BAYESIAN PREDICTIVE MODELS IN DETERMINING HEALTH BURDENS IN EMPLOYED POPULATIONS

A Dissertation
Presented to
The Academic Faculty

By

Leanne N. Metcalfe

In Partial Fulfillment
Of the Requirements for the Degree
Of Doctor of Philosophy in the
School of Bioengineering

Georgia Institute of Technology
December 2008

BAYESIAN PREDICTIVE MODELS IN DETERMINING HEALTH BURDENS IN EMPLOYED POPULATIONS

Approved by:

Dr. Brani Vidakovic, Advisor
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Stephen Sprigle
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Paul Griffin
School of Industrial Engineering
Georgia Institute of Technology

Dr. Arun Villivalam
Associate Medical Director, Hybrid Health and Primary Care
Chief Medical Information Officer
Comprehensive Health Services

Dr. Charlie Kemp
School of Biomedical Engineering
Georgia Institute of Technology

Date Approved: August 1, 2008

ACKNOWLEDGEMENTS

I would first like to thank my family for all their support through the years. To my parents, big brother Lee and sister Leisha, thank you for your patience. Thanks to my Uncle Anthony Lee Hing for providing me with a place to escape to in Atlanta. Thank you to my cousin Dr. Stacey DaCosta-Byfield for showing me the light at the end of the tunnel, and thanks to Grandma Metty and all my aunts, uncles, cousins and niece for providing the encouragement and comfort only family can give.

I would also like to thank my thesis committee Drs Brani Vidakovic, Paul Griffin, Sprigle, Charlie Kemp and Arun Villivalam who were able to provide excellent academic and professional guidance.

This research would not have been possible without the vision of Mr. Mel Hall and Mr. Ned Cooper of CHS, thank you, thank you, thank you. Thank you also to the rest of the team, Mr. Christopher Cooper, Mr. Steve Hinson, Ms. Trish Hook and Ms. Paula Bickford, who made sure that anything I needed was readily available and provided invaluable information every step of the way.

Thanks also to Dr. François Sainfort and Dr. Julie Jacko for their guidance through the years. Of course thank you to the HSI crew, David Huang, Sofia Espinoza, Brad Jones, Kevin Malony, Ji Soo Yi, Young Sang Choi, Erin Kinzel and Mahima Ashok for the added support. Thanks also to Akil Sutton, Sekou Remy, Michael Kettner, Yasmina Harmidy, Teresa Benincasa, Jennifer Chung, Chadrick LimSang and Dr. Wayne Johnson, when the going got tough, you pushed me further.

Most importantly I would like to thank God since there were many times I wanted to quit and needed strength to continue.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Absenteeism	Employees absent from work due to illness
BRFSS	Behavioral Risk Factor Surveillance System
CDH	Chronic Daily Headache
DE	Double Exponential
GDP	Gross Domestic Product
HPM	Health and Productivity Management
HRA	Health Risk Assessment
NCQA	National Committee for Quality Assurance
NHIS	National Health Interview Survey
Presenteeism	Condition where employees are at work but have affected productivity due to illness
QDA	Quality Divided Calculator
ROC	Receiver Operator Characteristic
STD	Short-Term Disability
UMHMRC	University of Michigan Health Management Research Center
WLQ	Work Limitations Questionnaire

SUMMARY

There has been an almost 60 percent increase in health care expenditures in the US in the past seven years. Employer-sponsored health coverage premiums have increased significantly (87 percent) in this same period. Besides the cost of care for chronic conditions such as migraine, arthritis and diabetes, absenteeism linked to these diseases also adds financial strain. Current health financial models focus on past spending instead of modeling based on current health burdens and future trends. This approach leads to suboptimal health maintenance and cost management.

Identifying the diseases which affect the most employees and are also the most costly (in terms of productivity, work-loss-days, treatment etc) is necessary, since this allows the employer to identify which combination of policies may best address the health burdens. The current predictive health model limits the amount of diseases it models since it ignores incomplete data sets. This research investigated if by using Bayesian methodology it will be possible to create a comprehensive predictive model of the health burdens being faced by corporations, allowing for health decision makers to have comprehensive information when choosing policies.

The first specific aim was to identify which diseases were the most costly to employers both directly and indirectly, and the pathogenesis of these diseases. Co-morbidity of diseases was also taken into account as in many cases these diseases are not treated independently. This information was taken into account when designing the models as the inference was disease specific.

One of the contributions of this thesis is coherent incorporation of prior information into the proposed expert model. The Bayesian models were able to estimate

the predicted disease burdens for corporations, including predicting the percentage of individuals with multiple diseases. The model was also comparable to, or better than current estimators on the market with limited input. The outputs of the model were also able to give further insight into the disease interactions which creates an avenue for further research in disease management.

CHAPTER 1 SPECIFIC AIMS

1.1 Problem Definition

Healthcare costs have risen steadily over the past few years. Data from the National Health Statistics Group shows that in 1997 13.6% of the Gross Domestic Product (GDP) was spent on healthcare. This number increased to 16.3% in 2007 and is expected to reach nearly 20% by 2017 [1]. In 2007 total national health expenditures were expected to rise 6.9%, twice the rate of inflation [2]. Business leaders also know that health care expenditures can negatively affect their organization's competitiveness and profitability. Corporations have been particularly affected by rising health care costs; in the past seven years, employer-sponsored health coverage premiums have increased by 87% [3]. Between spring of 2006 and spring of 2007 employers saw a healthcare premium increase of 6.1%. While this is lower than the 7.7% seen the year before, it continues to be higher than the growth in workers' earnings (3.7%) and inflation (2.6%) [2].

A company needs to have a highly productive and healthy workforce in order to stay competitive in the marketplace. Employers face additional costs in productivity due to workers being absent or impaired due to illness. Currently, workers average about four lost workdays per year due to illness, with absenteeism ranges from eight to twelve days for workers with migraine, arthritis, diabetes, or chronic obstructive pulmonary disease. Workers diagnosed with diabetes are also three times for likely to be limited in the work they do than workers without the disease [4].

There are ethical issues which are faced when trying to create new health policies. It is too important of an issue to remove human decision makers and simply rely on cost calculations. This does not mean that models cannot improve the decision making

process. Instead information on the health burdens of the company in addition to the financial modeling can aid in the policy making process [5].

Currently business leaders assess treatment and mitigation of health consequences rather than risk prevention and reductions. It has been shown, however, that health promotion and disease prevention programs can improve employee health, lower health risks and increase employee moral and day-to-day functioning, while effectively managing the rising costs of health care. Well designed and well-implemented programs have the potential to save money and even produce a positive return on investment [6]. Failure to define total health and safety costs in financial planning could have significant consequences for employers [7]. Analysis can be done to assess the health status and risk levels of employees, leading to more informed cost studies which would better show how resources are to be allocated [8].

The research uses Bayesian modeling techniques to predict the health burdens a company might experience for their employees and dependents over the age of 18. The model also predicts health risk factors which the employees of the company may be exhibiting in order to identify where resources on disease prevention should be concentrated.

1.2 Problem Statement

Clinicians, policymakers, employers and insurers are intently focused on the cost of healthcare [9, 10]. Health conditions increase work related absences and reduce workplace activity, creating a substantial economic burden for industry [11], which may extend to the society as a whole. The workplace is one of the most important settings

affecting the physical, mental, economic and social well-being of workers, and in turn the health of their families, communities and society [10]. For this reason it is important to focus on the improvement of predicting company health burdens. Proper modeling would lead to the creation of better health policies, improving the quality of life of the employees and by extension their communities.

Another issue is most of the diseases in this study do not exist in isolation and can have higher costs depending on the associated diseases and/or risk factors present. Many cost models assess the cost of each disease individually, without also considering the additional cost of secondary diseases and/or risk factors. Diet and lifestyle factors have been analyzed separately when considering their effect on disease, although behavioral factors are typically correlated with each other [12]. This research will use realistic Bayesian models, which are implemented by Markov Chain Monte Carlo (MCMC) methods in order to improve upon current techniques and allow for the investigation of disease combinations.

The diseases included in the model will be based on research identifying the most costly diseases for employers, based on extensive literature review. Multiple disease databases will be used in order to determine the risk factors associated with each of the diseases chosen, as well as test and train the model. The final model will be tested on actual company data.

1.3 Research Questions

1. Is it possible to create a model that will estimate the disease burden for populations not widely studied or not large enough to have enough data to analyze using traditional methods?

2. Is it possible to model co-morbidities which will allow for greater policy making for the treatment or management of diseases?

3. Will Bayesian methods improve prediction of health burdens?

CHAPTER 2 BACKGROUND AND SIGNIFICANCE

In this section we shall highlight current models in an effort to discuss the contributions of the developed model.

2.1 Review of Other Health Burden Models

Current health models tend to follow two formats

- HRA models: Screen the population for risks and then apply costs based on the information
- Disease Models: Predict the levels of disease based on available data.

The University of Michigan's Health Management Research Centre (UMHMRC) has designed a Medical Economics (MedEc) system to assess the risk levels of employees in a company [13]. Their model is based on research results they had obtained showing that there was a relationship between risk levels and health costs [14]. Their study involved employees from three major corporations completing a full Health Risk Assessment (HRA) form. This questionnaire requests data from individuals which gives a complete look at their current health status including the conditions they currently have. The HRA also asks about risk factors including smoking and drinking habits, along with how they feel about their job and so forth. The model then assigns a score based on their responses. A high wellness score indicated the employee was a low risk individual; if the employee had a low wellness score, they were considered high risk. Using this information they have developed a model which uses the results of a given HRA survey and groups employees into risk categories, and calculate the costs the employer is likely to face based on these calculations. The data for this model has been collected for over two decades and is used to either look directly at what a company in the database is

experiencing, or used to estimate the disease burden for a company in the same state and industry. Due to this wealth of information, the model does a good job at predicting the expected health burden. It uses either the HRA's obtained directly from the company, or HRA responses of the company it most closely resembles. This model may be a little intrusive to the company, and requires a large scale employee participation which could incur large costs due to man-power needed to collect the data, downtime of the employees submitting the questionnaire and employee participation incentives. One benefit of looking at the UMHMRC model is that it may be used in place of claims data in order to get an overview of the disease burden within a company.

D2Hawkeye has created their own modeling technique to assess the health needs of corporations. The model integrates information on the eligibility, medical claims, pharmacy claims, health risk assessment and case management data of the employees and processes the data according to a pre-defined set of rules [15]. Individuals are evaluated according to their risk information and assigned a risk index value based on their current disease status, as well as other risk factors. This risk index is used to associate an individual with cost buckets, and estimate their cost to the organization. D2Hawkeye faces the same limitations as UMHMRC as it also relies on information obtained from the company, as such, a corporation would not be able to get a quick overview of their health burdens. Both of these models also do not include dependent information as it would be difficult to obtain complete HRA forms from the dependents in addition to the employees. While dependants may not be eligible for onsite employee health management programs, knowledge of what the main health issues are for dependents would aid in the decision on which family plans to offer.

The Quality Dividend Calculator (QDC) proposed by the National Committee on Quality Assurance (NCQA) first estimates the disease levels a company faces, and then applies presenteeism and absenteeism costs to the estimated disease burden [16]. This system currently models alcohol abuse, asthma, hypertension, heart disease, chicken pox, depression, diabetes and smoking. This system uses data pulled from CDC's Behavioral Risk Factor Surveillance System (BRFSS) and the National Health Interview Survey (NHIS), and applies those prevalence rates directly to the number of employees in the company. They are able to make estimations on the genders of individuals in the company based on US Census bureau census information, but make no estimations on race. The limited amount of information on each modeled disease at the state level only allows the NCQA calculator to make estimates based on national or regional data. Limited data also means they do not model all of the top disease conditions.

The new model we developed improves upon the current models as it uses publicly available data, either through databases, or expert analysis in order to estimate expected health burdens, removing the need to pay to access databases, or conduct company wide health interviews.

2.2 Health Burdens for Corporations

The diseases modeled in this study were chosen based on how prevalent they were in employed populations, and how expensive they were to employers. This section will give a background on how the diseases were chosen, and discuss the techniques available to create the model.

As discussed by Riedel et al [17], most employers would prefer to institute a health plan which:

- Leads to faster recovery (and hence, reduced medical costs)
- Improves the quality of life
- Increases at work productivity
- Decreases absenteeism (days absent from work)
- Decreases presenteeism (impairment at work)

It has been argued that in order to achieve these goals while managing costs, one blanket healthcare policy may not be sufficient. Instead, the cost burden of certain health conditions should also consider the impact of those conditions on the individuals' productivity at work, in addition to the health and cost consequences [18]. Identifying which disease conditions affect the most employees and are the most costly (in terms of productivity, work-loss-days, treatment etc) may allow the employer to identify what areas need to be addressed, and which combination of policies may best address the health issues.

Analysis of medical costs and identification of the most costly conditions was conducted by Druss et al [5, 19]. Their study looked at the most costly conditions in the United States. They found that spending for the top fifteen most expensive conditions accounted for 44.2 percent of US healthcare spending. They also found that the rank of the conditions varied widely between cost and disability, and that the most costly diseases were not necessarily the most disabling ones. Table 1 below shows the most costly diseases in the United States as ranked by cost; the subsequent columns show the number of work-loss days and bed days. As can be seen in the table, the most expensive diseases to treat are no necessarily the ones that keep employees away from work the most.

Table 1: The Most Costly Conditions In The United States: National Expenditures, Bed Days [5]

	National Cost		Work-Loss Days		Bed Days	
Condition	Billions	Rank	Millions	Rank	Millions	Rank
Ischemic Heart Disease	21.5	1	21.8	9	70.1	10
Motor Vehicle Accidents	21.2	2	70	3	102.9	7
Acute Respiratory Infection	17.9	3	69.2	4	196.4	4
Arthropathies	15.9	4	67.2	5	359.7	1
Hypertension	14.8	5	12	11	61.1	12
Back Problems	12.2	6	83	1	191.6	5
Mood Disorders	10.2	7	78.2	2	227.3	2
Diabetes	10.1	8	27.5	8	210.5	3
Cerebrovascular Disease	8.3	9	5.2	13	97	13
Cardiac Dysrythmias	7.2	10	7.2	12	66.5	12
Peripheral Vascular Disorders	6.8	11	12.8	10	55.1	10
COPD	6.4	12	57.5	6	176.3	6
Asthma	5.7	13	31.4	7	102	7
Congestive Heart Faliure	5.2	14	1.1	15	48.7	15
Respiratory Malignancies	5	15	2.5	14	21.7	14

Goetzel et al have done extensive work in identifying what the most costly diseases are for employers in the United States [6, 18, 20-27]. They identified cost areas by looking at inpatient and outpatient health care claims for employers who contributed their data to Medstat's MarketScan Private Pay Fee-For-Service Database. This also included short term disability (STD) or absence claims. Pharmacy, STD, absence and medical claims data were all linked using a patient key unique to each employer. For the disease assessment, if the patient was linked to more than one chronic condition, then the cost was linked to the condition with the highest disease stage. Payments made by the employer were calculated on a per eligible employee basis. The disease conditions were then sorted according to payments and the top twenty most costly physical and mental health conditions were identified for each employer. They then compiled separate lists of the top ten physical illnesses per industry and overall [23]. The top twenty medical conditions identified by their study are shown in

Table 2 below.

Table 2: Table of Top 20 Diseases as reported by Goetzel et al [23]

			Per Eligible Employee				
Rank	Condition	Rate per 100	Total Healthcare Payments	Total Absence Payments	Total STD Payments	Total HPM Payments	Absence - STD% Share
1	Angina Pectoris, Chronic Maintenance	41.5	$205.39	$22.60	$7.70	$235.69	13%
2	Essential Hypertension, Chronic Maintenance	123.8	$91.44	$60.52	$8.27	$160.23	43%
3	Diabetes Mellitus, Chronic Maintenance	48.5	$74.76	$24.93	$4.63	$104.32	28%
4	Mechanical Low Back Disor.	52.7	$52.83	$22.53	$14.89	$90.24	41%
5	Acute Myocardial Infarction	4.1	$60.33	$2.29	$6.62	$69.23	13%
6	Chronic Obstructive Pulmonary Dis.	30.5	$37.24	$22.62	$5.21	$65.08	43%
7	Back Disor. Not Specified as Low Back	60.4	$39.49	$14.65	$9.37	$63.50	38%
8	Trauma to Spine & Spinal Cord	42.3	$28.99	$28.06	$5.11	$62.16	53%
9	Sinusitis	91.5	$32.64	$20.28	$7.26	$60.17	46%
10	Dis. of ENT or Mastoid Process NEC	134.8	$29.22	$15.33	$5.18	$49.72	41%
11	Pregnancy with Vaginal Delivery	4.2	$30.31	$2.79	$14.35	$47.46	36%
12	Osteoarthritis, Severe	14.6	$28.02	$6.99	$12.03	$47.04	40%
13	Renal Failure	2.4	$41.80	$1.54	$0.57	$43.92	5%
14	Injury to Semilunar Cartilages	10.5	$26.21	$5.92	$4.63	$36.76	29%
15	Cancer of Female Breast	5	$30.90	$3.28	$1.91	$36.09	14%
16	Disor. of Gastrointestinal Tract, NEC	44.1	$27.27	$4.50	$3.20	$34.97	22%
17	Cholecystitis & Cholelithiasis	4.8	$28.78	$2.50	$2.97	$34.24	16%
18	Nutritional, Immune, & Metabolic Disor., NEC	112.1	$24.45	$7.31	$2.31	$34.07	28%
19	Osteoarthritis, Chronic Maintenance	20.8	$18.18	$13.15	$2.61	$33.94	46%
20	Cancer of Colon & Rectum	2.8	$30.49	$2.50	$0.81	$33.81	10%
	Average	42.6	$46.94	$14.21	$5.98	$67.13	30%

The top twenty most costly diseases were found to be similar to the top ten illnesses nationally, although the rank varied. (For example, hypertension and diabetes are number 5 and 8 respectively on the national cost list, but number 2 and 3 on the employer list). The order the top 20 diseases fell in also varied by industry. While the top 10 looked similar between industries, there were some new entries which could possibly be attributed to the varying demographic mix between industries, as well as differing risk factors. Some examples of these differences are:

- Breast cancer appears in the top ten for the finance, insurance, and real estate industry

- Disorders and complications of pregnancy are in the top ten list for the finance, insurance, real estate and the service industry
- Chronic obstructive pulmonary disease is in the top ten list for manufacturers of durable goods
- Herniated inter-vertebral disks appears in the top ten list for the oil and gas extraction and mining industry
- Trauma to the spine or spinal cord is in the top ten list for the service industry

Goetzel et al worked on assessing the on-the-job productivity or "presenteeism" financial losses employers face for a number of diseases. There is an increasing interest on the part of employers and insurers to generate estimates of productivity decline related to common health conditions. This is particularly difficult due to the lack of standard measurement tools. In order to conduct their study, Goetzel et al started with the Health and Productivity Management (HPM) administrative claims database in order to identify the most prevalent and costly diseases. They then combined surveys methods used to measure productivity and health limitations and conducted calculations on the resulting data to obtain their presenteeism results. The conditions chosen were the top ten diseases most commonly found between their studies and other large-scale survey methods [6]. They then aggregated medical, absence, STD and average presenteeism costs, and developed costs by disease for the total expenditures per eligible employee. It was found that presenteeism costs accounted for about 61 percent of the total costs associated with the top ten diseases. The greater presenteeism costs were associated with conditions associated with seasonal changes and/or symptomatic "flare ups" such as allergies,

asthma, migraines, depression and arthritis, as opposed to chronic conditions such as cancer and heart disease.

Table 3 outlines the daily dollar impact for each of the top ten conditions affecting employers.

Table 3: Productivity Losses Due to Presenteeism Per Employee[6]

Condition	Range	Avg hours lost per day (assuming 8hr day)	Average Daily Dollar Impact	Low hours lost per day	Low Daily Dollar Impact	High house lost per day	High Daily Dollar Impact
Allergy	Low	0.9	$20	0.7	$15	1.2	$27
Arthritis	Low	0.9	$21	0.5	$12	1.3	$30
Asthma	Low	0.9	$20	0.6	$15	1.1	$26
Any Cancer	Med	0.7	$16	0.2	$4	1.2	$28
Depression/sadness/ mental illness	Med	1.2	$28	0.7	$17	2	$45
Diabetes	Med	0.9	$21	0.2	$4	1.7	$40
Heart disease	Med	0.5	$13	0	$ –	1.1	$25
Hypertension	Low	0.6	$13	0	$1	0.8	$19
Migraine/headache	Hi	1.6	$38	0.7	$15	2.3	$53
Respiratory disorders	Low	1.4	$32	1	$24	1.7	$38
Average Loss	Med	1	$22	0.5	$11	1.4	$33

A graph of the aggregate costs associated with each condition according to the findings of Goetzel et al. is shown in Figure 1 below.

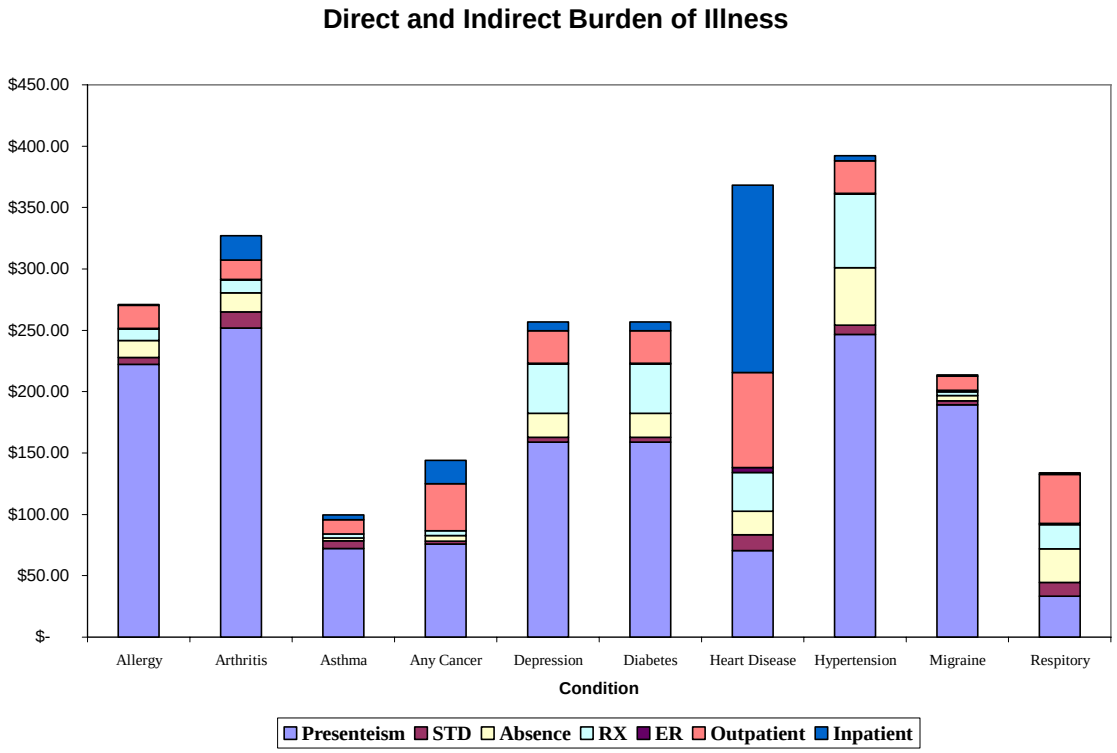

Figure 1: Aggregate Costs Associated with Each Condition [6]

It is important to note that the most costly diseases on the national list by Druss et al, and the most costly diseases identified in both studies by Goetzel et al vary in ranking, although the overall list generally agrees. Ignoring general disagreements in assessing cost, this could also be attributed to the difference in costs affecting employers, including accounting for productivity loss. One could infer from the differences between the two findings that national data including both employed and un-employed populations may not accurately represent the health needs of a given corporations. Employers will need to look at regional or local data whenever possible as risk factors and health outcomes vary by state and industry. In this study we have used mostly state data either as observed data or as prior information, only using national data when no state data was available.

2.3 Data sources

Medical information on the disease status and risk levels of individuals will be obtained from a number of national databases, as well as from peer-reviewed literature. These databases include:

- The National Heath Interview Survey (NHIS),
- The National Health and Nutrition Examination Survey (NHANES)
- The Centers for Disease Control (CDC) Behavioral Risk Factor Surveillance System (BRFSS)

Medical claims data will also be obtained from fortune 500 companies in order to gain greater insight on the health burdens of companies, as well as to serve as a method of testing the model. Information on the demographic breakdown of industries by state will be obtained from the US Census Bureau [28].

National Health Interview Survey (NHIS)

The NHIS is one of the main sources of information on the health of the US population, and one of the major data collection programs for the National Center for Health Statistics (NCHS) which is part of the CDC. The NHIS is a cross-sectional household interview survey. The interviews are conducted throughout each year. Every ten years the sampling plan is redesigned, with the current plan having been implemented in 2006 [29]. The data being used in this study is from the 2006 survey.

National Health and Nutrition Examination Survey (NHANES)

NHANES is a population-based survey designed to collect information on the health and nutrition of the US household population. There are two parts to the survey:

the home interview, which consists of asking participants questions about their health status, and the health examination, where many tests are performed which could give the participant more information about his/her health status [30]. The data being used in this study are the 2005-2006 survey results.

Behavioral Risk Factor Surveillance System (BRFSS)

The BRFSS is a state-based system of health surveys that collects information on health risk behaviors, preventive health practices, and health care access primarily related to chronic disease and injury [31]. Data are currently collected monthly in all 50 states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, and Guam. Over 350,000 adults are interviewed each year, making this the largest telephone survey system in the world. The survey consists of core questions, although individual states and territories had the option of adding additional questions. This study will use the results of the 2005 survey as the 2006 survey did not have information on some of the conditions we were interested in.

US Census Bureau

The Census Bureau serves as a source data about the people living in the United States. In addition to taking a census of the population every 10 years, the Census Bureau conducts censuses of economic activity and state and local governments every five years, and conducts more than 100 other surveys annually. The sole purpose of the censuses and surveys is to collect general statistical information from individuals and establishments in order to compile statistics.

The American Community Survey (ACS) is another nationwide survey conducted by the US census bureau. It is designed to provide communities with a look at how they

are changing annually. The ACS collects information such as age, race, income, commute time to work, home value, veteran status, and other important data from U.S. households. The ACS collects and produces population and housing information every year instead of every ten years. About three million households are surveyed each year, from across every county in the nation. Collecting data every year reduces the cost of the official decennial census, and provides more up-to-date information throughout the decade about trends in the U.S. population at the local community level. For the purposes of this research the information from the 2005 ACS survey will be used taking into account information about individuals by state, race, age, gender and industry.

2.4 Claims Data

Medical claims data contains information on the treatment received by an employee. It includes in-patient and out-patient procedures. Pharmaceutical claims data contains information on medication that was obtained by a patient and gives additional information about which conditions an individual may have. Patient diagnosis is reported on claims data according to the International Classification of Diseases (ICD). The ICD was designed by the World Health Organization to promote international comparability in the collections, processing, classification, and presentation of mortality statistics [32]. The diagnosed conditions and risk factors are translated into medical codes through a defined classification structure. The clinical modification (CM) version of the coding system is based on the original ICD, and is the official system of assigning codes to diagnoses and procedures associated with inpatient, outpatient, and physician office utilization in the United States. This classification scheme also provides additional morbidity detail.

The ICD has been revised periodically to incorporate changes in the medical field. There have so far been 10 modifications of the ICD. As the claims data available for this research uses mainly the coding scheme of the ninth revision, the ICD-9 codes will be used. This study used medical and pharmaceutical claims data in order to test the model. A full listing of the ICD-9 codes associated with the diseases of interest in this study can be found in Appendix A.

Medical claims data is sometimes limited in assessing the health burdens of individuals as physicians list what the patient had treated during that visit, and not necessarily other conditions. One way around this is to also look at the pharmacy claims data and to ensure that atleast a years worth of claims data is used in the study to capture any other conditions the individual may have.

2.5 Modeling Techniques

The previous section has outlined the need for good disease models which can accurately assess the health considerations for a corporation. Diseases to be modeled and their corresponding risk factors have also been identified as well as the main sources of information needed to develop the model. This section will cover basic modeling techniques which have been used, as well as those that have not been used, but will greatly improve on current practices. These include basic frequentist probability models, logistic regression and Bayesian analysis.

Probability Models

Current models to determine disease prevalence use basic probability. For example, the National Committee for Quality Assurance (NCQA) Quality Dividend Calculator uses the available prevalence rates for certain diseases based on the output

from BRFSS and applies those probabilities to the employed populations they are evaluating[16]. They are limited in their scope as there are typically not enough samples by state by disease in order to conduct these calculations for all diseases.

Logistic Regression

Logistic Regression is a model used for prediction of the probability of an event happening, by making use of several predictor variables that may be either numerical or categorical. For example, the probability that a person has diabetes might be predicted from knowledge about the person's age, gender, weight and race. Logistic regression analyzes binomially distributed data of the form

$$Y_i \approx B(n_i, p_i), \quad for\ i = 1, \ldots, m,$$

where the numbers of Bernoulli trials n_i is known and the probabilities of success p_i are unknown. The model proposes that for each trial, there is a set of explanatory variables that might inform the final probability. The logarithms of the odds ratio are modeled as a linear function of the X_i.

$$\log it(p_i) = \ln(\frac{p_i}{1-p_i}) = \beta_o + \beta_1 x_{1,i} + ... + \beta_k x_{k,i}.$$

The model has an equivalent formulation,

$$p_i = \frac{1}{1+e^{-(\beta_o + \beta_1 x_{1,i} + ... + \beta_k x_{k,i})}}$$

The β_j parameter may be interpreted as the additive effect on the log odds ratio of a unit change in x_j. For example, if β_j =0.383 and x_j goes from 0 to 1, then there is an increase of 0.383 in the log odds ratio. This property of logistic regression will be used for prior elicitation in section 3.4 (*Prior Selection*).

For this study, the dependent (response) variable will be the disease indicator, and the independent variables (x_j) will be the risk factors associated with that condition.

Logistic regression may be used to predict outcomes by looking at the resulting probability value. For example, in order to predict the presence of a disease, a discrimination threshold value would be assigned to the probability value. The presence of disease would then be indicated by higher than this threshold value, and consequently absence of disease would be indicated by values lower than the chosen threshold. The model would then be evaluated by its ability to correctly identify which individuals in a testing set were correctly classified. There are four possible outcomes for the model:

True Positives – The model correctly identifies and individual as having disease.

False Positive – The model incorrectly identifies and individual as having disease.

True Negatives - The model correctly identifies and individual as not having disease.

False Negatives - The model incorrectly identifies and individual as not having disease.

One of the metrics for tuning the model includes the sensitivity and specificity of the model. Sensitivity is the proportion of true positives of all diseased cases. It is used to assess how well the classification system correctly identifies the condition. For disease prediction high sensitivity is necessary in cases where outcomes could be costly. Specificity deals with the proportion of true negatives. It measures how well the model identifies individuals without disease. The model can be tuned by a standard ROC (receiver operating characteristic) analysis in which the threshold is set to optimize the ROC curve that depends on the sensitivity and specificity of model predictions.

Prior and Posterior Probabilities

In the context of the model, a prior probability distribution is the probability distribution that would express one's uncertainty about model parameters before the data is taken into account. There are informative and non-informative priors. An informative prior expresses specific information about model parameters. For example, in order to model for the temperature at noon tomorrow, one would make the prior a normal distribution, centered at today's noon-time temperature with variance guided by the day-to-day sample variance of the noon temperature.

Non-informative priors, otherwise known as *objective* priors give vague or general information about parameters. The prior can express some information, such as; the value is less than some limit and so forth. The use of non-informative priors typically yield results which are close to conventional statistical analysis, as the inference using the likelihood function only is close to the inference using the non-informative priors.

One common property among some priors is the ability for the posterior from one problem to become the prior for the next problem. Pre-existing information which has already been taken into account is part of the prior, and as more evidence accumulates, the prior is determined largely by evidence rather than the original assumption.

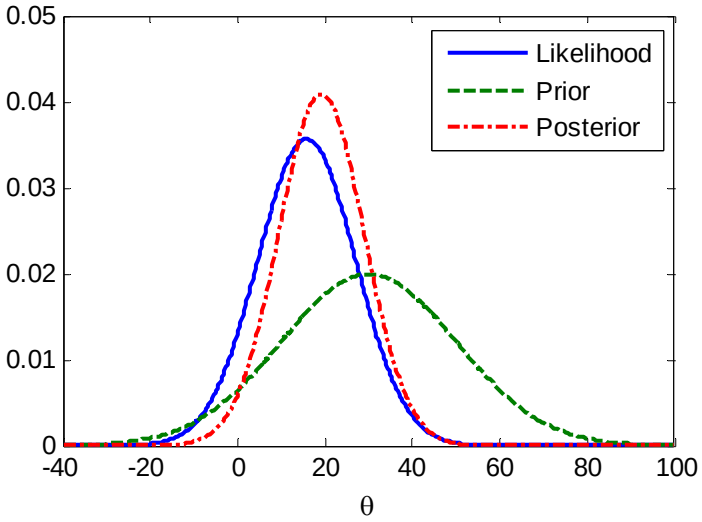

Figure 2: Image of Priors, Likelihood Function and Posterior

Markov Chain Monte Carlo

A Markov chain is a stochastic process with a countable number of possible steps for which the conditional distribution of any future state, X_{n+1}, given the past states X_0, X_1, … X_{n-1} and the present state, X_n, depends solely on the present state [33]. Monte Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution. The state of the chain after a large number of steps is then used as a sample from the desired distribution. The quality of the sample improves as a function of the number of steps. It is not usually hard to construct a Markov Chain with the desired properties. It is more difficult to determine how many steps are needed to converge to the stationary distribution with an acceptable error.

Many Markov chain Monte Carlo methods move around the equilibrium distribution in relatively small steps, with no tendency for the steps to proceed in the same direction. One such MCMC method is the Metropolis-Hastings algorithm which

generates a random walk using a proposal density and a method for rejecting proposed moves. Another method is Gibbs sampling, which is a special case of Metropolis-Hastings. This method requires that all the conditional distributions of the target distribution can be sampled exactly. There is an ever growing connection between Bayesian methods and simulation-based Monte Carlo techniques since complex models cannot be processed in closed form by a Bayesian analysis, while the graphical model structure inherent to statistical models, may allow for efficient simulation algorithms like the Gibbs sampling or Metropolis-Hastings algorithm schemes.

Combining regression analysis with Bayesian analysis and MCMC methods is possible [34], and may allow for better interpretation of the levels of risk factors associated with diseased individuals.

2.6 Chi Square Goodness of Fit Tests

Statistical methods are often used to check the validity of a model. One proposed test is the Chi-Square goodness of fit test. This test is particularly useful in determining how well a model fits observed data. In our context, the discrepancy statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(Np_i - n_i)^2}{Np_i} + \frac{(Nq_i - (N - n_i))^2}{Nq_i} \ ,$$

where N is the total number of employees, k is the number of diseases considered, p_i is the model prediction for disease i, q_i is the complimentary probability ($p_i + q_i = 1$), n_i is the number of people observed with the disease i.

The statistic χ^2 is approximately distributed as χ^2 distribution with *k-1* degrees of freedom. This is an approximate distribution since diseases are not independent and we assumed independence to utilized additive property of chi-squared distribution. The degrees of freedom is calculated as

$$df = 2k - 1 - k$$
$$= k - 1.$$

The goodness of fit can be tested in two ways:

i. Using all the data

ii. Using data split as training and validation samples.

In (i) we asses the quality of the model overall, in the sense of how well it describes the data used to establish the model parameters. In (ii) we assess the predictive power of the model using p_i's from the training set and n_i's from the validation set in the χ^2 statistic.

The test evaluates the null hypothesis Ho, that the data observed matches the predicted values of the model, against the alternative, Ha, that the observed data does not match the predicted models. Ho is rejected when the value exceeds the critical value of the $\chi^2_{(k-1)}$ distribution at the desired level of significance, α.

For comparison, the model median output value is used to compare against the observed value. The mean and mode are not used since the distribution is not assumed to always be symmetric. For a symmetric unimodal distribution, the median, mode and mean values are all the same. For skewed distributions, comparisons using mean or mode

would be questionable. Median values will always report the center of the distribution. Figure 3 below illustrates this concept.

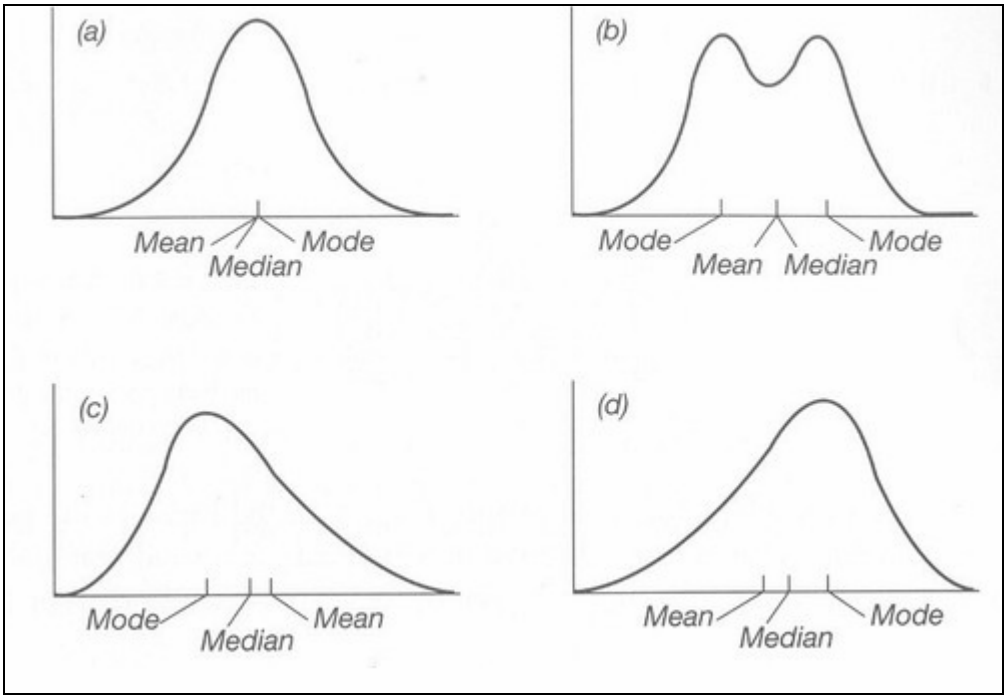

Figure 3: Distributions (a) and (b) are symmetrical, (c) is skewed left and (d) is skewed to the right

The example below (Example 1) shows how a typical comparison between the observed claims data and the model output would be compared.

Example 1

For a given state *S,* with sample size *N*, we consider diseases

$D_1, D_2, D_3, D_4,$

the number of affected individuals is

n_1, n_2, n_3, n_4.

The model would estimate the median probability of having each disease as

p_1, p_2, p_3, p_4

and consequently, probabilities of not having the disease

q_1, q_2, q_3, q_4.

The estimated number of people with a disease would then be given by

Np_1, Np_2, Np_3, Np_4.

The χ^2 distribution would have 3 degrees of freedom, and is calculated as

$$\chi^2 = \sum_{i=1}^{4} \frac{(Np_i - n_i)^2}{Np_i} + \frac{(Nq_i - (N - n_i))^2}{Nq_i}$$

The calculated value would need to be less than the χ^2 quantile value at three degrees of freedom which is 0.3518 for α=0.05.

CHAPTER 3 METHODOLOGY

The techniques described in the previous section were used to develop the disease analysis model. This section discusses the methods used in creating the model and gives examples on how some of the techniques were used.

3.1 Parameters of Interest

The size of the databases proved to be challenging to work with and so each database needed to be modified in order to make them easier to work with. For example, the BRFSS database contained over 300K data points, with 325 variables. Since many of these variables were unimportant, they were removed in order to make the file faster to open and easier to manipulate. In the case of the health conditions, the data files were first cleaned to remove unnecessary data and then the data pertaining to diseases were kept in one file, while the data pertaining to risk factors kept in another file.

For demographic data, the census bureau information also cleaned in order to have a file that was easy to use.

3.2 Data Cleaning

The parameters of interest taken from the health databases were:

- State
- Gender
    - Male
    - Female
- Age Groups
    - 1 = 18 to 24

    - 2 = 25 to 34
    - 3 = 35 to 44
    - 4 = 45 to 54
    - 5 = 55 to 64
    - 6 = > 64
- Race Groups
    - White, non-Hispanic
    - Black, non-Hispanic
    - Hispanic
    - Mixed
    - Other
- Disease Variable
    - Have disease, yes/no
- Risk Factor Variable
    - Currently have risk, yes/no

In the case where age categories were not already created, the age grouping was calculated from the listed age or year of birth. The same was done for databases that did not automatically code the race groups. For some health states, the presence or absence of a condition had to be computed giving a “1” for presence of disease and “0” for absence of a disease, counting the answer “don’t know” as an absence. For conditions where a range of answers was acceptable, then the parameter of interest was coded as a “1” with the other entries coded as “0” (example for smoking there were responses for heavy

smoker, occasional smoker, non-smoker and former smoker). Diseases and risk factors modeled are discussed in greater detail in section 3.8 (*Disease Pathology*).

Co-morbidity data was created by summing the condition indicators belonging to the health states of interest (for example diabetes + hypertension). A value of "2" indicated that that individual had both conditions while a value <2 indicated they only had one disease or none. The same method is applied when we were investigating three diseases.

In order to track how many people had multiple diseases, again the condition indicators were summed so we could see how many people had "no disease", "one disease" or "multiple diseases". The same was done for risk factors, with "presence of disease" being included as a risk factor. This allowed us to create an estimation of the number of people in low, medium and high risk categories. Low risk individuals were defined as those having less than 2 risks factors. Medium risk individuals had between 2 and 4 risk factors, while high risk individuals had greater than 4 risks.

Demographic data from the census bureau was also modified to only have the relevant information which included industry as well as state, age, race and gender.

As discussed in the background, the top ten diseases concerns in the workforce differed from the concerns of the nation as a whole. Literature review also showed that disease and risk levels varied according to level of education [35]. These two findings prompted an initial investigation on the national prevalence levels of certain diseases using the BRFSS data. The results showed that there was a difference between the disease levels of employed vs. unemployed individuals. Data relevant to employed populations was used to model the employee health burdens, while complete data was

used for the dependents since it could not be assumed that all the dependents were employed. Table 4 below shows a sample of the initial results comparing the prevalence of diabetes for the employed population with the prevalence when no filters were used.

Table 4: Prevalence for Diabetes: Employed only vs. Total National Population

	Diabetes Prevalence	
	Employed Only	All
National	5.30%	9.40%
Arizona	6.20%	10.10%
Florida	6.10%	11.00%
Georgia	5.80%	9.90%
Minnesota	4.70%	6.50%

In addition to data from the aforementioned databases, prevalence information was obtained from peer-reviewed literature so as to create more informative priors.

A file was created for each disease and risk factor category on a national level. This data was then cross tabbed in SPSS to obtain the prevalence of the disease by gender, age and race, to fit the format shown in the example below. A cross-tabulated result was also performed for the state location for Company A.

Data taken from the US Census Bureau went through a similar process. Data was cross tabbed by industry, age, race and gender. Since each major category of industry had sub-categories, those subcategories were first merged into the main category (e.g. Furniture industry was merged into the larger Manufacturing group).

Regression analysis models were built for each of the disease categories using the following risk factors: age, race, gender, lack of exercise, smoking, high cholesterol, obesity and high blood pressure. In this case, the age variables were adjusted to 1 if the individual was older than 45, and 0 if he was younger than 45. With each regression

model run, the risk factors were evaluated for their fit in the model, and irrelevant risk factors were removed. The results were then compared with the literature to validate the regression model, and also to get a better understanding of which risk factors to model.

The following section discusses the Bayesian applications used in creating the model.

3.3 Bayesian models

Bayesian data analysis makes inferences from data using probability models for quantities we observe and for quantities about which we wish to learn. The Bayesian approach provides some advantages over traditional data analysis in that it can avoid the assumption of infinite amounts of forthcoming data, recognize that fixed-point assumptions about human behaviour are dubious, and it provides a direct way to include existing expertise in the field.

The essentials of Bayesian modeling are contained in three general steps:

- Specify a probability model that includes some prior knowledge about the parameters if available for unknown parameter values.
- Update knowledge about the unknown parameters by conditioning this probability on observed data
- Evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions.

In this section there will be some general notation discussions and general discussion on Bayes theorem before getting into the detailed discussion of each model used in this research.

General Notation

Generally, θ, is used to denote unobservable vector quantities or population parameters of interest (e.g. the probability of contracting a disease). Observed data is denoted by y (e.g. the numbers of people contracting the disease in a certain population), and $\hat{y}$ denote the unknown but potentially observable quantities (e.g. disease levels among a similar population in a different region.)

In many statistical studies, data are gathered on each of a set of n objects or *units*. We can write the data as a vector, $\mathbf{Y} = (y_1, \ldots y_n)$. In the case of the disease models $y_i = 1$ if individual i shows presence of the disease, and $y_j = 0$ if individual j does not have the disease.

Bayesian Inference

Bayesian statistical conclusions about a parameter θ, or unobserved data, $\hat{y}$, are made in terms of probability statements. These probability statements are conditional on the observed value of y, and are simply written as $p(\theta|y)$ or $p(\hat{y}|y)$. They are also conditioned on the known values of any observed covariates, x. It is at the fundamental level of conditioning on observed data that Bayesian inference departs from the traditional statistical approach.

Priors, Posteriors and Likelihood Functions

The joint probability mass or density function can be written as a product of two densities that are often referred to as the prior distribution $p(\theta)$ and the sampling distribution (data model or likelihood) $p(y|\theta)$ respectively:

$$p(\theta, y) = p(\theta)\,p(\theta \mid y).$$

Conditioning on the known value of the data, *y*, using the basic property of conditional probability known as Bayes rule yields the posterior density:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y \mid \theta)}{p(y)}, \qquad \textbf{1}$$

Where

$$p(y) = \sum_{\theta} p(\theta)\mathrm{p}(\mathrm{y} \mid \theta) \ ,$$

and the sum is over all possible values of θ or:

$$p(y) = \int p(\theta)p(y \mid \theta)d\theta \ ,$$

in the case of continuous θ.

Modifying equation 1 will yield the un-normalized posterior density:

$$p(\theta \mid y) \propto p(\theta)p(y \mid \theta) \ .$$

To make predictive inferences, similar logic is utilized. Before *y* is considered, the distribution of the unknown but observable *y* is:

$$p(y) = \int p(y, \theta)d\theta = \int p(\theta)p(y \mid \theta)d\theta$$

This is called the prior predictive distribution: Prior because it is not conditional on a previous observation of the process and predictive because it is the distributions for a quantity that is potentially observable.

After the data *y* has been observed, an unknown observable $\hat{y}$ may be predicted by the same conditioning process.

The new distribution of $\hat{y}$ is called the posterior predictive distribution, because it is conditional on the observed *y* (posterior), and it is a prediction for an observable $\hat{y}$*:*

$$p(\hat{y} \mid y) = \int p(\hat{y}, \theta \mid y)d\theta$$

$$= \int p(\hat{y} \mid \theta, y) p(\theta \mid y) d\theta$$

$$= \int p(\hat{y} \mid \theta) p(\theta \mid y) d\theta .$$

The data y affect the posterior inference only through the function $p(y|\theta)$, which, when regarded as a function of θ, for fixed y, is called the *likelihood function*. The Bayesian inference in this method obeys what is sometimes called the likelihood principle, which states that for a given sample of data, any two probability models $p(y|\theta)$, that have the same likelihood function yield the same inference for θ.

Example 2

For example, suppose one observes x, where the model for x, is binomial

$$x \mid p \sim Bin(n, p) .$$

Now suppose one has prior information stating that p follows a beta distribution

$$p \sim Beta(\alpha, \beta) .$$

This is the prior information for the model. By incorporating observed data into the prior, the posterior distribution is obtained as:

$$p \mid x \sim Beta(\alpha + x, \beta + n) ,$$

with x having a beta-binomial distribution.

$$x \sim betabinomial .$$

From the posterior, all inference about p is made.

Simulation in Bayesian Analysis

Simulation is a central part of much applied Bayesian analysis due to the relative ease with which samples can often be generated from a probability distribution, which could be otherwise intractable. In performing simulations, it is helpful to consider the duality between a probability density function and a histogram of a set of random draws from the distribution: given a large enough sample, the histogram can provide practically complete information about the density, and in particular, various sample moments, percentiles and other summary statistics provide estimates of any aspect of the distribution to a level of precision that can be estimated.

Another benefit of simulation is that extremely large or small simulated values often flag a problem with the model specification or parameterization that might not be noticed if estimates and probability statements were obtained in analytic form.

Generating values from a probability distribution is often straightforward with modern computing techniques based on pseudo-random number sequences. The most common used simulation techniques are the Gibbs Sampler and Metropolis Hastings algorithm. For the purposes of this research these computations were conducted the Bayesian inference package, BUGS, in conjunction with the MatBUGS package in Matlab.

Next we describe the sampling schemes known as Markov Chain Monte Carlo methods where Gibbs sampling and Metropolis Hastings are specific approaches.

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) simulation is a general method based on drawing values of θ from approximate distributions and then correcting those draws to

better approximate the target posterior distribution, $p(\theta|y)$. The samples are drawn sequentially, with the distribution of the sampled draws depending on the last value drawn; hence the draws form a Markov chain. The key to the success of the method is that the approximate distributions are improved at each step in the simulation, in the sense of converging to the target distribution.

In MCMC, several independent sequences of simulation draws are created; each sequence, θ^t, t = 1, 2, 3… is produced by starting at some point θ^0 and then, for each t, drawing θ^t from a transition distribution, $T_t(\theta^t \mid \theta^{t-1})$ that depends on the previous draw, θ^{t-1}. It is often convenient to allow the transition distribution to depend on the iteration number t; hence the notation T_t. The transition probability distributions must be constructed so that the Markov Chain converges to a unique stationary distribution that is the posterior distribution, $p(\theta|y)$.

MCMC is used when it is not computationally efficient, or feasible to sample θ directly from $p(\theta|y)$. The samples are taken iteratively in such a way that at each step of the process it can be expected to draw from a distribution that becomes closer and closer to $p(\theta|y)$.

The key to MCMC is to create a Markov process whose stationary distribution is the specified $p(\theta|y)$ and run the simulation long enough that the distribution of the current draws is close enough to this stationary distribution. For any specific $p(\theta|y)$, or un-normalized density $p(\theta|y)$, a variety of Markov chains with the desired property can be constructed.

Once the simulation algorithm has been implemented and the simulations drawn, it is absolutely necessary to check the convergence of the simulated sequences. Examples of this will be shown in the following section.

Example 3

It is possible to run the simulation discussed in example 2 using the MCMC simulation techniques previously described. The model below is written in BUGS language as follows:

Model

```
{
y~dbin(p,10)
p~dbeta(1,1)
new.y~dbin(p,10)
}
```

Data

```
list(y=7)
```

For parameters we selected N=10, α=1, β=1; While y=7 was observed. The BUGS output is below. The graphs show the densities of new.y and p.

Table 5: Posterior Estimates for new.y and *p*

Node	mean	sd	MC error	2.50%	median	97.50%
new.y	6.626	1.972	0.06151	3	7	10
p	0.6663	0.1321	0.004284	0.3933	0.6733	0.898

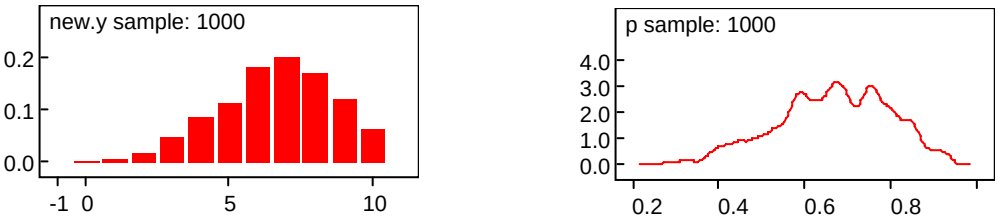

Figure 4: Kernel Density Estimates for new.y and p given by BUGS

Logistic Regression Model

The general formulation for logistic regression has been discussed previously in the background. Consider a vector of dichotomous observations **Y** = $(y_1, \ldots y_n)$, with $y_i \in \{0,1\}$. The relationship between the success probability $p_i = \Pr(y_i = 1)$ and explanatory variables was determined to be

$$\log it(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_i X_i$$

Formulation in our context may be written as

$$z_i \sim Binomial(p_i, n_i)$$

$$\log it(p_i) = \beta_0 + \beta_1 Gender_i + \beta_2 Age_i + \beta_3 R1_i + \beta_4 R2_i + \beta_5 R3_i + \beta_6 R4_i$$

A generic logistic model may be visualized using the doodle application which is a graphical interface within BUGS.

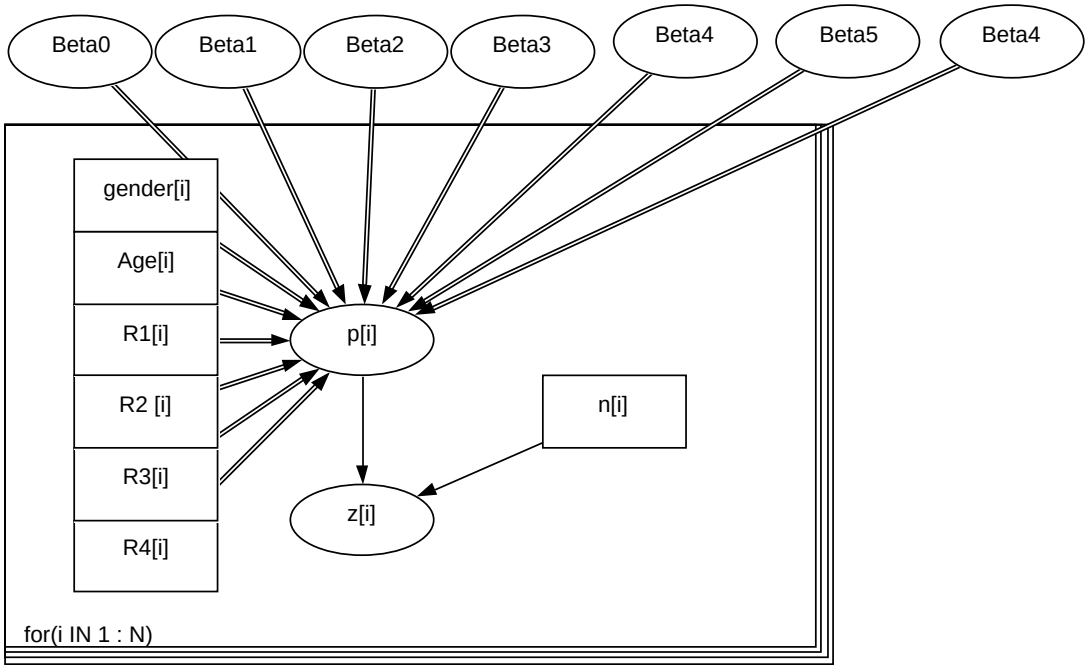

Figure 5: Graphical Representation of Regression Model

The R variables are used to indicate race. For example, if R1=1, then R2=R3=R4=0 means the individual is black. The remaining combinations for race are show the following table (

Table 6).

Table 6: Arrangement of indicators

Race	R_1	R_2	R_3	R_4
White	0	0	0	0
Black	1	0	0	0
Hispanic	0	1	0	0
Mixed	0	0	1	0
Other	0	0	0	1

3.4 Prior Selection

As discussed previously, Bayesian statistics is particularly powerful in that it allows for the coherent inclusion of expert opinion in the analysis. Prior knowledge is useful when sufficient data is not available to make final inferences. Certain parameters may also be difficult to analyze even with a reasonable amount of data. This expert opinion may consist of subjective inputs from experienced researchers or medical professionals or a summary of past research in the area. This section will discuss methods for selecting priors which will incorporate expert information.

When there is no other information known other than the observed data, the non-informative, or flat priors are used. If however, there is information available, then informed priors may be available. Suppose we have a logistic regression equation

$$\log it(p') = \beta_0 + \beta_1 x_1 + ... + \beta_{p\text{-}1} x_{p-1},$$

where β_i's are parameters for which we want to elicit priors. Suppose we are interested in the inference made by β_1. We can write

$$\log it(p'') = \beta_0 + \beta_1(x_1 + 1) + ... + \beta_{p-1} x_{p-1}$$

Taking the difference between both equations and manipulating using logarithmic rules will give the relationship between p', p'' and β_1.

$$\log it(p'') - \log it(p') = \beta_1$$

$$\log \frac{p''}{1-p''} - \log \frac{p'}{1-p'} = \beta_1$$

$$\log \frac{\frac{p''}{1-p''}}{\frac{p'}{1-p'}} = \beta_1$$

$$oddsratio(p'', p') = e^{\beta_1}.$$

This means that for $\beta_1 > 0$, odds ratio exceeds 1, at $\beta_1 < 0$ odd ratio is less than 1, and at $\beta_1 = 0$, odd ratio is just 1.

The next example shows that the prior may be selected with the same summaries (means, variances) but more robust with respect to possible outliers.

Example 4

Suppose that the parameter β in the logistic regression is set to have a prior mean 0 and prior variance 100. Then for β distributed normally, proper prior is

$$\beta \sim N(0, 10^2)$$

$$f(\beta) = \frac{1}{\sqrt{2\pi}10} e^{-\frac{\beta^2}{200}}$$

Another option is to use Laplace distributed methods instead of Gaussian. For a prior mean 0 and prior variance 100, the so-called corresponding Laplace prior is

$$\beta \sim DE(0, \lambda)$$
$$f(\beta) = \frac{1}{2\lambda} e^{-\frac{|\beta|}{\lambda}}$$

Parameter λ can be found as,

$$Var(\beta) = 2\lambda^2$$
$$\sqrt{\frac{100}{2}} = \lambda$$
$$\lambda = 5\sqrt{2}$$

Although these two priors share the mean and variance, the Laplace prior is producing more robust inference and is suitable for sparse modeling. In the case of auxiliary race variables, R, priors may need to be elicited simultaneously. The next example shows how this is done if we have information about prevalence of a disease in black and Hispanic groups.

Example 5

Suppose we were looking at the effect of race in predicting high blood pressure, then we could have a prior distribution on the β's set to incorporate available information

on prevalence of high blood pressure in a particular race group. Supposed we are interested in setting priors on β_1 and β_2, then,

$$\log it(p') = \beta_0 + \beta_1 Blacks(1) + \beta_2 Hispanics(0) + ... + \beta_{p-1} x_{p-1}$$

$$\log it(p'') = \beta_0 + \beta_1 Blacks(0) + \beta_2 Hispanics(1) + ... + \beta_{p-1} x_{p-1}$$

$$\log it(p'') - \log it(p') = \beta_2 - \beta_1$$

$$\frac{\frac{p''}{1-p''}}{\frac{p'}{1-p'}} = e^{\beta_2 - \beta_1}$$

Historically blacks have a higher odds ratio of hypertension than Hispanics, and so we expect that β_2-$\beta_1<0$. So, the following might be set as priors for β_2 and β_1.

$$\beta_2 \sim N(0,10^3)$$
$$\beta_1 \sim N(10,10^3)$$

Notice that 'variance' is large, suggesting that priors only focus on the sign of β_2-β_1, but not on its magnitude.

To create the disease and risk factor models, suitable probabilistic descriptions of the expert data were used to elicit the priors used in each model. Since prior information

varied from condition to condition separate priors had to be determined for each disease or risk factor state. After data was obtained and priors determined, the models needed to be run. As this is a computationally intensive process the model had to be automated using Bayesian software applications which will be discussed in the following section.

3.5 Model Automation

Smaller regression models were created according to the methods described in the previous section and run in BUGS to check their validity. More complex models were then designed and run in MatBUGS. The MatBUGS application was created to interface with BUGS and allows for Bayesian models to be run through Matlab. This software toolbox is useful in that it has more graphical capabilities than BUGS and simplifies the process of running multiple models. The program was set to monitor the particular parameters of interest, including the resulting β values and prevalence rates.

After obtaining a random sample from the slice sampler, it is important to investigate issues such as convergence and mixing, to determine whether the sample can reasonably be treated as a set of random realizations from the target posterior distribution. Looking at marginal trace plots is the simplest way to examine the output. The following is an example of trace plots obtained from BUGS.

Example 6

The model from example 1 was manipulated to included three chains of initial values. With new.y(1)=1, new.y(2) = 10, new.y(3)=5, p(1)=0.01, p(2)=1.00 and p(3) =0.50. The new model is as follows.

model

{

y~dbin(p,10)

p~dbeta(1,1)

new.y~dbin(p,10)

}

Data

list(y=7)

Init

list(p = 0.01, new.y = 1)

list(p = 1.00, new.y = 10)

list(p = 0.5, new.y = 5)

The three simulation chains are obtained. Figure 6 below shows the first 50 iterations for *new.y* and *p*. For *new.y*, the chains converge after the first 30 iterations, the chains for p seem to converge a little faster moving away from the initial values and converging about the mean.

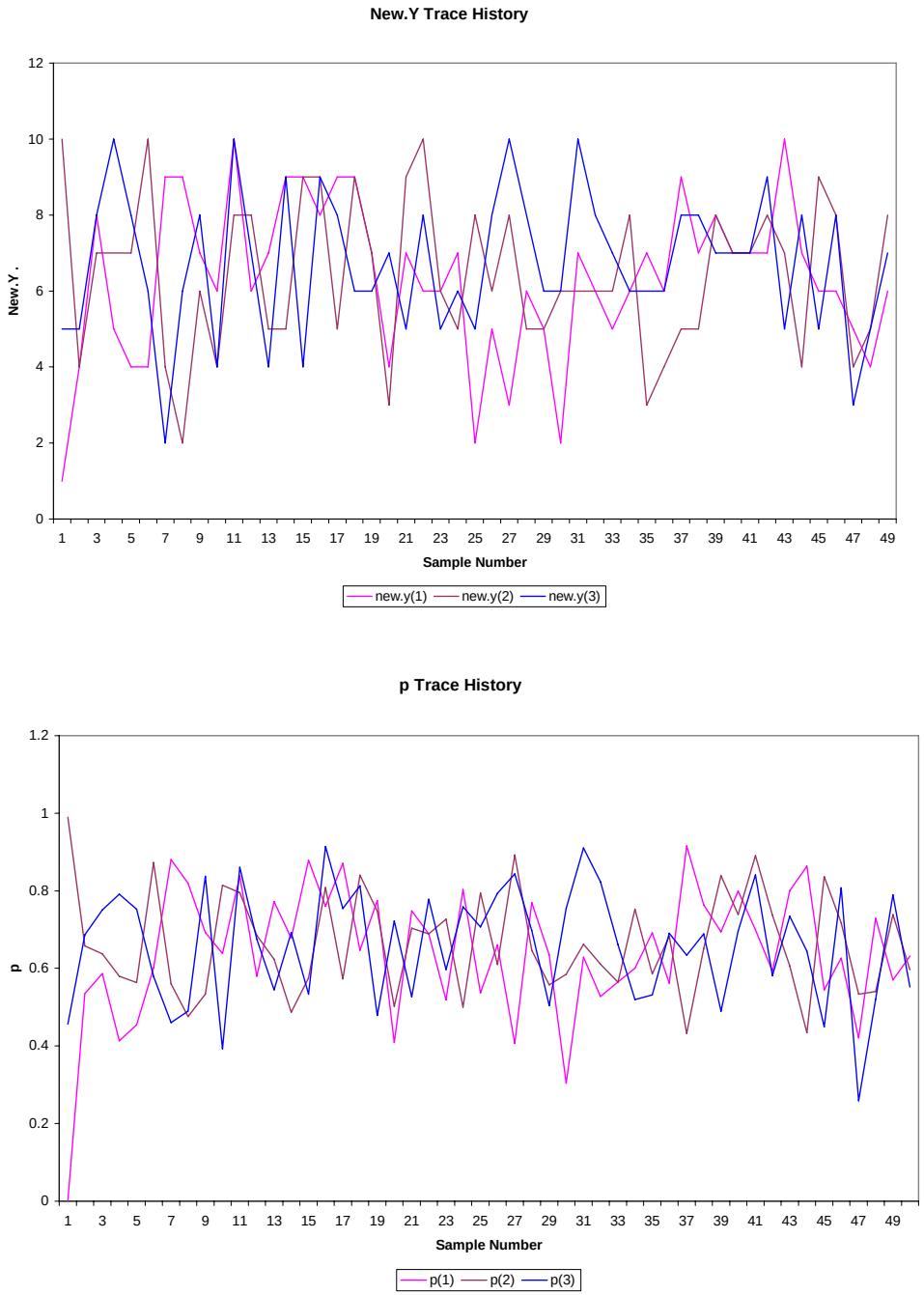

Figure 6: Trace History (a) new.y (b) p

Checking the chains is also helpful in determining how many iterations are needed for the model to converge to stationarity. After determining the optimal number of iterations and running the disease and risk factor models, the data was loaded into an

SQL server so that it could be applied to the company demographic data for ease of calculations. The necessary computations and queries were created in the SQL database.

3.6 Model Testing

We can test of model at two different levels.

- One level involves re-running the model, making the newly obtain posterior values the new prior values, then checking the results back with a testing portion of the original database. The theory behind this is that if the model is working correctly, then the output with the posterior values as inputs should lead to the same distribution as the original output.
- The second level simply uses χ^2 discrepancy statistics and obtained predicted values and tested them against an independent data set.

The second method simply uses the obtained posterior values and tests them against an independent data set.

For this model both methods of testing were employed. The model was tested against a small testing sample of the BRFSS and NIH databases, as well as against claims data for two companies. For confidentiality reasons the names of the companies cannot be released and so they will be called Company A, B and C.

Company A is a manufacturer located throughout the United States. For the purpose of testing, only four of their sites were used to test the model. The claims data from these locations were used to compare against the output from the model. Companies B and C are also a large manufacturing entities. Only one location for each of these companies was used to test the model. Outputs from the NCQA model were also used for comparison.

3.7 Demographic Data

The census bureau data was very comprehensive and so it was decided that traditional statistical methods could be used in creating the estimations for the demographic breakdown for the different industries. The demographic breakdown was obtained in general for the national population as well as by state. Also demographic breakdowns were obtained for the following industries.

1. Agriculture
2. Extracting (Mining)
3. Utilities
4. Manufacturing – Distribution
5. Manufacturing – Processing
6. Manufacturing – Instruments
7. Wholesalers
8. Retail
9. Transportation
10. Information
11. Finance
12. Professional – Consulting/lawyers/scientists/engineers
13. Professional – Services
14. Education
15. Medical
16. Military
17. Public Sector

The census bureau codes for these industries can be found in the appendix.

3.8 Disease Pathology

The early sections of this document discuss the costs associated with disease, and identify the most costly conditions to the nation and to employers. This section will give an overview of the diseases in the model. The diseases were chosen based on their frequency of appearance in the literature reviewed, and include:

- Allergies
- Arthritis
- Asthma
- Cancer
- Depression
- Diabetes
- Heart Disease
- Hypertension
- Migraine
- Respiratory Disease

Allergic Rhinitis

Allergic Rhinitis, commonly known as hay fever, may arise from exposure to a number of antigens including pollen, mold, or mite exposure. Instances are often seasonal, and are characterized by nasal inflammation, nasal blockage, and other cold like symptoms. Goetzel et al report that allergic rhinitis is the fifth most common chronic

condition in the United States[23]. It can significantly decrease the quality of life and impair social work functions either directly (due to reactions to allergens) or indirectly (due to effect of medications taken to relieve symptoms). As a result of these reactions, allergic rhinitis incurs very high presenteeism costs. Allergic rhinitis often co-occurs with asthma and sinusitis, perhaps exacerbating the severity of those diseases. Allergic Rhinitis is often associated with rhino sinusitis and asthma, occasionally aggravating those conditions, or occurring simultaneously.

Allergens may be divided into two groups: Seasonal and perennial.

- Perennial allergens provoke antigen responses that produce recurring symptoms year-round. Indoor allergies are more prominent during those months that individuals tend to stay indoors (e.g. winter, rainy seasons). Examples of perennial allergens are:

- Mold – Can be indoor or outdoor allergens. These are worse in cool, humid weather.

- Dust

- Pet dander – Dog dander is usually more problematic for the owner; however cat dander can cling to clothing and be carried into other public spaces. Dander from hamsters, rabbits and other furry animals may also cause an allergic reaction.

- Seasonal allergens are primarily pollens, and occur at certain times of the year. Seasonal pollen allergy may increase signs and symptoms of perennial rhinitis.

- Spring - Trees
- Summer – Grass
- Fall – Weeds

Family history generally determines allergies. If one parent has allergies, then the child has a 50% chance of also reacting to allergens. If both parents have an allergy the child's risk increases to 75%. Persons who have other allergies (e.g. eczema, food allergies, asthma) also tend to develop allergic sensitivity to some allergens[36].

Allergic Rhinitis is the most common atopic reaction which affects almost 35 million Americans. Due to the variability of studies, the prevalence rate of allergic rhinitis in the United States of America has been difficult to quantify, with most studies putting it between 17% and 25%. Finding quantifiable American studies discussing the different affects of allergens between races, genders or ages was difficult, with the most quantitative studies being conducted in Europe. For this study we used the results found in [37] and then the prior values were set using the given values from the Asthma and Allergy foundation of America, as well as references found in other papers[38-42]. The model used for this layout put a higher emphasis on the prior information (data reviewed from the papers). Due to the unavailability of data for specific states, the disease model for allergic rhinitis was only run for the national case. As more data becomes available, it will be possible to create more state specific models. This would be useful for states like Georgia which are known for having high levels of seasonal allergens, and so may incur higher incidence rates, and hence higher costs for this disease.

Table 7: Allergy Reference Table

Develops From	Disease	ICD9 - Codes	Leads to
Exposure to allergens Family History	Allergy	**477.0-477.9**	Asthma

Arthritis

Arthritis is the term applied to over 100 medical conditions affecting the musculoskeletal systems and the joints[44]. Joint related problems include pain, stiffness, inflammation and damage to joint cartilage and surrounding structures. Such damage can lead to joint weakness that may eventually affect daily activities including walking, using a computer, or brushing your teeth. It is the most common cause of disability in the United States, limiting the activities of nearly 19 million adults[45]. For many people joint pain is not the full extent of the diseases. Many forms of arthritis can affect the entire body. These forms of arthritis can cause damage to the heart, lungs and other organs. According to Goetzel et al, it is also the greatest contributor to presenteeism costs in the workplace. There are many different types of arthritis including, osteoarthritis, rheumatoid arthritis, gout and so forth. The more common ones are discussed here:

- Osteoarthritis – The most common form of arthritis. It is characterized by the breakdown of the joint's cartilage. Although incidence is highly correlated with age, osteoarthritis is no solely due to wear and tear, and may have other underlying causes. The main symptom of osteoarthritis is chronic pain which causes a loss of mobility and stiffness. It has been suggested that osteoarthritis is

hereditary due to the high prevalence rates within families. Risk factors for this type of arthritis include age, obesity, injury, genetics, muscle weakness and other types of arthritis. Currently there is no cure for osteoarthritis[46].

- Rheumatoid arthritis – This is an inflammatory autoimmune disorder which causes the body's immune system to attack the joints. It is also a systemic disease which can also affect other tissues throughout the body including the skin, blood vessels, heart, muscles and lungs. In a healthy immune system, white blood cells produce antibodies which protect the body against foreign substances. People with Rheumatoid arthritis develop antibodies which mistake the body's healthy tissue for a foreign particle and attack it. People with high levels of rheumatoid factor may have a malfunctioning immune system. All people with high levels of rheumatoid factor do not have RA, while not all RA individuals have high levels of rheumatoid factors. Rheumatoid arthritis affects women up to three times more than men. It is believed that some cases of RA are triggered by some form of infection, although RA itself is not contagious. It can develop and any age, unlike osteoarthritis which is more common in older individuals. There is no cure for rheumatoid arthritis. There are a variety of treatments which can minimize the effects and/or alleviate the symptoms[46].

- Septic arthritis – The invasion of a join by an infectious agent which produces arthritis. The usual cause is bacterial, but it may also be fungal and viral[46].

- Gout – This is a disease created by the build up of uric acid. This creates the formation of crystals on the cartilage of joints, on tendons and surrounding tissue.

It mostly affects men between the ages of 40 and 50. Typical suffers of gout are obese, prone to diabetes or hypertension and at a higher risk of heart disease[46].

Risk factors for Arthritis include age, gender, obesity, past joint injuries and family history. For this research we will assume all types of arthritis incur similar costs and so model the general category “arthritis” rather than attempting to distinguish between each type of disease. Data for this model was obtained from the CDCs BRFSS database, while also including prior information from literature[46-52].

Table 8: Arthritis Reference Table

Develops From	Disease	ICD9 - Codes	Leads to
Gender Race	Arthritis	**714.0–714.9** **715.00 - 715.98** **716.00-716.99** **719.40-719.59**	Back Pain

Asthma

Asthma is a chronic condition of the respiratory system where, in response to a trigger, the airway constricts, becomes inflamed, and is lined with excessive amounts of mucus. The condition cannot be cured, but can be controlled so individuals with the disease can live active lives. An asthma attack is when the muscles around the airways constrict making the airways narrower so that less air flows through. Cells in the airways may produce more mucus, and inflammation of the airways may occur, each process further reducing the amount of airflow. Severe attacks may lead to chest pain, loss of consciousness and even death.

The airways of asthmatics are sensitive to certain stimuli leading to the reactions discussed above. Inhaled allergens go into the inner airways and are ingested by antigen presenting cells[54]. In non-asthmatics the immune cells check and generally ignore the allergen molecules. In asthmatics, however, these immune cells transform into another type of cell which activate the humoral immune system. The humoral immune system produces antibodies against the inhaled allergen. When the asthmatic later inhales the same allergen, the antibodies activate a humoral response which results in inflammation, constriction and mucus production.

Examples of stimuli are:

- Allergens (inhaled and also food)
- Perfumes and perfumed products
- Medications
- Air pollutants (e.g. smog, ozone, nitrogen dioxide etc)
- Exercise

- Hormonal changes
- Emotional Stress

Treatment methods include using preventative medications such as an inhaled corticosteroid which will help to suppress inflammation and reduce swelling within the airways. This is generally recommended for individuals who present symptoms atleast twice a week. It is also recommended that asthmatics identify and avoid stimuli that could trigger an attack.

Asthma is much more common now than it was 20 years ago. Between 1980 and 1996, the number of self-reported asthma cases more than doubled. The economic cost of asthma is estimated to exceed $6.2 billion nationally, with hospital and emergency costs declining over the past ten years, but pharmaceutical costs have increased[55]. Risk factors for asthma include gender, race, smoking status and obesity. Asthma data was obtained from BRFSS as well as from expert opinion[56-60].

Table 9: Asthma Reference Table

Develops From	Disease	ICD9 - Codes	Leads to
Smoking Status Obesity Gender Race Family History	Asthma	**493**	Obesity Respiratory Diseases

Cancer

Cancer is the general term for a group of diseases which involve the abnormal growth of cells[61]. Cells are the building block of the human body. Within each cell are twenty three pairs of chromosomes which contain millions of messages telling the body how it should grow, behave and function. The chromosomes reproduce themselves every time the cell divides creating many opportunities for mistakes to occur, and a mutation to alter some of the genes. In the case of cancer, a damaged chromosome leads to rapid cell growth and multiplication until a malignant lump is formed.

Rapid cell growth is not always malignant. What differentiates cancer from other rapid growth cycles is the absence of instructions stopping the growth cycle. Malignant growths can invade and destroy adjacent tissues, and also spread to other locations within the body (metastasize). Malignant tumors put down roots and directly invade surrounding tissues. Bits of malignant cells fall off of a tumor and then travel to another tissue and start a similar growth. Benign tumors are generally limited in their growth and do not metastasize, although some benign tumors are capable of becoming malignant.

The term cancer encompasses over 200 diseases. There is no single cause for any of the types of cancers, although triggers for some types have been identified. Almost all cancers arise from atleast two “hits” to genes in the cell. These hits build up over time eventually triggering cancerous growth. The hits may come from chemical or foreign substances called “carcinogens”. These initiate the cancer process. Hits may also be promoters which accelerate the growth of abnormal cells.

Examples of Initiators:

- Tobacco and Tobacco Smoke
- Exposure to radiation
- Some immunosuppressive drugs
- Excessive exposure to sunlight.
- Industrial Agents or Toxic products
- Carcinogens in food or created during the cooking process

Examples of Promoters

- Asbestos
- Certain hormones (e.g. estrogen)
- Dietary Factors
- Alcohol
- Stress
- Obesity

Cancer causes about 13 percent of all deaths in the United States[46]. The exact risk factors for cancer have not been identified, although certain types of cancers have been associated with genetics and certain behaviors. For example smoking is a risk factor for lung cancer. In this study we will look at breast, cervical, prostate and skin cancers. Each of these will be model separately since treatment varies with each type. Data for the model was obtained from NHANES and NHIS as well as from the American Cancer Society and other expert sources [46, 49, 62-79].

Table 10: Cancer Reference Table

Develops From	Disease	ICD9 - Codes	Leads to
Alcohol Smoking Family History Exposure to Carcinogens Stress Obesity Diet	Cancer		Tumours

Depression

Depression is considered a down turn in mood which can be temporary, although clinical depression lasts for at least two weeks, and may interfere with day to day activities. Depression is considered a major public health problem [80]. Community based epidemiological studies have yielded varying prevalence estimates due to varying methods of case ascertainment. For the purposes of this study, we have chosen to study all forms of depression, which may lead to higher prevalence rates than others who only consider clinically diagnosed depression as the condition. Depression data was taken from NHIS as well as from literature [80-88].

Table 11: Depression Reference Table

Develops From	Disease	ICD9 - Codes	Leads to
Family History Obesity Marital Status Gender	Dysthymic Disorder (Includes anxiety depression, Depressive Reaction, Neurotic Depressive State, Depression with anxiety)	**300.4**	Migraine Obesity

Diabetes

Diabetes is a group of diseases involving issues with the insulin hormone [89]. Diabetes is characterized by inappropriately high blood sugar levels (hyperglycemia) and disordered metabolism resulting from either low levels of the insulin hormone, or from abnormal resistance to insulin's effects coupled with inadequate levels of insulin secretion to compensate. Diabetes can lead to many serious complications including cardiovascular disease, blindness and gangrene (which may lead to amputation). There are three main types of diabetes, Type 1, Type 2 and gestational diabetes (occurring during pregnancy). Type 1 diabetes is usually due to autoimmune destruction of the pancreatic beta cells. Type 2 diabetes is characterized by insulin resistance in target tissues. For the purpose of this research, we will consider all types of diabetes.

Several lifestyle factors affect the incidence of type 2 diabetes. Obesity and weight gain can dramatically increase the risk, and physical inactivity elevated the risk,

independently of obesity. Specific dietary fatty acids may differentially affect insulin resistance and the risk of diabetes [12]. Diabetes risk factors include high cholesterol, poor diet, obesity, race, age, smoking, hypertension and physical activity.

Diabetes is associated with a number of diseases including hypertension and heart disease. It can lead to diabetic comas, gangrene, hypoglycemia, diabetic retinopathy, diabetic neuropathy, diabetic nephropathy, stroke and other vascular diseases due to the elevation of blood glucose leading to damage of the blood vessels. These conditions may lead to blindness, loss of some motor function, amputation, other disabilities and death.

A quick overview of the risk factors and conditions associated with diabetes is shown below. Diabetes data was obtained from BRFSS as well as from literature [12, 45, 46, 89, 90].

Table 12: Diabetes Reference Table

Develops From	Disease	ICD9 - Codes	Leads to
Hypertension Coronary Heart Disease High Cholesterol Obesity Smoking Family History Age	Diabetes	**250**	Hypertension Coronary Heart Disease Coma Vascular Diseases Stroke

Heart Disease

Heart disease is one of the most expensive conditions for employers, and is present in almost every top ten most costly disease list investigated. It is also the leading cause of death within the United States [92]. Most individuals with coronary heart disease do not show symptoms for years before a sudden onset of symptoms, and so identification of high risk individuals may help with early identification and/or prevention of serious complications. Risk factors include, high blood pressure, cigarette smoking, cholesterol, obesity, family history of premature coronary heart disease and diabetes[93]. People with diabetes are particularly at risk of CHD, so much so that the National Cholesterol Education Program (NCEP) now recommends that diabetic patients do not need specific CHD risk assessment, but instead be managed as if they had CHD [92].

While Heart Disease is a term which covers a wide range of conditions, this study will only consider CHD conditions and excludes incidence of heart attack or stroke. Heart disease is closely associated with diabetes in that the risk factors are similar, and diabetes is also a risk factor for CHD. Data for the model was obtained from BRFSS as well as from heart disease studies [92-99].

Table 13: Heart Disease Reference Table

Develops From	Disease	ICD9 - Codes	Leads to
High Cholesterol Hypertension Diabetes Smoking Obesity Alcohol	Coronary Heart Disease	**402.0–404.9** **410.00-410.92** **411** **411.1** **411.0** **411.81 - 411.89** **412** **413** **414** **414.0** **414.00 - 414.05** **414.1** **414.10** **414.8 - 414.9**	Hypertension Diabetes High Cholesterol

Hypertension

Hypertension, commonly known as high blood pressure is the condition where the blood pressure is chronically elevated[100]. Presence of the disease is generally confirmed when the blood pressure readings of an individual are persistently high. Usually three consistent measurements are taken one week apart. Hypertension shares risk factors with CHD and diabetes, and is often seen as a risk factor for those diseases.

Hypertension is usually found during routine checkups although some people report headaches, fatigue, blurred vision, difficulty sleeping and facial flushing. Blacks are at a higher risk of being hypertensive and also, at the same blood pressure as Caucasians have more severe organ complications. This group needs to be especially careful when it comes to managing their hypertension due to these effects[46].

A person is said to be hypertensive when their systolic (contraction of heart chambers driving blood out of chambers) pressure exceeds 140 mmHg, and their diastolic (ventricles are relaxing) pressure exceeds 90 mmHg.

Table 14: Blood Pressure Levels

	Blood Pressure Level (mmHg)	
Category	**Systolic**	**Diastolic**
Normal	< 120	< 80
Pre-hypertension	120-139	80-89
High Blood Pressure		
Stage 1 Hypertension	140–159	90–99
Stage 2 Hypertension	greater than or equal to160	> 100

Data for the model was obtained from BRFSS as well as from literature [100, 101].

Table 15: Hypertension Reference Table

Develops From	Disease	ICD9 – Codes	Leads to
High Cholesterol Heart Disease Diabetes Smoking Obesity Alcohol	Hypertension	**401** **405**	Diabetes Migraine Heart Disease

Migraine

Migraine is a neurological disease which is commonly associated with severe headaches. Some analysts have considered patients that report having frequent episodes of headaches [102]. Others consider chronic daily headache (CDH) which is defined as headache occurring at least 15 days per month, which account for about 2.4 percent of the population [103]. Another popular migraine assessment is to look at strict migraine, which refer to migraine with aura [104]. These varying definitions make it difficult to predict the prevalence of migraine. For this model we had opted to use frequent headache as the modeling criteria, this includes all headache episodes including CDH and strict migraine. Data for the model was obtained from headache studies [102-110].

Table 16: Migraine Reference Table

Develops From	Disease	ICD9 - Codes	Leads to
Depression Asthma Cancer Hypertension	Migraine Headache Frequent Headache	**346.0–346.9** **307.81** **784.0**	

Respiratory Disease

Respiratory diseases are diseases of the lungs, bronchial tubes, trachea and throat. So far for this study we are concentrating on Chronic Obstructive Pulmonary Diseases (COPD), most specifically, emphysema and bronchitis. Emphysema is often associated with exposure to toxic chemicals, or long term exposure to cigarette smoke. Emphysema is caused by loss of elasticity (increased compliance) of the lung tissue, from destruction of structures supporting the alveoli, and destruction of capillaries feeding the alveoli. The result is that the small airways collapse during exhalation (although alveolar collapsibility has increased), leading to an obstructive form of lung disease.

Bronchitis is an inflammation of the bronchi in the lungs. It may be caused by an infection and last several days or weeks. Chronic cases of the disease are not always caused by a virus or bacteria, and may last between three months and two years.

Risk factors for respiratory diseases include age, gender, race, smoking, obesity, genetic factors, allergies or asthma. Respiratory data was obtained from NHANES as well as from literature [45, 46, 49, 111-113].

Table 17: Respiratory Reference Table

Develops From	Disease	ICD9 – Codes	Leads to
Asthma Age Gender Race Family History Allergies	Respiratory Disease	**490** **491.0-491.9** **492.0-492.8**	Cancer

3.9 Risk Factors

The actions individuals take can greatly affect their disease status. Many of the diseases listed in the previous section are affected by what people do, or not do, on a daily basis. Common risk factors, such as smoking, poor eating habits and lack of physical activity are major contributors to heart disease and cancer [45]. Risk factors not only affect employees, but they also lead to increased health costs for employers. Goetzel et al found that risk factors for some diseases were associated with significantly higher medical costs over a short term period. Employees with high risk for depression, high stress, high blood glucose, obesity, smoking, high blood pressure, and inactive lifestyles incurred higher medical expenditures than those lacking the risks [21]. Anderson et al found that about 25% of total health care expenditures were attributable to ten modifiable health risks [114]. The relationship between risk factors and mortality and morbidity rates, as well as increased healthcare costs should motivate corporations to not just identify diseased individuals, but also individuals who are at risk for contracting a disease.

Risk factors are broken up into two categories: modifiable and non-modifiable risks. The non-modifiable risk factors are included in each of the disease and risk factor models and include:

- Age
- Race
- Gender

Table 18 below shows the non-modifiable risk factors and the conditions that are most affected by them.

Table 18: Non-Modifiable Risk Factors and Associated Conditions

Risk Factor	Related Condition
Age	Arthritis
	High cholesterol
	Diabetes
	Coronary Heart Disease
Race	Diabetes
	Asthma
	Hypertension
	Coronary Heart Disease
Gender	Arthritis
	Asthma
	High cholesterol
	Migraine
	Coronary heart disease
	Depression

Modifiable risk factors include:

- Obese/Overweight
- High Cholesterol
- Lifestyle Choices
    - Poor eating habits

- Alcohol Consumption
- Activity Level
- Smoking Status

Models were created for each of these risk factors. Data for obese, overweight or smoking individuals were used for co-morbidity models.

Obesity/Overweight

The CDC reports that about one third of American's are obese [45]. Obesity is considered the condition where the amount of stored fatty tissue exceeds healthy limits, and is usually defined as having a body mass index (BMI) of 30kg/m^2 [115]. Obesity was classified as a disease by the CDC in the 1980s, however, for the purposes of these models, obesity is treated as a risk factor since it affects so many other diseases, and is easier to prevent, treat or manage than the other conditions.

Being overweight is considered to be less severe, and is defined as having more body fat than is optimally healthy. An overweight individual is one that has a BMI between 25 and 29.99 kg/m^2 [116]. About 64% of the US population is currently overweight.

These obese and overweight states are associated with a number of health conditions which are outlined in

Table 19 below.

Table 19: Obesity and Overweight with related conditions

Risk Factors	ICD-9 Code	Related Condition
Obesity	278.0 - 278.01	Arthritis
		Asthma
		Back pain (general)
		Cancer
		Coronary heart disease
		Occupational injuries
		Diabetes
		High cholesterol
		Hypertension
Overweight	278.02	Arthritis
		Depression
		Asthma
		Occupational injuries
		Cancer
		Diabetes
		High cholesterol
		Hypertension

High Cholesterol

Hypercholesterolemia, otherwise known as "High Cholesterol" is the presence of high levels of cholesterol in the blood. Cholesterol is a lipid found in the blood. Most of the cholesterol in the body is synthesized by the body, with tissues that have more abundant densely-packed membranes having more cholesterol present. These include the liver, spinal cord and brain. Since it is insoluble in blood it is transported within the circulatory system by lipoproteins. There is a large range of lipoproteins in the blood range:

- VLDL – Very low density lipoprotein
- IDL – Intermediate density lipoprotein
- LDL – Low density lipoprotein
- HDL – High density lipoprotein

LDL transports cholesterol and triglycerides from the liver to the tissues. It also regulates cholesterol synthesis at these sites. It is considered “bad cholesterol” since it is responsible for carrying cholesterol to the heart where it can be retained by proteoglycans starting the formation of plaques. This can lead to atherosclerosis and hence stroke, heart attack and other vascular diseases. LDL Measurements are as follows [49]:

LDL cholesterol = total cholesterol – HDL cholesterol – (0.20 × triglycerides)

Recommended Levels:

Table 20: Interpretation of LDL Levels

Level mg/dL	Level mmol/L	Interpretation
<100	<2.6	Optimal LDL cholesterol, corresponding to reduced, but not zero, risk for heart disease
100 to 129	2.6 to 3.3	Near optimal LDL level
130 to 159	3.3 to 4.1	Borderline high LDL level
160 to 189	4.1 to 4.9	High LDL level
>190	>4.9	Very high LDL level, corresponding to highest increased risk of heart disease

HDL proteins transport cholesterol from the tissues to the liver. HDL is considered "good cholesterol" since it is believed that it is able to remove cholesterol from within the arteries, thus reducing the risk of heart related problems.

Recommended Levels:

Table 21: Interpretation of HDL Levels

Level mg/dL	Level mmol/L	Interpretation
<40	<1.03	Low HDL cholesterol, heightened risk for heart disease, <50 is the value for women
40–59	1.03–1.52	Medium HDL level
>60	>1.55	High HDL level, optimal condition considered protective against heart disease

Men tend to have lower HDL levels than women. Men are also shown to have higher incidence of atherosclerotic heart disease.

Cholesterol may be regulated through the use of pharmaceutical products aimed at reducing LDL levels and/or increasing HDL levels. It has also been suggested that proper diet may also help manage cholesterol levels. Table 22 below highlights the conditions associated with high cholesterol.

Table 22: High Cholesterol and Associated Conditions

Risk Factor	ICD-9 Code	Related Condition
High cholesterol/High LDL/Low HDL/High Triglycerides	272.2 - 272.4	Coronary heart disease Occupational injuries Diabetes

Lifestyle Choices

Lifestyle choices which are also risk factors include smoking, poor eating habits, activity levels and alcohol consumption.

For this study we considered smokers to be those that currently smoke cigarettes daily, and not casual (or social) or former smokers. This might be a strict definition of smoking and could be relaxed in the future as more data is available.

Poor eating habits have been linked to a number of conditions, for example, diabetes, hypertension and heart disease. Poor diet has also been linked as a cause of other risk factors mentioned previously, including obesity and high cholesterol.

Activity levels have been shown to affect the weight of an individual hence reducing their risk of becoming overweight and/or developing weight related conditions. For this study we considered individuals who were not active at all. That is, individuals reporting that their jobs are not physically intensive and they do not participate in fitness programs or exercise on their own. This is a very strict definition and could be relaxed to "minimal" activity (include individuals who do some exercise, but still less than recommended) as more data becomes available.

Alcohol consumption can lead to a number of chronic conditions including cirrhosis of the liver, pancreatitis, cancer and high blood pressure. Alcohol consumption can also lead to accidental injury due to falling, car accidents and so forth, which can also be costly to employers. For this study we chose to look at individuals who drink approximately one drink per day (5 drinks per week). This might be a lose definition since individuals responding to an HRA may consider 5 drinks in a week a regular weekend consumption level.

Table 23 below shows the relationship between the aforementioned risk factors and some health conditions.

Table 23: Lifestyle Risk Factors and Associated Conditions

Risk Factor	Related Condition
Smoking	Hypertension
	Occupational injuries
	Peptic ulcer
	Coronary heart disease
	Asthma
	Back pain (general)
	Diabetes
	Cancer
Poor diet	Diabetes
	Cancer
	High cholesterol
	Hypertension
	Obesity
	Heart Disease
Lack of physical activity	Coronary heart disease
	High cholesterol
	Asthma
	Diabetes
Alcohol abuse	Hypertension
	Peptic ulcer
	Cancer

3.10 Risk Factor Outline

Risk factors may affect each other as well as other diseases. Disease may also be considered risk factors for other conditions, for example diabetes and heart disease.

Table 24 below provides and easy reference to the more common risk factors and the diseases associated with them. This list includes the risk factors and conditions previously mentioned as being in the model as well as other conditions that may be considered for future iterations of the model. Items with a star are used as inputs to the model. Items just in bold are currently model outputs.

Table 24: Risk Factors Quick Reference

Risk Factors	Attribute	Related Condition
* **Age**	Non-modifiable	Arthritis
		Back pain (general)
		High cholesterol
		Diabetes
		Coronary heart disease
Alcohol abuse	Modifiable	Hypertension
		Peptic ulcer
		Cancer
Diabetes	Modifiable	Coronary heart disease
* **Race**	Non-modifiable	Diabetes
		Asthma
		Coronary heart disease
		Hypertension
Family history/Genetics	Non-modifiable	Cancer
		Coronary heart disease
		Arthritis
		High cholesterol
		Hypertension
		Diabetes
* **Gender**	Non-modifiable	Arthritis
		Asthma
		High cholesterol
		Migraine
		Coronary heart disease
		Depression

Table 24: Risk Factors Quick Reference Cont.

Risk Factors	Attribute	Related Condition
High cholesterol/High LDL/Low HDL/High Triglycerides	Modifiable	Coronary heart disease
		Occupational injuries
		Diabetes
Hypertension	Modifiable	Coronary heart disease
		Diabetes
Lack of physical activity	Modifiable	Coronary heart disease
		High cholesterol
		Asthma
		Diabetes
Obesity	Modifiable	Arthritis
		Asthma
		Back pain (general)
		Cancer
		Coronary heart disease
		Occupational injuries
		Diabetes
		High cholesterol
		Hypertension
* **Occupation**	Modifiable	Asthma
		Occupational injuries
		Arthritis

Table 24: Risk Factors Quick Reference Cont.

Risk Factors	Attribute	Related Condition
Overweight	Modifiable	Arthritis
		Depression
		Asthma
		Occupational injuries
		Cancer
		Diabetes
		High cholesterol
		Hypertension
Physical disability	Non-modifiable	Occupational injuries
		Back pain (general)
		Cancer
		Diabetes
		Hypertension
Poor diet	Modifiable	Diabetes
		Cancer
		Obesity
		Coronary heart disease
		High cholesterol
		Hypertension
Smoking	Modifiable	Hypertension
		Occupational injuries
		Peptic ulcer
		Coronary heart disease
		Asthma
		Back pain (general)
		Diabetes
		Cancer

Table 24: Risk Factors Quick Reference Cont.

Risk Factors	Attribute	Related Condition
Environmental Exposure/Pollution	Modifiable	Asthma
		Cancer
Stress/Sleeping Disorders	Modifiable	Hypertension
		Occupational injuries
		Coronary heart disease
Using aspirin, ibuprofen or naproxen	Modifiable	Peptic ulcer
Been divorced	Non-modifiable	Depression
Change occupations	Non-modifiable	Depression
Contaminated injections in health care settings	Modifiable	Cancer
Ergonomic exposure (lifting heavy weights, bending back and forward)/Yard Work and Gardening	Modifiable	Back pain (general)
Hay fever	Modifiable	Asthma
Perceived body part discomfort	Modifiable	Occupational injuries
Phychological distress	Modifiable	Back pain (general)
Indoor smoke from household use of solid fuels	Modifiable	Cancer
Infection (passt)	Non-modifiable	Arthritis
Joint injuries (past)	Non-modifiable	Arthritis

CHAPTER 4 RESULTS

4.1 Estimation of Unknown Values

One of the main roadblocks to doing comprehensive disease analysis is the limited availability of data on certain disease conditions. Compound that with even less data on the number of people with two or more diseases and then it becomes harder to look at the diseases or risk factors that greatly affect certain diseases. The first hurdle the Bayesian model had to overcome was estimation of the missing variables. Since Bayesian simulation methods easily estimate missing variables this was not an issue for any of the disease or risk factor models run. The model took into account other input values within the same row (age) and column (race) and made estimations on what that missing value should be. An example of this is shown in

Table 25 and Table 26 below. According to the observed data in

Table 25 using traditional statistical methods, the prevalence of diabetes for white males 55-64 would be 19.2 %, blacks in the same age group would be at 8% and Hispanic males would either be 0% or inconclusive inaccurate results would be obtained. We know from previous studies that Hispanics and blacks are at a higher risk for diabetes than their white counterparts, and so even without missing data, traditional calculations may not correctly capture real observations. In this dataset, using classical statistical methods, some demographic groups would never be at risk for developing diabetes.

Table 25: Diabetes data with incomplete information

AGE	Diabetes	White	Black	Hispanic	Mixed	Other
18-24	0	135	13	22	8	2
	1	2	**0**	**0**	**0**	**0**
		137	13	22	8	2
25 - 34	0	445	53	111	44	19
	1	16	3	2	3	2
		461	56	113	47	21
35-44	0	661	61	85	51	17
	1	38	**0**	13	6	5
		699	69	98	57	22
45-54	0	690	90	54	34	17
	1	101	13	16	8	5
		791	103	70	42	22
55-64	0	303	63	13	14	5
	1	72	5	**0**	6	2
		375	26	14	20	7
>65	0	22	2	3	2	1
	1	9	1	**0**	1	**0**
		31	3	3	3	1

The following table (Table 26) shows that the Bayesian model, taking into account prior information, was able to estimate for the areas where observed data was incomplete. This allows for detailed observations by demographic group to be made.

Table 26: Bayesian Estimation

Age	Race	Lower	Median	Upper
18-24	White	0.005202	0.02856	0.1088
	Black	0.006516	0.01068	0.1837
	Hispanic	0.007409	0.01574	0.284
	Mixed	0.005396	0.01639	0.03863
	Other	0.02132	0.01519	0.05415
25-34	White	0.01681	0.05239	0.1238
	Black	0.007453	0.02349	0.05805
	Hispanic	0.01985	0.06128	0.143
	Mixed	0.02001	0.0782	0.2099
	Other	0.02519	0.03748	0.05345
35-44	White	0.03926	0.05418	0.07226
	Black	0.05096	0.1061	0.1884
	Hispanic	0.07072	0.1234	0.1965
	Mixed	0.04086	0.09435	0.1815
	Other	0.06755	0.1768	0.3479
45-54	White	0.05845	0.0737	0.09131
	Black	0.1047	0.1265	0.1507
	Hispanic	0.06794	0.1185	0.1884
	Mixed	0.1278	0.2091	0.3121
	Other	0.07893	0.164	0.2868
55-64	White	0.06973	0.1746	0.3445
	Black	0.1177	0.138	0.1597
	Hispanic	0.1513	0.1889	0.2295
	Mixed	0.04474	0.1227	0.2635
	Other	0.01307	0.06301	0.2016
>65	White	0.09768	0.2292	0.4217
	Black	0.03305	0.1504	0.4096
	Hispanic	0.1547	0.1892	0.228
	Mixed	0.1239	0.2432	0.4027
	Other	0.01992	0.1189	0.4333

4.2 Sub-Group Risk Assessment

Along with this estimation, the Bayesian model was also able to rank the specific combinations of diseases and so provide an easy way to identify the most affected populations using graphics. For a fixed disease and fixed sub-group it is possible to obtain a distribution of the rank; the ranking application in BUGS gives the probability of belonging to a particular class. The graphical representation is a histogram in which for a particular rank on the horizontal axis, the corresponding probability is given on the y-axis. The lower the rank placement, the "healthier" the group is, or the less likely to have a disease. The narrower the ranking distribution, the more likely it is for the group to fall in those ranks.

The figures 7 - 12 below show the rank outputs for one diabetes model. There are 276 subgroups separated by age, race, gender and smoking and obesity status. For a particular subgroup, the rank is a random variable and is estimated by the model. For the groups, rank 1 is the rank of the group which is the least susceptible to the disease, while the rank 276 denotes the group with the highest chance of developing the disease, therefore increasing ranks means increasing susceptibility to the disease.

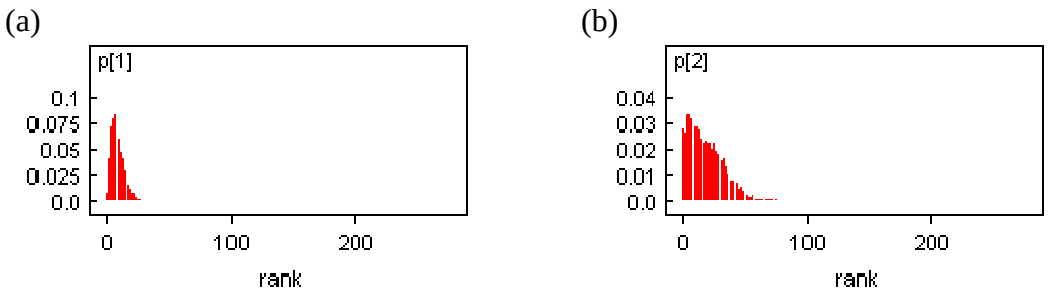

Figure 7: Rankings in Bugs, (a) Non-Obese, Non-Smoking White Males 18-24 (b) Non-Obese, Non-Smoking Black Males 18-24

(a) (b)

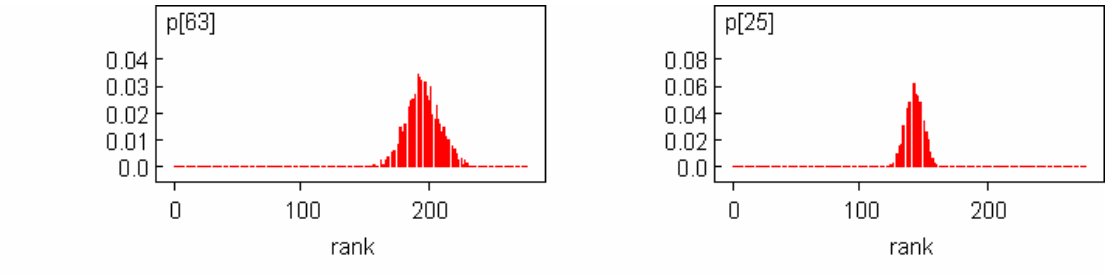

Figure 8: Ranking in BUGS, (a) Non-Obese, Non-Smoking Hispanic Females 55-64 (b) Non-Obese Non-Smoking White Males 55-64

(a) (b)

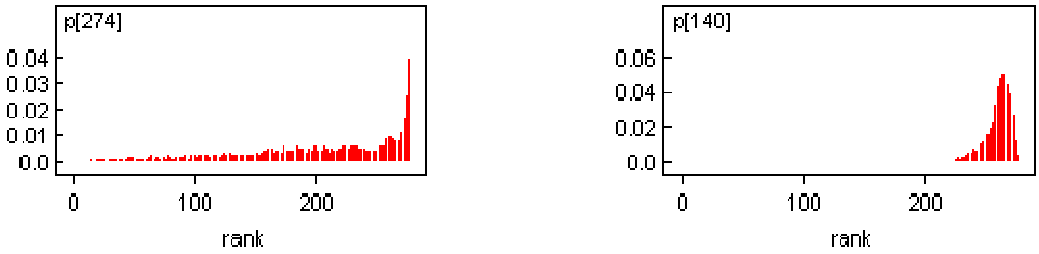

Figure 9: Ranking in BUGS, (a) Obese, Smoking Hispanic Females 55-64 (b) Obese Smoking White Males 55-64

The rankings show exactly what is expected in literature [49], with non-obese, non-smoking, young white males with the lowest chance of being susceptible to the disease. Non-obese, non-smoking young black males, while still at the lower end of the rank show a higher chance of being susceptible to disease as their ranking distribution is wider. The Bayesian model gives an alternate method of identifying the most at risk categories.

4.3 Comparison with Claims Data

Demographic information and claims data for 2005 and 2006 was obtained and used to test the performance of the model. The demographic information contained birthday, age, and health coverage eligibility information. The claims information had employee identifier numbers which allowed for the identification of individuals that made claims, but this information could not be tied back to the demographic file. It was then assumed that the claims information corresponded to the list of eligible employees given. Diseases were identified using the ICD-9 coding system. This information was needed to test the preliminary model's effectiveness.

The model was tested numerically and graphically in order to see how well the predicted values fit the observed data. Numerical methods involved the use of chi-squared tests according to the methods described earlier in section 2.6. Graphical methods involved comparing the claims value with the predicted 95% credible set obtained from the model. This method was also necessary since, as discussed in section 2.6, chi-squared tests assume independence of each condition, and it is not always the case that the diseases behave independently.

Chi-Squared Test

For the state of Texas, the calculated χ^2 value was 3.7949. There were 11 diseases used in the comparison. These included allergy, arthritis, asthma, diabetes, heart disease, hypertension, migraine, prostate, breast and skin cancer, and respiratory disease. Since there were 11 diseases compared, the inverse χ^2 value with with α=0.05 and $k\text{-}1$ = 10 was 3.9403. This value is larger than the 3.79 calculated.

For the state of Michigan, the calculated χ^2 value was 11.477. There were 11 diseases used in the comparison. The disease which contributed to the greatest disparity was migraine. When this disease was removed, and the χ^2 test re-run for 10 diseases, the value obtained was 3.054, which is less than the 3.3251 value for 9 degrees of freedom.

For the state of Mississippi, the calculated χ^2 value was 54.51. There were 12 diseases used in the comparison. These included allergy, arthritis, asthma, diabetes, heart disease, hypertension, migraine, prostate, colon, lung and skin cancer, and respiratory disease. This is much larger than the inverse χ^2 value calculated for 11 degrees of freedom. Heart disease, migraine and hypertension were the largest disparities in the group. Removing those diseases, the χ^2 value with 9 degrees of freedom became 2.155 lower than the inverse χ^2 value of 2.732.

For the state of Tennessee, the calculated χ^2 value was 105.243. There were 13 diseases used in the comparison. These included allergy, arthritis, asthma, diabetes, heart disease, hypertension, migraine, prostate, breast, colon and skin cancer, and respiratory disease. This value is much larger than the inverse value of 5.226. The diseases which contributed the most to this large number were allergy, asthma, migraine, heart disease and hypertension.

Graphical Comparison

In comparing the data to the claims data for Company A, the following results were obtained. The first graph shows the overall comparison between the model and the total claims data. The second set of graphs shows the comparison by disease for each of the four states. The NCQA quality calculator value added to the results for comparison

for the diseases it also modeled for (asthma, diabetes, heart disease, hypertension).

The final set of graphs for Company A. looks at the claims count by disease for each state and compares it with the model prediction.

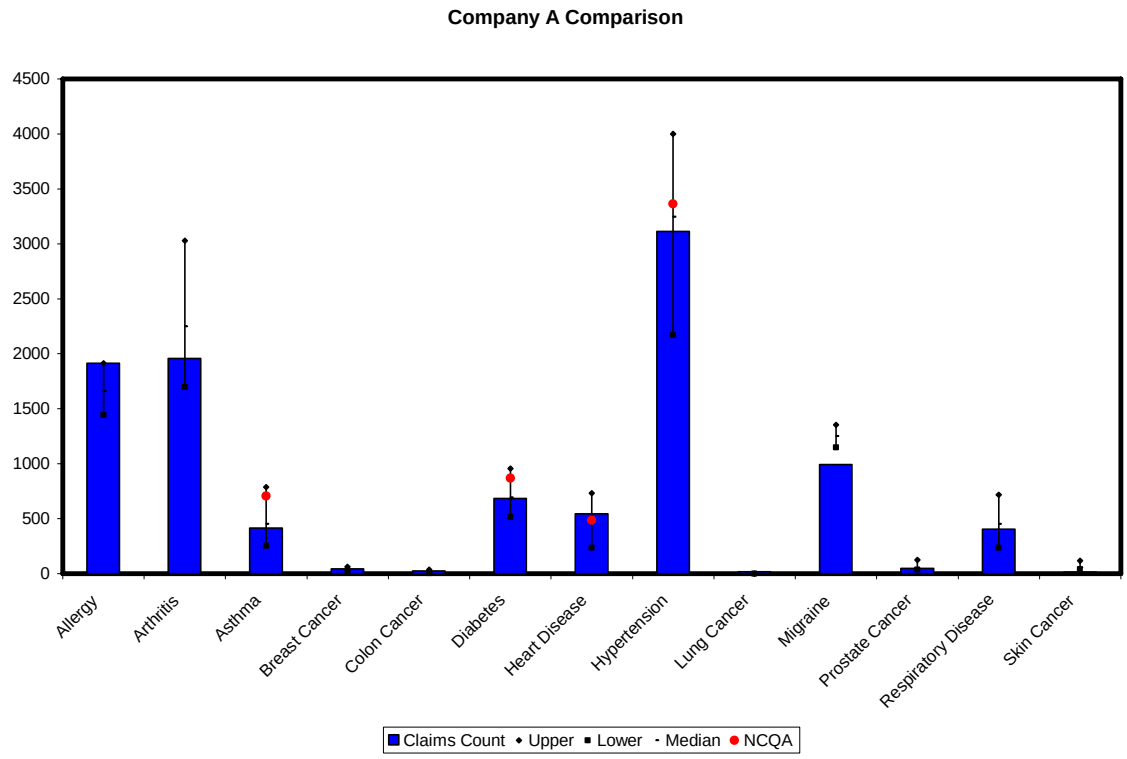

Figure 10: Comparison of Company A with claims and NQCA data

The above graph shows most of the conditions falling within the confidence intervals, with the exception of migraine. In the case of allergy, the claims value is very close to the upper bound of the model estimation. For the other disease states, the claims count fell closer to the median value. The following graphs will show each disease in

more detail. They show the comparison between each condition and the claims count by state.

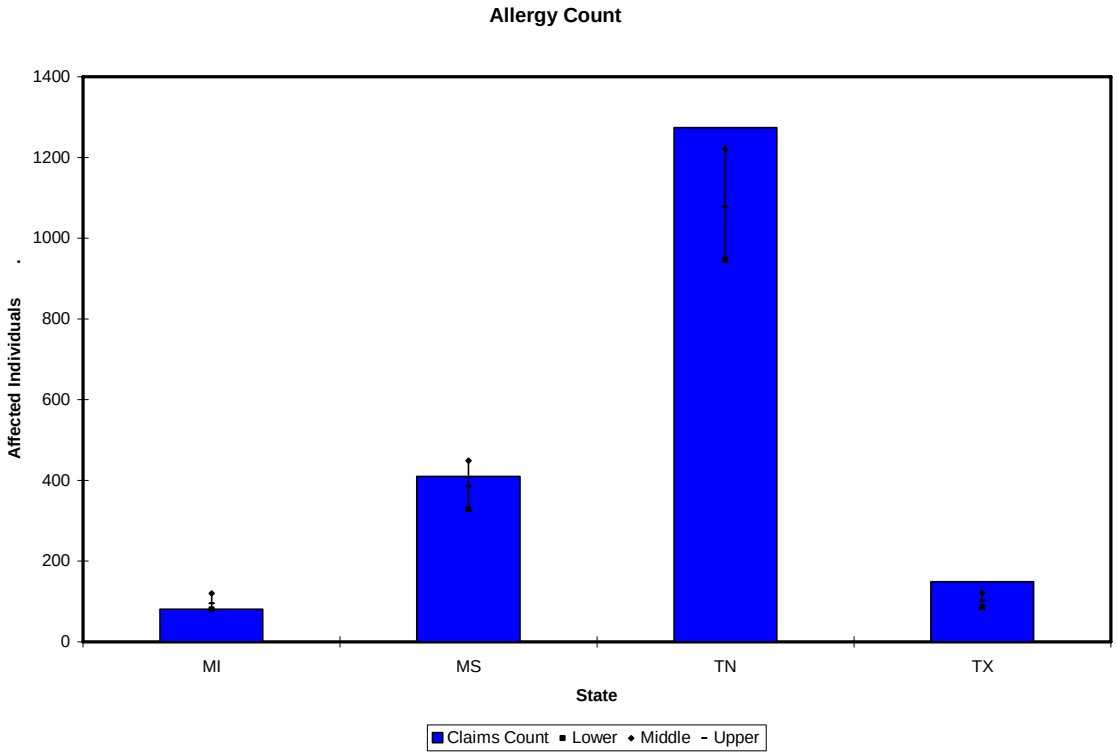

Figure 11: Company A. Allergy Claims Count vs. Model Output

The allergy graph shows the claims data falling within the confidence interval for two states and then being greater than the predicted value in two other states.

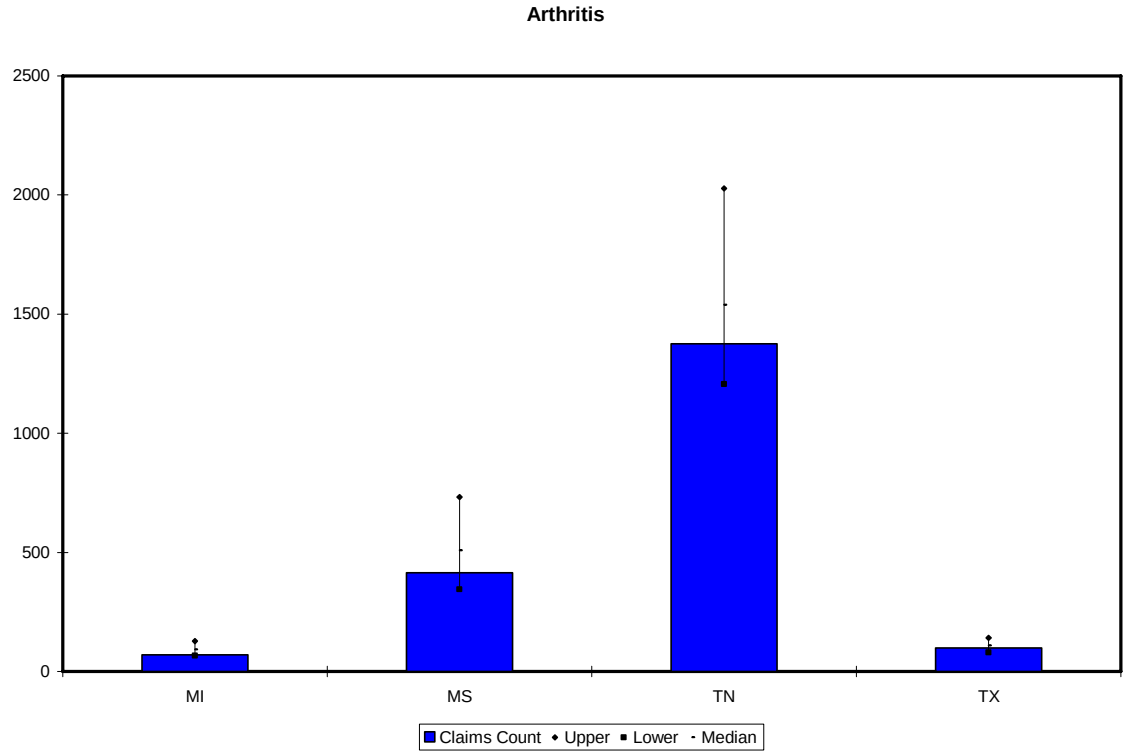

Figure 12: Company A. Arthritis Claims Count vs. Model Output

The arthritis graph shows the actual data falling within the credible set of the model's predicted value, but generally less than the median value.

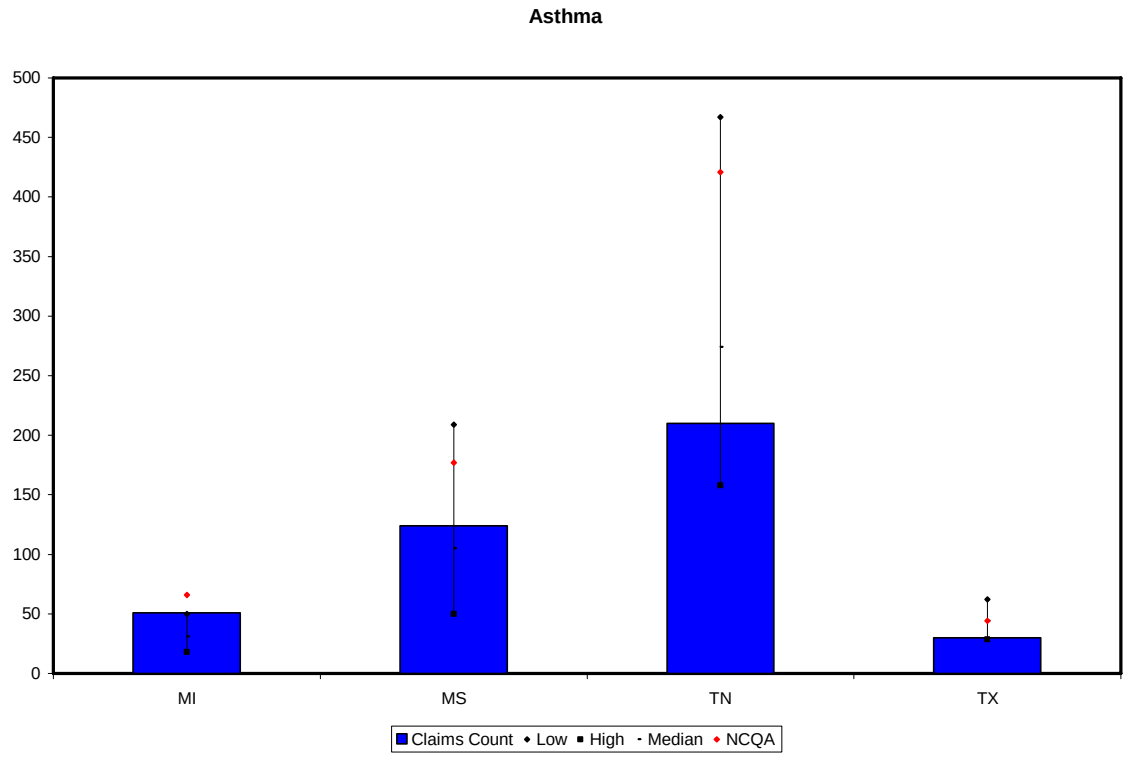

Figure 13: Company A. Asthma Claims Count vs. Model Output

In general the asthma model falls within the credible set. For one state the claims count is just at the upper bound of the model. The model predicted mean falls below the claims data for two states and above for two states.

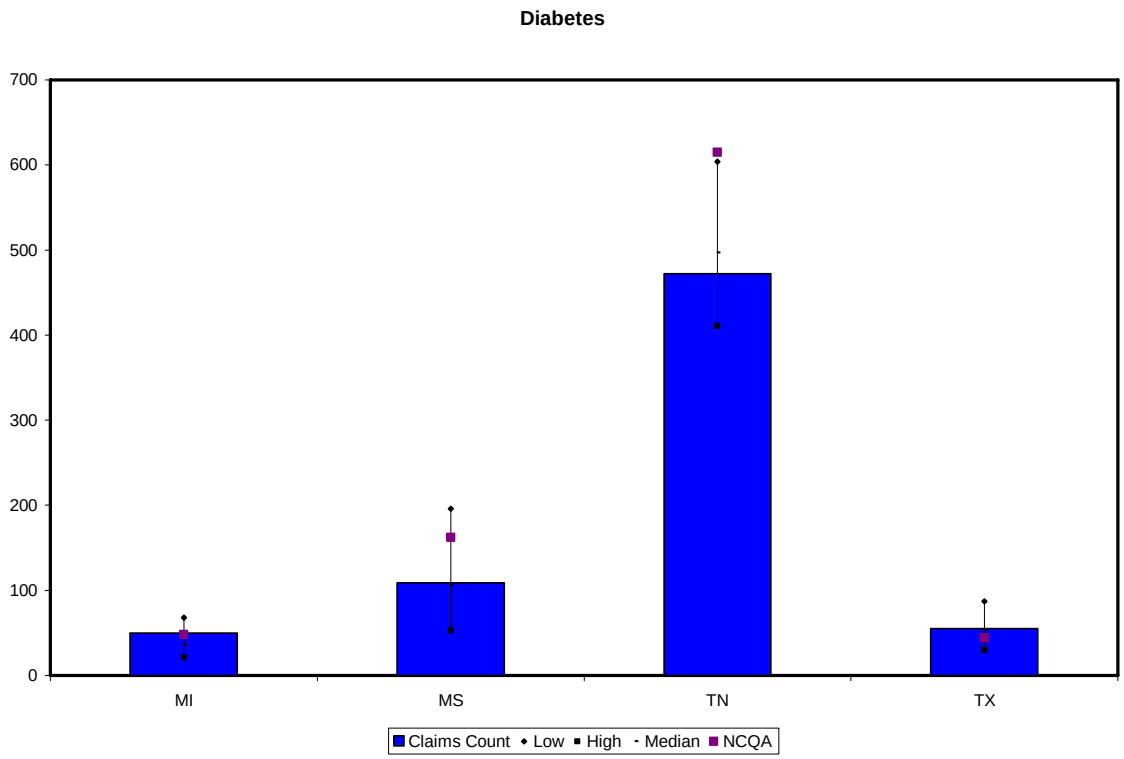

Figure 14: Company A. Diabetes Claims Count vs. Model Output

The claims data falls within the predicted credible set for the model. The median value is close to the observed value, falling higher for Tennessee.

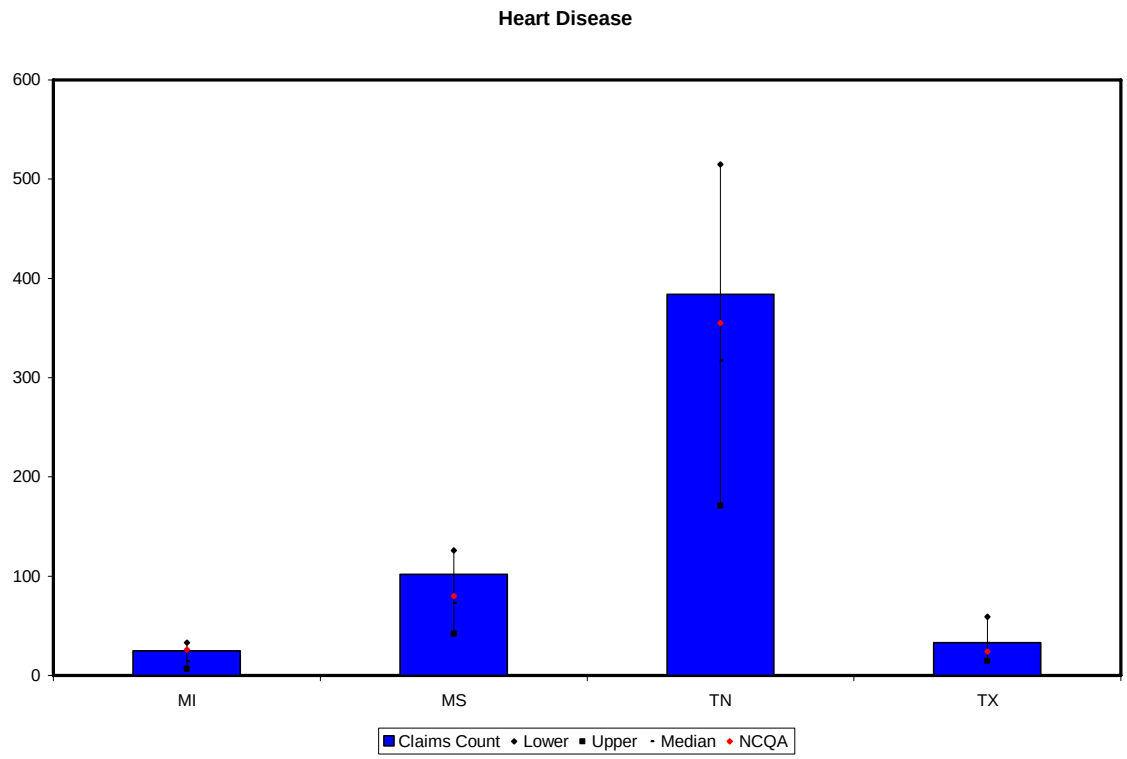

Figure 15: Company A. Heart Disease Claims Count vs. Model Output

The heart disease comparisons show heart disease falling generally lower than the claims data, however it as always within the credible set.

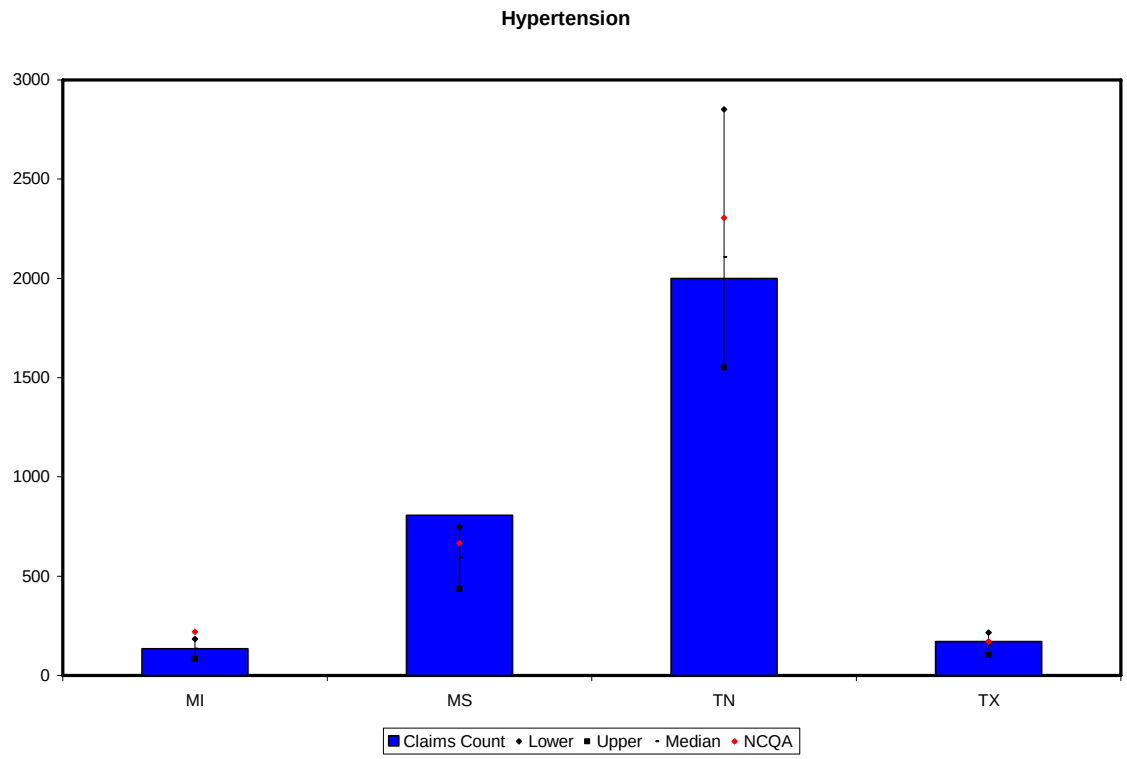

Figure 16: Company A. Hypertension Claims Count vs. Model Output

The hypertension graph shows the claims values falling between the credible set for three of the states. For one state the claims estimation is greater than the predicted upper bound.

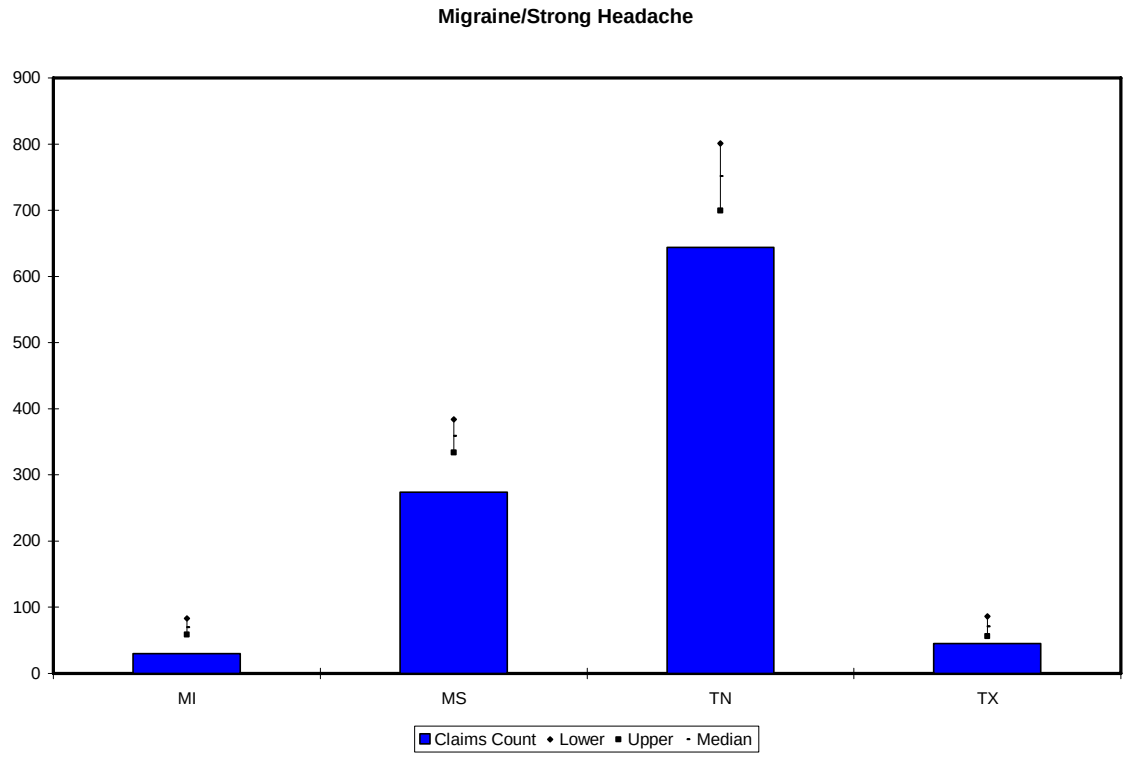

Figure 17: Company A. Migraine Claims Count vs. Model Output

For migraine, the predicted values were larger than the actual migraine claims data.

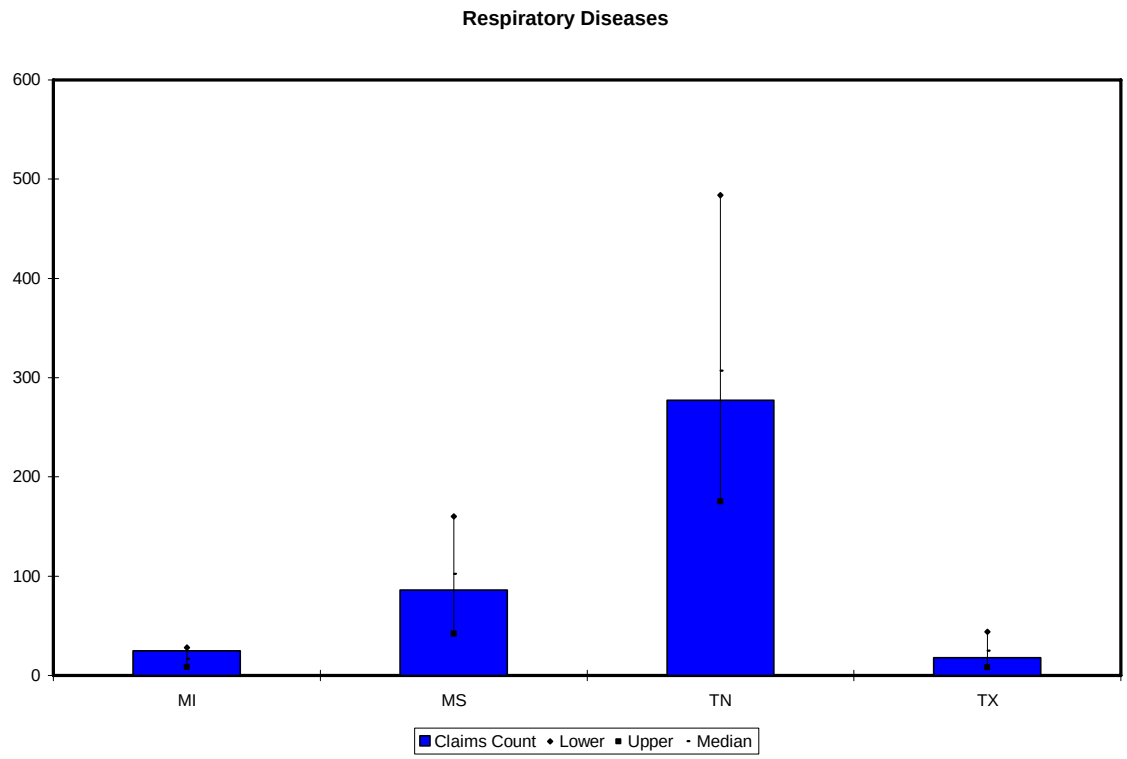

Figure 18: Company A. Respiratory Claims Count vs. Model Output

For respiratory diseases, the claims data falls within the credible set for all states. The median value is greater than the claims data value in three instances, but lower for one state.

The following graphs show the comparison of the model output with the claims data from each state by gender male (M) and female (F).

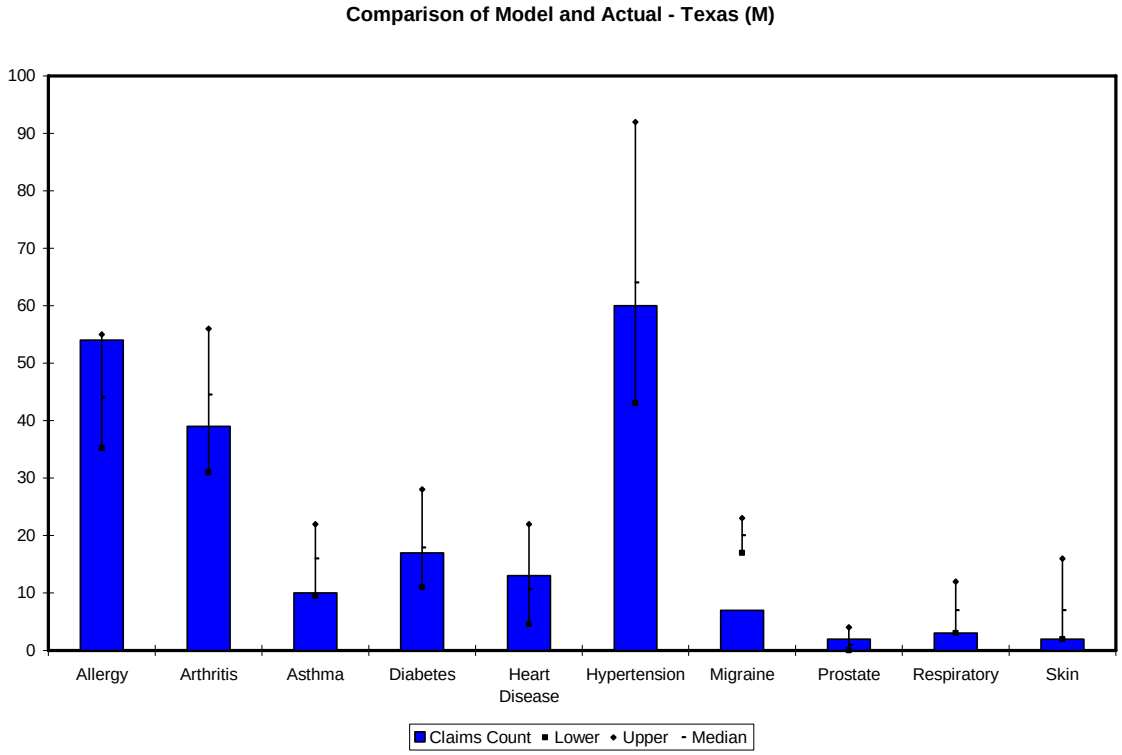

Figure 19: Company A. TX (M) Claims Count vs. Model Output

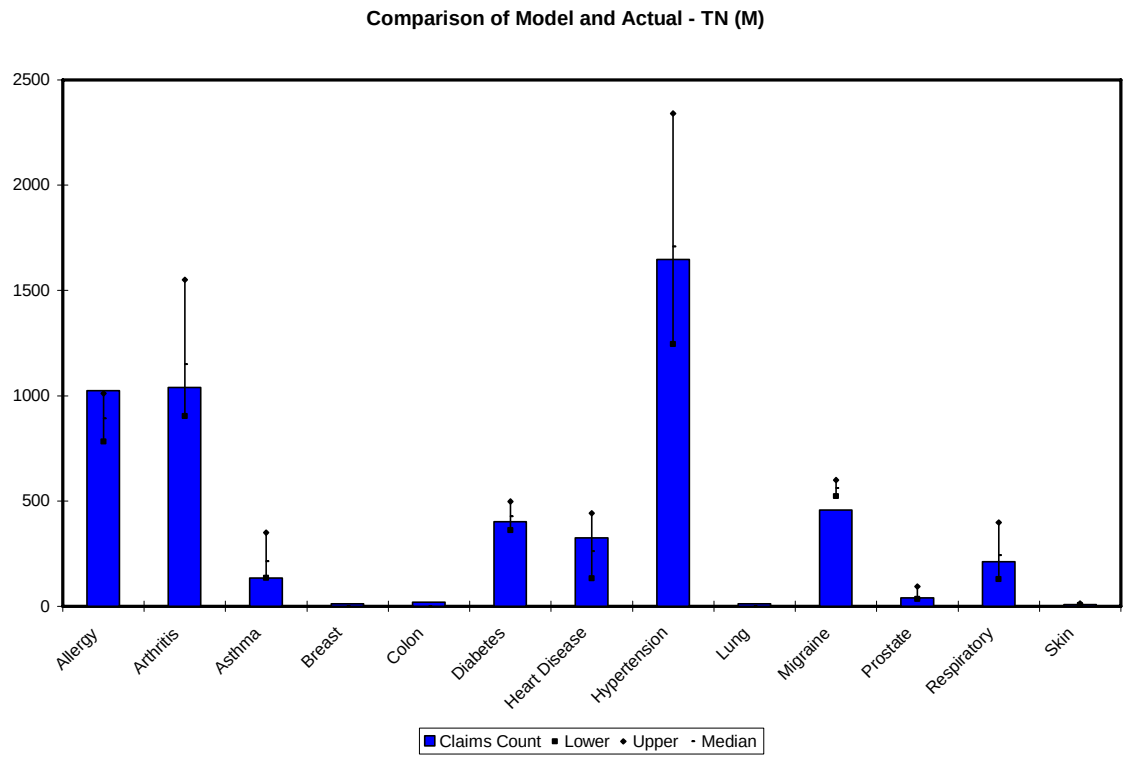

Figure 20: Company A. TN (M) Claims Count vs. Model Output

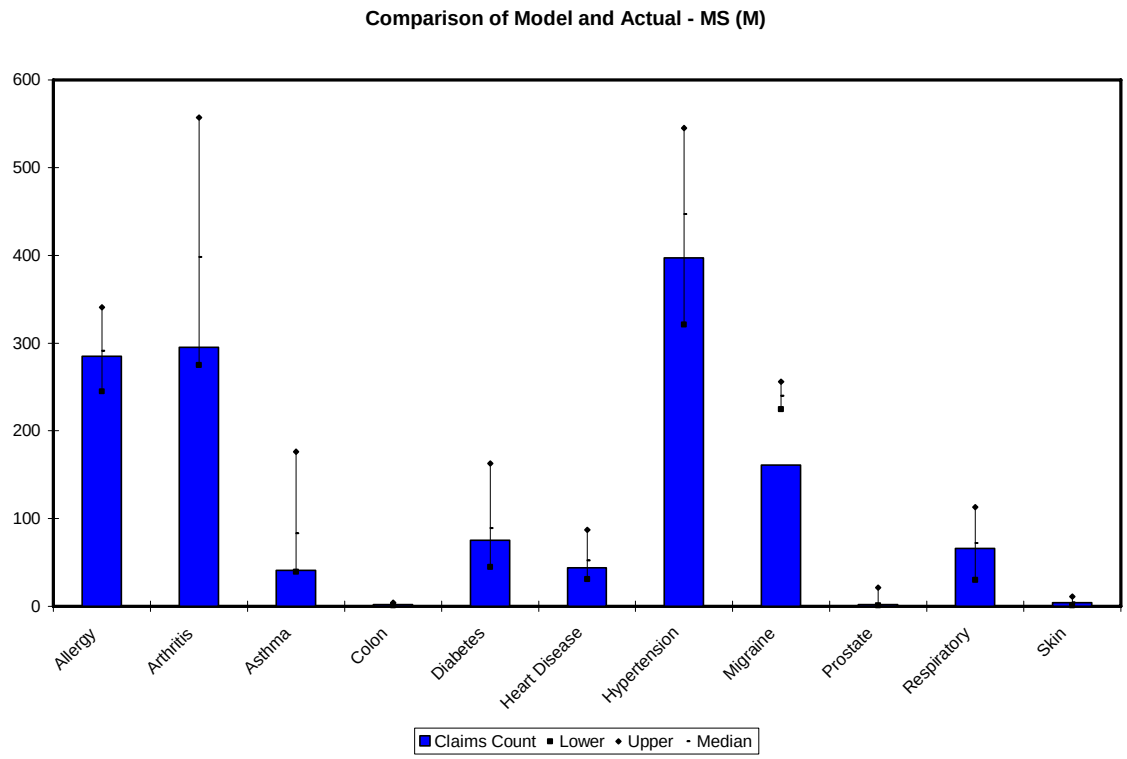

Figure 21: Company A. MS (M) Claims Count vs. Model Output

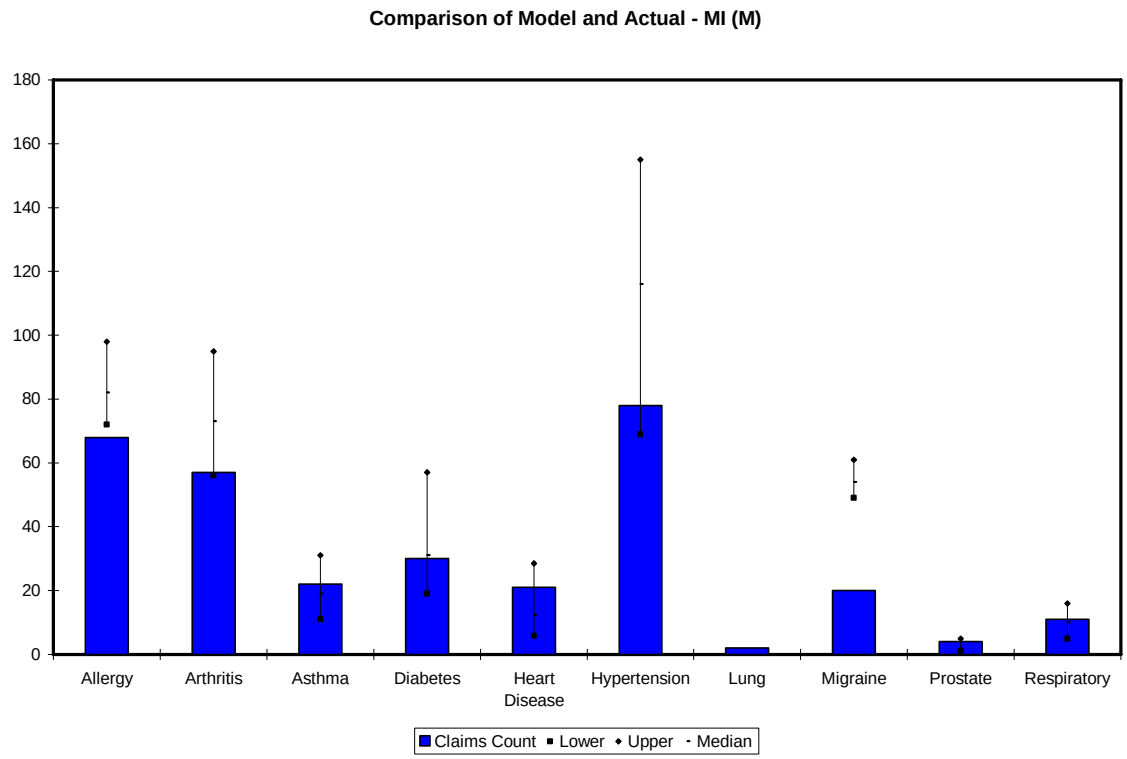

Figure 22: Company A. MI (M) Claims Count vs. Model Output

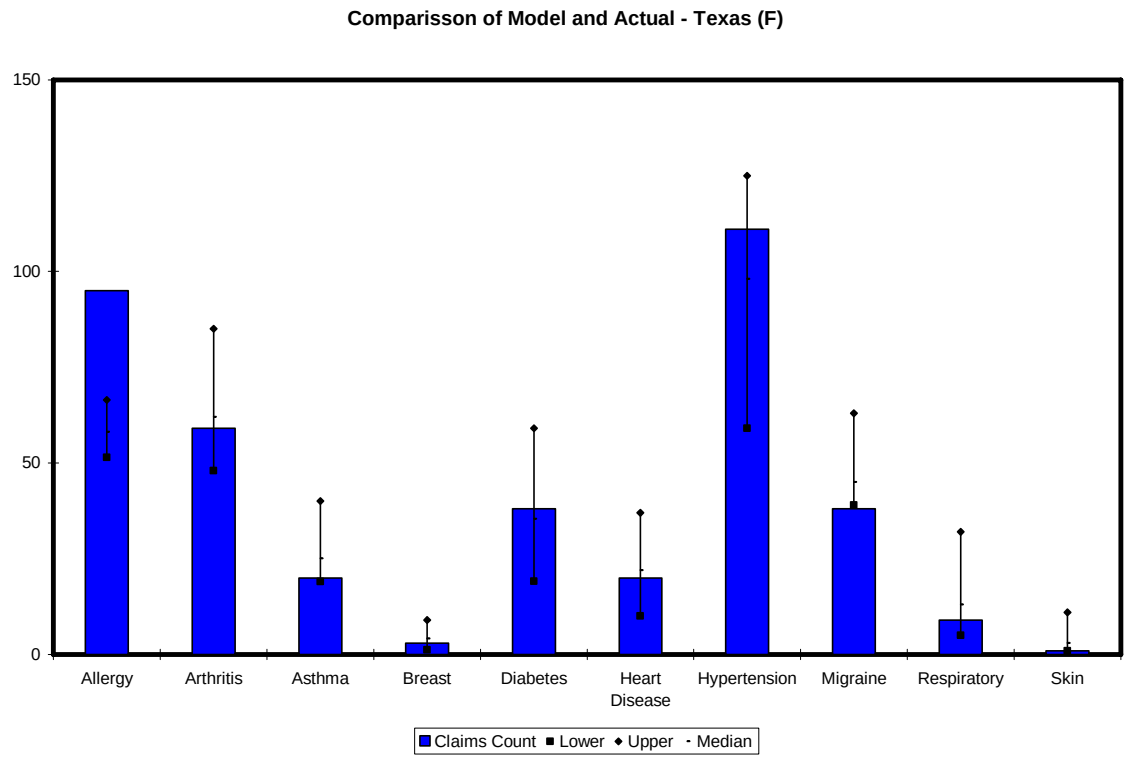

Figure 23: Company A. TX (F) Claims Count vs. Model Output

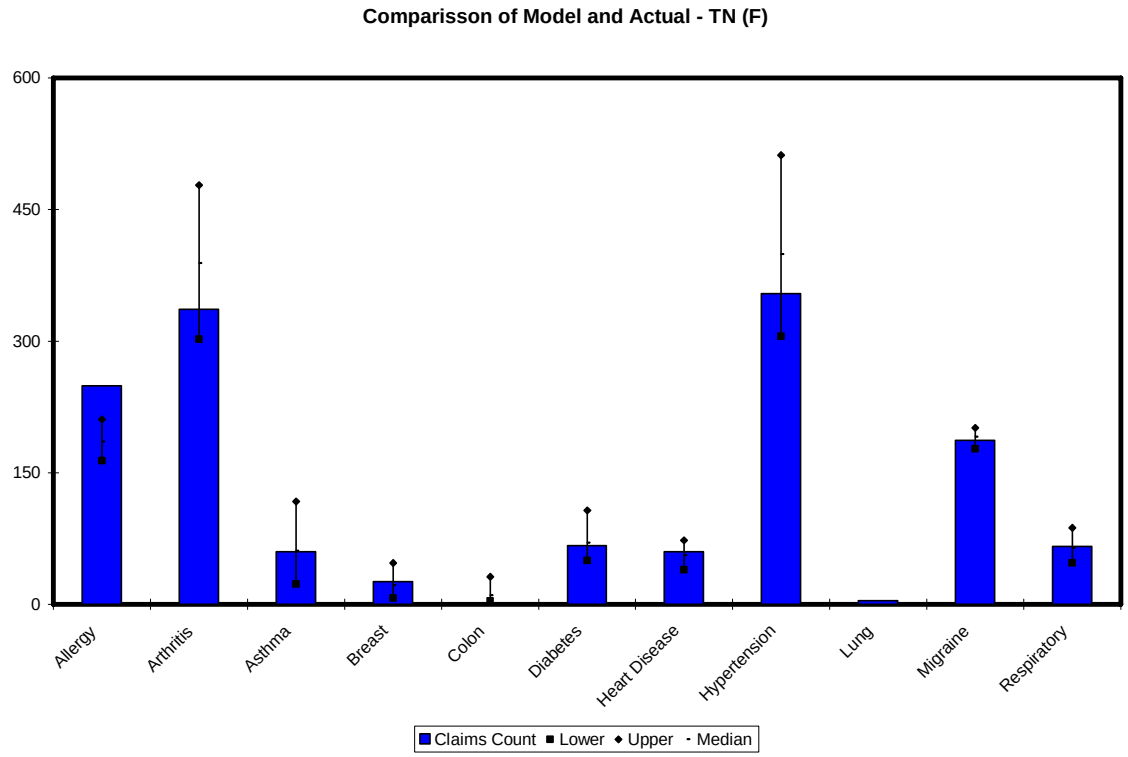

Figure 24: Company A. TN (F) Claims Count vs. Model Output

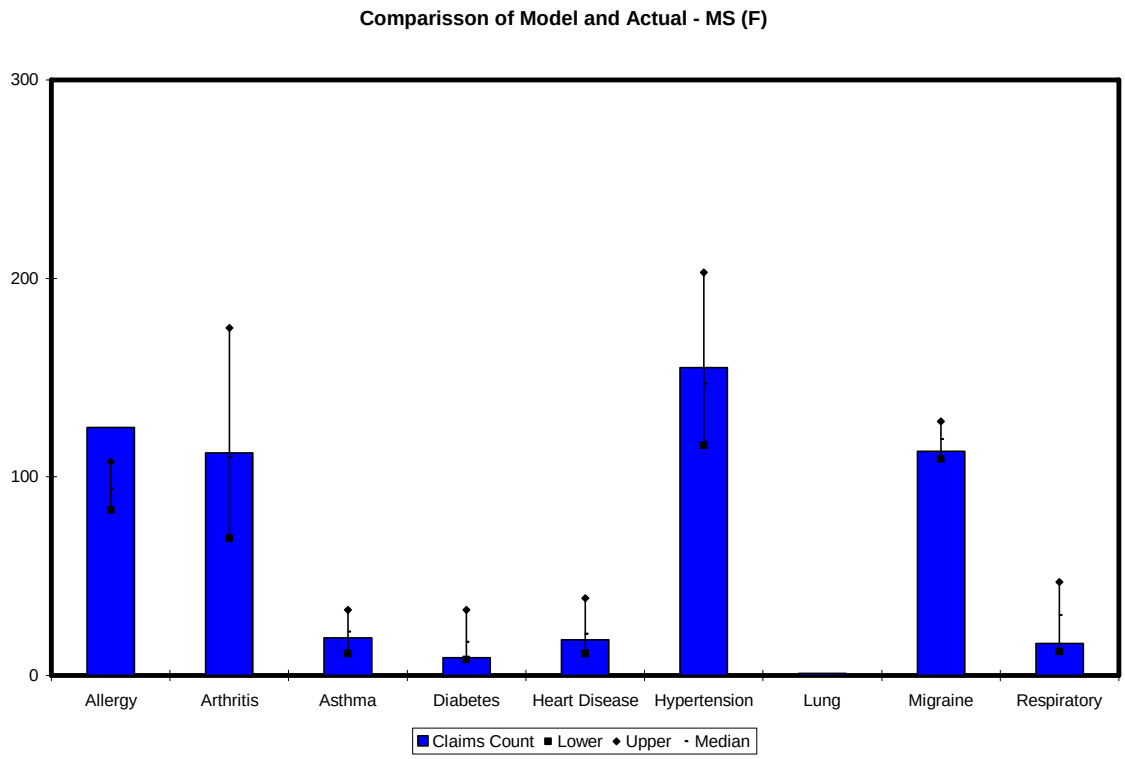

Figure 25: Company A. MS (F) Claims Count vs. Model Output

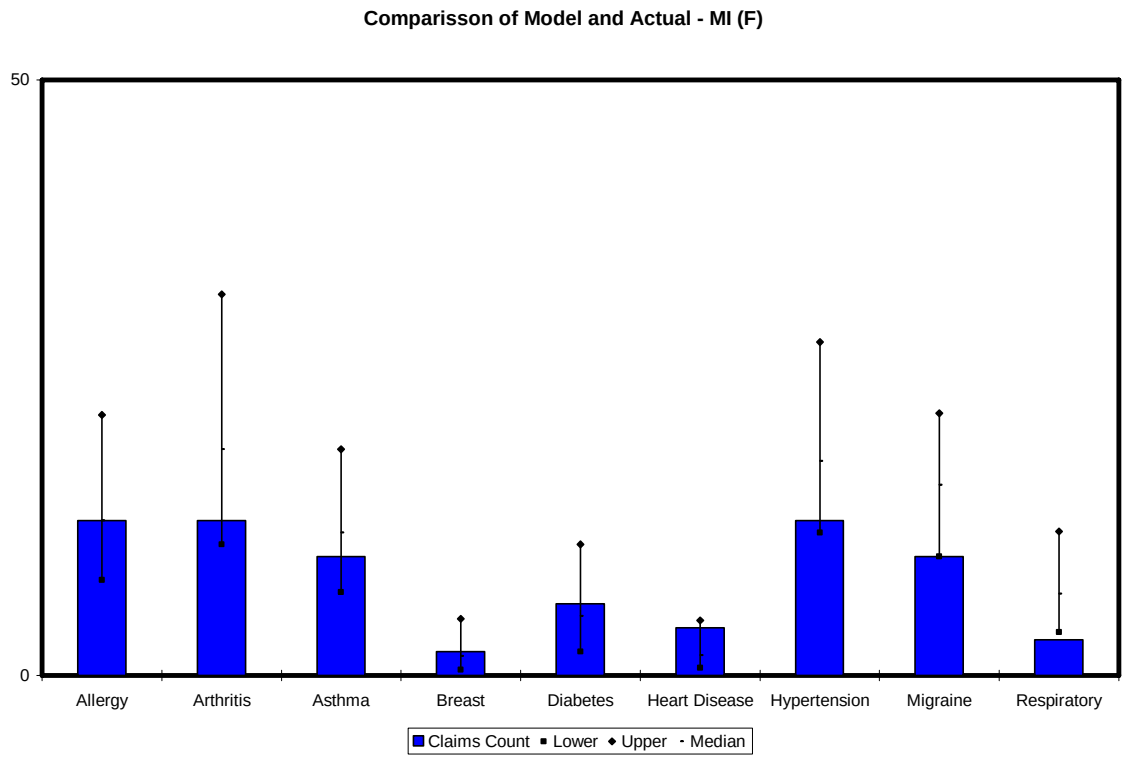

Figure 26: Company A. MI (F) Claims Count vs. Model Output

The following graphs show the output comparisons for Company B.

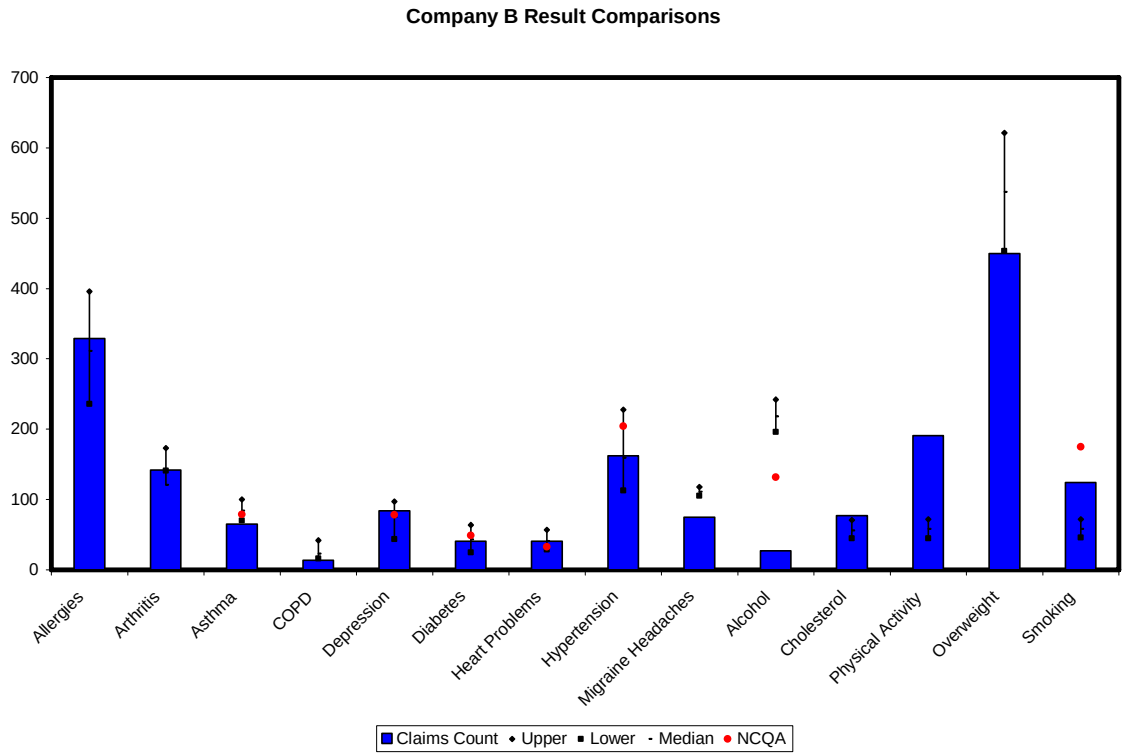

Figure 27: Company B. Company Data and NCQA Estimations vs. Model Output

The results from each of the models are fairly consistent with the disease estimations. The risk factor values however fall outside of the credible set. These disparities will be addressed in the discussion section of this document.

The following graph shows the comparison between the claims data of Company C and the model output. The NCQA predicted values are also included.

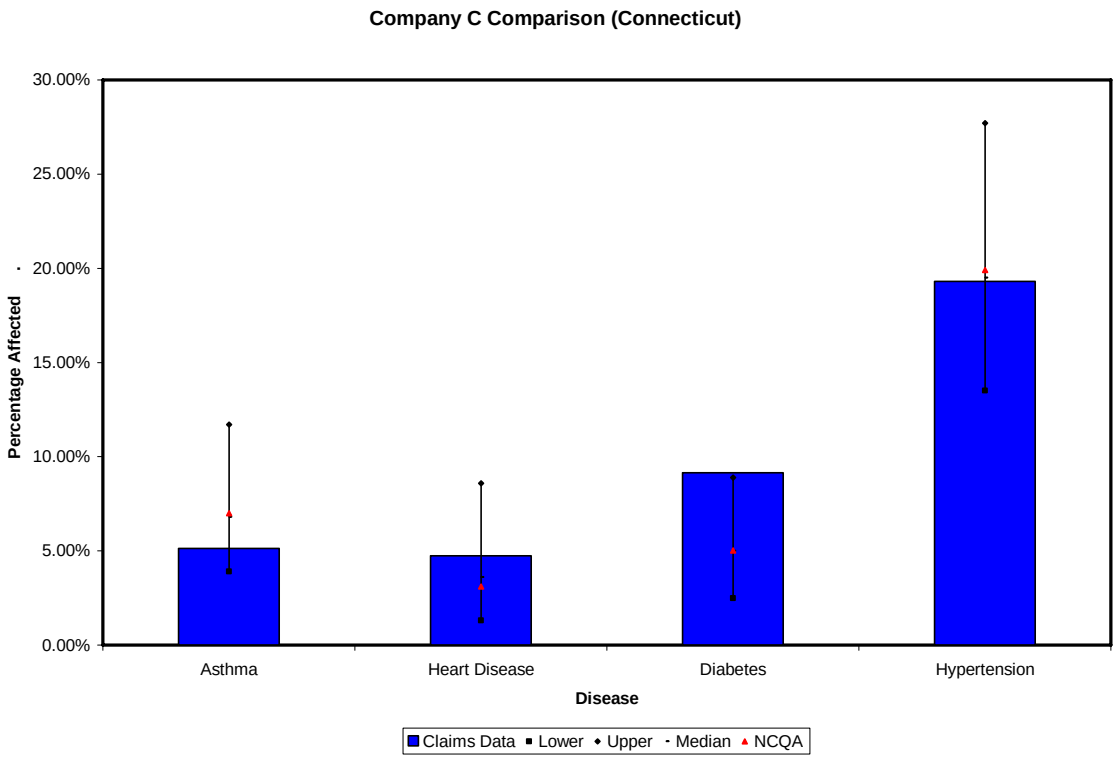

Figure 28: Company C. Claims Count vs. Model Output

The claims data falls within the credible set for most of the diseases. For diabetes, the claims data is larger than the upper bound of the credible set.

4.4 Co-Morbidity

The data were run to see the co-morbidity between certain diseases. The disease interactions that were investigated were:

- Diabetes and Heart Disease
- Diabetes and Hypertension
- Diabetes, Heart Disease and Hypertension

These were chosen due to the availability of data. Another disease combination that would have been interesting to analyze would have been asthma and allergies, but data pertaining to allergies was limited and so could not be investigated at this time.

In terms of risk factors, the risk factor/disease combinations which were investigated were:

- Obesity and diabetes
- Obesity and heart disease
- Obesity and hypertension

Disease combinations have even less available data than single diseases and so running the model with co-morbidities allowed us to demonstrate further the capabilities of Bayesian simulation.

It was decided when running this model that we would look at the percentage of people with a given condition that also had another condition, rather than looking at the percentage of people in the total population that had both conditions. For example, in the case of diabetes and hypertension, the model was designed to estimate the number of diabetics that also had hypertension, and not the number of people overall that currently had both conditions. This was due to the usefulness of the data for disease management policy as it would more clearly show the connection between the two conditions.

As with the previous single disease and risk factor models, the co-morbidity prevalence rates were compared to a given test set and the results were comparable. Unfortunately we were not able to obtain claims data that would allow us to check how the model predicted an actual company's disease burden.

The table below shows the rate of hypertension in diabetics nationally. In general the rate is high, with over 50% of diabetics also having hypertension once they are in the higher age risk category. When compared with the general prevalence rate of hypertension we see that the rate for diabetics is significantly higher than those in the general population (p-value <.001). This confirms what is generally known about the two diseases, however, the Bayesian model allows for further investigation to take place.

Table 27: Prevalence rates of Hypertension in Diabetics

			Diabetes/Hypertension			Hypertension		
Age	Sex	race	2.50%	median	97.50%	2.50%	median	97.50%
18-24	male	white	0.1893	0.2152	0.2403	0.0537	0.05566	0.0575
18-24	male	black	0.1985	0.2234	0.2487	0.05738	0.05938	0.06131
18-24	male	Hispanic	0.2055	0.2319	0.2594	0.06097	0.06332	0.06567
18-24	male	mixed	0.2112	0.2406	0.2715	0.06464	0.06752	0.07053
18-24	male	other	0.2155	0.2494	0.2855	0.06833	0.07198	0.07587
25-34	male	white	0.2808	0.3065	0.3308	0.09588	0.09841	0.1008
25-34	male	black	0.2926	0.3168	0.3415	0.1021	0.1047	0.1072
25-34	male	Hispanic	0.3012	0.3274	0.3541	0.1081	0.1113	0.1146
25-34	male	mixed	0.3066	0.3379	0.3695	0.114	0.1183	0.1227
25-34	male	other	0.3113	0.3491	0.3881	0.1201	0.1256	0.1315
35-44	male	white	0.3937	0.4159	0.4371	0.1651	0.1682	0.171
35-44	male	black	0.4068	0.4276	0.4487	0.1748	0.178	0.1811
35-44	male	Hispanic	0.415	0.4396	0.464	0.184	0.1883	0.1927
35-44	male	mixed	0.4201	0.4516	0.4826	0.193	0.199	0.2053
35-44	male	other	0.4233	0.4639	0.5026	0.2021	0.2102	0.2188
45-54	male	white	0.5163	0.5345	0.5521	0.2687	0.2724	0.2761
45-54	male	black	0.5288	0.5465	0.564	0.2822	0.2863	0.2903
45-54	male	Hispanic	0.5355	0.5585	0.5807	0.2948	0.3006	0.3064
45-54	male	mixed	0.5395	0.5707	0.5989	0.3071	0.3152	0.3237
45-54	male	other	0.5427	0.5825	0.6189	0.3193	0.3302	0.3416
55-64	male	white	0.6322	0.6494	0.6658	0.4042	0.4096	0.4148
55-64	male	black	0.6431	0.66	0.6775	0.4205	0.4263	0.4319
55-64	male	Hispanic	0.6487	0.6709	0.6924	0.4356	0.4432	0.4508
55-64	male	mixed	0.6524	0.6814	0.7089	0.4501	0.4602	0.4705
55-64	male	other	0.6551	0.692	0.7254	0.4644	0.4773	0.4907
>65	male	white	0.7312	0.7488	0.767	0.5551	0.5623	0.5696
>65	male	black	0.7399	0.7576	0.7765	0.5717	0.5792	0.5865
>65	male	Hispanic	0.7453	0.7664	0.7877	0.587	0.5958	0.6045
>65	male	mixed	0.7491	0.775	0.8001	0.6012	0.6123	0.6232
>65	male	other	0.7514	0.7836	0.8128	0.6151	0.6285	0.6418

The same observations are made with the rest of the combinations. As with the earlier model outputs, the Bayesian outputs also have a graphical ranking of the demographic groups to show which group was the most at risk for the having both conditions.

4.5 Highlights of Possible Research Areas

An interesting by product of being able to run models resulting in data for small demographic groupings is the ability to see results which may lead to further investigations. One interesting finding after looking at the obesity-diabetes results, we find that obese individuals have a prevalence of diabetes their non-obese counterpart. Across the demographic groups we see that for non-obese individuals, the rate of diabetes in Hispanic females is higher than their white male counterpart somewhat uniformly. However, when you look at the rates between obese groups, the gap between the prevalence of diabetes shrinks with increasing age group. This could probably mean that age is a greater risk factor than race, and seems to magnify the effect of obesity as the individual ages.

Table 28: Small comparison of groups with Diabetes

	Hispanic Females		White Males	
Age	**% Non-Obese**	**% Obese**	**% Non-Obese**	**% Obese**
18-24	3%	3.29%	1%	1.10%
24-34	4%	7.20%	1%	2.60%
35-44	5.50%	14.70%	1.60%	6.30%
45-54	7.30%	16%	3.17%	13%
55-64	11.25%	20%	6.54%	21%

CHAPTER 5 DISCUSSION

This section will discuss the results and the effectiveness of the model. Section 4.1 showed that the model was able to predict the missing values, and also predict values for prevalence which are consistent with expert data.

Section 4.2 shows the risk assessment properties of Bayesian analysis and the ability to create a distribution for the rank of the susceptibility of an individual to a particular disease. One feature that classical models cannot provide is the posterior distribution of the health ranking of the particular subgroups.

Section 4.3 shows the predicted values from the model fit the claims data for some diseases with the computed χ^2 value falling below the inverse χ^2 value. Migraine was the condition which did not align with three of the four states tests. Hypertension and heart disease were also outside the χ^2 range for Mississippi and Tennessee. This could be due to errors in either under-reporting for the company, or errors in the training data. When the model was tested for these states against the BRFSS database the model was consistent with observations made through surveys. The model values were closer to the observed office values than the NCQA estimates. The NCQA model uses similar datasets and so this would indicate that there may be some disparity between what is observed in claims data and what is observed through surveys for some of these diseases. As more claims data becomes available, this theory could be tested.

For the graphical comparisons, the credible sets appear to vary by state. This could be alleviated by:

(i) Observing data from more states to see if this trend continues across all states.

(ii) Flat priors were used in many cases which would affect the confidence intervals.

(iii) Given more data there would be more information to create informative priors which would narrow confidence intervals

In general, for the graphical comparisons, the observed data fell within the credible set for the model. One notable exception is migraine headaches. In Figure 17 and Figure 27, the predicted values for migraine are larger than the observed values. In pulling the claims data, the migraine information recorded was for strict migraine, and did not take into account frequent or severe headache, so the predicted values were expected to be higher than the observed value for the migraine comparisons.

As mentioned in section 3.8, the data used for allergic rhinitis was national data and so this affected how well the model prediction fit the observed data. It is expected that certain states will have higher allergic incidences than others. Mississippi, Texas and Tennessee both have cities in the top 20 list of worst spring allergy cities [117]. This information can be seen in the model results with the median predicted values falling below the observed value for the high allergy states.

The asthma predictive model run consistently higher than the claims data obtained from Company A. The result however, is still within the confidence interval, and also is lower than the estimations from NCQA. The difference in claims information could be attributed to a lack of consistency between data sources. Calculations of a person's current asthma status may include individuals that were at one time diagnosed, but no longer experience symptoms of the condition. For some states, the claims data only included the medical claims and not the pharmaceutical claims data as well, and so may

not have included individuals who have asthma, but did not see a professional about their condition that year.

Differences in definition could account for the disparity between the model and the obtained values. Our model assumed individuals that are at sedentary or very low levels of physical activity, whereas the given data takes into account all individuals that are below the recommended levels of activity. A similar statement could be made for the comparison of heavy drinkers. The data received considers heavy drinkers that regularly drink more than 2 drinks a day. Our model looks at individuals that have more than 5 drinks a week. Since this value could be exceed by drinking a glass of wine a day, or a weekend of social activity, then many individual qualify as heavy drinkers. This definition may need to be revisited.

The data received classes an individual as being overweight if they have a BMI of 27.5 or higher. For our estimations, obese individuals are classed as having a BMI >30 and overweight if they have a BMI between 25 and 30. Since our BMI value for overweight is less than the given values it would be expected that our model overestimates the number of overweight individuals.

In comparison to the NCQA calculator, the median values obtained by the model were closer to or the same as the observed values than the predicted values of the NCQA calculator. The exception to this was the estimate of heavy drinkers and smokers in company B (Figure 27).

At the end of section 4 it was also seen that it was possible to model co-morbidities. This took advantage of the ability of Bayesian modeling to estimate missing data since these are usually smaller data sets. The results in section 4.4, show that the

number of individuals with hypertension is significantly higher amongst diabetics than the general prevalence rate. Similar information can be garnered from the co-morbidity studies which can give actual prevalence rates across demographic groups.

As discussed in the background, employers are looking at ways to reduce healthcare costs and improve the productivity of their workforce. This model is able to predict the disease burden and risk factors of company employees. The ranking tools and co-morbidity capabilities allow employers to see how risk factors directly affect their health and wellness of their employees, which in turn may be used to institute company healthcare policies.

CHAPTER 6 CONCLUSION

This research has shown that using Bayesian methods, models can be created to estimate the disease burdens for corporations. These models, which are capable of looking at the disease burdens in small data sets, also allow for more insightful analysis of disease effect and disease interactions. The disease models can help in creating start points for future disease research, and also help in the analysis of diseases which occur less frequently and so where wrong inferences could be made due to missing or incomplete information. The model is useful in its ability to predict health burdens. This research has shown it is able to:

1. Estimate missing data
    a. The Bayesian models were able to estimate prevalence rates for diseases and risk factors with even limited data input allowing for estimations to be made for even lesser studied diseases and risk factors. This allowed for more diseases to be modeled making this method more comprehensive than the current NCQA model.
2. Incorporate any available prior information
    a. This was useful for the point discussed above; in the case where there was no prior information, this was still fine as non-informative priors could be used.
3. Create confidence bounds on the predicted values
    a. It is difficult for statistical models to precisely predict the expected value; however, the credible sets give a range that the observed value could fall into.

4. Assess subgroup risk
    a. The ability to create a distribution for the rank of how likely a group is to develop a disease and graphically review this give a simple method for comparing groups with certain characteristics.
5. Include additional predictors
    a. The models are open and so addition of new predictors is easy.

6.1 Limitations

All modeling techniques are subject to some limitations. There are two specific limitations unique to the methods employed in this research.

(i) Accuracy of prior information – Non-informative

Although using previous data and updating priors, there are possibilities of being wrong if nature of company is changed (e.g. from post office to asbestos production) and so not necessarily have valid priors from Alaska valid in Arizona. Need to care about prior information and one should pay attention to non-informative unless sure that the prior information is accurate.

(ii) This is connected with logistic regression. Since the regression produces estimator of probability which generates the mean and the variance, it may happen that observed mean and variance do not match predicted, the reason is that single parameter influences two types of statistics. In such a case called over- or under-dispersion, there is a possibility to introduce random effect factor that will absorb excess or insufficiency of variance. In our data we did not encounter under- or over-dispersion, but it is possible for a particular state or type of company such effect may exist.

CHAPTER 7 FUTURE WORK

This chapter outlines future areas of research which are of interest. The first are some modifications which could be made to improve the current model. The second discusses modifications which could be made to give the model other applications. The others would be further research in health systems as related to disease burden research.

7.1 Modifications

Improvements could me made to the model in the following areas:

- Prior Elicitation
- Statistical Power
- Conditional Inference

Prior Elicitation

As discussed in this documentation, the key benefit of using Bayesian methodology is in incorporating the available prior information by eliciting so called informative priors on the model parameters. A prior elicitor could be developed to distill the available information into sensible informative prior distributions on coefficients/parameters in the model. This could be achieved not only by matching empirical moments with theoretical distributions of priors, but by an expert system that evaluates various conditional relationships.

Statistical Power

Investigating the *Bayesian Power* as discussed by Joseph et al [118], will show the interplay between required sample size and the power of inference. This feature may translate to economic benefits in the sense of more lean data collection and analysis.

Conditional Inference

Precise modeling may be improved by inputting more information about symptoms and other useful measurements and incorporating conditional analysis into the modeling process. This is, condition the output of interest (likelihood of disease, the predicted number of affected workers, or economic burden) on conditions that can be defined either as new covariates or simply as a subset of observations.

7.2 Extending the Model

This model looks exclusively at diseases which are most costly to employers. The next step would be to model diseases which are most costly to government entities such as the state, Medicare and Medicaid. The results of those models could be used in the formation of government health policies which would improve health conditions while managing costs.

The model also only takes into account diseases which affect adults. Diseases affecting children could probably also be modeled using the same methods to pediatric policies and also state children's programs (e.g. Peachcare, GA).

7.3 Disease Distribution

This research looked at creating a model to estimate the disease burden of corporations using HRA data. The problem with this is that sometimes and individual does not want to acknowledge that they have a disease or have a current risk factor [119-121]. Another factor is that the individual may not know that they actually suffer from that disease. This leads to the question of what the distribution is of people who know they have a disease versus those that have the disease but are unaware. Bayesian methods

may be able to model this disparity using data obtained from a series of interviews and medical exams.

7.4 Disease Cost

The above work may also lead into deciding on the monetary cost of not knowing or not treating a particular disease. Further investigations into the pathology of these diseases and possible effects (for example gangrene or blindness caused by diabetes) could lead to cost models which would look at the effect on the employer and the society for individuals that are not treated.

APPENDIX A: ICD-9 CODES

ALLERGIES	
Code	**Definition**
477.0-477.9	**Allergic rhinitis** Includes: allergic rhinitis (nonseasonal) (seasonal) hay fever spasmodic rhinorrhea Excludes: allergic rhinitis with asthma (bronchial) (493.0)
ARTHRITIS	
Code	**Definition**
714.0–714.9	**Rheumatoid arthritis and other inflammatory polyarthropathies** Excludes: rheumatic fever (390) rheumatoid arthritis of spine NOS (720.0)
715.00 - 715.98	**Osteoarthrosis and allied disorders** Note: Localized, in the subcategories below, includes bilateral involvement of the same site. Includes: arthritis or polyarthritis: degenerative hypertrophic degenerative joint disease osteoarthritis
716.00-716.99	**Other and unspecified arthropathies** Excludes: cricoarytenoid arthropathy (478.79)
719.40-719.59	**719.4Pain in joint** [0-9] Arthralgia
	719.5 Stiffness of joint, not elsewhere classified [0-9]
ASTHMA	
Code	**Definition**
493	**Asthma** Excludes: wheezing NOS (786.07) The following fifth-digit subclassification is for use with category 493.0-493.2, 493.9: 0 unspecified 1 with status asthmaticus 2 with (acute) exacerbation

CANCER	
Code	**Definition**
174-175	**174 Malignant neoplasm of female breast** Includes: breast (female) connective tissue soft parts Paget's disease of: breast nipple Excludes: skin of breast (172.5, 173.5)
	175 Malignant neoplasm of male breast Excludes: skin of breast (172.5, 173.5)
180	**180 Malignant neoplasm of cervix uteri**
185	**Malignant neoplasm of prostate** (Excludes seminal vesicles - 187.8)
153-154	**153 Malignant neoplasm of colon**
	154 Malignant neoplasm of rectum, rectosigmoid junction, and anus
172	**Malignant melanoma of skin** Includes: melanocarcinoma melanoma (skin) NOS Excludes: skin of genital organs (184.0-184.9, 187.1-187.9) sites other than skincode to malignant neoplasm of the site
DIABETES	
250	**Diabetes mellitus** Excludes: gestational diabetes (648.8) hyperglycemia NOS (790.6) neonatal diabetes mellitus (775.1) nonclinical diabetes (790.29) The following fifth-digit subclassification is for use with category 250: 0 type II or unspecified type, no
HEART DISEASE	
Code	**Definition**
402.0–404.9	402 Hypertensive heart disease
	403 Hypertensive chronic kidney disease
	404 Hypertensive heart and chronic kidney disease

410.00-410.92	**410 Acute myocardial infarction** Includes: cardiac infarction coronary (artery): embolism occlusion rupture thrombosis infarction of heart, myocardium, or ventricle rupture of heart, myocardium, or ventricle ST elevation (STEMI) and non-ST elevation (NSTEM
411	**Other acute and subacute forms of ischemic heart disease**
411.1	**Intermediate coronary syndrome** Impending infarction Preinfarction angina Preinfarction syndrome Unstable angina Excludes: angina (pectoris) (413.9) decubitus (413.0)
411.0	**Postmyocardial infarction syndrome** Dressler's syndrome
411.81 - 411.89	**411.81 Acute coronary occlusion without myocardial infarction** Acute coronary (artery): embolism without or not resulting in myocardial infarction obstruction without or not resulting in myocardial infarction occlusion without or not resulting in myocardia
	411.89 Other Coronary insufficiency (acute) Subendocardial ischemia
412	**Old myocardial infarction** Healed myocardial infarction Past myocardial infarction diagnosed on ECG [EKG] or other special investigation, but currently presenting no symptoms
413	**Angina pectoris**
414	**Other forms of chronic ischemic heart disease** Excludes: arteriosclerotic cardiovascular disease [ASCVD] (429.2) cardiovascular: arteriosclerosis or sclerosis (429.2) degeneration or disease (429.2)

414.0	**Coronary atherosclerosis** Arteriosclerotic heart disease [ASHD] Atherosclerotic heart disease Coronary (artery): arteriosclerosis arteritis or endarteritis atheroma sclerosis stricture Excludes: embolism of graft (996.72) occlusion NOS of graft (996.72) th
414.00 - 414.05	
414.1	**Aneurysm and dissection of heart**
414.10	**Aneurysm of heart (wall)** Aneurysm (arteriovenous): mural ventricular
414.8 - 414.9	**414.8 Other specified forms of chronic ischemic heart disease** Chronic coronary insufficiency Ischemia, myocardial (chronic) Any condition classifiable to 410 specified as chronic, or presenting with symptoms after 8 weeks from date of infarction Excludes:
	414.9 Chronic ischemic heart disease, unspecified Ischemic heart disease NOS
HYPERTENSION	
Code	**Definition**
401	**Essential hypertension** Includes: high blood pressure hyperpiesia hyperpiesis hypertension (arterial) (essential) (primary) (systemic) hypertensive vascular: degeneration disease Excludes: elevated blood pressure without diagnosis of hypertension (796.2) p
405	**Secondary hypertension**
MIGRAINE/HEADACHE	
346.0–346.9	**Migraine**

307.81	**Tension headache** Excludes: headache: NOS (784.0) migraine (346.0-346.9)
784.0	**Headache** Facial pain Pain in head NOS Excludes: atypical face pain (350.2) migraine (346.0-346.9) tension headache (307.81)
RESPIRATORY DISEASES	
Code	**Definition**
490	**Bronchitis, not specified as acute or chronic** Bronchitis NOS: catarrhal with tracheitis NOS Tracheobronchitis NOS Excludes: bronchitis: allergic NOS (493.9) asthmatic NOS (493.9) due to fumes and vapors (506.0)
491.0-491.9	**491 Chronic bronchitis** Excludes: chronic obstructive asthma (493.2)
492.0-492.8	**492 Emphysema**

APPENDIX B: BRFSS SURVEY QUESTIONS

State FIPS Code
Section: 0.1 Record Identification Type: Num
Column: 1-2 SAS Variable Name: _STATE
Prologue:
Description: State FIPS Code

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Alabama	3,197	0.90	1.53
2	Alaska	2,813	0.79	0.21
4	Arizona	4,710	1.32	1.88
5	Arkansas	5,280	1.48	0.92
6	California	6,134	1.72	11.91
8	Colorado	5,979	1.68	1.55
9	Connecticut	5,254	1.48	1.20
10	Delaware	4,192	1.18	0.28
11	District of Columbia	3,743	1.05	0.20
12	Florida	8,190	2.30	6.03
13	Georgia	6,064	1.70	2.93
15	Hawaii	6,416	1.80	0.44
16	Idaho	5,734	1.61	0.46
17	Illinois	5,077	1.43	4.25
18	Indiana	5,635	1.58	2.08
19	Iowa	5,051	1.42	1.01
20	Kansas	8,626	2.42	0.91
21	Kentucky	6,628	1.86	1.41
22	Louisiana	2,936	0.82	1.29
23	Maine	3,960	1.11	0.46
24	Maryland	8,632	2.42	1.88
25	Massachusetts	8,906	2.50	2.22
26	Michigan	12,136	3.41	3.40
27	Minnesota	2,829	0.79	1.72
28	Mississippi	4,439	1.25	0.96
29	Missouri	5,164	1.45	1.94
30	Montana	4,983	1.40	0.32
31	Nebraska	8,332	2.34	0.59
32	Nevada	3,161	0.89	0.78
33	New Hampshire	6,038	1.70	0.45
34	New Jersey	13,663	3.84	2.94
35	New Mexico	5,585	1.57	0.62
36	New York	7,796	2.19	6.59
37	North Carolina	17,261	4.85	2.88
38	North Dakota	4,010	1.13	0.22
39	Ohio	7,498	2.11	3.87
40	Oklahoma	13,707	3.85	1.19

41	Oregon	12,015	3.37	1.23		
42	Pennsylvania	13,378	3.76	4.29		
44	Rhode Island	3,976	1.12	0.38		
45	South Carolina	8,440	2.37	1.42		
46	South Dakota	6,915	1.94	0.26		
47	Tennessee	4,749	1.33	2.01		
48	Texas	6,512	1.83	7.29		
49	Utah	5,137	1.44	0.74		
50	Vermont	6,763	1.90	0.22		
51	Virginia	5,493	1.54	2.55		
53	Washington	23,302	6.54	2.11		
54	West Virginia	3,553	1.00	0.64		
55	Wisconsin	4,900	1.38	1.87		
56	Wyoming	5,009	1.41	0.17		
72	Puerto Rico	3,789	1.06	1.27		
78	Virgin Islands	2,422	0.68	0.03		

Ever Told by Doctor You Have Diabetes
Section: 5.1 Diabetes Type: Num
Column: 85 SAS Variable Name: DIABETE2
Prologue:
Description: Have you ever been told by a doctor that you have diabetes (If "Yes" and respondent is female, ask "Was this only when you were pregnant?". If Respondent says pre-diabetes or boderline diabetes, use response code 4.)

Value	Value Label	Frequency	Percentage	Weighted Percentage		
1	Yes	33,320	9.36	7.77		
2	Yes, but female told only during pregnancy	2,986	0.84	0.94		
3	No	315,599	88.62	90.38		
4	No, pre-diabetes or boarderline diabetes	3,853	1.08	0.83		
7	Don’t know/Not Sure	220	0.06	0.05		
9	Refused	133	0.04	0.03		
BLANK	Not asked or Missing	1				

Ever Told Blood Pressure High
Section: 6.1 Hypertension Awareness Type: Num
Column: 86 SAS Variable Name: BPHIGH4
Prologue:
Description: Have you ever been told by a doctor, nurse or other health professional that you have high blood pressure? (If "Yes" and respondent is female, ask "Was this only when you were pregnant?".)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	112,687	31.64	26.13
2	Yes, but female told only during pregnancy - Go to Section 07.01 BLOODCHO	3,271	0.92	1.06
3	No - Go to Section 07.01 BLOODCHO	234,108	65.74	71.41
4	Told borderline high or pre-hypertensive - Go to Section 07.01 BLOODCHO	5,425	1.52	1.26
7	Don´t know/Not Sure - Go to Section 07.01 BLOODCHO	517	0.15	0.12
9	Refused - Go to Section 07.01 BLOODCHO	102	0.03	0.02
BLANK	Not asked or Missing - Go to Section 07.01 BLOODCHO	2		

Ever Told Blood Cholesterol High
Section: 7.3 Cholesterol Awareness Type: Num
Column: 90 SAS Variable Name: TOLDHI2
Prologue:
Description: Have you ever been told by a doctor, nurse or other health professional that your blood cholesterol is high?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	114,166	39.11	35.69
2	No	175,448	60.10	63.63
7	Don’t know/Not Sure	2,191	0.75	0.63
9	Refused	137	0.05	0.04
BLANK	Not asked or Missing			

Notes: Section 7.01, BLOODCHO, is coded 2, 7, 9, or Missing 64,170

Ever Diagnosed with Heart Attack
Section: 8.1 Cardiovascular Disease Type: Num
Column: 91 SAS Variable Name: CVDINFR3
Prologue: Has a doctor, nurse, or other health professional ever told you that you had any of the following? For each, tell me "Yes", "No", or you're "Not sure":
Description: A heart attack, also called a myocardial infraction?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	18,700	5.25	4.08
2	No	335,562	94.23	95.44
7	Don't know/Not sure	1,759	0.49	0.45
9	Refused	90	0.03	0.02
BLANK	Not asked or Missing	1		

Angina or coronary heart disease
Section: 8.2 Cardiovascular Disease Type: Num
Column: 92 SAS Variable Name: CVDCRHD3
Prologue:
Description: [Has a doctor, nurse, or other health professional ever told you that you had any of the following? For each, tell me "Yes", "No", or you're "Not sure":] Angina or coronary heart disease.

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	19,610	5.51	4.45
2	No	333,378	93.62	94.89
7	Don't know/Not sure	3,035	0.85	0.63
9	Refused	87	0.02	0.03
BLANK	Not asked or Missing	2		

Ever Diagnosed with a Stroke
Section: 8.3 Cardiovascular Disease Type: Num
Column: 93 SAS Variable Name: CVDSTRK3
Prologue:
Description: [Has a doctor, nurse, or other health professional ever told you that you had any of the following? For each, tell me "Yes", "No", or you're "Not sure":] (Ever told) you had a stroke.

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	12,079	3.39	2.61
2	No	343,051	96.33	97.17
7	Don't know/Not sure	917	0.26	0.21
9	Refused	64	0.02	0.02

BLANK Not asked or Missing 1

Ever Told Had Asthma
Section: 9.1 Asthma Type: Num
Column: 94 SAS Variable Name: ASTHMA2
Prologue:
Description: Have you ever been told by a doctor, nurse, or other health professional that you had asthma?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	45,333	12.73	12.53
2	No - Go to Section 10.01 FLUSHOT3	309,948	87.04	87.26
7	Don’t know/Not Sure - Go to Section 10.01 FLUSHOT3	784	0.22	0.20
9	Refused - Go to Section 10.01 FLUSHOT3	46	0.01	0.01
BLANK	Not asked or Missing	1		

Still Have Asthma
Section: 9.2 Asthma Type: Num
Column: 95 SAS Variable Name: ASTHNOW
Prologue:
Description: Do you still have asthma?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	30,323	66.89	62.42
2	No	13,754	30.34	34.98
7	Don’t know/Not Sure	1,247	2.75	2.59
9	Refused	9	0.02	0.01
BLANK	Not asked or Missing			

Notes: Section 9.01, ASTHMA2, is coded 2, 7, 9, or Missing 310,779

Reported Age in Years
Section: 13.1 Demographics Type: Num
Column: 112-113 SAS Variable Name: AGE
Prologue:

Description: What is your age?

Value	Value Label	Frequency	Percentage	Weighted Percentage
7	Don't know/Not sure	180	0.05	0.04
9	Refused	2,514	0.71	0.52
18 - 24	Age 18 - 24			
Notes: _ _	Code age in years	18,290	5.14	13.11
25 - 34	Age 25 - 34	46,613	13.09	18.02
35 - 44	Age 35 - 44	63,425	17.81	19.73
45 - 54	Age 45 - 54	73,297	20.58	18.48
55 - 64	Age 55 - 64	64,441	18.10	13.39
65 - 99	Age 65 or older	87,351	24.53	16.71
BLANK	Not asked or Missing	1		

Employment Status
Section: 13.8 Demographics Type: Num
Column: 126 SAS Variable Name: EMPLOY
Prologue:
Description: Are you currently: (employment status)

Value	Value Label	Frequency	Percentage	Weighted Percentage	
1	Employed for wages	169,448	47.59	52.49	
2	Self-employed	31,347	8.80	8.72	
3	Out of work for more than 1 year	6,213	1.75		2.04
4	Out of work for less that 1 year	8,153	2.29		2.98
5	A homemaker	28,787	8.09	7.97	
6	A student	7,839	2.20	4.46	
7	Retired	82,102	23.06	16.01	
8	Unable to work	21,182	5.95	4.94	
9	Refused	954	0.27	0.39	
BLANK	Not asked or Missing	87			

Told Have Arthritis
Section: 16.4 Arthritis Burden Type: Num
Column: 156 SAS Variable Name: HAVARTH2
Prologue:
Description: Have you ever been told by a doctor or other health professional that you have some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia? (Arthritis

diagnoses include: rheumatism, polymyalgia rheumatica; osteoarthritis (not osteporosis); tendonitis, bursitis, bunion, tennis elbow; carpal tunnel syndrome, tarsal tunnel syndrome; joint infection, etc. (See Questionnaire for Complete List))

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	119,485	33.98	26.52
2	No	230,651	65.59	73.08
7	Don't know/Not Sure	1,417	0.40	0.36
9	Refused	100	0.03	0.04
BLANK	Not asked or Missing	4,459		

Age When Told Diabetic
Module: 1.1 Diabetes Type: Num
Column: 201-202 SAS Variable Name: DIABAGE2
Prologue:
Description: How old were you when you were told you have diabetes?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1 - 97	Age in years			
Notes: _ _ Code age in years, 97 = 97 or older		24,397	95.52	96.16
98	Don't know/Not sure	1,059	4.15	3.54
99	Refused	85	0.33	0.30
BLANK	Not asked, Module not used, or Missing			
Notes: Section 5.01, DIABETE2, is coded 2, 3, 4, 7, 9, or Missing		330,571		

How many days depressed in past 30 days
Module: 4.2 Healthy Days (Symptoms) Type: Num
Column: 241-242 SAS Variable Name: QLMENTL2
Prologue:
Description: During the past 30 days, for about how may days have you felt sad, blue, or depressed?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1 - 30	Number of days	3,034	44.41	46.58
88	None	3,665	53.64	51.44
77	Don't know/Not sure	115	1.68	1.63
99	Refused	18	0.26	0.34

BLANK	Not asked, Module not used or Missing	349,280

Age at Asthma Diagnosis
Module: 9.1 Adult Asthma History Type: Num
Column: 278-279 SAS Variable Name: ASTHMAGE
Prologue: Previously you said you were told by a doctor, nurse, or other health professional that you had asthma.
Description: How old were you when you were first told by a doctor, nurse or other health professional that you had asthma?

Value	Value Label	Frequency	Percentage	Weighted Percentage
11 - 96	Age 11 or older			
Notes: 96=96 and older		11,920	67.47	61.13
97	Age 10 or younger	4,854	27.47	34.57
98	Don't know/Not sure	866	4.90	4.03
99	Refused	27	0.15	0.26
BLANK	Not asked, Module not used, or Missing			
Notes: Section 9.01, ASTHMA2, is coded 2, 7, 9, or Missing				338,445

Ever Told You Had Prostate Cancer
Module: 14.5 Prostate Cancer Screening Type: Num
Column: 325 SAS Variable Name: PROSTATE
Prologue:
Description: Have you ever been told by a doctor, nurse, or other health professional that you had prostate cancer?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	100	4.14	3.52
2	No	2,309	95.65	96.39
7	Don't know/Not sure	5	0.21	0.09
BLANK	Not asked or Missing			
Notes: Section 13.01, AGE, is less than 40; or Section 13.17, SEX, is coded 2				353,698

Hlth pro ever said have osteoporosis
Module: 16.1 Osteoporosis Type: Num
Column: 330 SAS Variable Name: OSTPROS
Prologue:
Description: Have you EVER been told by a doctor, nurse, or other health professional that you have osteoporosis?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	8,403	8.03	4.75
2	No	95,746	91.51	94.94
7	Don't know/Not Sure	475	0.45	0.31
BLANK	Not asked, Module not used, or Missing	251,488		

Age group codes used in post-stratification.
Weighting: 1.16 Weighting And Stratification Variables Type: Num
Column: 937-938 SAS Variable Name: _AGEG_
Prologue:
Description: Age groups used in post-stratification (_AGEG_ is calculated by _REGION. For states using more than one _REGION, there could be more than one response for a given _AGEG_ value.)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Age 18 to 24	17,516	4.92	12.85
2	Age 25 to 34	44,541	12.51	17.52
3	Age 35 to 44	63,442	17.82	19.73
4	Age 45 to 54	75,458	21.19	18.90
5	Age 55 to 64	64,792	18.19	13.45
6	Age 65 to 74	45,878	12.88	8.15
7	Age 75 or older	38,572	10.83	7.70
8	Age 18 to 34	2,817	0.79	0.75
9	Age 35 to 54	126	0.04	0.07
11	Age 18 to 44	69	0.02	0.02
13	Age 65 or older	2,901	0.81	0.86

Gender group codes used in post-stratification.
Weighting: 1.17 Weighting And Stratification Variables Type: Num
Column: 939 SAS Variable Name: _SEXG_
Prologue:

Description: Gender categories used in post-stratification (_SEXG_ is calculated by _REGION. For states using more than one _REGION, there could be more than one response for a given _SEXG_ value.)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Male	136,201	38.25	48.52
2	Male	219,911	61.75	51.48

High Blood Pressure Calculated Variable
Calculated: 6.1 Calculated Variables Type: Num
Column: 1157 SAS Variable Name: _RFHYPE5
Prologue:
Description: Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No			
Notes: BPHIGH4 = 2,3,4		242,804	68.18	73.72
2	Yes			
Notes: BPHIGH4 = 1		112,687	31.64	26.13
9	Don’t know/Not Sure/Refused/Missing			
Notes: BPHIGH4 = .,7,9		621	0.17	0.15

Cholesterol Checked Calculated Variable
Calculated: 7.1 Calculated Variables Type: Num
Column: 1158 SAS Variable Name: _CHOLCHK
Prologue:
Description: Cholesterol check within past five years.

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Had cholesterol checked in past 5 years.			
Notes: BLOODCHO = 1 and CHOLCHK = 1,2,3		274,107	76.97	71.07
2	Did not have cholesterol checked in past 5 years			
Notes: BLOODCHO = 1 and CHOLCHK = 4		13,936	3.91	3.46
3	Have never had cholesterol checked			
Notes: BLOODCHO = 2		57,016	16.01	22.31
9	Don’t know/Not Sure Or Refused/Missing			
Notes: BLOODCHO = .,7,9 or CHOLCHK = .,7,9		11,053	3.10	3.16

High Cholesterol Calculated Variable
Calculated: 7.2 Calculated Variables Type: Num
Column: 1159 SAS Variable Name: _RFCHOL
Prologue:
Description: Adults who have had their cholesterol checked and have been told by a doctor, nurse, or other health professional that it was high

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No			
Notes: BLOODCHO=1 and TOLDHI2 = 2		175,448	60.10	63.63
2	Yes			
Notes: BLOODCHO=1 and TOLDHI2 = 1		114,166	39.11	35.69
9	Don’t know/Not Sure Or Refused/Missing			
Notes: BLOODCHO=1 and TOLDHI2 = .,7,9		2,328	0.80	0.68
BLANK	Missing			
Notes: BLOODCHO=.,2,7,9		64,170		

Lifetime Asthma Calculated Variable
Calculated: 9.1 Calculated Variables Type: Num
Column: 1160 SAS Variable Name: _LTASTHM
Prologue:
Description: Adults who have ever been told they have asthma

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No			
Notes: ASTHMA2 = 2		309,948	87.04	87.26
2	Yes			
Notes: ASTHMA2 = 1		45,333	12.73	12.53
9	Don’t know/Not Sure Or Refused/Missing			
Notes: ASTHMA2 = 7 or 9 or missing		831	0.23	0.21

Current Asthma Calculated Variable
Calculated: 9.2 Calculated Variables Type: Num
Column: 1161 SAS Variable Name: _CASTHMA
Prologue:
Description: Adults who have been told they currently have asthma

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No			

Notes: ASTHMA2 = 2 or ASTHMA2 = 1 and ASTHNOW = 2 323,702
90.90 91.64
2 Yes
Notes: ASTHMA2 = 1 and ASTHNOW = 1 30,323 8.52 7.82
9 Don’t know/Not Sure Or Refused/Missing
Notes: ASTHMA2 = 7 or 9 or ASTHNOW = 7 or 9 2,087 0.59 0.54

Computed Asthma Status
Calculated: 9.3 Calculated Variables Type: Num
Column: 1162 SAS Variable Name: _ASTHMST
Prologue:
Description: Computed asthma status
Value Value Label Frequency Percentage Weighted Percentage
1 Current
Notes: ASTHMA2 = 1 and ASTHNOW = 1 30,323 8.52 7.82
2 Former
Notes: ASTHMA2 = 1 and ASTHNOW = 2 13,754 3.86 4.38
3 Never
Notes: ASTHMA2 = 2 309,948 87.04 87.26
9 Don’t know/Not Sure Or Refused/Missing
Notes: ASTHMA2 = 7 or 9 or ASTHNOW = 7 or 9 2,087 0.59 0.54

Computed Smoking Status
Calculated: 11.1 Calculated Variables Type: Num
Column: 1165 SAS Variable Name: _SMOKER3
Prologue:
Description: Four-level smoker status: Everyday smoker, Someday smoker, Former smoker, Non-smoker
Value Value Label Frequency Percentage Weighted Percentage
1 Current smoker - now smokes every day
Notes: (SMOKE100 = 1) AND (SMOKDAY2 = 1) 52,303 14.69
14.79
2 Current smoker - now smokes some days
Notes: (SMOKE100 = 1) AND (SMOKDAY2 = 2) 17,135 4.81
5.56
3 Former smoker
Notes: (SMOKE100 = 1) AND (SMOKDAY2 = 3) 99,209 27.86
24.23

4 Never smoked
Notes: (SMOKE100 = 2) 185,935 52.21 54.97
9 Don’t know/Refused/Missing
Notes: (SMOKE100 = 1) AND ((SMOKDAY2 = 9) OR (SMOKE100 = (. OR 7 OR 9))
1,530 0.43 0.45

Current Smoking Calculated Variable
Calculated: 11.2 Calculated Variables Type: Num
Column: 1166 SAS Variable Name: _RFSMOK3
Prologue:
Description: Adults who are current smokers

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No			
Notes: (_SMOKER2 = (3 OR 4))		285,144	80.07	79.19
2	Yes			
Notes: (_SMOKER2 = (1 OR 2))		69,438	19.50	20.36
9	Don’t know/Refused/Missing			
Notes: (_SMOKER2 = 9)		1,530	0.43	0.45

Binge Drinking Calculated Variable
Calculated: 12.2 Calculated Variables Type: Num
Column: 1170 SAS Variable Name: _RFBING3
Prologue:
Description: Binge drinkers (adults having five or more drinks on one occasion)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No			
Notes: ALCDAY4 < 300 and DRNK2GE5 = 0 or ALCDAY4 = 888		314,690	88.37	84.85
2	Yes			
Notes: ALCDAY4 < 300 and DRNK2GE5 = 1		37,922	10.65	14.08
9	Don’t know/Refused/Missing			
Notes: DRNK2GE5 = 7 or 9 or Missing or ALCDAY4 = 777 or 999 or Missing		3,500	0.98	1.07

Heavy Alcohol Consumption Calculated Variable
Calculated: 12.5 Calculated Variables Type: Num
Column: 1179 SAS Variable Name: _RFDRHV3
Prologue:
Description: Heavy drinkers (adult men having more than two drinks per day and adult women having more than one drink per day)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No			
Notes: SEX = 1 and _DRNKDY3 <= 2 or SEX = 2 and _DRNKDY3 <= 1 or ALCDAY4 = 888		334,060	93.81	93.07
2	Yes			
Notes: SEX = 1 and _DRNKDY3 > 2 or SEX = 2 and _DRNKDY3 > 1		15,845	4.45	5.02
9	Don’t know/Refused/Missing			
Notes: _DRNKDY3 = 900		6,207	1.74	1.90

Computed race groups used for internet prevalence tables
Calculated: 13.8 Calculated Race Variables Type: Num
Column: 1201 SAS Variable Name: _RACE_G
Prologue:
Description: Race groups used for internet prevalence tables

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	White - Non-Hispanic			
Notes: _RACEGR2 = 1		278,672	79.01	69.27
2	Black - Non-Hispanic			
Notes: _RACEGR2 = 2		27,735	7.86	9.51
3	Hispanic			
Notes: _RACEGR2 = 5		25,539	7.24	14.96
4	Other race only, Non-Hispanic			
Notes: _RACEGR2 = 3		14,280	4.05	4.79
5	Multiracial, Non-Hispanic			
Notes: _RACEGR2 = 4		6,470	1.83	1.47
BLANK	Don’t know/Not sure/Refused component question			
Notes: _RACEGR2 = 9 or missing		3,416		

Reported age in five-year age categories calculated variable
Calculated: 13.11 Calculated Variables Type: Num

Column: 1204-1205 SAS Variable Name: _AGEG5YR
Prologue:
Description: Fourteen-level age category

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Age 18 to 24			
Notes: 18 LE AGE LE 24		18,290	5.14	13.11
2	Age 25 to 29			
Notes: 25 LE AGE LE 29		20,602	5.79	8.04
3	Age 30 to 34			
Notes: 30 LE AGE LE 34		26,011	7.30	9.98
4	Age 35 to 39			
Notes: 35 LE AGE LE 39		29,844	8.38	9.41
5	Age 40 to 44			
Notes: 40 LE AGE LE 44		33,581	9.43	10.32
6	Age 45 to 49			
Notes: 45 LE AGE LE 49		36,288	10.19	9.56
7	Age 50 to 54			
Notes: 50 LE AGE LE 54		37,009	10.39	8.93
8	Age 55 to 59			
Notes: 55 LE AGE LE 59		35,078	9.85	7.56
9	Age 60 to 64			
Notes: 60 LE AGE LE 64		29,363	8.25	5.84
10	Age 65 to 69			
Notes: 65 LE AGE LE 69		25,530	7.17	4.78
11	Age 70 to 74			
Notes: 70 LE AGE LE 74		22,012	6.18	3.87
12	Age 75 to 79			
Notes: 75 LE AGE LE 79		18,794	5.28	4.08
13	Age 80 or older			
Notes: 80 LE AGE LE 99		21,015	5.90	3.97
14	Don’t know/Refused/Missing			
Notes: 7 LE AGE LE 9		2,695	0.76	0.56

Reported age in two age groups calculated variable
Calculated: 13.12 Calculated Variables Type: Num
Column: 1206 SAS Variable Name: _AGE65YR
Prologue:
Description: Two-level age category

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Age 18 to 64			
Notes: 18 LE AGE LE 64		266,066	74.71	82.74
2	Age 65 or older			
Notes: 65 LE AGE LE 99		87,351	24.53	16.71

3 Don’t know/Refused/Missing
Notes: 7 LE AGE LE 9 2,695 0.76 0.56

Imputed age in six groups
Calculated: 13.13 Calculated Variables Type: Num
Column: 1207 SAS Variable Name: _AGE_G
Prologue:
Description: Six-level imputed age category

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Age 18 to 24			
Notes: 18 LE AGE LE 24		18,290	5.14	13.11
2	Age 25 to 34			
Notes: 25 LE AGE LE 34		46,613	13.09	18.02
3	Age 35 to 44			
Notes: 35 LE AGE LE 44		63,530	17.84	19.76
4	Age 45 to 54			
Notes: 45 LE AGE LE 54		75,536	21.21	18.95
5	Age 55 to 64			
Notes: 55 LE AGE LE 64		64,792	18.19	13.45
6	Age 65 or older			
Notes: AGE GE 65		87,351	24.53	16.71

Computed body mass index categories
Calculated: 13.18 Calculated Variables Type: Num
Column: 1223 SAS Variable Name: _BMI4CAT
Prologue:
Description: Three-categories of Body Mass Index (BMI)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Neither overweight nor obese			
Notes: _BMI4 < 2500 (_BMI4 has 2 implied decimal places)		129,513	36.37	36.95
2	Overweight			
Notes: 2500 <= _BMI4 < 3000		123,692	34.73	35.14
3	Obese			
Notes: 3000 <= _BMI4 < 9999		86,463	24.28	23.36
9	Don’t know/Refused/Missing			
Notes: _BMI4 = 9999		16,444	4.62	4.55

Overweight or obese calculated variable
Calculated: 13.19 Calculated Variables Type: Num
Column: 1224 SAS Variable Name: _RFBMI4
Prologue:
Description: Adults who have a body mass index greater than 25.00 (Overweight or Obese)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No			
Notes: 0 <= _BMI4 < 2500 (_BMI4 has 2 implied decimal places)		129,513	36.37	36.95
2	Yes			
Notes: 2500 <= _BMI4 < 9999		210,155	59.01	58.49
9	Don’t know/Refused/Missing			
Notes: _BMI4 = 9999		16,444	4.62	4.55

Respondents diagnosed with arthritis
Calculated: 16.1 Calculated Variables Type: Num
Column: 1228 SAS Variable Name: _DRDXART
Prologue:
Description: Respondents that have had a doctor diagnose them as having some form of arthritis.

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Diagnosed with arthritis	119,485	34.13	26.63
2	Not diagnosed with arthritis	230,651	65.87	73.37
BLANK	Don´t know/Not Sure/Refused/Missing	5,976		

Computed Moderate Physical Activity Categories
Calculated: 18.3 Calculated Variables Type: Num
Column: 1260 SAS Variable Name: MODCAT_
Prologue:
Description: 3 level moderate physical activity category.

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Meet recommendations for moderate physical activity			

Notes: MODPACT = 1 and _MODPAMN >= 30 and MODPADAY >= 5; 121,578
34.14 33.23
2 Insufficient activity to meet moderate recommendations
Notes: MODPACT = 1 and _MODPAMN < 30 or MODPADAY < 5; 151,860
42.64 43.79
3 No moderate physical activity
Notes: MODPACT = 2 or _MODPAMN=0 58,401 16.40 16.07
9 Don't know/Not sure/Refused/Missing
Notes: MODPACT=., 7, 9; or MODPACT=1 and MODPADAY=., 77, 99; or
MODPATIM=., 777, 999 24,273 6.82 6.90

Computed Vigorous Physical Activity Categories
Calculated: 18.4 Calculated Variables Type: Num
Column: 1261 SAS Variable Name: VIGCAT_
Prologue:
Description: 3 level vigorous physical activity category.
Value Value Label Frequency Percentage Weighted Percentage
1 Meet recommendations for vigorous physical activity
Notes: VIGPACT = 1 and _VIGPAMN >= 20 and VIGPADAY >= 3 81,216
22.81 25.95
2 Insufficient activity to meet vigorous recommendations
Notes: VIGPACT = 1 and _VIGPAMN < 20 or VIGPADAY < 3 64,098
18.00 19.67
3 No vigorous physical activity
Notes: VIGPACT = 2 or _VIGPAMN=0 193,958 54.47 48.95
9 Don't know/Not sure/Refused/Missing
Notes: VIGPACT=., 7, 9; or VIGPACT=1 and VIGPADAY=., 77, 99; or VIGPATIM=.,
777, 999 16,840 4.73 5.42

Computed Overall Physical Activity Categories
Calculated: 18.5 Calculated Variables Type: Num
Column: 1262 SAS Variable Name: PACAT_
Prologue:
Description: 5 level physical activity category.
Value Value Label Frequency Percentage Weighted Percentage
1 Meet recommendations for moderate and vigorous physical activity
Notes: MODCAT_ = 1 and VIGCAT_ = 1 47,376 13.30 14.52
2 Meet recommendations for vigorous physical activity
Notes: VIGCAT_ = 1 and MODCAT_ not equal to 1 33,840 9.50
11.44

3 Meet recommendations for moderate physical activity
Notes: MODCAT_ = 1 and VIGCAT_ not equal to 1 74,202 20.84 18.72
4 Insufficient activity to meet moderate or vigorous recommendations
Notes: MODCAT_ = 2 and VIGCAT_ = 2 or 3 or VIGCAT_ = 2 and MODCAT_ = 3 125,166 35.15 35.04
5 No moderate or vigorous physical activity
Notes: MODCAT_ = 3 and VIGCAT_ = 3 50,399 14.15 13.15
9 Don't know/Not sure/Refused/Missing
Notes: MODCAT_=9 or VIGCAT_=9 25,129 7.06 7.15

Moderate Physical Activity Calculated Variable
Calculated: 18.6 Calculated Variables Type: Num
Column: 1263 SAS Variable Name: _RFPAMOD
Prologue:
Description: Adults that have reported participating in either moderate physical activity defined as 30 or more minutes per day for 5 or more days per week, or vigorous activity for 20 or more minutes per day on 3 or more days

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes			
Notes: PACAT_ = 1,2,3		155,418	43.64	44.67
2	No			
Notes: PACAT_ = 4,5		175,565	49.30	48.18
9	Don't know/Not Sure/Refused/Missing			
Notes: PACAT_ = 9		25,129	7.06	7.15

Vigorous Physical Activity Calculated Variable
Calculated: 18.7 Calculated Variables Type: Num
Column: 1264 SAS Variable Name: _RFPAVIG
Prologue:
Description: Adults that have reported participating in vigorous activity for 20 or more minutes per day on 3 or more days

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes			
Notes: PACAT_ = 1,2		81,216	22.81	25.95
2	No			
Notes: PACAT_ = 3,4,5		258,056	72.46	68.62
9	Don't know/Not Sure/Refused/Missing			
Notes: PACAT_ = 9		16,840	4.73	5.42

Recommended Physical Activity Calculated Variable
Calculated: 18.8 Calculated Variables Type: Num
Column: 1265 SAS Variable Name: _RFPAREC
Prologue:
Description: Adults self reported physical activity level status

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Meet physical activity recommendations			
Notes: PACAT_ = 1,2,3		155,418	43.64	44.67
2	Insufficient physical activity			
Notes: PACAT_ = 4		125,166	35.15	35.04
3	No physical activity			
Notes: PACAT_ = 5		50,399	14.15	13.15
9	Don’t know/Not Sure/Refused/Missing			
Notes: PACAT_ = 9		25,129	7.06	7.15

No Physical Activity or Exercise Calculated Variable
Calculated: 18.9 Calculated Variables Type: Num
Column: 1266 SAS Variable Name: _RFNOPA
Prologue:
Description: Adults that have reported participating in physical activity or exercise

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes			
Notes: _RFPAREC = 1,2 or _TOTINDA = 1		312,033	87.62	88.53
2	No			
Notes: _RFPAREC = 3 and _TOTINDA = 2		35,310	9.92	8.99
9	Don’t know/Not Sure/Refused/Missing			
Notes: _RFPAREC = 9 and _TOTINDA = 9		8,769	2.46	2.47

APPENDIX C: CENSUS BUREAU CODES

INDP 4

Industry recode

bbbb .N/A (less than 16 years old/unemployed who
.never worked/NILF who last worked more than
.5 years ago)
0170 .AGR-CROP PRODUCTION
0180 .AGR-ANIMAL PRODUCTION
0190 .AGR-FORESTRY EXCEPT LOGGING
0270 .AGR-LOGGING
0280 .AGR-FISHING, HUNTING, AND TRAPPING
0290 .AGR-SUPPORT ACTIVITIES FOR AGRICULTURE AND FORESTRY
0370 .EXT-OIL AND GAS EXTRACTION
0380 .EXT-COAL MINING
0390 .EXT-METAL ORE MINING
0470 .EXT-NONMETALLIC MINERAL MINING AND QUARRYING
0480 .EXT-NOT SPECIFIED TYPE OF MINING
0490 .EXT-SUPPORT ACTIVITIES FOR MINING
0570 .UTL-ELECTRIC POWER GENERATION, TRANSMISSION AND
.DISTRIBUTION
0580 .UTL-NATURAL GAS DISTRIBUTION
0590 .UTL-ELECTRIC AND GAS, AND OTHER COMBINATIONS
0670 .UTL-WATER, STEAM, AIR CONDITIONING, AND IRRIGATION SYSTEMS
0680 .UTL-SEWAGE TREATMENT FACILITIES
0690 .UTL-NOT SPECIFIED UTILITIES
0770 .CON-CONSTRUCTION, INCL CLEANING DURING AND IMM AFTER
1070 .MFG-ANIMAL FOOD, GRAIN AND OILSEED MILLING
1080 .MFG-SUGAR AND CONFECTIONERY PRODUCTS
1090 .MFG-FRUIT AND VEGETABLE PRESERVING AND SPECIALTY FOODS
1170 .MFG-DAIRY PRODUCTS
1180 .MFG-ANIMAL SLAUGHTERING AND PROCESSING
1190 .MFG-RETAIL BAKERIES
1270 .MFG-BAKERIES, EXCEPT RETAIL
1280 .MFG-SEAFOOD AND OTHER MISCELLANEOUS FOODS, N.E.C.
1290 .MFG-NOT SPECIFIED FOOD INDUSTRIES
1370 .MFG-BEVERAGE
1390 .MFG-TOBACCO
1470 .MFG-FIBER, YARN, AND THREAD MILLS
1480 .MFG-FABRIC MILLS, EXCEPT KNITTING
1490 .MFG-TEXTILE AND FABRIC FINISHING AND COATING MILLS
1570 .MFG-CARPETS AND RUGS
1590 .MFG-TEXTILE PRODUCT MILLS, EXCEPT CARPETS AND RUGS
1670 .MFG-KNITTING MILLS
1680 .MFG-CUT AND SEW APPAREL
1690 .MFG-APPAREL ACCESSORIES AND OTHER APPAREL
1770 .MFG-FOOTWEAR
1790 .MFG-LEATHER TANNING AND PRODUCTS, EXCEPT FOOTWEAR
1870 .MFG-PULP, PAPER, AND PAPERBOARD MILLS
1880 .MFG-PAPERBOARD CONTAINERS AND BOXES

1890 .MFG-MISCELLANEOUS PAPER AND PULP PRODUCTS
1990 .MFG-PRINTING AND RELATED SUPPORT ACTIVITIES
2070 .MFG-PETROLEUM REFINING
2090 .MFG-MISCELLANEOUS PETROLEUM AND COAL PRODUCTS
2170 .MFG-RESIN, SYNTHETIC RUBBER AND FIBERS, AND FILAMENTS
2180 .MFG-AGRICULTURAL CHEMICALS
2190 .MFG-PHARMACEUTICALS AND MEDICINES
2270 .MFG-PAINT, COATING, AND ADHESIVES
2280 .MFG-SOAP, CLEANING COMPOUND, AND COSMETICS
2290 .MFG-INDUSTRIAL AND MISCELLANEOUS CHEMICALS
2370 .MFG-PLASTICS PRODUCTS
2380 .MFG-TIRES
2390 .MFG-RUBBER PRODUCTS, EXCEPT TIRES
2470 .MFG-POTTERY, CERAMICS, AND RELATED PRODUCTS
2480 .MFG-STRUCTURAL CLAY PRODUCTS
2490 .MFG-GLASS AND GLASS PRODUCTS
2570 .MFG-CEMENT, CONCRETE, LIME, AND GYPSUM PRODUCTS
2590 .MFG-MISCELLANEOUS NONMETALLIC MINERAL PRODUCTS
2670 .MFG-IRON AND STEEL MILLS AND STEEL PRODUCTS
2680 .MFG-ALUMINUM PRODUCTION AND PROCESSING
2690 .MFG-NONFERROUS METAL, EXCEPT ALUMINUM, PRODUCTION AND
.PROCESSING
2770 .MFG-FOUNDRIES
2780 .MFG-METAL FORGINGS AND STAMPINGS
2790 .MFG-CUTLERY AND HAND TOOLS
2870 .MFG-STRUCTURAL METALS, AND TANK AND SHIPPING CONTAINERS
2880 .MFG-MACHINE SHOPS; TURNED PRODUCTS; SCREWS, NUTS AND BOLTS
2890 .MFG-COATING, ENGRAVING, HEAT TREATING AND ALLIED ACTIVITIES
2970 .MFG-ORDNANCE
2980 .MFG-MISCELLANEOUS FABRICATED METAL PRODUCTS
2990 .MFG-NOT SPECIFIED METAL INDUSTRIES
3070 .MFG-AGRICULTURAL IMPLEMENTS
3080 .MFG-CONSTRUCTION, MINING AND OIL FIELD MACHINERY
3090 .MFG-COMMERCIAL AND SERVICE INDUSTRY MACHINERY
3170 .MFG-METALWORKING MACHINERY
3180 .MFG-ENGINES, TURBINES, AND POWER TRANSMISSION EQUIPMENT
3190 .MFG-MACHINERY, N.E.C.
3290 .MFG-NOT SPECIFIED MACHINERY
3360 .MFG-COMPUTER AND PERIPHERAL EQUIPMENT
3370 .MFG-COMMUNICATIONS, AUDIO, AND VIDEO EQUIPMENT
3380 .MFG-NAVIGATIONAL, MEASURING, ELECTROMEDICAL, AND CONTROL
.INSTRUMENTS
3390 .MFG-ELECTRONIC COMPONENTS AND PRODUCTS, N.E.C.
3470 .MFG-HOUSEHOLD APPLIANCES
3490 .MFG-ELECTRICAL LIGHTING, EQUIPMENT, AND SUPPLIES, N.E.C.
3570 .MFG-MOTOR VEHICLES AND MOTOR VEHICLE EQUIPMENT
3580 .MFG-AIRCRAFT AND PARTS
3590 .MFG-AEROSPACE PRODUCTS AND PARTS
3670 .MFG-RAILROAD ROLLING STOCK
3680 .MFG-SHIP AND BOAT BUILDING

3690 .MFG-OTHER TRANSPORTATION EQUIPMENT
3770 .MFG-SAWMILLS AND WOOD PRESERVATION
3780 .MFG-VENEER, PLYWOOD, AND ENGINEERED WOOD PRODUCTS
3790 .MFG-PREFABRICATED WOOD BUILDINGS AND MOBILE HOMES
3870 .MFG-MISCELLANEOUS WOOD PRODUCTS
3890 .MFG-FURNITURE AND RELATED PRODUCTS
3960 .MFG-MEDICAL EQUIPMENT AND SUPPLIES
3970 .MFG-TOYS, AMUSEMENT, AND SPORTING GOODS
3980 .MFG-MISCELLANEOUS MANUFACTURING, N.E.C.
3990 .MFG-NOT SPECIFIED INDUSTRIES
4070 .WHL-MOTOR VEHICLES PARTS AND SUPPLIES MERCHANT WHOLESALERS
4080 .WHL-FURNITURE AND HOME FURNISHING MERCHANT WHOLESALERS
4090 .WHL-LUMBER AND OTHER CONSTRUCTION MATERIALS MERCHANT
.WHOLESALERS
4170 .WHL-PROFESSIONAL AND COMMERCIAL EQUIPMENT AND SUPPLIES
.MERCHANT WHOLESALERS
4180 .WHL-METALS AND MINERALS, EXCEPT PETROLEUM, MERCHANT
.WHOLESALERS
4190 .WHL-ELECTRICAL GOODS MERCHANT WHOLESALERS
4260 .WHL-HARDWARE, PLUMBING AND HEATING EQUIPMENT, AND SUPPLIES
.MERCHANT WHOLESALERS
4270 .WHL-MACHINERY, EQUIPMENT, AND SUPPLIES MERCHANT WHOLESALERS
4280 .WHL-RECYCLABLE MATERIAL MERCHANT WHOLESALERS
4290 .WHL-MISCELLANEOUS DURABLE GOODS MERCHANT WHOLESALERS
4370 .WHL-PAPER AND PAPER PRODUCTS MERCHANT WHOLESALERS
4380 .WHL-DRUGS, SUNDRIES, AND CHEMICAL AND ALLIED PRODUCTS
.MERCHANT WHOLESALERS
4390 .WHL-APPAREL, FABRICS, AND NOTIONS MERCHANT WHOLESALERS
4470 .WHL-GROCERIES AND RELATED PRODUCTS MERCHANT WHOLESALERS
4480 .WHL-FARM PRODUCT RAW MATERIALS MERCHANT WHOLESALERS
4490 .WHL-PETROLEUM AND PETROLEUM PRODUCTS MERCHANT
WHOLESALERS
4560 .WHL-ALCOHOLIC BEVERAGES MERCHANT WHOLESALERS
4570 .WHL-FARM SUPPLIES MERCHANT WHOLESALERS
4580 .WHL-MISCELLANEOUS NONDURABLE GOODS MERCHANT WHOLESALERS
4585 .WHL-ELECTRONIC MARKETS AGENTS AND BROKERS
4590 .WHL-NOT SPECIFIED TRADE
4670 .RET-AUTOMOBILE DEALERS
4680 .RET-OTHER MOTOR VEHICLE DEALERS
4690 .RET-AUTO PARTS, ACCESSORIES, AND TIRE STORES
4770 .RET-FURNITURE AND HOME FURNISHINGS STORES
4780 .RET-HOUSEHOLD APPLIANCE STORES
4790 .RET-RADIO, TV, AND COMPUTER STORES
4870 .RET-BUILDING MATERIAL AND SUPPLIES DEALERS
4880 .RET-HARDWARE STORES
4890 .RET-LAWN AND GARDEN EQUIPMENT AND SUPPLIES STORES
4970 .RET-GROCERY STORES
4980 .RET-SPECIALTY FOOD STORES
4990 .RET-BEER, WINE, AND LIQUOR STORES
5070 .RET-PHARMACIES AND DRUG STORES
5080 .RET-HEALTH AND PERSONAL CARE, EXCEPT DRUG, STORES

5090 .RET-GASOLINE STATIONS
5170 .RET-CLOTHING AND ACCESSORIES, EXCEPT SHOE, STORES
5180 .RET-SHOE STORES
5190 .RET-JEWELRY, LUGGAGE,AND LEATHER GOODS STORES
5270 .RET-SPORTING GOODS, CAMERA, AND HOBBY AND TOY STORES
5280 .RET-SEWING, NEEDLEWORK AND PIECE GOODS STORES
5290 .RET-MUSIC STORES
5370 .RET-BOOK STORES AND NEWS DEALERS
5380 .RET-DEPARTMENT AND DISCOUNT STORES
5390 .RET-MISCELLANEOUS GENERAL MERCHANDISE STORES
5470 .RET-FLORISTS
5480 .RET-OFFICE SUPPLIES AND STATIONARY STORES
5490 .RET-USED MERCHANDISE STORES
5570 .RET-GIFT, NOVELTY, AND SOUVENIR SHOPS
5580 .RET-MISCELLANEOUS STORES
5590 .RET-ELECTRONIC SHOPPING
5591 .RET-ELECTRONIC AUCTIONS
5592 .RET-MAIL-ORDER HOUSES
5670 .RET-VENDING MACHINE OPERATORS
5680 .RET-FUEL DEALERS
5690 .RET-OTHER DIRECT SELLING ESTABLISHMENTS
5790 .RET-NOT SPECIFIED TRADE
6070 .TRN-AIR TRANSPORTATION
6080 .TRN-RAIL TRANSPORTATION
6090 .TRN-WATER TRANSPORTATION
6170 .TRN-TRUCK TRANSPORTATION
6180 .TRN-BUS SERVICE AND URBAN TRANSIT
6190 .TRN-TAXI AND LIMOUSINE SERVICE
6270 .TRN-PIPELINE TRANSPORTATION
6280 .TRN-SCENIC AND SIGHTSEEING TRANSPORTATION
6290 .TRN-SERVICES INCIDENTAL TO TRANSPORTATION
6370 .TRN-POSTAL SERVICE
6380 .TRN-COURIERS AND MESSENGERS
6390 .TRN-WAREHOUSING AND STORAGE
6470 .INF-NEWSPAPER PUBLISHERS
6480 .INF-PUBLISHING, EXCEPT NEWSPAPERS AND SOFTWARE
6490 .INF-SOFTWARE PUBLISHING
6570 .INF-MOTION PICTURES AND VIDEO INDUSTRIES
6590 .INF-SOUND RECORDING INDUSTRIES
6670 .INF-RADIO AND TELEVISION BROADCASTING AND CABLE
6675 .INF-INTERNET PUBLISHING AND BROADCASTING
6680 .INF-WIRED TELECOMMUNICATIONS CARRIERS
6690 .INF-OTHER TELECOMMUNICATION SERVICES
6692 .INF-INTERNET SERVICE PROVIDERS
6695 .INF-DATA PROCESSING, HOSTING, AND RELATED SERVICES
6770 .INF-LIBRARIES AND ARCHIVES
6780 .INF-OTHER INFORMATION SERVICES
6870 .FIN-BANKING AND RELATED ACTIVITIES
6880 .FIN-SAVINGS INSTITUTIONS, INCLUDING CREDIT UNIONS
6890 .FIN-NON-DEPOSITORY CREDIT AND RELATED ACTIVITIES

6970 .FIN-SECURITIES, COMMODITIES, FUNDS, TRUSTS, AND OTHER
.FINANCIAL INVESTMENTS
6990 .FIN-INSURANCE CARRIERS AND RELATED ACTIVITIES
7070 .FIN-REAL ESTATE
7080 .FIN-AUTOMOTIVE EQUIPMENT RENTAL AND LEASING
7170 .FIN-VIDEO TAPE AND DISK RENTAL
7180 .FIN-OTHER CONSUMER GOODS RENTAL
7190 .FIN-COMMERCIAL, INDUSTRIAL, AND OTHER INTANGIBLE ASSETS
.RENTAL AND LEASING
7270 .PRF-LEGAL SERVICES
7280 .PRF-ACCOUNTING, TAX PREPARATION, BOOKKEEPING AND PAYROLL
.SERVICES
7290 .PRF-ARCHITECTURAL, ENGINEERING, AND RELATED SERVICES
7370 .PRF-SPECIALIZED DESIGN SERVICES
7380 .PRF-COMPUTER SYSTEMS DESIGN AND RELATED SERVICES
7390 .PRF-MANAGEMENT, SCIENTIFIC AND TECHNICAL CONSULTING
.SERVICES
7460 .PRF-SCIENTIFIC RESEARCH AND DEVELOPMENT SERVICES
7470 .PRF-ADVERTISING AND RELATED SERVICES
7480 .PRF-VETERINARY SERVICES
7490 .PRF-OTHER PROFESSIONAL, SCIENTIFIC AND TECHNICAL SERVICES
7570 .PRF-MANAGEMENT OF COMPANIES AND ENTERPRISES
7580 .PRF-EMPLOYMENT SERVICES
7590 .PRF-BUSINESS SUPPORT SERVICES
7670 .PRF-TRAVEL ARRANGEMENTS AND RESERVATION SERVICES
7680 .PRF-INVESTIGATION AND SECURITY SERVICES
7690 .PRF-SERVICES TO BUILDINGS AND DWELLINGS, EX CONSTR CLN
7770 .PRF-LANDSCAPING SERVICES
7780 .PRF-OTHER ADMINISTRATIVE, AND OTHER SUPPORT SERVICES
7790 .PRF-WASTE MANAGEMENT AND REMEDIATION SERVICES
7860 .EDU-ELEMENTARY AND SECONDARY SCHOOLS
7870 .EDU-COLLEGES AND UNIVERSITIES, INCLUDING JUNIOR COLLEGES
7880 .EDU-BUSINESS, TECHNICAL, AND TRADE SCHOOLS AND TRAINING
7890 .EDU-OTHER SCHOOLS, INSTRUCTION, AND EDUCATIONAL SERVICES
7970 .MED-OFFICES OF PHYSICIANS
7980 .MED-OFFICES OF DENTISTS
7990 .MED-OFFICE OF CHIROPRACTORS
8070 .MED-OFFICES OF OPTOMETRISTS
8080 .MED-OFFICES OF OTHER HEALTH PRACTITIONERS
8090 .MED-OUTPATIENT CARE CENTERS
8170 .MED-HOME HEALTH CARE SERVICES
8180 .MED-OTHER HEALTH CARE SERVICES
8190 .MED-HOSPITALS
8270 .MED-NURSING CARE FACILITIES
8290 .MED-RESIDENTIAL CARE FACILITIES, WITHOUT NURSING
8370 .SCA-INDIVIDUAL AND FAMILY SERVICES
8380 .SCA-COMMUNITY FOOD AND HOUSING, AND EMERGENCY SERVICES
8390 .SCA-VOCATIONAL REHABILITATION SERVICES
8470 .SCA-CHILD DAY CARE SERVICES
8560 .ENT-INDEPENDENT ARTISTS, PERFORMING ARTS, SPECTATOR SPORTS

.AND RELATED INDUSTRIES
8570 .ENT-MUSEUMS, ART GALLERIES, HISTORICAL SITES, AND SIMILAR
.INSTITUTIONS
8580 .ENT-BOWLING CENTERS
8590 .ENT-OTHER AMUSEMENT, GAMBLING, AND RECREATION INDUSTRIES
8660 .ENT-TRAVELER ACCOMMODATION
8670 .ENT-RECREATIONAL VEHICLE PARKS AND CAMPS, AND ROOMING AND
.BOARDING HOUSES
8680 .ENT-RESTAURANTS AND OTHER FOOD SERVICES
8690 .ENT-DRINKING PLACES, ALCOHOLIC BEVERAGES
8770 .SRV-AUTOMOTIVE REPAIR AND MAINTENANCE
8780 .SRV-CAR WASHES
8790 .SRV-ELECTRONIC AND PRECISION EQUIPMENT REPAIR AND
.MAINTENANCE
8870 .SRV-COMMERCIAL AND INDUSTRIAL MACHINERY AND EQUIPMENT
.REPAIR AND MAINTENANCE
8880 .SRV-PERSONAL AND HOUSEHOLD GOODS REPAIR AND MAINTENANCE
8970 .SRV-BARBER SHOPS
8980 .SRV-BEAUTY SALONS
8990 .SRV-NAIL SALONS AND OTHER PERSONAL CARE SERVICES
9070 .SRV-DRYCLEANING AND LAUNDRY SERVICES
9080 .SRV-FUNERAL HOMES, CEMETERIES AND CREMATORIES
9090 .SRV-OTHER PERSONAL SERVICES
9160 .SRV-RELIGIOUS ORGANIZATIONS
9170 .SRV-CIVIC, SOCIAL, ADVOCACY ORGANIZATIONS, AND GRANTMAKING
.AND GIVING SERVICES
9180 .SRV-LABOR UNIONS
9190 .SRV-BUSINESS, PROFESSIONAL, POLITICAL AND SIMILAR
.ORGANIZATIONS
9290 .SRV-PRIVATE HOUSEHOLDS
9370 .ADM-EXECUTIVE OFFICES AND LEGISLATIVE BODIES
9380 .ADM-PUBLIC FINANCE ACTIVITIES
9390 .ADM-OTHER GENERAL GOVERNMENT AND SUPPORT
9470 .ADM-JUSTICE, PUBLIC ORDER, AND SAFETY ACTIVITIES
9480 .ADM-ADMINISTRATION OF HUMAN RESOURCE PROGRAMS
9490 .ADM-ADMINISTRATION OF ENVIRONMENTAL QUALITY AND HOUSING
.PROGRAMS
9570 .ADM-ADMINISTRATION OF ECONOMIC PROGRAMS AND SPACE RESEARCH
9590 .ADM-NATIONAL SECURITY AND INTERNATIONAL AFFAIRS
9670 .MIL-U.S. ARMY
9680 .MIL-U.S. AIR FORCE
9690 .MIL-U.S. NAVY
9770 .MIL-U.S. MARINES
9780 .MIL-U.S. COAST GUARD
9790 .MIL-U.S. ARMED FORCES, BRANCH NOT SPECIFIED
9870 .MIL-MILITARY RESERVES OR NATIONAL GUARD
9920 .UNEMPLOYED, WITH NO WORK EXPERIENCE IN THE LAST 5 YEARS **

REFERENCES

1. Centers for Medicare and Medicaid Services, O.o.t.A., National Health Statistics Group, *2006 National Health Care Expenditures Data*. 2008.

2. Kaiser Family Foundation and Health Research and Educational Trust, T., *Employer Health Benefits: 2007 Summary of Findings.* Kaiser Family Foundation and Health Research and Educational Trust, 2008.

3. Kaiser Family Foundation and Health Research and Educational Trust, T., *Employer Health Benefits 2006 Annual Survey*. 2007.

4. Pharmaceuticals, P., *The Health Status of the United States Workforce: Findings from the National Health Interview Survey (2005).* 2007(2007).

5. Druss, B.G., et al., *The most expensive medical conditions in America.* Health Affairs, 2002. **21**(4): p. 105-111.

6. Goetzel, R.Z., et al., *Health, absence, disability, and presenteeism cost estimates of certain physical and mental health conditions affecting US employers.* Journal of Occupational and Environmental Medicine, 2004. **46**(4): p. 398-412.

7. Brady, W., et al., *Defining total corporate health and safety costs - Significance and impact - Review and recommendations.* Journal of Occupational and Environmental Medicine, 1997. **39**(3): p. 224-231.

8. Segel, J.E., *Cost-of-Illness Studies—A Primer*. 2006, RTI International, RTI-UNC Center of Excellence in Health Promotion Economics.

9. Collins, J., *The Assessment of Chronic Health Conditions on Work Performance, Absence, and Total Economic Impact for Employers.* Journal of Occupational & Environmental Medicine, 2005. **47**(6): p. 547-557.

10. Chu, C., et al., *Health-promoting workplaces - international settings development.* Health Promotion International, 2000. **15**(2): p. 155-167.

11. Danna, K. and R.W. Griffin, *Health and well-being in the workplace: A review and synthesis of the literature.* Journal of Management, 1999. **25**(3): p. 357-384.

12. Hu, F., et al., *Diet, Lifestyle, and the Risk of Type 2 Diabetes Mellitus in Women.* The New England Journal of Medicine, 2001. **345**(11): p. 790-797.

13. University of Michigan, T., *Company A: Estimated Medical Economics Report*. 2006, University of Michigan.

14. Yen, L., et al., *Association Between Wellness Score from a Health Risk Appraisal and Prospective Medical Claims Costs.* Journal of Occupational & Environmental Medicine, 2003. **45**(10): p. 1049-1057.

15. Kryder, C. and M.V. Bjamadottir. *New Methods in Predictive Modelling*. in *INFORMS*. 2007.

16. National Committee for Quality Assurance, T., *Quality Dividend Calculator*. 2001.

17. Riedel, J.E., et al., *The effect of disease prevention and health promotion on workplace productivity: A literature review.* American Journal of Health Promotion, 2001. **15**(3): p. 167-+.

18. Goetzel, R.Z., et al., *Pharmaceuticals - Cost or investment? An employer's perspective.* Journal of Occupational and Environmental Medicine, 2000. **42**(4): p. 338-351.

19. Druss, B.G., et al., *Comparing the national economic burden of five chronic conditions.* Health Affairs, 2001. **20**(6): p. 233-241.

20. Goetzel, R.Z., *The financial impact of health promotion and disease prevention programs - Why is it so hard to prove value? Introduction.* American Journal of Health Promotion, 2001. **15**(5): p. 277-280.

21. Goetzel, R.Z., et al., *The relationship between modifiable health risks and health care expenditures - An analysis of the multi-employer HERO health risk and cost database.* Journal of Occupational and Environmental Medicine, 1998. **40**(10): p. 843-854.

22. Goetzel, R.Z., et al., *Health and productivity management: Establishing key performance measures, benchmarks, and best practices.* Journal of Occupational and Environmental Medicine, 2001. **43**(1): p. 10-17.

23. Goetzel, R.Z., et al., *The health and productivity cost burden of the "top 10" physical and mental health conditions affecting six large US employers in 1999.* Journal of Occupational and Environmental Medicine, 2003. **45**(1): p. 5-14.

24. Goetzel, R.Z., et al., *Estimating the return-on-investment from changes in employee health risks on the Dow Chemical Company's health care costs.* Journal of Occupational and Environmental Medicine, 2005. **47**(8): p. 759-768.

25. Goetzel, R.Z., et al., *The business case for quality mental health services: Why employers should care about the mental health and well-being of their employees.* Journal of Occupational and Environmental Medicine, 2002. **44**(4): p. 320-330.

26. Goetzel, R.Z., et al., *Return on investment in disease management: A review.* Health Care Financing Review, 2005. **26**(4): p. 1-19.

27. Goossens, M., et al., *The cost diary: a method to measure direct and indirect costs in cost-effectiveness research.* Journal of Clinical Epidemiology, 2000. **53**(7): p. 688-695.

28. US Census Bureau, T., *2006 American Community Survey*. 2006.

29. National Center For Health Statistics, T., *National Health Interview Survey*. 2008.

30. National Center For Health Statistics, T., *National Health and Nutrition Examination Survey*. 2008.

31. Center For Disease Control, T., *Behavioral Risk Factor Surveillance System*. 2007.

32. National Center For Health Statistics, T., *International Classification of Diseases, Ninth Revision (ICD-9)*. 2007.

33. Ross, S.M., *Stochastic Processes*. 2nd ed. 1996, New York, NY: John Wiley and Sons Inc.

34. Berger, J.O., *Statistical Decision Theory and Bayesian Analysis*. 2 ed. Soringer Series in Statistics. 1985, New York: Springer-Verlage New York Inc. 598.

35. N.V. Patel, M.E.B., K.B. Kolodner, C. Leotta, J.E. Lafata, R.B. Lipton, *Prevalence and Impact of Migraine and Probable Migraine in a health plan.* Neurology, 2004. **2004**(63): p. 1432-1438.

36. Mygind, N. and G. Scadding, *Allergic Rhinitis*. 2000: Health Press. 52.

37. Bachert, C., et al., *Prevalence, classification and perception of allergic and nonallergic rhinitis in Belgium.* Allergy, 2006. **61**(6): p. 693-8.

38. Turjanmaa, K. and S. Makinen-Kiljunen, *Latex allergy: prevalence, risk factors, and cross-reactivity.* Methods, 2002. **27**: p. 10-14.

39. Sakurai, Y., et al., *Prevalence and risk factors of allergic rhinitis and cedar pollinosis among Japanese men.* Prev Med, 1998. **27**(4): p. 617-22.

40. Rosenwasser, L.J., *Treatment of allergic rhinitis.* American Journal of Medicine, 2002. **113**(9A): p. 17S-24S.

41. Jones, N., *Allergic rhinitis: Aetiology, predisposing and risk factors.* Rhinology, 2004. **42**(2): p. 49-56.

42. Hepner, D.L. and M.C. Castells, *Latex allergy: An update.* Anesthesia and Analgesia, 2003. **96**(4): p. 1219-1229.

43. Mayo Foundation for Medical Education and Research, T. 2007.

44. Foltz-Gray, D., *The Arthritis Foundation's Guide to Good Living with Rheumatoid Arthritis*. 1 ed. 2006: Arthritis Foundation.

45. CDC, *The burden of chronic diseases and their risk factors*, CDC, Editor. 2004, DHS.

46. Center For Disease Control, T., *National Center for Chronic Disease Prevention and Health Promotion*. 2008, Center for Disease Control.

47. Weir, P.T., et al., *The incidence of fibromyalgia and its associated comorbidities - A population-based retrospective cohort study based on International Classification of Diseases, 9th Revision codes.* Journal of Clinical Rheumatology, 2006. **12**(3): p. 124-128.

48. Sharma, L., D. Kapoor, and S. Issa, *Epidemiology of osteoarthritis: an update.* Current Opinion in Rheumatology, 2006. **18**(2): p. 147-156.

49. National Institutes of Health, T., *US National Library of Medicine*. 2008, National Institutes of Health.

50. Hootman, J.M. and C.G. Helmick, *Projections of US prevalence of arthritis and associated activity limitations.* Arthritis and Rheumatism, 2006. **54**(1): p. 226-229.

51. Alamanos, Y. and A.A. Drosos, *Epidemiology of adult rheumatoid arthritis.* Autoimmunity Reviews, 2005. **4**(3): p. 130-136.

52. Aho, K. and M. Heliovaara, *Risk factors for rheumatoid arthritis.* Annals of Medicine, 2004. **36**(4): p. 242-251.

53. Health, N.I.o. 2008.

54. Kaliner, M.A., P.J. Barnes, and G.A. Persson, *Asthma: It's Pathology and Treatment*. 1991: Informa Health Care. 779.

55. Redd, S.C., *Asthma in the United States: Burden and Current Theories.* Environmental Health Perspectives, 2002. **110**: p. 557-560.

56. Koh, Y.Y. and C.K. Kim, *The development of asthma in patients with allergic rhinitis.* Curr Opin Allergy Clin Immunol, 2003. **3**(3): p. 159-64.

57. King, M.E., D.M. Mannino, and F. Holguin, *Risk factors for asthma incidence - A review of recent prospective evidence.* Panminerva Medica, 2004. **46**(2): p. 97-110.

58. Gwynn, R.C., *Risk factors for asthma in US adults: Results from the 2000 behavioral risk factor surveillance system.* Journal of Asthma, 2004. **41**(1): p. 91-98.

59. Guerra, S., et al., *Rhinitis as an independent risk factor for adult-onset asthma.* Journal of Allergy and Clinical Immunology, 2002. **109**(3): p. 419-425.

60. Asthma and Allergy Foundation of America, T., *Asthma and Allergy Foundation of America*. 2008, AAFA.

61. Morris, D., J.H. Kearsley, and C. Williams, *Cancer: A Comprehensive Clinical Guide*. 1998: Taylor and Francis. 321.

62. Bostwick, D.G., et al., *Human prostate cancer risk factors.* Cancer, 2004. **101**(10): p. 2371-2490.

63. Brekelmans, C.T.M., *Risk factors and risk reduction of breasts and ovarian cancer.* Current Opinion in Obstetrics and Gynecology, 2003. **15**: p. 63-68.

64. Cocco, P., E.F. Heineman, and M. Dosemeci, *Occupational risk factors for cancer of the central nervous system (CNS) among US women.* American Journal of Industrial Medicine, 1999. **36**(1): p. 70-74.

65. Coups, E.J., et al., *Multiple behavioral risk factors for colorectal cancer and colorectal cancer screening status.* Cancer Epidemiology Biomarkers & Prevention, 2007. **16**(3): p. 510-516.

66. Fitzgibbons, P., et al., *Prosnostic Factors in Breast Cancer - College of American Pathologists Consensus Statement.* Archives of pathology & laboratory medicine, 2000. **124**(7): p. 966-978.

67. Gerber, B., et al., *Nutrition and lifestyle factors on the risk of developing breast cancer.* Breast Cancer Research and Treatment, 2003. **79**(2): p. 265-276.

68. Giovannucci, E., *Modifiable risk factors for colon cancer.* Gastroenterology Clinics of North America, 2002. **31**(4): p. 925-+.

69. Gronberg, H., *Prostate cancer epidemiology.* Lancet, 2003. **361**(9360): p. 859-864.

70. McTiernan, A., *Behavioral risk factors in breast cancer: Can risk be modified?* Oncologist, 2003. **8**(4): p. 326-334.

71. Robertson, I., R. Bound, and L. Segal, *Colorectal cancer, diet and lifestyle factors: opportunities for prevention.* Health Promotion International, 1998. **13**(2): p. 141-150.

72. Setiawan, V.W., et al., *Racial/ethnic differences in endometrial cancer risk: The multiethnic cohort study.* American Journal of Epidemiology, 2007. **165**(3): p. 262-270.

73. Singletary, S.E., *Rating the risk factors for breast cancer.* Annals of Surgery, 2003. **237**(4): p. 474-482.

74. Stein, C.J. and G.A. Colditz, *Modifiable risk factors for cancer.* British Journal of Cancer, 2004. **90**(2): p. 299-303.

75. Tominaga, S., *Major avoidable risk factors of cancer.* Cancer Letters, 1999. **143**: p. S19-S23.

76. van Loon, A.J.M., et al., *Lifestyle risk factors for cancer: the relationship with psychosocial work environment.* International Journal of Epidemiology, 2000. **29**(5): p. 785-792.

77. Weiland, S.K., et al., *Workplace risk factors for cancer in the German rubber industry: part 1. Mortality from respiratory cancers.* Occupational and Environmental Medicine, 1998. **55**(5): p. 317-324.

78. Whittemore, A.S., et al., *Prostate-Cancer in Relation to Diet, Physical-Activity, and Body-Size in Blacks, Whites, and Asians in the United-States and Canada.* Journal of the National Cancer Institute, 1995. **87**(9): p. 652-661.

79. Wrensch, M., et al., *Risk factors for breast cancer in a population with high incidence rates.* Breast Cancer Research, 2003. **5**(4): p. R88-R102.

80. Blazer, D.G., et al., *The Prevalence and Distribution of Major Depression in a National Community Sample - the National Comorbidity Survey.* American Journal of Psychiatry, 1994. **151**(7): p. 979-986.

81. Barkow, K., et al., *Risk factors for depression at 12-month follow-up in adult primary health care patients with major depression: an international prospective study.* Journal of Affective Disorders, 2003. **76**(1-3): p. 157-169.

82. Harlow, B.L., *Demographic, family, and occupational characteristics associated with major depression: the Harvard study of moods and cycles.* Acta Psychiatrica Scandinavica, 2002. **105**(3): p. 209-217.

83. Horwath, E., et al., *Depressive Symptoms as Relative and Attributable Risk-Factors for 1st-Onset Major Depression.* Archives of General Psychiatry, 1992. **49**(10): p. 817-823.

84. Katon, W. and H. Schulberg, *Epidemiology of Depression in Primary Care.* General Hospital Psychiatry, 1992. **14**(4): p. 237-247.

85. Kessler, R.C., et al., *Lifetime and 12-Month Prevalence of Dsm-Iii-R Psychiatric-Disorders in the United-States - Results from the National-Comorbidity-Survey.* Archives of General Psychiatry, 1994. **51**(1): p. 8-19.

86. Lorant, V., et al., *Socioeconomic inequalities in depression: A meta-analysis.* American Journal of Epidemiology, 2003. **157**(2): p. 98-112.

87. Salokangas, R.K.R. and O. Poutanen, *Risk factors for depression in primary care - Findings of the TADEP project.* Journal of Affective Disorders, 1998. **48**(2-3): p. 171-180.

88. Weich, S., et al., *Mental health and the built environment: cross-sectional survey of individual and contextual risk factors for depression.* British Journal of Psychiatry, 2002. **180**: p. 428-433.

89. Hilding, A., et al., *The impact of familly history of diabetes and lifestyle on abnormal glucose regulation in middle-aged Swedish men and women.* Diabetologia, 2006. **49**: p. 2589-2598.

90. Boyle, J.P., et al., *Projection of Diabetes Burden Through 2050: Impact of changing demography and disease prevalence in the U.S.* Diabetes Care, 2001. **24**(11): p. 1936-1940.

91. Labs, S. 2007.

92. Wilson, P., et al., *Prediction of Coronary Heart Disease Using Risk Factor Categories.* Journal of the American Heart Association, 1998. **97**: p. 1837-1847.

93. Folsom, A., et al., *Prediction of Coronary Heart Disease in Middle-Aged Adults with Diabetes.* DIabetes Care, 2003. **26**: p. 2777.

94. Chang, M.-h., et al., *Multiple risk factors and population attributable risk for ischemic heart disease mortality in the United States, 1971-1992.* Journal of Clinical Epidemiology, 2001. **54**(6): p. 634-644.

95. Greenland, P., et al., *Major risk factors as antecedents of fatal and nonfatal coronary heart disease events.* Jama-Journal of the American Medical Association, 2003. **290**(7): p. 891-897.

96. Hackam, D.G. and S.S. Anand, *Emerging risk factors for atherosclerotic vascular disease - A critical review of the evidence.* Jama-Journal of the American Medical Association, 2003. **290**(7): p. 932-940.

97. Kannel, W.B., *Risk stratification in hypertension: new insights from the Framingham study.* American Journal of Hypertension, 2000. **13**(1, Supplement 1): p. S3-S10.

98. Khot, U.N., et al., *Prevalence of conventional risk factors in patients with coronary heart disease.* Jama-Journal of the American Medical Association, 2003. **290**(7): p. 898-904.

99. Thorpy, J., *Risk Factors for Heart Disease.* JAMA, 2003. **290**(7): p. 280.

100. Skarfors, E.T., H.O. Lithell, and I. Selinus, *Risk-Factors for the Development of Hypertension - a 10-Year Longitudinal-Study in Middle-Aged Men.* Journal of Hypertension, 1991. **9**(3): p. 217-223.

101. Laurenzi, M., et al., *Multiple Risk-Factors in Hypertension - Results from the Gubbio Study.* Journal of Hypertension, 1990. **8**: p. S7-S12.

102. Scher, A.I., et al., *Prevalence of frequent headache in a population sample.* Headache, 1998. **38**(7): p. 497-506.

103. Katsarava, Z., et al., *Incidence and predictors for chronicity of headache in patients with episodic migraine.* Neurology, 2004. **62**(5): p. 788-790.

104. Patel, N., et al., *Prevalence and Impact of Migraine and Probable Migraine in a health plan.* Neurology, 2004. **2004**(63): p. 1432-1438.

105. Bigal, M.E., J.N. Liberman, and R.B. Lipton, *Obesity and migraine - A population study.* Neurology, 2006. **66**(4): p. 545-550.

106. Bigal, M.E. and R.B. Lipton, *Obesity is a risk factor for transformed migraine but not chronic tension-type headache.* Neurology, 2006. **67**(2): p. 252-257.

107. Bigal, M.E. and R.B. Lipton, *Modifiable risk factors for migraine progression (or for chronic daily headaches) - Clinical lessons.* Headache, 2006. **46**: p. S144-S146.

108. Bigal, M.E. and R.B. Lipton, *Modifiable risk factors for migraine progression.* Headache, 2006. **46**(9): p. 1334-1343.

109. Scher, A.I., W.F. Stewart, and R.B. Lipton, *Caffeine as a risk factor for chronic daily headache - A population-based study.* Neurology, 2004. **63**(11): p. 2022-2027.

110. Scher, A.I., et al., *Factors associated with the onset and remission of chronic daily headache in a population-based study.* Pain, 2003. **106**(1-2): p. 81-89.

111. Viegi, G. and C. Di Pede, *Chronic obstructive lung diseases and occupational exposure.* Curr Opin Allergy Clin Immunol, 2002. **2**(2): p. 115-21.

112. Trupin, L., et al., *The occupational burden of chronic obstructive pulmonary disease.* Eur Respir J, 2003. **22**(3): p. 462-469.

113. Hnizdo, E., et al., *Association between Chronic Obstructive Pulmonary Disease and Employment by Industry and Occupation in the US Population: A Study of Data from the Third National Health and Nutrition Examination Survey.* Am. J. Epidemiol., 2002. **156**(8): p. 738-746.

114. Anderson, D.R., et al., *The relationship between modifiable health risks and group-level health care expenditures.* American Journal of Health Promotion, 2000. **15**(1): p. 45-52.

115. Blackburn, G.L. and B.S. Kanders, *Obesity: Pathophysiology, Physcology and Treatment*. 1994: Jones & Bartlett. 368.

116. World Health Organization, T., *Obesity: Preventing and Managing the Global Epidemic*. 2000, Singapore: World Health Organization. 253.

117. WebMD, T., *100 Worst Spring Allergy Cities*. 2008, WebMD.

118. Joseph, L., D.B. Wolfson, and R. Du Berger, *Sample size calculations for binomial proportions via highest posterior density intervals.* The Statistician, 1995. **44**: p. 143-154.

119. Elena Losinaa, b., *, Jane Barrettd,e, John A. Baronc,d,e, Jeffrey N. Katzb, *Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients.* Journal of Clinical Epidemiology, 2003. **56**: p. 515-519.

120. Fredric D. Wolinsky, P., *† Thomas R. Miller, MBA,† Hyonggin An, PhD,† John F. Geweke, PhD,†, et al., *Hospital Episodes and Physician Visits:The Concordance Between Self-Reports and Medicare Claims.* Medical Care, 2007. **45**(4): p. 300-307.

121. Diana M. Tisnado, J.L.A., Honghu Liu, F.A.H. Cheryl L. Damberg, Wen-Pin Chen, and C.M.M. David M. Carlisle, Katherine L. Kahn, *Does the concordance between medical records and patient self-report vary with patient characteristics?* Health Serv Outcomes Res Method, 2006. **2006**(6): p. 157-175.

VITA

Leanne Metcalfe

Leanne Metcalfe was born and raised in Kingston, Jamaica. She received her B.E. with High Honours in Electrical Engineering from Stevens Institute of Technology in Hoboken NJ in 2000. She then received her MS in Industrial Engineering from Georgia Institute of Technology in the summer of 2002. Leanne then stayed on at Georgia Tech to pursue her PhD in Biomedical Engineering. Ms. Metcalfe has been involved in health and wellness for most of her career. In undergrad she was a member of three varsity sports teams, and in her summers interned with the pharmaceutical company Merck and Company. She obtained her personal training certification when she became a student at Georgia Tech. Ms. Metcalfe is also interested in diversity of thought and enjoys attending cultural events and traveling.