following tables show the number of times each approach placed first, second, third, or fourth. Table 3.4 shows the results for leaf A, where Approach 3 performed best. Table 3.5 shows the results for leaf B, where again Approach 3 did the best overall. Table 3.6 shows the results for the root. This is where Approach 4 was more advantageous. Approach 3 fared poorly.

In conclusion, in the single-period setting all three of our approaches demonstrated the ability to outperform Approach 1, the Bonferroni approach. The two integer optimization-based approaches, Approaches 3 and 4, showed the strongest performance. Approach 4 is the most nuanced, allowing the significance levels to be adjusted over a range, rather than the more crude all-or-nothing characteristic of Approaches 2 and 3. Approach 4 also has the most appealing objective function because it explicitly minimizes a value of concern, missed detections. Therefore, we now turn our attention to adapting Approach 4 to the multi-period setting.

	Approach			
Rank	1	2	3	4
1	0	0	12	0
2	0	1	0	0
3	2	11	0	0
4	10	0	0	12

Table 3.5: Ranking of approaches, leaf B

	Approach			
Rank	1	2	3	4
1	0	1	0	20
2	3	20	0	1
3	18	0	0	0
4	0	0	21	0

Table 3.6: Ranking of approaches, root node

3.4 The Multi-Period Setting

We now consider the more realistic setting in which surveillance is conducted over time. Typically, new claims data would become available monthly. In this setting, it is possible to test a different set of null hypotheses each month. We Approach 4 from the single-period setting to dynamically update the set of tests performed each month. As in [14], we specify in advance the number of months over which we will monitor the drug.

3.4.1 Determination of p-values

When there was only one period, we specified the family-wise significance level α. With multiple periods, we again specify α, but in this case the Type I error rate must be controlled not only across the multiple tests performed in a single period but across all tests performed in all periods. Therefore, in each period a significance level less than α must be used. Let α_t be the significance level used in period t and let T denote the total number of periods. (Note that α_t will be the family-wise error rate

for the set of tests performed in period t.) The α_t's must be set in such a way as to control the overall error rate at α. We can determine the α_t's using simulation.

Furthermore, it is not necessary that $\alpha_1 = \alpha_2 = \cdots = \alpha_T$. How the individual α_t's are set can affect the probability of detection and the time until detection. There are trade-offs involved in having larger α_t's at the beginning or end of the time horizon. For example, suppose we set $\alpha_1 = .05$, and $\alpha_2 = \alpha_3 = \cdots = \alpha_T = 0$. Then, all the power to reject the null hypothesis is allocated to the first period and we might hope to reject the null hypothesis very quickly. The drawback is that in the first period we have the least data available to us, which mitigates the benefit of the larger α_t. Setting $\alpha_1 = \alpha_2 = \cdots = \alpha_{T-1} = 0$ and $\alpha_T = .05$ would allow us to use the most power when we have the full set of gathered data but would preclude the possibility of rejecting the null hypothesis at any point prior to the last period. Between these two extremes there lies a set of α_t's that balances this trade-off.

We conducted simulations to compare six strategies: constant α_t's (i.e., $\alpha_1 = \alpha_2 = \cdots = \alpha_T = c$), linearly increasing α_t's and linearly decreasing α_t's. For each of these three variations, observations were either accumulated (that is, in period t all observations gathered up to and including period t were used in the test) or treated one period at a time (that is, in period t only observations gathered in period t were used in the test).

The α_t's were determined via simulation such that the overall significance level was controlled at $\alpha = 0.05$ for all six strategies. These strategies, and the corresponding α_t's are illustrated in Figure 3-2.

To compare the power of the strategies, we simulated a situation in which the null hypothesis was false and compared the ability of the six strategies to detect this.

Table 3.7 shows the percent of time that the null hypothesis was rejected using each strategy and the number of periods it took to reject it (when it was rejected).

The three strategies that used accumulation (strategies 4-6) had detection rates more than double the strategies that didn't (strategies 1-3). A constant α_t led to the highest detection rate by a 10% margin when data wasn't accumulated. When data was accumulated, constant α_t's were tied for the highest detection rate. Whether

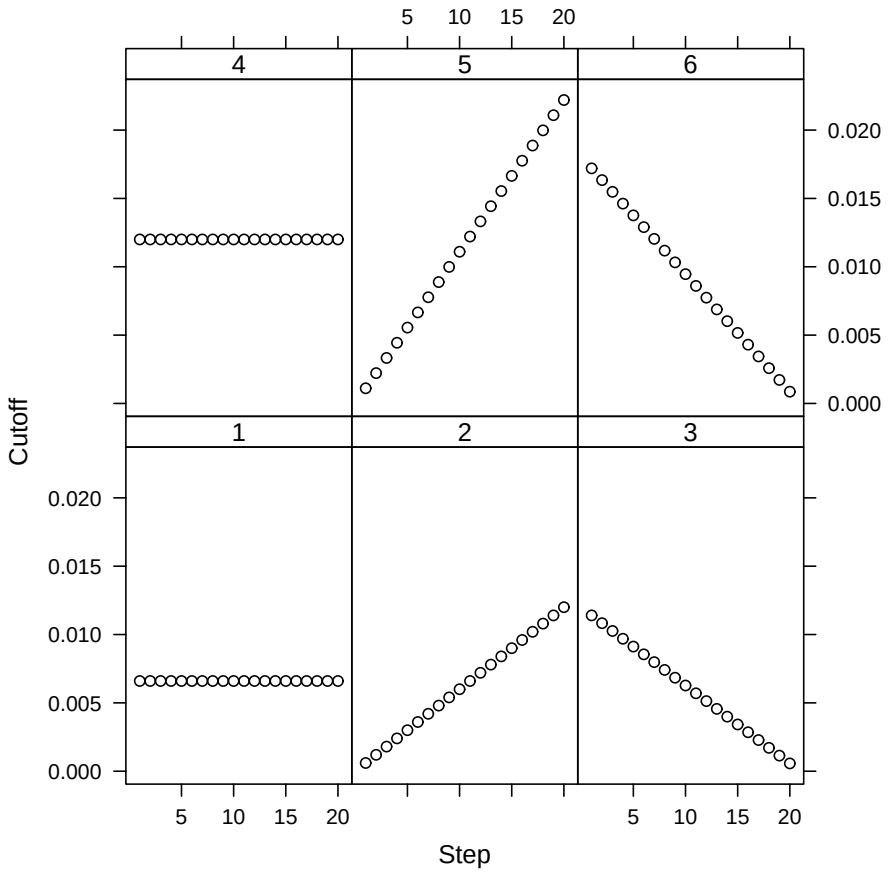


Figure 3-2: The six strategies for setting significance levels. The upper three are with data accumulation and the lower three are without.

	Percent rejected	Periods until rejection
1	48	9.5
2	38	12.5
3	37	7.0
4	99	6.5
5	99	7.4
6	96	6.3

Table 3.7: Rejection rate and rejection time for the six p-value strategies.

data was accumulated or not, a strategy of decreasing α_t led to the fastest detections, and increasing α_t led to the slowest, with constant α_t's somewhere in between. For the implementation of the dynamic algorithm we will use constant α_t's.

3.4.2 The Dynamic Algorithm

Suppose we are in period t. We want to set the levels α_i^t at which to test node i in period t, $\forall i \in \mathcal{N}$. We use the data collected in periods $1, \ldots, t-1$ to estimate the rate of occurrence of i in the treatment group, $\hat{r}_{i,t}^T$, and the control group, $\hat{r}_{i,t}^C$. The estimate is simply the rate at which i has occurred thus far in each group. It is calculated using the following formulas, where $x_{i,j}^T$ and $x_{i,j}^C$ are the number of occurrences of i in period j in the treatment and control groups respectively and N_j^T and N_j^C are the number of people in the treatment and control groups, respectively, in period j:

$$\begin{aligned} \hat{r}_{i,t}^T &= \sum_{j=1}^{t-1} x_{i,j}^T / \sum_{j=1}^{t-1} N_i^T \\ \hat{r}_{i,t}^C &= \sum_{j=1}^{t-1} x_{i,j}^C / \sum_{j=1}^{t-1} N_i^C \end{aligned}$$

For each i we use the estimated rates to compute an estimated power curve for its test. The power curves are calculated using the method described in [34] and implemented in the `bpower` function in the `Hmisc` library in R. For each possible value of α_i^t the curve gives the probability of detecting an effect. As in the single-period setting, to maintain linearity in our optimization model we then compute a piecewise-linear approximation

of the estimated power curve, using n pieces ($n = 100$ in our implementation). The curve only needs to be computed over the interval $[0, \alpha_t]$.

To determine which tests to perform and at which significance levels, we solve the following optimization problem. Let t be the current period. For each node i we approximate the power curve for its test with n linear pieces with slopes $m_j^{i,t}$ and intercepts $b_j^{i,t}$ for $j = 1, ..., n$. We assign a weight w_i to each test based on its height in the tree. The decision variables $\alpha_{i,t}$ are the significance levels at which to perform each test. The variables $r_{i,t}$ represent the power of each test and binary variables $p_{i,t}$ indicate whether each test is performed or not. We also keep track of nodes that have been rejected (as there is no need to test them again) using an indicator variable R_i that is updated every period. R_i is set to 1 if the null hypothesis corrresponding to node i has been rejected.

$$
\begin{aligned}
\text{maximize} \quad & \sum_{i \in \mathcal{N}} w_i r_{i,t} \\
\text{subject to} \quad & \sum_{i \in \mathcal{N}} \alpha_{i,t} = \alpha_t \\
& \alpha_{i,t} \leq p_{i,t} && \forall i \in \mathcal{N} \\
& r_{i,t} \leq \alpha_{i,t} m_j^{i,t} + b_j^{i,t} && \forall i \in \mathcal{N},\ j = 1, \ldots, n \\
& \alpha_{i,t} \leq 1 - R_i && \forall i \in \mathcal{N} \\
& p_{k,t} + p_{l,t} \leq 1 && \forall k, l \in \mathcal{N} \text{ where } k \text{ is a descendant of } l \\
& p_{i,t} \in \{0, 1\}, \alpha_{i,t} \geq 0 && \forall i \in \mathcal{N}.
\end{aligned}
$$

The only new constraint is the fourth one, which ensures that we don't perform a test if it has been performed previously and resulted in the null hypothesis being rejected.

When solving the MIP the full tree of ICD-9 codes we relax the integrality constraints to speed up the solution. We do this by removing the constraints that prevent a node and one of its descendants from being tested simultaneously. Once these constraints are removed, the remaining maximization problem is an LP. The solution of the LP can be transformed into an approximately optimal solution to the original

MIP by using a simple heuristic that "pushes down" the power to the bottom of the tree.

The heuristic works as follows. For each leaf of the tree that has been assigned a positive significance level by the LP, check to see if any nodes above it have also been assigned a positive significance level. If they have, decrease their significance level to 0 and increase the significance level of one of their children by the same amount. (This can be done in several ways: the significance level can always be given to the left-most child, it can be divided evenly between all children, it can be allocated to the child with the highest current significance level, etc.) Repeat this process until no leaf that has a positive significance level has an ancestor with a positive significance level. Note that the "power" is "pushed down" the tree one level at a time. Next, repeat this process for the second level of the tree, and so on up the tree until the root node has been reached. At this point, the solution will be feasible for the MIP. An alternative approach would be to "push up" the power from the bottom of the tree to the top.

There is also the issue of initializing the algorithm in the first period, when we have no previous data with which to estimate the rates. There are several ways to initialize the algorithm. In our implementation, we set the estimated rates for the control group to the base rates in the whole population and the rates for the treatment group to double those in the control group. Information gathered during the clinical trials or from experts in the field can also be used to specify initial estimated rates for some or all of the i.

There are two variants of the approach which we also tested in our simulations. In the first variant, the observations are allowed to accumulate from period to period. That is, when a particular hypothesis is tested, all data gathered so far is used in the test, not just the data gathered in the current period. This can lead to more powerful tests as time goes on, but introduces the complication of using data to decide whether to test a hypothesis and including the same data when testing the hypothesis. This may increase the rate of false detections. The second variant is to allow hypotheses to be retested after they have been rejected. This allows for stronger confirmation

of a result but also leaves open the possibility of a reversal in outcomes which could be difficult to interpret. Also, there is a trade-off to consider between detecting the largest number of possible side effects or, once a single side effect has been detected, concentrating effort on confirming that side effect.

3.5 Simulation of the Multi-Period Setting

We performed three small-scale simulations of six variations of this approach to better understand its behavior and performance. In all three simulations there were four diagnoses grouped into two categories and one large category of all four diagnoses. The corresponding tree structure is shown in Figure 3-3. Nodes D, E, F, and G represent the four diagnosis. Category B consists of diagnoses D and E. Category C consists of diagnoses F and G. Category A consists of all four diagnoses.

There were 100 people in the treatment group and 100 people in the control group. They were followed for 20 months. In each month, each patient incurred diagnosis x with the appropriate probability (described below), independent of any other diagnoses incurred in the month. Thus, a given patient in a given month could have from zero to all four of the possible diagnoses. Diagnoses incurred in one month were independent from those incurred in other months. 1000 trials were performed.

The six variations of the algorithms that were tested in the simulations were:

1. $\alpha_t = .05$ in each month.

2. $\alpha_t = .0066$ in each month so that the overall $\alpha = .05$ across the 20 months.

3. $\alpha_t = .0066$ and hypotheses that have been rejected can be retested. The disposition of the hypothesis is taken as the result of the last test if more than one test was performed.

4. Observations are allowed to accumulate over time. To control the overall significance level at .05, $\alpha_t = .012$ was used in each month.

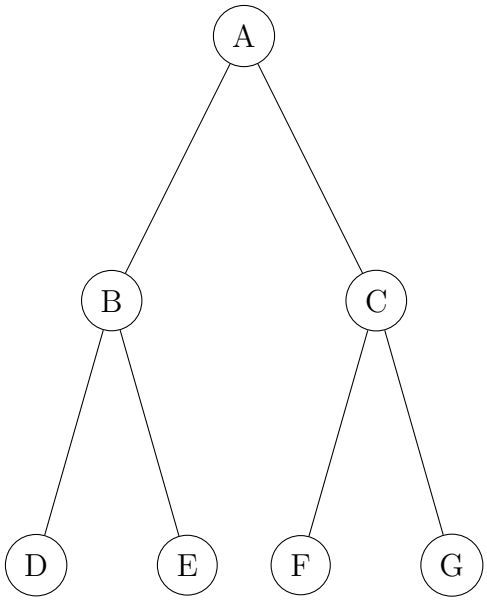


Figure 3-3: The structure of the tree used in the simulations.

5. Observations are allowed to accumulate and a smaller significance level, $\alpha_t = .0066$, was used in each month.

6. Observations are allowed to accumulate and retesting of rejected hypotheses is allowed. An $\alpha_t = .0066$ was used in each month.

The Bonferonni Approach was also included in the simulation for comparison and is labeled variation 7 in the results.

3.5.1 Simulation of a single effect

In the first set of simulations, diagnoses E, F, and G each occurred at a rate of 10% in both the treatment and control groups (i.e., the "treatment" had no effect on these diagnoses). Diagnosis D occurred at a rate of 20% in the treatment group and 10% in the control group (i.e., the treatment doubled the rate of occurrence of this diagnosis).

Figure 3-4 illustrates what our approach (using Variation 1) did in each period in one trial of the simulation. In the first period, when no prior information was available, all the leaves were tested and it happened that the effect at node D was immediately detected. Node B was tested in the next three periods (along with some

of the leaves) and the effect at node B was detected. Then testing focused on node A, where four periods later the effect was detected. After that point, testing focused on the remaining leaves but no additional effects were detected (as, in fact, there were none).

Table 3.8 shows the percent of trials in which each variation detected the increased rate of diagnosis D in the treatment group. Variations 1, 4, 5, and 6 detected the effect all the time. Variation 2 detected the effect almost all of the time. All five of these variations performed better than the Bonferroni approach, which only detected the effect 66% of the time. Variation 3, which allowed retesting but not accumulation of the data, performed poorly.

1	2	3	4	5	6	7
100	97	18	100	100	100	66

Table 3.8: Perecent of trials in which the increased rate of diagnosis D was detected by each variation.

Table 3.9 shows the average number of months until the increased rate of diagnosis D was detected by each variation (when it was detected). The variations using accumulation – Variations 4,5, and 6 – all detected the effect in approximately three months on average. Variation 1 took four months. Variations 2 and 3 took over twice as long as the variations that used accumulation. All the variations outperformed the Bonferroni approach, which took nine months on average.

1	2	3	4	5	6	7
3.1	6.4	6.5	2.7	2.8	2.9	9.2

Table 3.9: Average number of months until the increased rate of diagnosis D was detected.

Table 3.10 shows the percent of trials in which each variation detected the increased rate of diagnosis D's parent (node B) and grandparent (node A) in the tree. (Note that for the variations that actually detected the effect at node D, whether the effect at its ancestors is detected is really of secondary importance.) Variations 4 and 5 performed the best. The variations that used retesting – 2 and 6 – performed

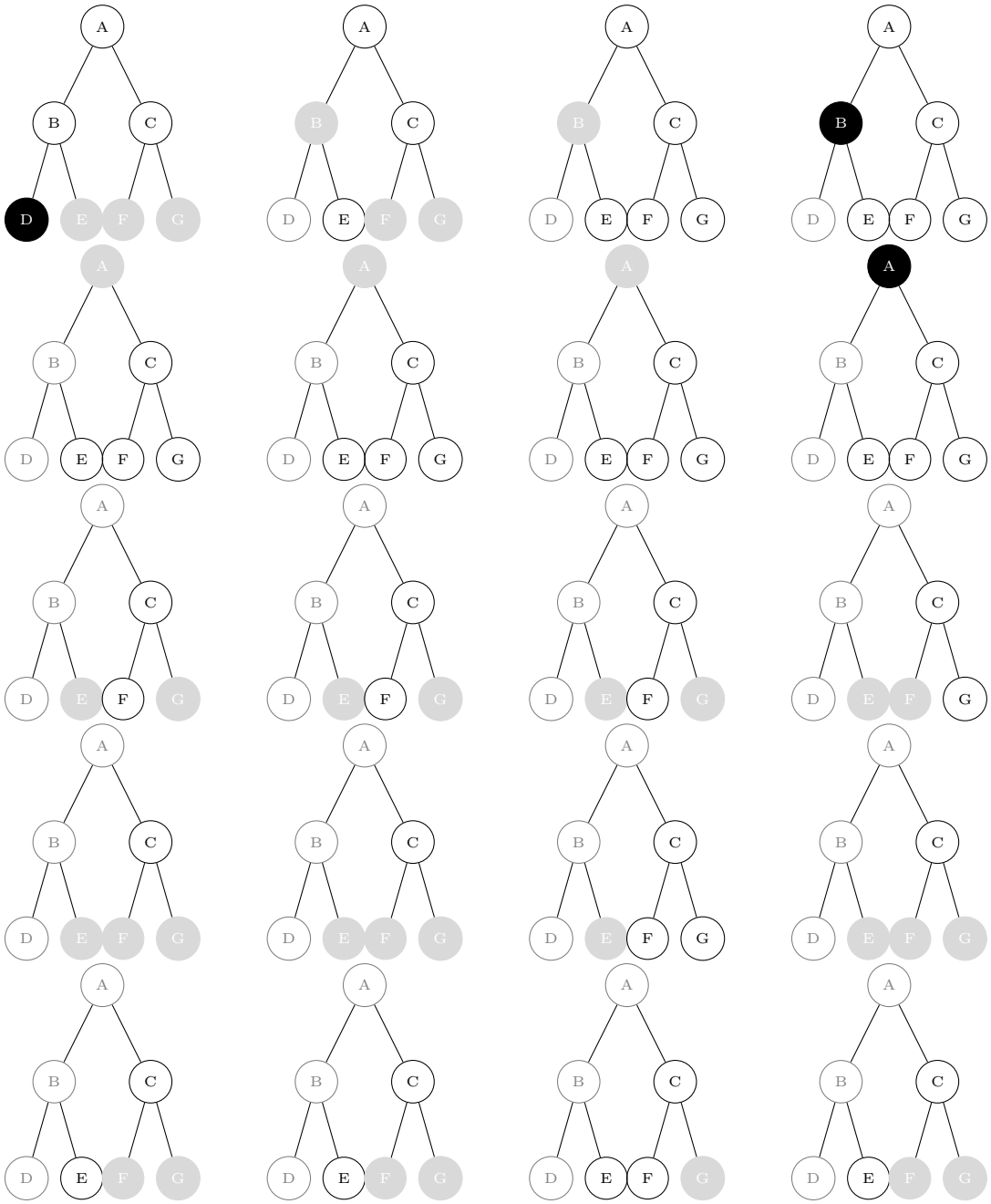


Figure 3-4: The behavior of the algorithm. A tree is shown for each month. Solid black nodes are those that were tested and rejected. Solid gray nodes were tested but not rejected. Nodes that were rejected in earlier periods are shown with a gray outline. The figure reads across from top to bottom.

quite poorly. They most likely spent every month retesting the leaves, rather than testing nodes A and B, which would explain their poor performance. Variation 1 was the third best. Variation 2 was a distant fourth, although it did outperform the Bonferroni approach.

	1	2	3	4	5	6	7
A	85	22	0	100	100	2	22
B	100	67	0	100	100	9	37

Table 3.10: Percent of trials in which the increased rates at nodes A and B were detected.

Table 3.11 shows the percent of trials in which effects of the treatment were found at other nodes in the tree. These are false detections. Variations 2, 3, and 6 had the lowest false detection rates, comparable to the Bonferroni approach. Note that the retesting used by Variations 3 and 6 can only reduce the rejection rate, which would explain why they had so few false detections. Variations 4 and 5 had higher false detection rates, with Variation 5 performing slightly better than Variation 4. Variation 1 had the highest false detections rate, which makes sense because we made no attempt to account for the number of time periods when setting α_t.

	1	2	3	4	5	6	7
C	5	1	1	4	2	0	2
E	11	2	0	5	5	2	0
F	5	0	0	1	0	0	1
G	9	0	0	6	2	1	0

Table 3.11: Percent of trials in which there were false detections at the other nodes in the tree.

Looking across all of these measures of performance, Variation 2 never performed worse than the Bonferroni approach and substantially outperformed it on nearly all of the measures. Of the other variations, many did dramatically better than the Bonferroni in terms of detections and time until detection, though not in terms of false detections.

3.5.2 Simulation of two effects on same branch

We next simulated a scenario in which there were effects on two diagnosis that share the same parent node in the tree. Diagnoses F and G each occurred at a rate of 10% in both the treatment and control groups as before. Diagnoses D and E each occurred at a rate of 10% in the treatment group and 5% in the control group.

Figure 3-5 illustrates what our approach (using Variation 1) did in each period in one trial of the simulation. The algorithm spent several periods testing leaves on the tree. By period four it began to focus on the correct half of the tree and by period seven it detected the effect at node B. Note that it detected the combined effect at node B before it detected the smaller individual effects at nodes D and E. The effect at node E was actually detected in the following period. The algorithm then focused on node A, but failed to detect the effect following nine successive tests. Finally in period 18 it returned to node D and detected the effect there. In period 20 it detected the effect at node A, though since the effects at B, D, and E had already been detected, detecting the effect at node A is really of little importance.

Table 3.12 shows the percent of trials in which each variation detected the increased rates of diagnoses D and E in the treatment group. The variations that use accumulation detected both effects all of the time. Variation 1 was close behind. Variation 2 didn't perform as well, but still outperformed the Bonferroni approach. Variation 3, with its retesting without accumulation, performed very poorly. Overall, the relative order of the variations was the same as in the previously simulation. The gap between Variations 4, 5, and 6 and Variations 2 and 7 widened.

	1	2	3	4	5	6	7
D	91	36	2	100	100	100	32
E	90	27	0	100	100	100	22

Table 3.12: Perecent of trials in which the increased rate of diagnoses D and E were detected.

Table 3.13 shows the average number of months until the increased rates of diagnoses D and E were detected by each variation (when detected). Variation 3 had

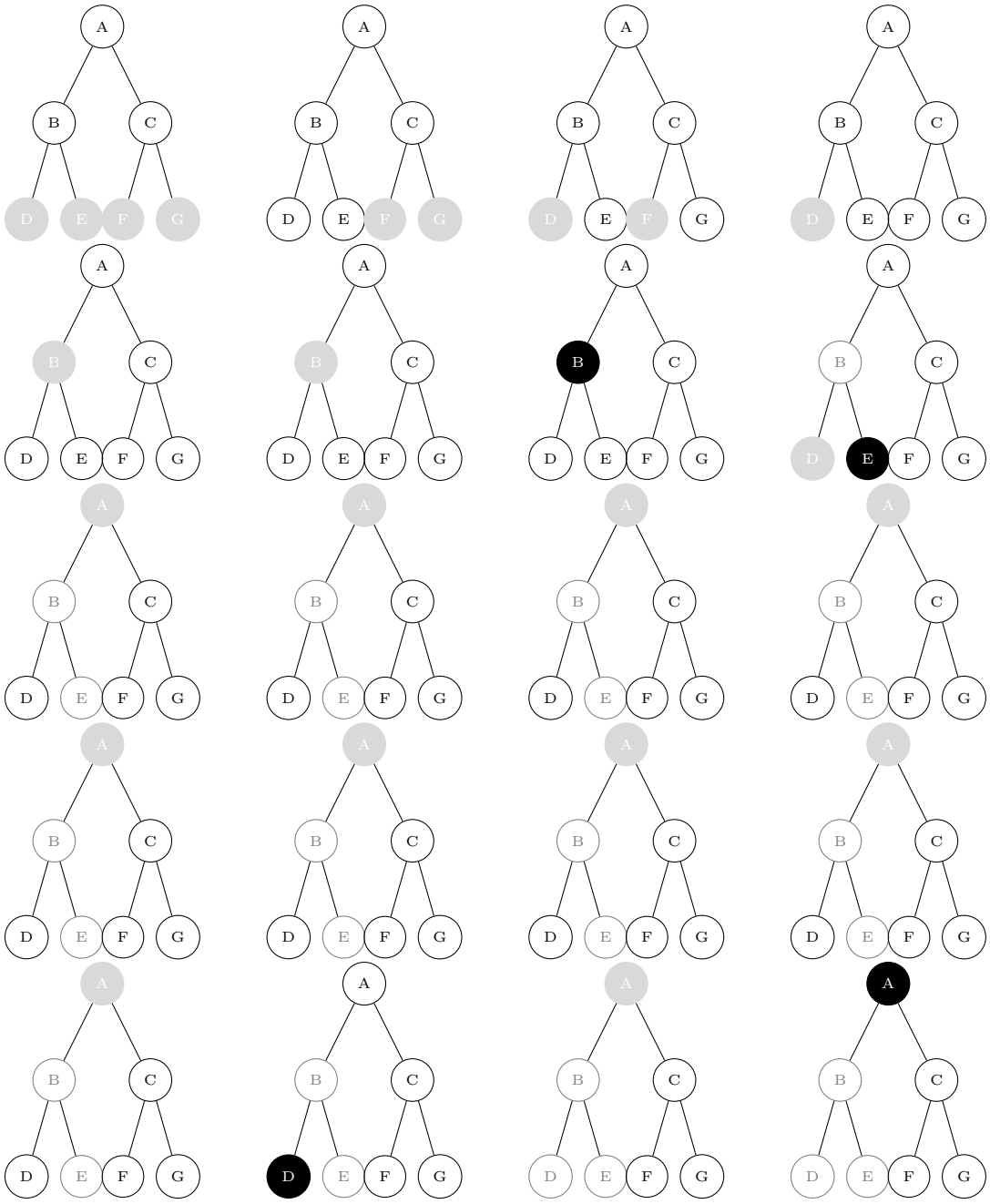


Figure 3-5: The behavior of the algorithm. A tree is shown for each month. Solid black nodes are those that were tested and rejected. Solid gray nodes were tested but not rejected. Nodes that were rejected in earlier periods are shown with a gray outline. The figure reads across from top to bottom.

the quickest detections, but this figure is misleading because these detections were always reversed by later re-testing. Variations 4, 5, and 6 each took approximately six months. Variation 1 took nine months, Variation 7 took 11 months, and Variation 2 took a little more than 11 months. Note that the variations took longer in general to detect the effects than in the previous simulation because the rates at which the diagnoses occurred were smaller.

	1	2	3	4	5	6	7
D	8.6	11.4	3.4	5.5	6.2	6.3	10.7
E	9.1	12.0	5.5	5.8	6.4	6.7	10.9

Table 3.13: Average number of months until the increased rates of diagnoses D and E were detected.

Table 3.14 shows the percent of trials in which each variation detected the increased rate of diagnoses D and E's parent (node B) and grandparent (node A). For node B, all of the variations outperformed the Bonferroni approach with the exception of the variations that used retesting. For node A, the Bonferroni approach did better than Variation 2, but again this is of little concern because detecting the effect at node B is more valuable and Variation 2 did better there. The retesting variations had very low detection rates at node A because they devoted their efforts to retesting nodes further down on the tree.

	1	2	3	4	5	6	7
A	68	4	0	98	98	1	18
B	100	93	19	100	100	58	64

Table 3.14: Percent of trials in which the increased rates at nodes A and B were detected.

Table 3.15 shows the percent of trials in which false detections occurred at the other nodes. As before, Variations 1, 4, and 5 had much higher false detection rates. The variations that use retesting had no false detections. Variation 2 also had no false detections, slightly better than the Bonferroni.

In summary, again Variation 2 outperformed the Bonferroni approach while the other variations had mixed results. Note that Variation 2's performance on the leaves

wasn't exceptional, but it detected the effect at node B over 90% of the time, which demonstrates the advantage of using the hierarchy of diagnoses rather than just testing individual diagnoses.

The retesting variation with accumulation did very well at detecting the effects at the leaves and had no false detections. It also detected the effects faster than the Bonferroni approach. Although its detection rate at nodes A and B was worse than the Bonferroni, this really isn't important since it made the detections at the leaves. The variations that used accumulation without retesting had the highest detection rates and quickest detections, but had unacceptably high false detection rates.

	1	2	3	4	5	6	7
C	1	0	0	7	4	0	1
F	3	0	0	4	2	0	0
G	5	0	0	6	4	0	1

Table 3.15: Percent of trials in which there were false detections at the other nodes in the tree.

3.5.3 Simulation of two effects on different branches

Lastly, we simulated a scenario in which there were effects on two diagnoses that have different parent nodes. Diagnoses E and G each occurred at a rate of 10% in both the treatment and control groups. Diagnoses D and F each occurred at a rate of 10% in the treatment group and 5% in the control group.

Figure 3-6 illustrates what our approach (using Variation 1) did in each period in one trial of the simulation. The testing was generally spread across three of the four leaves, and occasionally included one of the mid-level nodes. Node A was never tested. The effect at node D was detected in the 11th month. After that, testing was still spread across nodes B, F, and G, until the 16th period when the focus started to narrow to nodes B and F. However, by the 20th period the effect at node F still hadn't been detected.

Table 3.16 shows the percent of trials in which each variation detected the increased rates of diagnoses D and F in the treatment group. Even with the effects

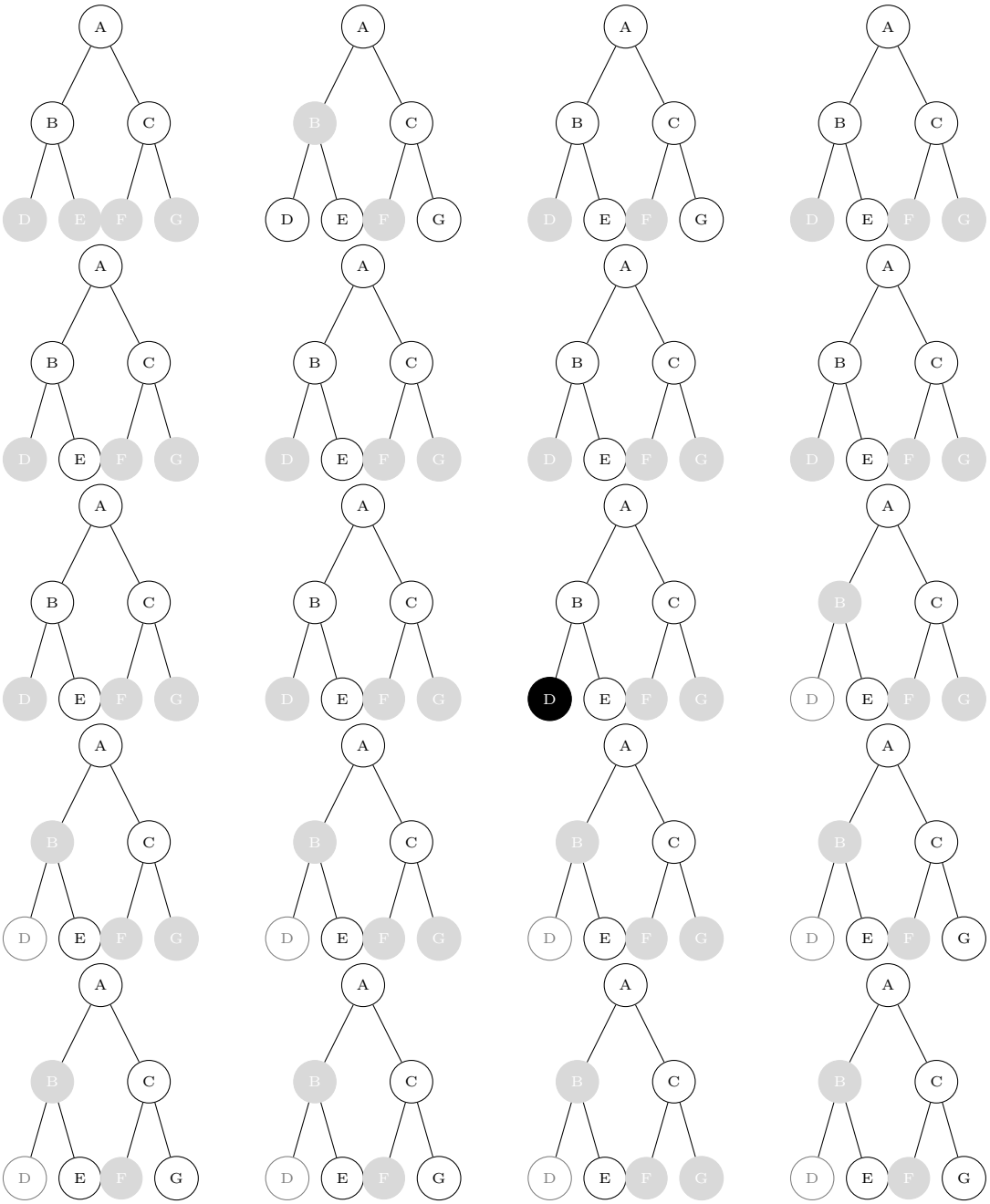


Figure 3-6: The behavior of the algorithm. A tree is shown for each month. Solid black nodes are those that were tested and rejected. Solid gray nodes were tested but not rejected. Nodes that were rejected in earlier periods are shown with a gray outline. The figure reads across from top to bottom.

separated into different halves of the tree, the variations that used accumulation had no problem detecting them. Variation 2's performance was markedly stronger than the Bonferroni approach.

	1	2	3	4	5	6	7
D	88	37	2	100	100	100	25
F	94	35	4	100	100	100	22

Table 3.16: Perecent of trials in which the increased rate of diagnoses D and E were detected.

Table 3.17 shows the average number of months until the increased rates of diagnoses D and F were detected by each variation (when detected). The results are similar to the previous simulations.

	1	2	3	4	5	6	7
D	6.8	10.5	10.3	5.1	5.7	5.7	10.9
F	8.3	10.6	10.2	5.5	6.1	6.2	9.8

Table 3.17: Average number of months until the increased rates of diagnoses D and F were detected.

Table 3.18 shows the percent of trials in which each variation detected the increased rate of diagnosis D and F's parents (nodes B and C, respectively) and grandparent (node A) in the tree. The variations that used accumulation without re-testing performed strongly here and the variations that used retesting performed poorly. The Bonferroni approach did much better at detecting effects at nodes A and C than Variation 2 did. As illustrated in Figure 3-6, these nodes were rarely tested by Variations 1 and 2.

	1	2	3	4	5	6	7
A	71	5	1	100	99	5	19
B	32	2	0	93	85	6	6
C	31	0	0	92	85	2	13

Table 3.18: Percent of trials in which the increased rates at nodes A, B, and C were detected.

Table 3.19 shows the percent of trials in which false detections occurred at the other nodes. Variations 4 and 5 continued to have higher false detection rates than the rest. Between the remaining variations the differences are much smaller than in the other two simulations, with very low false detection rates. Even Variation 1, which had a higher false detection rate in the other two simulations, has a low false detection rate here. This is most likely because there are true effects at five of the seven nodes on the tree, so it spent very little power testing for the non-existent effects.

Overall, the story was more mixed in this scenario. This is to be expected – when the true effects are dispersed the hierarchical approach doesn't lead to stronger tests. For example, in this scenario a test at node B is really no more likely to discover an effect than a test at node D – it only has the added noise from node E, where there is no effect.

While Variation 2 performed better on the leaves, it didn't spend much time testing nodes A and C, where the Bonferroni approach was able to detect effects. Variation 1 dominated the Bonferroni approach in this scenario, having much higher detection rates at all the nodes and it wasn't hurt by a high false detection rate as it was in the other scenarios.

	1	2	3	4	5	6	7
E	0	0	0	5	3	0	0
G	1	0	0	4	3	0	1

Table 3.19: Percent of trials in which there were false detections at the other nodes in the tree.

Across these three sets of simulations, we have tested our approach against a range of situations that might arise in practice: a single side effect, multiple related side effects, and multiple unrelated side effects. In all three simulations our approach outperformed the Bonferroni approach, often by a considerable margin. Consistently, "plain vanilla" Variation 2 and the accumulation/retesting Variation 6 performed best. Variation 2 had higher detection rates than the Bonferroni approach while maintaining a comparable false detection rate. In particular, when there were two related side effects, it detected the effect on their common parent node 93% of the time com-

pared with 64% for the Bonferroni approach. When accumulation is combined with retesting of rejected hypotheses, as in Variation 6, the performance is very strong. This combination led to detection rates up to four times higher than the Bonferroni approach and detection times that were twice as fast. The variations that used data accumulation without retesting, on the other hand, led to false positive rates that were unacceptably high. A possible extension would be to consider a variation that uses a fraction of the data collected each month to estimate the power curves and the remaining fraction to perform the hypothesis tests.

3.6 Conclusion

We have demonstrated a promising algorithm for conducting large scale drug surveillance. The algorithm allows the rates of all diagnoses to be monitored, yet controls the rate of false detections. Based on our simulations, side effects were generally detected more quickly and more often than the Bonferroni approach. The integer optimization formulation also entails great flexibility, allowing diagnoses of particular concern to be prioritized.

Chapter 4

Depression and Cost of Health Care

4.1 Background

Chronic diseases constitute a growing proportion of total global disease burden [63], and are projected to increase to 60% of global disease burden by the year 2020 [66]. Depression is currently ranked fourth of all causes of global disease burden, and is projected to rise to second by 2020 [66]. However, in spite of its global importance, the interaction between depression and chronic comorbid diseases remains incompletely understood with regard to prevalence, severity of disease, and potential causative factors mediating this interaction [32].

Over the past 25 years, a growing body of evidence has established an association between depression and high utilization of general medical services. Recent studies of this issue have used cost of services as a measure of utilization of care, and have quantified the increased cost of general medical services associated with depression in several different medical settings.

Simon [97] found the per capita annual cost for primary care patients diagnosed with depression was $4246, compared to $2371 for nondepressed primary care patients. Mental health care accounted for only 20% of increased cost in depressed individuals. Henk et al [43], in a study of high utilizers of care, found that depressed patients

had per capita annual cost of $5,764, compared to $4227 for nondepressed patients. Unutzer [105] found that median annual healthcare costs in Medicare recipients who were diagnosed with depression was $2147, compared to $1461 for Medicare recipients who were not depressed. Druss [28], in a Veteran's Administration cohort of medical and surgical inpatients, found that the average annual cost of the most depressed patients was $9,408, compared to $5,290 for the least depressed patients. Thomas et al [102], in a study of Medicaid beneficiaries, found that depressed patients had total annual cost of $7,284, while nondepressed patients had total annual cost of $2649.

Studies of specific chronic medical illnesses have found that depression is associated with significantly greater annual per capita cost of care. Ciechanowski [20] found that the median annual cost of care for patients with diabetes and depression was 1.86 times the cost of care for diabetic patients without depression. Egede [30] found the annual cost of care in depressed diabetics was 4.5 times the cost of care for non-depressed diabetics. Sullivan [100], in a study of patients with congestive heart failure, found that depressed patients had median annual cost of care 1.29 times the cost of nondepressed patients.

Although there is some consistency in the magnitude of cost differences reported in these studies, the relative magnitude between various chronic comorbid diseases remains unclear. Thus far there has been no attempt to measure and compare the cost differences associated with depression in the most prevalent chronic comorbid diseases in a primary care population.

Administrative data sets allow accurate measurement of medical costs across large populations and a wide range of treatment settings [65]. They also sometimes reflect real-world patterns of utilization and medical practice more accurately than data from randomized trials [11]. For these reasons, a large administrative data set was considered the optimal basis for measuring the cost of healthcare in different disease states.

The objectives of this study were:

1. To examine the relationship between depression and cost of non-mental health care in 11 chronic comorbid diseases.

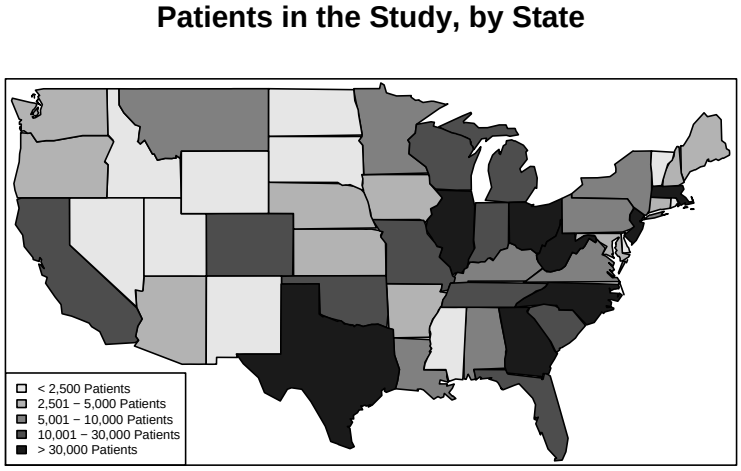


Figure 4-1: Geographic distribution of the research cohort.

2. To examine patterns of utilization of services, comorbidity, and prevalence, for evidence of causation between depression and 11 chronic comorbid diseases.

4.2 Methods

The database for this study consisted of de-identified medical claims data for 618,780 patients in self-insured plans. Only patients enrolled in an insurance plan for the entire 12-month study interval were included. The geographical distribution of patients is shown in Figure 4-1. No Medicare or Medicaid recipients were included. The study interval was September 1, 2004 to August 31, 2005.

Codes for chronic comorbid diseases were selected by the criteria of prevalence and chronicity in order to capture the maximum number of patients with chronic comorbidity. Previously published comorbidity indices [18, 31, 81, 25] were considered but not used, because they were developed to measure mortality risk, not cost of care.

All diagnostic codes were ranked in order of 12-month prevalence in the study

cohort. In order to select comorbidities with high prevalence, codes with a 12-month prevalence lower than 8.0 per 1000 were not included in the study. Of the codes satisfying prevalence criteria, only those which capture mainly chronic diseases, defined as diseases requiring care for years or decades, were selected for the study. Individual diagnostic codes were then grouped into 11 chronic diseases (see Table C.1). To further exclude non-chronic diseases, study subjects were not assigned a diagnosis of a chronic comorbid condition unless they had at least two outpatient visits or one inpatient admission under any one of these ICD-9 codes during the 12-month study interval. The chronic comorbid disease did not have to be the primary diagnosis.

Patients were designated as depressed by either of two criteria:

1. If during the study interval they received at least two outpatient codes or one inpatient code for any one of the 21 ICD-9 codes for depression listed in Table C.2. Depression did not have to be the primary diagnosis. Patients designated as depressed may or may not have had pharmacy claims for antidepressant medication. Patients who received a diagnostic code for depression but no antidepressant medication were included in the designation of depressed.

2. If during the study interval they filled two or more prescriptions for an antidepressant medication. It was not necessary for all prescriptions to be for the same antidepressant.

Patients designated as depressed were then assigned to one of three categories:

1. Patients not having received a coded diagnosis of depression, but taking antidepressants.

2. Patients having received a coded diagnosis of depression, and taking antidepressants.

3. Patients having received a coded diagnosis of depression, but not taking antidepressants.

Patients who did not meet study criteria for designation as depressed were designated as not depressed.

For each chronic comorbid disease, the median annual per patient cost of care for the study interval was calculated for patients designated as not depressed, and for each of the three categories of patients designated as depressed. Costs were then compared between depressed and not depressed patients, and the differences calculated in both absolute dollar amounts and as ratios.

Most patients in the study cohort were under separate mental health management arrangements, which made the cost data of all inpatient and some outpatient mental health care unavailable for this study. Because of this fragmentation of mental health administrative data, median annual per patient mental health expenditure could not be calculated in this study, and was not part of the final cost analysis. However, mental health pharmaceutical charges and some outpatient mental health service charges were present in the medical administrative database. These were subtracted from the median annual per patient cost to calculate median per patient non-mental health cost.

Median annual per patient cost for inpatient, outpatient, emergency room and pharmaceutical services was calculated for each chronic comorbid disease, both in the depressed and not depressed cohort, and the difference was calculated both as an absolute dollar amount, and as a ratio.

In order to control for number of comorbidities as an independent cause of increased cost, the relationship between the number of chronic comorbid diseases per patient and median annual per patient cost was calculated.

The prevalence of each chronic comorbid disease was calculated in the depressed and not depressed cohorts in the study population, and the difference in prevalence between depressed and not depressed was expressed as a ratio.

The prevalence of depression in patients with each of the 11 chronic comorbid diseases was calculated, and compared to the prevalence of depression in patients without that comorbid disease. The difference in prevalence was expressed as a ratio.

Each chronic comorbid cohort was then divided into cost deciles, and prevalence

of depression in each cost decile was determined.

Because of separate mental health management arrangements, median annual per patient cost for specialty mental health care could not be calculated, and consequently the relationship between prevalence of chronic comorbid diseases and cost of mental health services could not be examined.

Longitudinal relationships between depression and chronic comorbid diseases were calculated for an extended 24 month study interval. The incidence of first diagnosis of each chronic comorbid disease subsequent to a diagnosis of depression or treatment with antidepressant was calculated for all 11 chronic comorbid diseases and compared to the incidence in not depressed individuals. The incidence of first diagnosis of depression or first antidepressant prescription subsequent to a diagnosis of each chronic comorbid disease was calculated for all chronic comorbid diseases, and compared to the incidence in individuals without prior diagnosis of that chronic comorbid disease.

4.2.1 Statistics

Median costs were used rather than mean costs because the cost distributions are skewed to the right and the mean cost is highly sensitive to a few very expensive patients. Therefore, the median gives a better sense of the location of the center of the distribution. We performed a sensitivity of analysis by comparing means, first quartiles, and third quartiles, and in each case the general pattern was the same.

The bootstrap was used to construct confidence intervals for the differences between median costs of the depressed and not depressed patients. To compare prevalences, standard chi-square tests were used. 95% confidence intervals are provided for all of the comparisons. All analyses were performed using R [78].

4.3 Results

Selection criteria yielded a study cohort of 618,780 subjects with a mean age of 41 (s.d. 12), of whom 53% were female and 47% were male (Table 4.1). 14.3 % of the study cohort had one or more of the 11 chronic comorbid diseases selected for the

	Subjects	Percent Female
Research cohort	618,780	53%
One or more chronic comorbid diseases	88,687	56%
No diagnosis of depression, on antidepressants	55,945	72%
Diagnosis of depression	14,005	72%
Diagnosis of depression on antidepressants	9,208	74%
Diagnosis of depression not on antidepressants	4,797	68%

Table 4.1: Summary of research cohort.

study. 11.3% of the study cohort were designated as depressed, but only 2.3% of the study cohort received a coded diagnosis of depression. In the patients designated as depressed, 72% were female. 11% of the study cohort were prescribed an antidepressant without receiving a coded diagnosis of depression during the study interval. The most prevalent chronic comorbid diseases were, in order of prevalence, diabetes, hypertension, pain in joint, back pain, and intravertebral disc disease (Table 4.2).

	Total	Non-depressed	On antidepressants, not diagnosed	Diagnosed, on antidepressants	Diagnosed, not on antidepressants
Asthma	5,406	3,988	1,087	237	94
Back Pain	13,434	9,942	2,673	604	215
CHF	1,131	826	250	39	16
CAD	5,758	4,609	981	120	48
Diabetes	20,843	16,632	3,499	523	189
Epilepsy	1,597	1,202	253	87	55
Headache	9,133	5,909	2,541	507	176
Hypertension	20,624	16,553	3,502	395	174
IVDD	13,158	9,623	2,739	584	212
Obesity	1,341	896	290	101	54
Pain in Joint	15,575	11,882	2,906	559	228

Table 4.2: Number of members in each disease category in the study.

The distribution of diagnostic codes for depression was concentrated in codes 296 and 311, which respectively comprised 50.6% and 23.2% of all depression codes assigned (Table C.2).

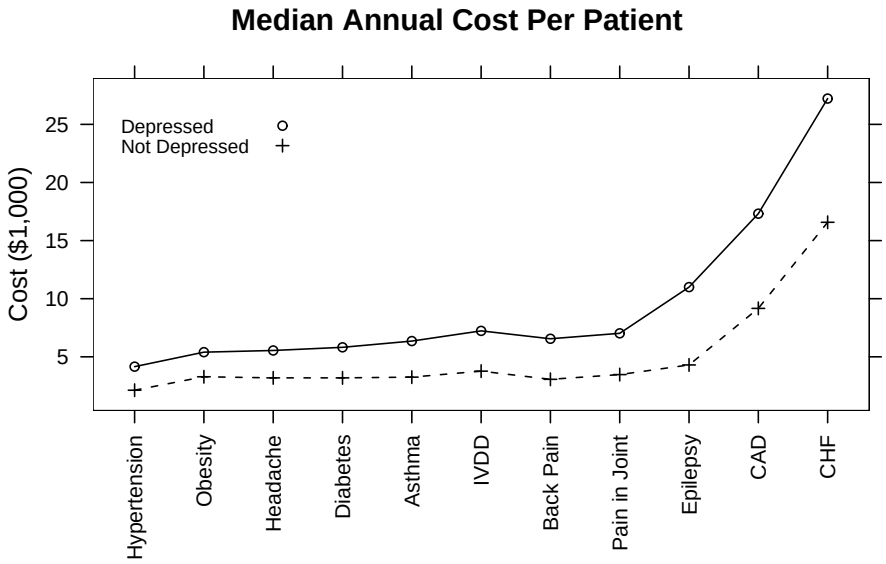


Figure 4-2: Annual per patient cost with and without depression. Costs of antidepressant prescriptions and mental health treatment are excluded.

Depressed patients had higher non-mental health costs than not depressed patients in all 11 comorbid diseases studied (Figure 4-2). The per-patient difference in non-mental health cost between depressed and not depressed patients ranged from $2027 in hypertension to $10,644 in CHF. Diseases with higher median annual per patient costs tended to have larger absolute dollar differences between depressed and not depressed patients. The ratio of cost between depressed and not depressed patients ranged from 1.64 in both CHF and obesity to 2.56 in epilepsy.

Patients in all three categories of depression consistently had higher costs than not depressed patients (Figure 4-3). Of the three depression categories, the patients diagnosed with depression and on antidepressants tended to have the largest cost differences, and patients not diagnosed with depression but taking antidepressants had the smallest cost differences.

Median annual pharmaceutical costs of depressed patients were consistently higher than the costs of not depressed patients (Figures 4-4 and 4-5). For most chronic comorbid diseases pharmaceutical cost was the largest component of total cost difference. The per-patient difference in pharmaceutical cost between depressed and

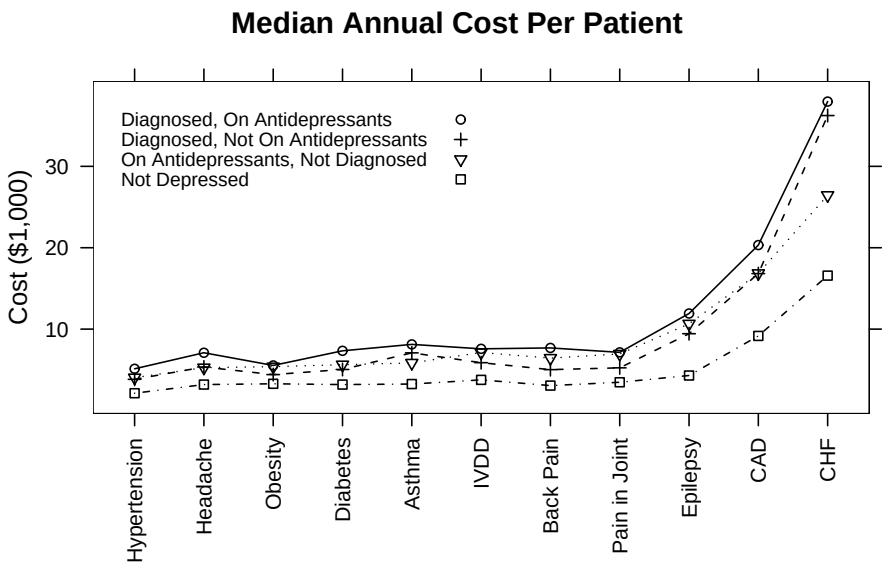


Figure 4-3: Annual per patient cost by depression subgroup. Costs of antidepressant prescriptions and mental health treatment are excluded.

not depressed patients ranged from $928 in obesity to $1911 in CHF. The ratio of depressed to not depressed pharmaceutical costs ranged from 2.17 in CHF to 7.55 in pain in joint.

Median annual outpatient costs of depressed patients were also consistently higher than the costs of not depressed patients (Figures 4-4 and 4-5). The difference in outpatient cost ranged from $964 in obesity to $1785 in CHF. The ratio of depressed to not depressed outpatient costs ranged from 1.32 in CHF to 2.04 in epilepsy.

Inpatient cost differences were a significant component only in CAD and CHF, in which the inpatient cost differences were $2312 and $4519 respectively. Emergency room cost differences were $95 for CHF and $317 for epilepsy, but for other chronic comorbid diseases, emergency room cost was not a significant component of cost increases associated with depression.

Depressed patients have a higher number of comorbidities than non-depressed patients (Table 4.3). The mean number of comorbidities in the depressed cohort was 0.8, and in the not depressed cohort was 0.34.

When controlled for number of comorbidities, depressed patients still had higher

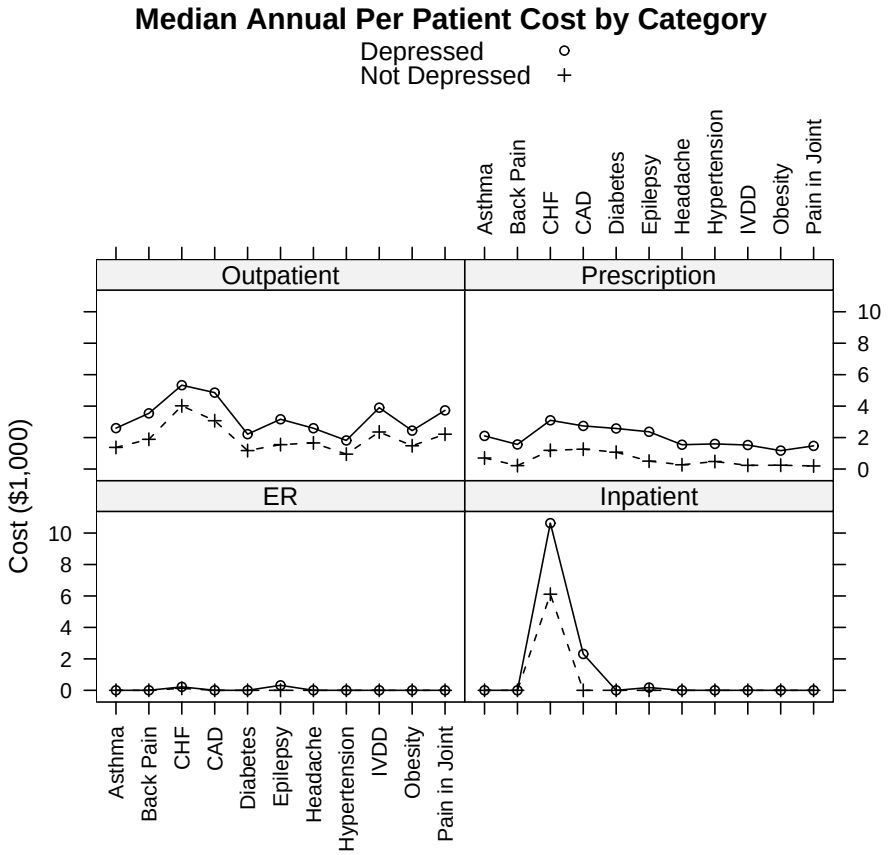


Figure 4-4: Median annual per patient cost by type of service. Costs of antidepressants and mental health treatment are excluded.

Comorbidities	0	1	2	3+
Not Depressed	77% (423,236)	16% (86,756)	5% (27,024)	2% (11,814)
Depressed	41% (29,082)	34% (23,507)	15% (10,271)	10% (7,090)

Table 4.3: Number of comorbidities versus depression status. Each row sums to 100%.

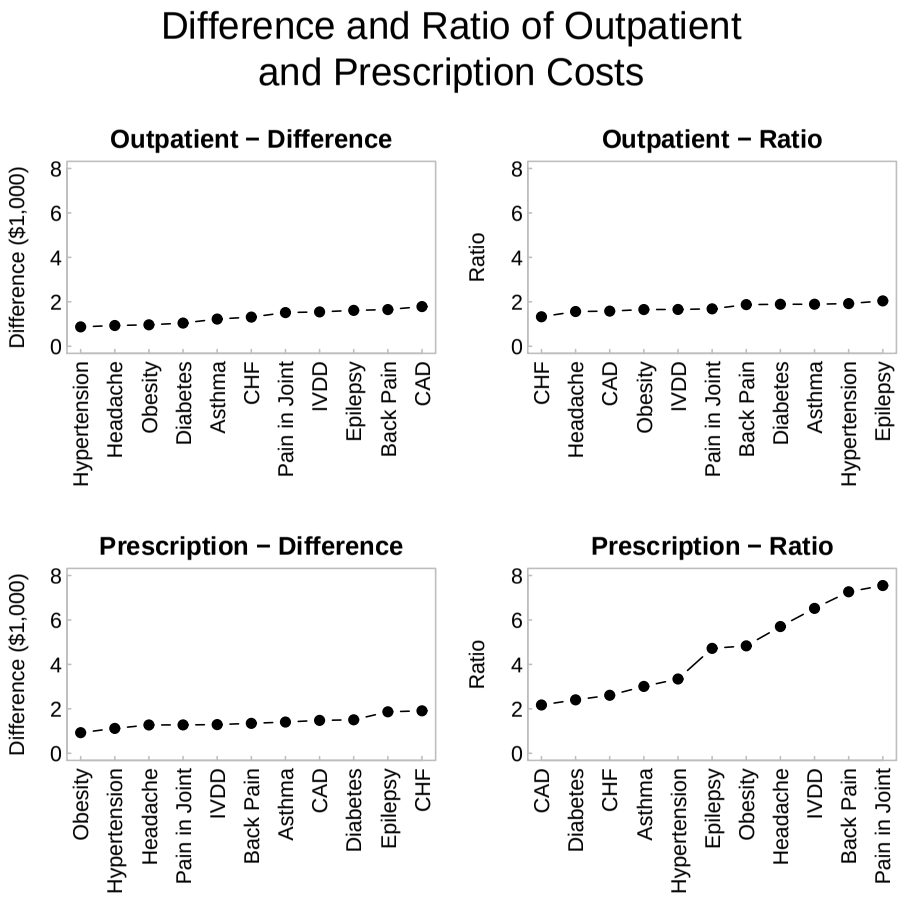


Figure 4-5: Outpatient and prescription costs. The ratio of the median cost of the depressed members to the median cost of the not depressed members and the median cost of the depressed members minus the median cost of the not depressed members. Costs of antidepressants and mental health treatment are excluded.

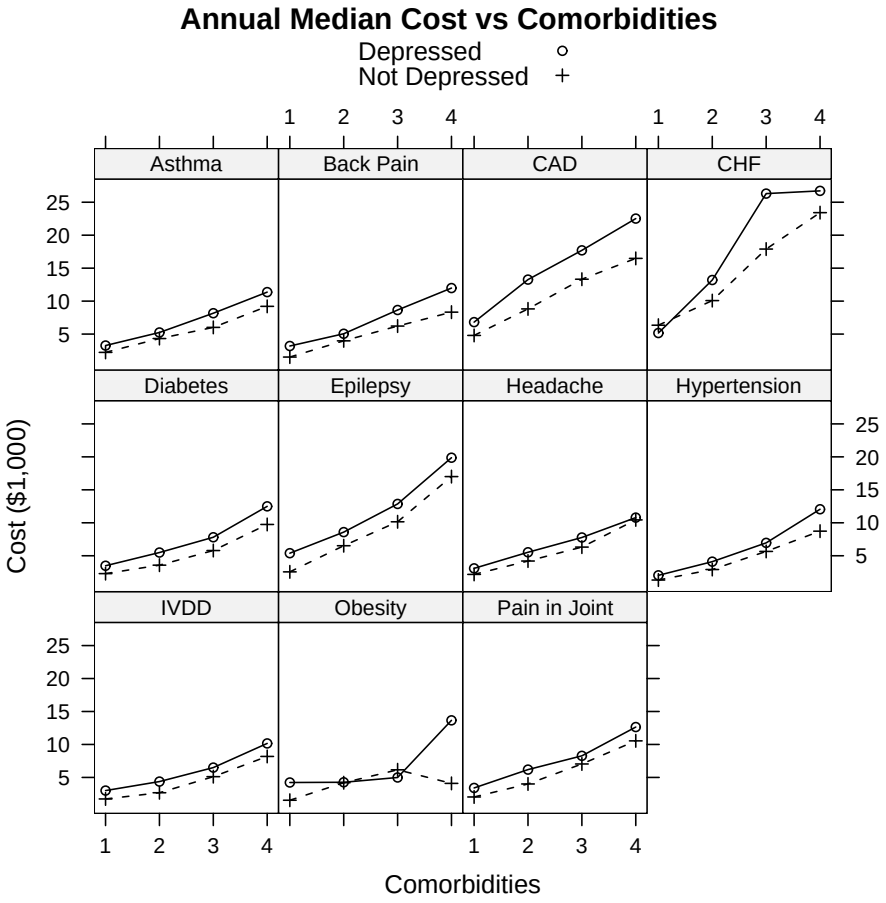


Figure 4-6: Annual cost vs. number of comorbidities.

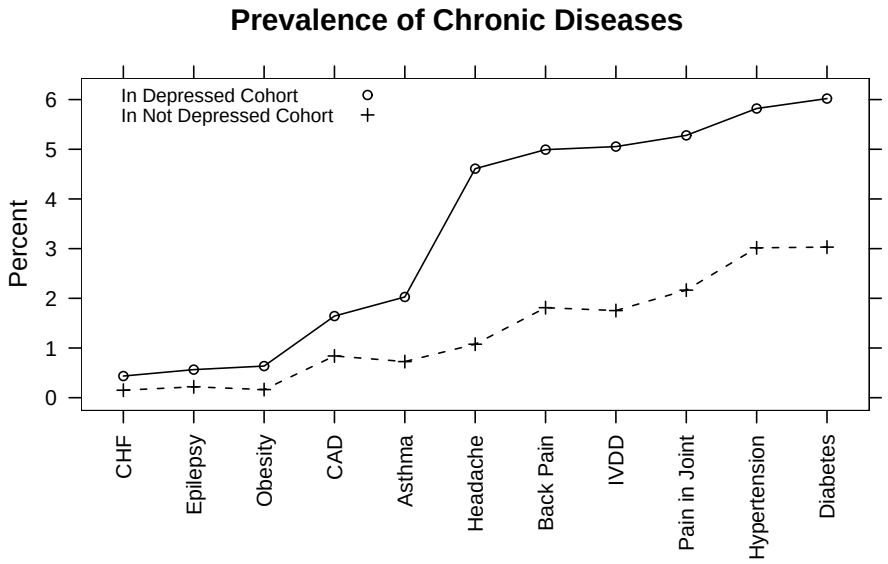


Figure 4-7: Prevalence of chronic diseases vs depression status.

costs than non-depressed patients (Figure 4-6). In each disease graph, the number of comorbidities increases from 1 to 4. For most patients with only one chronic disease, the difference in cost between depressed and not depressed patients was low. With rising number of comorbidities, the cost difference between depressed and not depressed patients increased in some comorbid diseases, but remained fairly constant in others.

Each of the 11 chronic comorbid diseases was more prevalent in the depressed cohort than in the total study cohort (Figure 4-7). The ratio of prevalence between depressed and not depressed patients ranged from 1.93 in hypertension to 4.28 in headache.

Depression is more prevalent in each of the 11 comorbid diseases than in the total study cohort (Figure 4-8). The ratio of depression prevalence between those with one particular chronic comorbid disease and those without it ranged from 1.78 in CAD to 3.22 in headache.

For each disease, patients were divided into non-mental health cost deciles (Figure 4-9). Within each cost decile, the 12 month prevalence of depression was calculated. In almost all of the diseases, the prevalence of depression increases linearly with

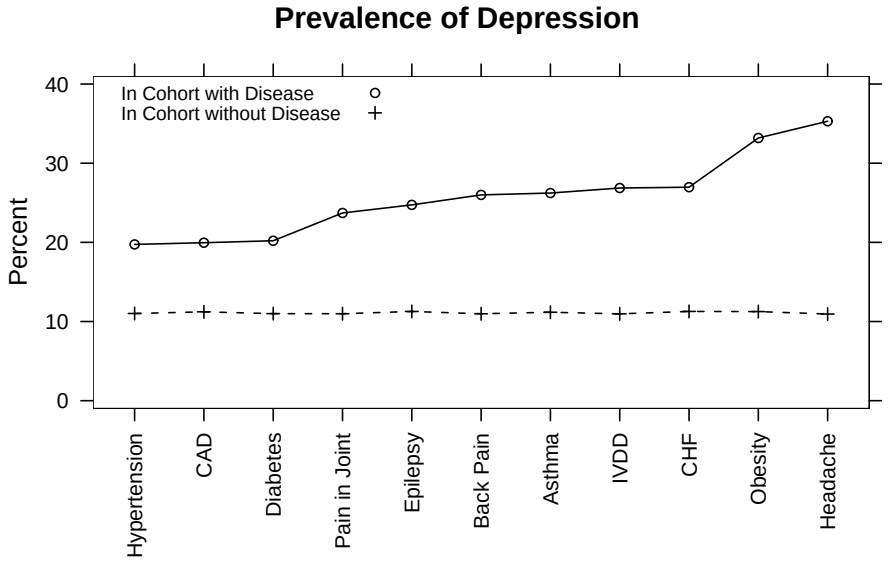


Figure 4-8: Prevalence of depression vs chronic disease status.

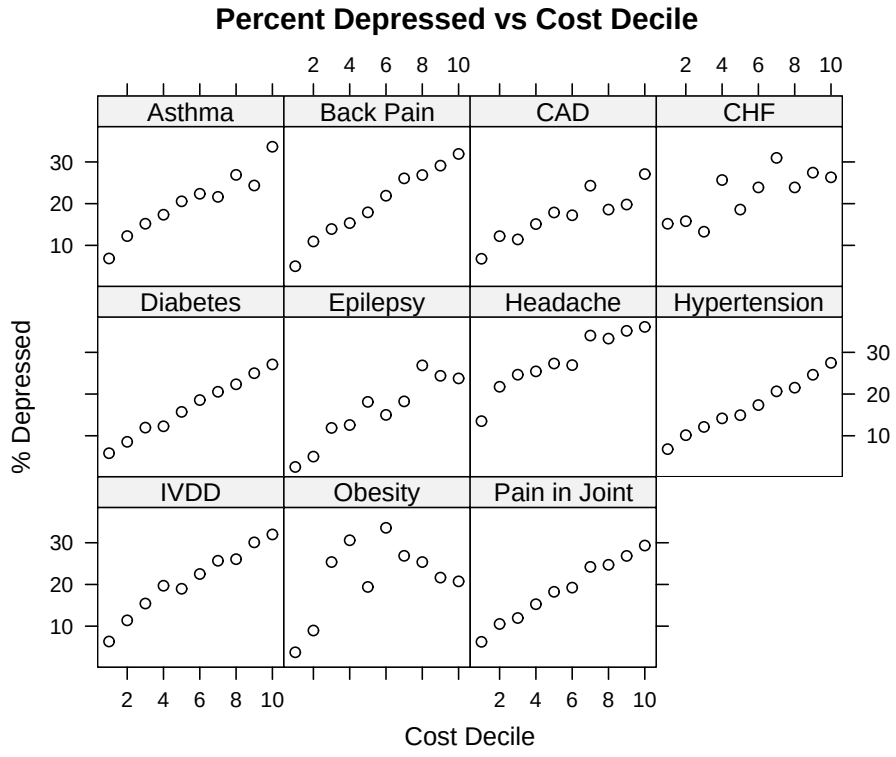


Figure 4-9: Prevalence of depression vs cost. Each disease has been broken into ten equal-sized cost strata. These are labeled along the horizontal axis. The vertical axis shows the percent of members in each stratum who are depressed.

	Percent in Not Depressed Cohort (%)	Percent in Depressed Cohort (%)	Ratio
Asthma	0.61	1.1	1.8
Back Pain	1.1	3.1	3.9
CHF	0.17	0.38	2.2
CAD	0.55	1.0	1.8
Diabetes	0.98	1.9	1.9
Epilepsy	0.14	0.31	2.2
Headache	0.74	2.4	3.3
Hypertension	1.6	3.2	2.0
IVDD	1.0	3.0	3.0
Obesity	0.10	0.35	3.4
Pain in Joint	1.7	4.2	2.4

Table 4.4: Members without disease in Year 1 who are diagnosed with disease in Year 2.

non-mental health cost. The exception is obesity, for which prevalence is biphasic.

Longitudinal measurement of incidence revealed that depressed individuals had a higher incidence of subsequent onset of all 11 chronic comorbid diseases than did not depressed individuals (Table 4.3). The odds ratio ranged from 1.8 in both asthma and CAD to 3.9 in back pain. Furthermore, individuals with any of the 11 chronic comorbid diseases had higher incidence of subsequent onset of depression than did individuals with none of the 11 chronic comorbid diseases (Table 4.3). The odds ratio ranged from 1.5 in asthma to 3.6 in headache.

4.4 Comment

The central finding of the study is that depression is associated with markedly greater cost of non-mental health care in all 11 chronic comorbid diseases studied. Even when controlled for number of chronic comorbid diseases, depressed patients had significantly higher costs than nondepressed patients. The magnitude of the cost difference is similar to that reported in prior studies, but the consistency of the magnitude across 11 chronic comorbid diseases is a finding not previously reported. These cost differentials are unlikely to be an artifact of methodology or data set,

	Percent in Cohort without Disease (%)	Percent in Disease Cohort (%)	Ratio
Asthma	2.7	4.2	1.5
Back Pain	2.7	6.8	2.6
CHF	2.7	6.6	2.4
CAD	2.7	5.0	1.9
Diabetes	2.7	4.6	1.7
Epilepsy	2.7	5.0	1.8
Headache	2.7	9.7	3.6
Hypertension	2.7	4.5	1.7
IVDD	2.7	7.3	2.8
Obesity	2.7	7.0	2.6
Pain in Joint	2.7	5.5	2.1

Table 4.5: Members without depression in Year 1 who are depressed in Year 2.

since similar cost differentials have been reported with survey-based methodology [43, 105, 28, 20, 30], and with different administrative data sets [102, 97].

In this study, the most important components of higher cost in depression were higher pharmaceutical and outpatient costs. Emergency room costs were not a significant factor in most chronic comorbid diseases, and inpatient costs were significant only in CAD and CHF. This pattern of increased utilization in depressed individuals was consistent with prior studies [20, 30]. Outpatient and pharmaceutical differences follow similar patterns when graphed as absolute dollar amounts, but follow very different patterns when expressed as ratios (Figure 4-5). Further research is necessary to determine the underlying reasons for these variations. However, these differences in utilization do indicate that depressed patients not only saw doctors more often, but also were prescribed non-mental health drugs at higher cost or in greater quantity than not depressed patients.

Only 14.1% of those patients taking antidepressants received a coded diagnosis for depression during the study interval. There are four possible reasons for this low rate of coding of depression. First, in the primary care setting, 36-65% of depressed patients are not recognized as depressed, and therefore are never coded as such [22, 96, 70]. Second, over 50% of primary care physicians intentionally assign alternative diagnoses to depressed patients due to diagnostic uncertainty or barriers to

reimbursement [84]. Third, an unknown number of depressed patients were treated by mental health professionals who bill a mental health management companies for their services, through administrative databases separate from the medical database. Antidepressants prescribed by these mental health professionals are recorded in the medical database, but services and diagnostic codes for depression care are not. Fourth, some patients may have been receiving antidepressant medication under a diagnosis other than depression, such as anxiety disorder.

The absolute prevalences of chronic comorbid diseases in both the depressed and not depressed cohorts were somewhat lower than reported in survey based studies, probably because of lower sensitivity of administrative data for diagnostic codes. However, the prevalence of chronic comorbid diseases in depressed individuals ranged from 1.93 to 4.28 times the prevalence in not depressed individuals. The principal importance of this finding is the relative prevalence of the 11 chronic comorbid diseases in depressed individuals. As has been previously reported [55], the largest differences in prevalence tended to occur in chronic comorbid diseases characterized by pain as the primary symptom, with the exception of obesity.

In this study there was an increased prevalence of depression associated with all 11 chronic comorbid diseases, relative to the prevalence in the not depressed cohort. The prevalence of depression in these chronic comorbid diseases was similar to that reported in previous studies [32, 4, 50, 85, 99, 5].

Prior research has shown an association between number of comorbid illnesses and increased cost [92]. In our data, number of comorbid illnesses had a linear relationship to annual cost, with the exception of patients with a diagnosis of obesity. However, in all chronic comorbid diseases, the cost differences associated with depression persisted even after controlled for number of comorbidities, a result which is consistent with the prior finding that higher annual costs associated with depression persisted even after controlled for severity of comorbid illness [97].

The longitudinal analysis of patterns of onset demonstrates that depression is associated with a greater than expected subsequent incidence of all chronic comorbid diseases in the 12-month period. Each chronic comorbid disease is also associated

with higher than expected subsequent incidence of depression. These data are consistent with the hypothesis that depression and chronic comorbid disease are reciprocally causative. However, in both cases, administrative data cannot ascertain that coding of a diagnosis occurs at first onset of that diagnosis. Therefore, replication of this analysis using either medical record data, or a longer study interval, would be necessary to more accurately quantify the increased risk of incidence in both directions.

While the causes of the association between depression and increased non-mental health cost are unclear, several possibilities should be considered.

The first consideration is whether it is an artifact of administrative data. In light of the findings regarding sensitivity of administrative data discussion in Section 1.1, it is likely that the methodology of this study created a bias towards identification of the more symptomatic patients in each chronic comorbid condition, while patients who were less symptomatic were identified at a lower frequency. However, it is unlikely that this selection bias would affect the main findings regarding cost, since it would apply equally to depressed and nondepressed cohorts. It would increase the median medical costs in both cohorts, but it would not affect the ratio between cost of depressed patients and cost of not depressed patients, nor would it affect the relative size of cost differences between the 11 chronic comorbid conditions.

The cohort designated as depressed was probably not selected with bias towards greater severity to the same degree as the cohorts of chronic comorbid diseases. The 12 month prevalence of depression in primary care populations has previously been reported as 5-10% [53]. Diagnostic coding for depression in this primary care database identified only 2.3% of the study population, but combining prescription and diagnostic codes raised the identification of depressed individuals to 11.3 % of the study cohort, a level consistent with that reported in prior survey-based studies. This methodology appears to have accurately identified a high percent of the depressed individuals in the study population. However, it may also have identified some patients who were receiving antidepressants for indications other than major depression.

In the cohort of patients on antidepressants, 1,977 (3.5% of this cohort) carried a diagnosis of anxiety or panic disorder in the absence of a diagnostic code for de-

pression. This subgroup was not removed from the cohort designated as depressed, because anxiety is a common presenting symptom of depression, and an unknown proportion of these patients on antidepressants were probably suffering from depressive pathophysiology. Whether this subgroup is included in or excluded from the designation of depressed, it is not large enough to affect the main findings of the study.

Low sensitivity for comorbid diagnoses is another potential source of artifact. Survey-based studies of individuals with depression report an average of approximately 3 comorbid conditions per patient [106, 103], but the average in this study is 0.8 chronic comorbid conditions per patient. Prior research has found that the sensitivity of administrative data is under 30% for second diagnoses, and even lower for third diagnoses [73]. Therefore, the low number of chronic comorbid conditions per patient in this study can be attributed to this artifact of administrative data. Nevertheless, this artifact has equal effect on the depressed and nondepressed cohorts and consequently is unlikely to affect relative cost of the two cohorts.

A second possible cause of greater cost associated with depression is patient behavior. Self-neglect is a documented behavior of depressed patients with comorbid disease. In diabetes and heart disease, depressed patients are less compliant with care than are non-depressed patients, and this behavior is correlated with higher utilization of emergency room, outpatient, inpatient, and specialty services [20, 27, 114, 16]. Depression is also associated with higher rates of harmful lifestyle factors such as smoking, overeating, and lack of physical activity [83, 39].

A third possible cause of greater cost is more severe pathophysiology of comorbid disease when it occurs in association with depression. There is a growing body of evidence that depression is associated not just with increased prevalence of comorbid disease but also with more severe pathophysiology of that disease. Compared to nondepressed cardiac patients, depressed cardiac patients have increased incidence and severity of ventricular arrhythmias, higher mortality and readmission rates [16, 36], decreased heart rate variability [67, 98], and increased platelet reactivity [68]. Compared to nondepressed diabetics, depressed diabetics have more complications of

diabetes, including retinopathy, neuropathy, nephropathy, and vascular disease, and have more severe glycemic dyscontrol [24].

Finally, this study raises the question of whether the overall cost of health care could be reduced by creating systems of care that treat depression more effectively. Prior studies have reported that the proportion of patients receiving antidepressant treatment in the primary care setting was between 60% and 80% [91, 71], and the fastest-growing segment of patients treated for depression in the United States are those treated in primary care settings [108]. Studies of patients treated for depression in the primary care setting have generally found that they are less likely to receive adequate doses and duration of antidepressant drugs, and have lower rates of response or remission, compared to patients receiving specialty mental health care [62, 58, 109, 56]. Because a large majority of depressed patients receive care exclusively from non-mental health professionals, there has been significant interest in the question of whether disease management programs of depression in the primary care setting have the ability to improve outcomes or lower total medical costs. In the last decade, a number of studies have compared "usual care" of these patients with disease management that integrates the functioning of mental health professionals with primary care [52, 54, 26, 9]. Features of these innovative approaches to the treatment of depression include diagnostic screening, physician education, patient education, availability of mental health consultants either on site or off-site, close monitoring of patients, adherence to best practices, and increased use of telephone both for consultation between clinicians and for direct patient management [72, 38].

Studies of disease management of depression in primary care have reported lower depression scores and higher response rates compared to patients receiving usual care in the primary care setting [92, 8]. However, the findings with regards to cost have been variable. The cost of treating depression in these innovative programs was consistently higher than in usual care [107, 60, 93, 94, 87]. Most studies did not find a reduction in non-mental health costs, but in those that did, the reduction was not enough to offset the higher mental health costs associated with disease management of depression [95]. Disease management programs which target depression unequivocally

reduce psychiatric morbidity, but it remains unclear to what extent they have an impact on severity and cost of chronic comorbid conditions. Consequently, there is growing concern that meaningful reduction of the total disease burden of depression will require fundamental restructuring of the healthcare system, in a way that more effectively integrates the treatment of mental diseases with the delivery of primary care [108, 75].

4.4.1 Limitations

Research using administrative data is vulnerable to problems of accuracy, as well as unforeseen flaws in internal, construct, or external validity. Extrapolation of the study results to different settings and populations should be done with caution. For instance, the findings of this study cannot be generalized to all patients, but apply only to patients who are diagnosed with the codes used here. Patients with alternative diagnostic codes may or may not demonstrate the cost effects and comorbidities found in this patient population.

The use in this study of all ICD-9 depression codes fails to discriminate between unipolar major depression and other types of depression. However, in daily practice, medical billing professionals frequently are unable to accurately make this discrimination, and administrative data therefore cannot be as diagnostically precise as clinical interviews of individual patients. Nevertheless, the more inclusive set of depression codes used in this study probably results in a study cohort of depressed patients who are symptomatically similar to that encountered in primary medical practice.

4.4.2 Conclusions

Depression was associated with significantly greater non-mental health cost in all comorbid diseases studied, and the increase cannot be explained solely on the basis of artifact or number of comorbidities. Greater non-mental health cost is driven mainly by greater pharmaceutical and outpatient utilization. Our findings provide evidence for reciprocal causation between depression and a wide range of comorbid diseases,

although they do not reveal mechanisms by which this may occur. Efforts to improve treatment of depression in the general medical setting have not yet yielded significant reduction in non-mental health costs of chronic comorbid diseases in patients suffering from depression.

Appendix A

Full List of Variables used in Modeling Quality

A.1 Diabetes Treatment

The following variables are based on the guidelines for the treatment of diabetes or otherwise related to diabetes care.

EyeExam The number of eye exams. Note that this variable is hampered by the fact that some people may have visits with eye doctors that are not covered by their insurance or are covered by a different insurance plan.

HemoglobinTest The number of glycated hemoglobin tests.

LipidProfile The number of lipid profiles.

GenericLab The number of times unspecified lab work was performed (generally at a hospital where the details of the lab work are not recorded as carefully as for outpatient lab work). Multiple tests performed on the same day are considered one occurrence.

AnyLab the number of times any of the following lab work was performed: glycated hemoglobin tests, lipid profiles, or unspecified lab work.

DiabetesLab This variable is similar to AnyLab with two differences: the diagnosis recorded with the lab work must be diabetes, and a broader range of lab work is included. The included lab work is: glycated hemoglobin tests, lipid profiles, unspecified lab work, hemoglobin tests, metabolic panels, urine microalbumin tests, and serum creatinine test.

GlucoseSupplies The number of times glucose testing supplies were ordered. Note that supplies are sometimes stored in the claims database as medical claims and sometimes as drug claims.

AceInhibitors The number of prescriptions for ace inhibitors that the patient had. Ace inhibitors are a class of drugs used to lower blood pressure and which can slow damage to kidneys in diabetes patients.

AceInhibitorDays The number of days for which the patient had ace inhibitors prescribed. This may be a more accurate measure than the number of prescriptions, since prescriptions can vary in length.

ARBs The number of prescriptions for angiotensin II Receptor blockers that the patient had.

A.2 Patient

The following variables are demographic information about the patient or about the patient's claims data.

Age the patient's age.

Female 1 if the patient is a female, 0 otherwise.

Diabetic 1 if the patient is diabetic. After reviewing each patient's claims data, in the physician's opinion a handful of patients in the study were not actually diabetics but were included due to spurious coding.

DrugsMissing An indicator variable which is 1 if the pharmacy claims for the patient were unavailable.

DiseaseCount The number of chronic diseases that the patient had.

Anxiolytics The number of prescriptions that the patient had for anxiolytics.

Antidepressants The number of prescriptions that the patient had for antidepressants.

Pain 1 if the patient had any coding for pain, 0 otherwise.

MedianMonthlyCost The patient's median monthly cost over the study period.

CostDerivative The slope of the patient's monthly costs over the study period. This was calculated for each patient by fitting a linear regression of monthly cost versus time (as an index of $1 \ldots 24$ for the 24 months in the study period) and taking the coefficient for time.

CostSecondDerivative The second derivative of the patient' s monthly costs over the study period. This was calculated for each patient by fitting a linear regression of monthly cost versus time, including a quadratic term, and taking the coefficient for the quadratic term.

A.3 Utilization

InpatientDays The number of days spent in the hospital.

ERVisits The number of visits to the emergency room.

OfficeVisits The number of visits coded as 99213 or 99214 and taking place in an office or outpatient hospital setting.

InpatientPerOffice The ratio of inpatient days to office visits.

ERPerOffice The ratio of emergency room visits to office visits.

TotalVisits The sum of the number of office visits, emergency room visits, and inpatient days.

ERVisits.normalized ERVisits divided by TotalVisits.

InpatientDays.normalized InpatientDays divided by TotalVisits.

OfficeVisits.normalized OfficeVisits divided by TotalVisits.

ER.outpatient When an emergency room visit occurs, the percent of time that the next visit is an outpatient visit.

ER.inpatient When an emergency room visit occurs, the percent of time that the next visit is an inpatient visit.

ER.ER When an emergency room visit occurs, the percent of time that the next visit is another emergency room visit.

ER.other When an emergency room visit occurs, the percent of time that the next visit is any other type of visit than outpatient, inpatient, or emergency .

DaysSinceLastERVisit The number of days between the patient's last emergency room visits and the end of the study period. For patients who didn't have any emergency room visits this was set equal to the length of the study period.

PhysicalTherapy The number of days on which the patient had physical therapy performed. When a patient has physical therapy it often lasts a large number of days and this may drive up some of the quantity of care measures.

Chiropractic the number of days in which the patient had chiropractic services performed. As with physical therapy, chiropractic is usually performed a large number of times and this may drive up some of the quantity of care measures.

A.4 Ratios

GenericLabsPerOffice GenericLab divided by OfficeVisits

CostDrugRatio The patient's median monthly cost divided by the average number of "chronic" drugs they were on.

InpatientDrugRatio The number of days the patients spent in the hospital divided by the average number of "chronic" drugs they were on.

DiseaseVisitsRatio The number of chronic diseases divided by the number of visits.

DiseaseRegularityRatio The number of chronic diseases divided by VisitRegularity.

A.5 Markers of good care

The following variables correspond to aspects of the patient's care that are considered to be markers of good care.

Mammogram The number of mammograms.

BinaryMammogram 1 if the patient had at least one mammogram, 0 otherwise.

VisitRegularity The entire span of the patient's claims history is broken up into three month intervals and VisitRegularity is the fraction of those intervals in which there is a claim.

OfficeVisitRegularity The same as visit.regularity except that only office visits are counted, not all claims.

LongestOfficeGap The longest gap, in days, between successive office visits.

A.6 Markers of poor care

Narcotics The number of prescriptions for narcotics that the patient had.

NarcoticsDays The number of days for which the patient had narcotics prescribed. This may be a more accurate measure than the number of prescriptions, since prescriptions can vary in length.

B12 The number of prescriptions or injections of vitamin B12 that the patient had. Over-the-counter use of vitamin B12 would not be included if the patient paid for out of their own pocket.

Polypharmacy An indicator variable which is 1 if the patient's pharmaceutical treatment for diabetes was initiated with a combination of drugs at once. By default, the indicator is set to 0 for patients who were already on diabetes drugs at the beginning of the study period.

A.7 Providers

ProviderCount The number of providers (i.e. doctors) that served the patient. Note that a hospital or a lab or a clinic can be counted as a provider.So what will anesthesiologists, pathologists, and so on.

PrescriberCount The number of doctors who prescribed drugs for the patient.

DiabetesProviders The number of providers who treated the patient's diabetes. We include all providers who had a claim for which diabetes was listed as the diagnosis.

PrescribersPerProvider The number of prescribers divided by the number of providers.

A.8 Claims

MedicalClaims The number of days on which the patient had a medical claim (i.e. all claims except prescriptions).

ClaimLines The number of medical claims that the patient had. This differs from MedicalClaims in that multiple claims on the same date are each counted.

ClaimsPerDate A patient may have multiple claims on any given date. This is the total number of claims a patient had divided by the number of dates on which the patient had claims. This variable is an attempt to get at the "complexity" of

a visit – presumably the more claims that occurred on a date the more complex the encounter.

A.9 Prescriptions

DrugsStarted The number of drugs started during the time period. (Any drug for which the first prescription occurs within the first 90 days of the study period is not included since it is likely that patient was already on the drug and is renewing their prescription.)

DrugsEnded The number of drugs stopped during the time period. (Any drug for which the last prescription occurs within the last 90 days of the study period is not included because the patient may have continued on the drug after the study period ended.)

DrugsAtBeginning The number of drugs the patient is on at the beginning of the time period. (Any drug for which the first prescription occurs within the first 90 days of the study period is included here.)

MaxDrugs The maximum number of drugs the patient is on at one time.

AverageDrugs The average number of drugs the patient is on at a time

UniqueDrugs The number of distinct drugs that the patient is one over the course of the study period.

DrugGapNone The fraction of refills which occurred immediately after the previous prescription ran out (i.e. there was no gap before the refill).

DrugGapSmall The fraction of refills which were preceded by a small gap (between 1 and 30 days) after the previous prescription ran out.

DrugGapMedium The fraction of refills which were preceded by a medium gap (between 31 and 90 days) after the previous prescription ran out.

DrugGapLarge The fraction of refills which were preceded by a large gap (more than 90 days) . These likely aren't refills at all but indicate that the patient went off of the medication for a while.

We included three versions of the prescription variables: one set that applies to all drugs, one that applies only to chronic drugs, and one that applies only to acute drugs. The versions applying to chronic drugs are prefixed "Chronic" and those applying to a cute drugs are prefixed "Acute."

Appendix B

The 21 single-period scenarios

	Control A	Treatment A	Control B	Treatment B
1	0.02	0.04	0.02	0.04
2	0.04	0.08	0.04	0.08
3	0.08	0.16	0.08	0.16
4	0.02	0.04	0.02	0.02
5	0.04	0.08	0.02	0.02
6	0.08	0.16	0.02	0.02
7	0.02	0.04	0.04	0.04
8	0.04	0.08	0.04	0.04
9	0.08	0.16	0.04	0.04
10	0.02	0.04	0.08	0.08
11	0.04	0.08	0.08	0.08
12	0.08	0.16	0.08	0.08
13	0.02	0.01	0.02	0.04
14	0.04	0.02	0.02	0.04
15	0.08	0.04	0.02	0.04
16	0.02	0.01	0.04	0.08
17	0.04	0.02	0.04	0.08
18	0.08	0.04	0.04	0.08
19	0.02	0.01	0.08	0.16
20	0.04	0.02	0.08	0.16
21	0.08	0.04	0.08	0.16

Table B.1: The rate of occurrence of each diagnosis in the 21 single-period scenarios.

Appendix C

ICD-9 Codes Used in Depression Study

Asthma	49300 extrinsic asthma 49310 intrinsic asthma 49390 asthma unspecified 49391 asthma unspecified with status asthmaticus 49392 asthma unspecified with acute exacerbation
Back Pain	7242 pain low back 7244 thoracic or lumbosacral neuritis or radiculitis unspecified
CHF	4280 congestive heart failure, unspecified
CAD	41400 coronary atherosclerosis of unspecified vessel 41401 coronary atherosclerosis of native coronary artery
Diabetes	25000 diabetes without mention of complication 25001 diabetes with ketoacidosis 25002 diabetes with hyperosmolality
Epilepsy	34510 tonic-clonic seizure 34590 epilepsy NOS 78039 non-febrile convulsions
Headache	34600 classic migraine 34601 classic migraine with intractable migraine 34610 common migraine 34611 common migraine with intractable migraine 34690 migraine unspecified 34691 migraine unspecified with intractable migraine 7840 headache
Hypertension	4010 malignant hypertension 4011 benign hypertension
Intervertebral Disc Disease (IVDD)	7220 displacement of cervical intervertebral disc 7221 displacement of thoracic or lumbar intervertebral disc 7224 degeneration of cervical intervertebral disc 72252 degeneration of lumbar intervertebral disc 7231 cervicalgia
Obesity	27800 obesity unspecified 27801 morbid obesity
Pain in Joint	71940 site unspecified 71941 shoulder 71942 upper arm 71943 forearm 71944 hand 71945 pelvic region and hip 71946 lower leg 71947 ankle and foot 71949 multiple sites

Table C.1: Diagnostic Codes.

ICD 9 Code	Description	Members
311	Depressive disorder, not elsewhere classified	3003
3004	Dysthymic disorder	1452
29632	Major depressive disorder, recurrent episode – moderate	1226
3090	Adjustment disorder with depressed mood	871
29633	Major depressive disorder, recurrent episode – severe, without mention of psychotic behavior	677
29630	Major depressive disorder, recurrent episode – unspecified	666
29622	Major depressive disorder, single episode – moderate	484
29620	Major depressive disorder, single episode – unspecified	422
29623	Major depressive disorder, single episode – severe, without mention of psychotic behavior	325
29621	Major depressive disorder, single episode – mild	132
2963	Major depressive disorder, recurrent episode	127
29631	Major depressive disorder, recurrent episode – mild	126
29634	Major depressive disorder, recurrent episode – severe, specified with psychotic behavior	101
29635	Major depressive disorder, recurrent episode – in partial or unspecified remission	70
2962	Major depressive disorder, single episode	54
29636	Major depressive disorder, recurrent episode – in full remission	51
29624	Major depressive disorder, single episode – severe, specified with psychotic behavior	27
3091	Prolonged depressive reaction	26
29625	Major depressive disorder, single episode – in partial or unspecified remission	26
29626	Major depressive disorder, single episode – in full remission	21
2980	Depressive type psychosis	2
	More Than One Code	4116

Table C.2: ICD-9 codes used to identify members diagnosed with depression.

Bibliography

[1] Tobias Achterberg. *Constraint Integer Programming*. PhD thesis, Technische Universität Berlin, 2007. `http://opus.kobv.de/tuberlin/volltexte/2007/1611/`.

[2] Agency for Healthcare Research and Quality. National healthcare quality report, 2006. `http://www.ahrq.gov/qual/nhqr06/nhqr06.htm`.

[3] American Diabetes Association. Diabetes statistics. `http://www.diabetes.org/diabetes-statistics.jsp`.

[4] RJ Anderson, KE Freedland, RE Clouse, and PJ Lustman. The prevalence of comorbid depression in adults with diabetes: A meta-analysis. *Diabetes Care*, 24:1069–1078, 2001.

[5] BA Arnow, EM Hunkeler, CM Blasey, et al. Comorbid depression, chronic pain, and disability in primary care. *Psychosom Med*, 68:262–268, 2006.

[6] Steven M. Asch, Elizabeth A. McGlynn, Mary M. Hogan, Rodney A. Hayward, Paul Shekelle, Lisa Rubenstein, Joan Keesey, John Adams, and Eve A. Kerr. Comparison of quality of care for patients in the veterans health administration and patients in a national sample. *Ann Intern Med*, 141:938–945, December 2004.

[7] American Diabetes Association. Standards of medical care in diabetes–2007. *Diabetes Care*, 30:S4–41, 2007.

[8] E Badamgarav, SR Weingarten, JM Henning, et al. Effectiveness of disease management programs in depression: A systematic review. *Am J Psychiatry*, 160:2080–2090, 2003.

[9] SJ Bartels, KM Miles, Van AD Citters, BP Forester, MJ Cohen, and H Xie. Improving mental health assessment and service planning practices for older adults: A controlled comparison study. *Ment Health Serv Res*, 7:213–223, 2005.

[10] Donald M. Berwick. A user's manual for the iom's 'quality chasm' report. *Health Aff*, 21:80–90, May 2002.

[11] HG Birnbaum, PY Cremieux, PE Greenberg, J LeLorier, JA Ostrander, and L Venditti. Using healthcare claims data for outcomes research and pharmacoeconomic analyses. *Pharmacoeconomics*, 16:1–8, 1999.

[12] EA McGlynn RH Brook and PD Cleary. Measuring quality of care - part 2. *The New England Journal of Medicine*, 335:966–970, September 1996.

[13] Robert H. Brook, Elizabeth A. McGlynn, and Paul G. Shekelle. Defining and measuring quality of care: a perspective from us researchers. *Int J Qual Health Care*, 12:281–295, August 2000.

[14] Jeffrey S. Brown, Martin Kulldorff, K. Arnold Chan, Robert L. Davis, David Graham, Parker T. Pettus, Susan E. Andrade, Marsha A. Raebel, Lisa Herrinton, Douglas Roblin, Denise Boudreau, David Smith, Jerry H. Gurwitz, Margaret J. Gunter, and Richard Platt. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiology and Drug Safety*, 16(12):1275–1284, 2007.

[15] MF Bullano, S Kamat, VJ Willey, S Barlas, DJ Watson, and SK Brenneman. Agreement between administrative claims and the medical record in identifying patients with a diagnosis of hypertension. *Med Care*, 44:486–490, 2006.

[16] RM Carney, KE Freedland, GE Miller, and AS Jaffe. Depression as a risk factor for cardiac mortality and morbidity: A review of potential mechanisms. *J Psychosom Res*, 53:897–902, 2002.

[17] Centers for Disease Control and Prevention. *National diabetes fact sheet: general information and national estimates on diabetes in the United States, 2005.* U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA, 2005.

[18] ME Charlson, P Pompei, KL Ales, and CR MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J Chronic Dis*, 40:373–383, 1987.

[19] DB Christensen, B Williams, HI Goldberg, DP Martin, R Engelberg, and JP LoGerfo. Comparison of prescription and medical records in reflecting patient antihypertensive drug therapy. *Ann Pharmacother*, 28:99–104, 1994.

[20] PS Ciechanowski, WJ Katon, and JE Russo. Depression and diabetes: Impact of depressive symptoms on adherence, function, and costs. *Arch Intern Med*, 160:3278–3285, 2000.

[21] JM Corrigan. *Crossing the Quality Chasm.* Washington, DC, National Academy Press, 2001.

[22] JC Coyne, TL Schwenk, and S Fechner-Bates. Nondetection of depression by primary care physicians reconsidered. *Gen Hosp Psychiatry*, 17:3–12, 1995.

[23] K. Davis, C. Schoen, S. C. Schoenbaum, M. M. Doty, A. L. Holmgren, J. L. Kriss, and K. K. Shea. Mirror, mirror on the wall: An international update on the comparative performance of american health care, May 2007.

[24] M de Groot, R Anderson, KE Freedland, RE Clouse, and PJ Lustman. Association of depression and diabetes complications: A meta-analysis. *Psychosom Med*, 63:619–630, 2001.

[25] RA Deyo, DC Cherkin, and MA Ciol. Adapting a clinical comorbidity index for use with icd-9-cm administrative databases. *J Clin Epidemiol*, 45:613–619, 1992.

[26] AJ Dietrich, TE Oxman, JW Williams, Jr, et al. Re-engineering systems for the treatment of depression in primary care: Cluster randomised controlled trial. *BMJ*, 329:602, 2004.

[27] MR DiMatteo, HS Lepper, and TW Croghan. Depression is a risk factor for noncompliance with medical treatment: Meta-analysis of the effects of anxiety and depression on patient adherence. *Arch Intern Med*, 160:2101–2107, 2000.

[28] BG Druss, RM Rohrbaugh, and RA Rosenheck. Depressive symptoms and health costs in older medical patients. *Am J Psychiatry*, 156:477–479, 1999.

[29] XL Du, CR Key, L Dickie, R Darling, JM Geraci, and D Zhang. External validation of medicare claims for breast cancer chemotherapy compared with medical chart reviews. *Med Care*, 44:124–131, 2006.

[30] LE Egede, D Zheng, and K Simpson. Comorbid depression is associated with increased health care use and expenditures in individuals with diabetes. *Diabetes Care*, 25:464–470, 2002.

[31] A Elixhauser, C Steiner, DR Harris, and RM Coffey. Comorbidity measures for use with administrative data. *Med Care*, 36:8–27, 1998.

[32] DL Evans, DS Charney, L Lewis, et al. Mood disorders in the medically ill: Scientific review and recommendations. *Biol Psychiatry*, 58:175–189, 2005.

[33] ES Fisher, FS Whaley, WM Krushat, et al. The accuracy of medicare's hospital claims data: Progress has been made, but problems remain. *Am J Public Health*, 82:243–248, 1992.

[34] JL Fless, A Tytun, and HK Ury. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, 36(2):343–346, 1980.

[35] JB Fowles, EJ Fowler, and C Craft. Validation of claims diagnoses and self-reported conditions compared with medical records for selected chronic diseases. *J Ambulatory Care Manage*, 21:24–34, 1998.

[36] N Frasure-Smith, F Lesperance, and M Talajic. Depression and 18-month prognosis after myocardial infarction. *Circulation*, 91:999–1005, 1995.

[37] Tejal K Gandhi, E. Francis Cook, Ann Louise Puopolo, Helen R Burstin, Jennifer S Haas, and Troyen A Brennan. Inconsistent report cards: assessing the comparability of various measures of the quality of ambulatory care. *Medical Care*, 40:155–165, February 2002.

[38] J Gensichen, M Beyer, C Muth, FM Gerlach, M Von Korff, and J Ormel. Case management to improve major depression in primary health care: A systematic review. *Psychol Med*, 36:7–14, 2006.

[39] E Goodman and RC Whitaker. A prospective study of the role of depression in the development and persistence of adolescent obesity. *Pediatrics*, 110:497–504, 2002.

[40] David J Graham, David Campen, Rita Hui, Michele Spence, Craig Cheetham, Gerald Levy, Stanford Shoor, and Wayne 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. *The Lancet*, 365:475–481, 2005.

[41] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[42] RA Hayward. Performance measurement in search of a path. *New England Journal of Medicine*, 356:951, 2007.

[43] HJ Henk, DJ Katzelnick, KA Kobak, JH Greist, and JW Jefferson. mdical costs attributed to depression among patients with a history of high medical expenses in a health maintenance organization. *Arch Gen Psychiatry*, 53:899–904, 1996.

[44] Takahiro Higashi, Neil S. Wenger, John L. Adams, Constance Fung, Martin Roland, Elizabeth A. McGlynn, David Reeves, Steven M. Asch, Eve A. Kerr, and Paul G. Shekelle. Relationship between number of medical conditions and quality of care. *N Engl J Med*, 356:2496–2504, June 2007.

[45] Timothy P. Hofer, Rodney A. Hayward, Sheldon Greenfield, Edward H. Wagner, Sherrie H. Kaplan, and Willard G. Manning. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA*, 281:2098–2105, June 1999.

[46] KH Humphries, JM Rankin, RG Carere, CE Buller, FM Kiely, and JJ Spinelli. Co-morbidity data in outcomes research: Are clinical data derived from administrative databases a reliable alternative to chart review? *J Clin Epidemiol*, 53:343–349, 2000.

[47] LI Iezzoni. Assessing quality using administrative data. *Ann Intern Med*, 127:666–674, 1997.

[48] LI Iezzoni, SM Foley, J Daley, J Hughes, ES Fisher, and T Heeren. Comorbidities, complications, and coding bias. does the number of diagnosis codes matter in predicting in-hospital mortality? *JAMA*, 267:2197–2203, 1992.

[49] JG Jollis, M Ancukiewicz, ER DeLong, DB Pryor, LH Muhlbaier, and DB Mark. Discordance of databases designed for claims payment versus clinical information systems. implications for outcomes research. *Ann Intern Med*, 119:844–850, 1993.

[50] AM Kanner. Depression in epilepsy: Prevalence, clinical semiology, pathogenic mechanisms, and treatment. *Biol Psychiatry*, 54:388–398, 2003.

[51] T. Michael Kashner. Agreement between administrative files and written medical records. a case of the department of veterans affairs. *Medical Care*, 36:1324–1336, 1998.

[52] W Katon, P Robinson, M Von Korff, et al. A multifaceted intervention to improve treatment of depression in primary care. *Arch Gen Psychiatry*, 53:924–932, 1996.

[53] W Katon and H Schulberg. Epidemiology of depression in primary care. *Gen Hosp Psychiatry*, 14:237–247, 1992.

[54] W Katon, M Von Korff, E Lin, et al. Stepped collaborative care for primary care patients with persistent symptoms of depression: A randomized trial. *Arch Gen Psychiatry*, 56:1109–1115, 1999.

[55] WJ Katon. Clinical and health services relationships between major depression, depressive symptoms, and general medical illness. *Biol Psychiatry*, 54:216–226, 2003.

[56] WJ Katon, G Simon, J Russo, et al. Quality of depression care in a population-based sample of patients with diabetes and major depression. *Med Care*, 42:1222–1229, 2004.

[57] EA Kerr, SL Krein, S Vijan, TP Hofer, and RA Hayward. Avoiding pitfalls in chronic disease quality measurement: a case for the next generation of technical quality measures. *The American journal of managed care*, 7:1033–43, November 2001.

[58] KA Kobak, L Taylor, DJ Katzelnick, N Olson, P Clagnaz, and HJ Henk. Antidepressant medication management and health plan employer data information set (hedis) criteria: Reasons for nonadherence. *J Clin Psychiatry*, 63:727–732, 2002.

[59] Thorsten Koch. *Rapid Mathematical Programming*. PhD thesis, Technische Universität Berlin, 2004. ZIB-Report 04-58.

[60] JR Lave, RG Frank, HC Schulberg, and MS Kamlet. Cost-effectiveness of treatments for major depression in primary care practice. *Arch Gen Psychiatry*, 55:645–651, 1998.

[61] ST Leatherman and D McCarthy. *Quality of Health Care in the United States: A Chartbook*. Commonwealth Fund, 2002.

[62] EH Lin, WJ Katon, GE Simon, et al. Low-intensity treatment of depression in primary care: Is it problematic? *Gen Hosp Psychiatry*, 22:78–83, 2000.

[63] AD Lopez, CD Mathers, M Ezzati, DT Jamison, and CJ Murray. Global and regional burden of disease and risk factors, 2001: Systematic analysis of population health data. *Lancet*, 367:1747–1757, 2006.

[64] CH MacLean, R Louie, PG Shekelle, et al. Comparison of administrative data and medical records to measure the quality of medical care provided to vulnerable older patients. *Med Care*, 44:141–148, 2006.

[65] BR Motheral and KA Fairman. The use of claims databases for outcomes research: Rationale, challenges, and strategies. *Clin Ther*, 19:346–366, 1997.

[66] CJ Murray and AD Lopez. Alternative projections of mortality and disability by cause 1990-2020: Global burden of disease study. *Lancet*, 349:1498–1504, 1997.

[67] DL Musselman, DL Evans, and CB Nemeroff. The relationship of depression to cardiovascular disease: Epidemiology, biology, and treatment. *Arch Gen Psychiatry*, 55:580–592, 1998.

[68] DL Musselman and CB Nemeroff. Depression and endocrine disorders: Focus on the thyroid and adrenal system. *Br J Psychiatry Suppl*, pages 123–128, 1996.

[69] KM Newton, EH Wagner, SD Ramsey, et al. The use of automated data to identify complications and comorbidities of diabetes: A validation study. *J Clin Epidemiol.*, 52:199–207, 1999.

[70] GS Norquist and DA Regier. The epidemiology of psychiatric disorders and the de facto mental health care system. *Annu Rev Med*, 47:473–479, 1996.

[71] M Olfson and GL Klerman. Trends in the prescription of psychotropic medications. the role of physician specialty. *Med Care*, 31:559–564, 1993.

[72] TE Oxman, AJ Dietrich, and HC Schulberg. Evidence-based models of integrated management of depression in primary care. *Psychiatr Clin North Am*, 28:1061–1077, 2005.

[73] JW Peabody, J Luck, S Jain, D Bertenthal, and P Glassman. Assessing the accuracy of administrative data in health information systems. *Med Care*, 42:1066–1072, 2004.

[74] Hoangmai H. Pham, Deborah Schrag, Ann S. O'Malley, Beny Wu, and Peter B. Bach. Care patterns in medicare and their implications for pay for performance. *N Engl J Med*, 356:1130–1139, March 2007.

[75] HA Pincus. Depression and primary care: Drowning in the mainstream or left on the banks? *J Manage Care Pharm*, 12:3–9, 2006.

[76] H Powell, LL Lim, and RF Heller. Accuracy of administrative data to assess comorbidity in patients with heart disease. an australian perspective. *J Clin Epidemiol*, 54:687–693, 2001.

[77] H Quan, GA Parsons, and WA Ghali. Validity of procedure codes in international classification of diseases, 9th revision, clinical modification administrative data. *Med Care*, 42:801–809, 2004.

[78] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[79] John Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, 2nd ed edition, 1995.

[80] Patrick S Romano and Ryan Mutter. The evolving science of quality measurement for hospitals: implications for studies of competition and consolidation. *Int J Health Care Finance Econ*, 4:131–157, June 2004.

[81] PS Romano, LL Roos, and JG Jollis. Adapting a clinical comorbidity index for use with icd-9-cm administrative data: Differing perspectives. *J Clin Epidemiol*, 46:1075–1079, 1993.

[82] PS Romano, LL Roos, HS Luft, JG Jollis, and K Doliszny. A comparison of administrative versus clinical data: Coronary artery bypass surgery as an example. *J Clin Epidemiol*, 47:249–260, 1994.

[83] MC Rosal, JK Ockene, Y Ma, et al. Behavioral risk factors among members of a health maintenance organization. *Prev Med*, 33:586–594, 2001.

[84] K Rost, R Smith, DB Matthews, and B Guise. The deliberate misdiagnosis of major depression in primary care. *Arch Fam Med*, 3:333–337, 1994.

[85] B Rudisch and CB Nemeroff. Epidemiology of comorbid coronary artery disease and depression. *Biol Psychiatry*, 54:227–240, 2003.

[86] S Schneeweiss and J Avorn. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*, 58:323–337, 2005.

[87] M Schoenbaum, J Unutzer, C Sherbourne, et al. Cost-effectiveness of practice-initiated quality improvement for depression: Results of a randomized controlled trial. *JAMA*, 286:1325–1330, 2001.

[88] Stephen Schoenbaum, Douglas McCarthy, and Cathy Schoen. The agency for healthcare research and quality's 2006 national healthcare quality report, March 2007.

[89] Mark A Schuster, Elizabeth A McGlynn, and Robert H Brook. How good is the quality of health care in the united states? *Milbank Q*, 76:517–563, 1998.

[90] JP Shaffer. Multiple hypothesis testing. *Annu. Rev. Psychol.*, 46:561–584, 1995.

[91] S Shapiro, EA Skinner, LG Kessler, et al. Utilization of health and mental health services. three epidemiologic catchment area sites. *Arch Gen Psychiatry*, 41:971–978, 1984.

[92] GE Simon. Social and economic burden of mood disorders. *Biol Psychiatry*, 54:208–215, 2003.

[93] GE Simon, WJ Katon, M Von Korff, et al. Cost-effectiveness of a collaborative care program for primary care patients with persistent depression. *Am J Psychiatry*, 158:1638–1644, 2001.

[94] GE Simon, WG Manning, DJ Katzelnick, SD Pearson, HJ Henk, and CS Helstad. Cost-effectiveness of systematic depression treatment for high utilizers of general medical care. *Arch Gen Psychiatry*, 58:181–187, 2001.

[95] GE Simon, D Revicki, J Heiligenstein, et al. Recovery from depression, work productivity, and health care costs among primary care patients. *Gen Hosp Psychiatry*, 22:153–162, 2000.

[96] GE Simon and M Von Korff. Recognition, management, and outcomes of depression in primary care. *Arch Fam Med*, 4:99–105, 1995.

[97] GE Simon, M Von Korff, and W Barlow. Health care costs of primary care patients with recognized depression. *Arch Gen Psychiatry*, 52:850–856, 1995.

[98] PK Stein, RM Carney, KE Freedland, et al. Severe depression is associated with markedly reduced heart rate variability in patients with stable coronary heart disease. *J Psychosom Res*, 48:493–500, 2000.

[99] AJ Stunkard, MS Faith, and KC Allison. Depression and obesity. *Biol Psychiatry*, 54:330–337, 2003.

[100] M Sullivan, G Simon, J Spertus, and J Russo. . depression-related costs in heart failure care. *Arch Intern Med*, 162:1860–1866, 2002.

[101] R Tamblyn, G Lavoie, L Petrella, and J Monette. The use of prescription claims databases in pharmacoepidemiological research: The accuracy and comprehensiveness of the prescription claims database in quebec. *J Clin Epidemiol*, 48:999–1009, 1995.

[102] MR Thomas, JA Waxmonsky, PA Gabow, G Flanders-McGinnis, R Socherman, and K Rost. Prevalence of psychiatric disorders and costs of care among adult enrollees in a medicaid hmo. *Psychiatr Serv*, 56:1394–1401, 2005.

[103] MH Trivedi, AJ Rush, SR Wisniewski, AA Nierenberg, D Warden, L Ritz, G Norquist, RH Howland, B Lebowitz, PJ McGrath, K Shores-Wilson, MM Biggs, GK Balasubramani, M Fava, and STAR*D Study Team. Evaluation of outcomes with citalopram for depression using measurement-based care in star*d: Implications for clinical practice. *Am J Psychiatry*, 163:28–40, 2006.

[104] United States Department Of Health & Human Services. Hospital compare. `http://www.hospitalcompare.hhs.gov/`.

[105] J Unutzer, DL Patrick, G Simon, et al. Depressive symptoms and the cost of health services in hmo patients aged 65 years and older. a 4-year prospective study. *JAMA*, 277:1618–1623, 1997.

[106] M Valenstein, T Ritsema, L Green, et al. Targeting quality improvement activities for depression. implications of using administrative data. *J Fam Pract*, 49:721–728, 2000.

[107] M Von Korff, W Katon, T Bush, et al. Treatment costs, cost offset, and cost-effectiveness of collaborative management of depression. *Psychosom Med*, 60:143–149, 1998.

[108] PS Wang, O Demler, M Olfson, HA Pincus, KB Wells, and RC Kessler. Changing profiles of service sectors used for mental health care in the united states. *Am J Psychiatry*, 163:1187–1198, 2006.

[109] JB Weilburg, KM O'Leary, JB Meigs, J Hennen, and RS Stafford. Evaluation of the adequacy of outpatient antidepressant treatment. *Psychiatr Serv*, 54:1233–1239, 2003.

[110] JP Weiner, ST Parente, DW Garnick, J Fowles, AG Lawthers, and RH Palmer. Variation in office-based quality. a claims-based profile of care provided to medicare patients with diabetes. *JAMA*, 273:1503–1508, May 1995.

[111] JE 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*, 257:933–936, February 1987.

[112] Roland Wunderling. *Paralleler und objektorientierter Simplex-Algorithmus*. PhD thesis, Technische Universität Berlin, 1996. `http://www.zib.de/Publications/abstracts/TR-96-09/`.

[113] S Yasmeen, PS Romano, ME Schembri, JM Keyzer, and WM Gilbert. Accuracy of obstetric diagnoses and procedures in hospital discharge data. *Am J Obstet Gynecol*, 194:992–1001, 2006.

[114] RC Ziegelstein, JA Fauerbach, SS Stevens, J Romanelli, DP Richter, and DE Bush. Patients with depression are less likely to follow recommendations to reduce cardiac risk during recovery from a myocardial infarction. *Arch Intern Med*, 160:1818–1823, 2000.