

ARE THERE ASSOCIATIONS BETWEEN
PARAESOPHAGEAL HERNIA AND COLLAGEN
DISORDERS OR OTHER DISEASES?
--- A DATA MINING APPROACH

by

Jianji Yang, MA

A MASTER'S THESIS

Presented to the Department of Medical Informatics & Clinical Epidemiology

And the Oregon Health & Science University School of Medicine

In partial fulfillment of the requirements for the degree of

Master of Science

September 2003

School of Medicine
Oregon Health & Science University

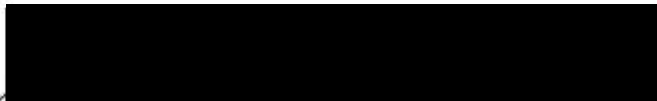
Certificate of Approval

This is to certify that the Masters thesis of

Jianji Yang

*“Are there associations between paraesophageal hernia
and collagen disorders or other diseases?
--A data mining approach”*

Has been approved



Professor in charge of thesis



Member



Member



Member

TABLE OF CONTENTS

Acknowledgements	iii
Abstract.....	iv
1. Introduction.....	1
<i>1.1 Hiatal Hernia.....</i>	<i>1</i>
<i>1.2 Collagens and Related Diseases.....</i>	<i>4</i>
<i>1.3 Previous Work.....</i>	<i>5</i>
<i>1.4 Goal of This Study.....</i>	<i>7</i>
<i>1.5 Data Mining.....</i>	<i>8</i>
<i>1.6 Data Source</i>	<i>12</i>
2. Methodology	14
<i>2.1 Problem Definition.....</i>	<i>14</i>
<i>2.2 Data Preprocessing</i>	<i>14</i>
<i>2.3 Data Mining.....</i>	<i>19</i>
<i>2.4 Post-Data Mining Analysis.....</i>	<i>22</i>
3. Results	24
<i>3.1 Descriptive Statistics.....</i>	<i>24</i>
<i>3.2 Association Rule Mining.....</i>	<i>25</i>
<i>3.3 Classification</i>	<i>25</i>
4. Discussion.....	31
<i>4.1 Major Findings</i>	<i>31</i>
<i>4.2 The Advantage of Using Data Mining Technology.....</i>	<i>32</i>
<i>4.3 Results from the Classification Process.....</i>	<i>33</i>
<i>4.4 Limitations of the Study</i>	<i>34</i>

<i>4.5 Future Work</i>	36
5. Conclusion	39
6. References	40
7. Appendices	43

Acknowledgements

I would like to express my sincere thanks to my committee members whose assistance made this thesis possible: Dr. Judith Logan, my advisor, for her guidance, inspiration, and unwavering support; Dr. Dale Kraemer, for his encouragement and generous help in statistics; Dr. Cynthia Morris, for her direction and strong support; and Dr. Blair Jobe, for giving me the opportunity to work on the project and sharing his expertise in gastroenterology.

Thanks are also due to Dr. Aaron Cohen for his elegant computer program that made the control selection process less laborious.

Furthermore, I would like to express my gratitude to my friends in the OHSU Department of Medical Informatics and Clinical Epidemiology, whose friendships were invaluable and kept me going through the whole project.

Finally, special thanks to my husband and sons, Yizhi, Bill, and Kevin Wang, for their support during these two years of graduate study.

Abstract

A hiatal hernia occurs when a portion of the stomach prolapses through the diaphragmatic esophageal hiatus. Even though it is believed that the major cause of hiatal hernias is loss of elasticity of supporting structures [1,2], the detailed mechanisms have not been fully studied. There are reports of a possible association between hiatal hernias and connective tissue weakness [6]. Recently, studies of inguinal and incisional hernias have shown their connection to collagen matrix defects [9, 10]. Since collagen is the major stress-bearing component of connective tissue, it is speculated that collagen defects may also cause hiatal hernias. The study reported here used data mining techniques to explore the associations between paraesophageal hernia (PEH), a severe type of hiatal hernia, and other disorders, including collagen. The data source used in this study was the 1999 National Inpatient Sample dataset of the Healthcare Cost and Utilization Project, the largest all-payer inpatient discharge dataset in the United States. Association rule mining (Apriori algorithm implemented in the program Classification Based on Associations) and classification (CART®) were used in the data mining process. This two-step analysis failed to detect associations between collagen diseases and PEH. Instead, the results suggest that gall bladder and bile duct diseases and peritoneal adhesions are possible PEH-associated factors. A decision tree was also built in the classification step to differentiate sliding and paraesophageal hernias.

1. Introduction

1.1 Hiatal Hernia

A hiatal hernia is a herniation of part of the stomach into the thoracic cavity through the esophageal hiatus of the diaphragm. The existence of this disease has been acknowledged for more than 400 years and it is recently getting attention because of its association with gastroesophageal reflux disease [30-32] and because of the life-threatening condition caused by strangulation and perforation of paraesophageal hernias, a severe form of hiatal hernia (Figure 1, 2 and 3).

It is believed that causes of hiatal hernia include the following [1,2]:

1. Excessive contraction of the longitudinal muscle of the esophagus.
2. Loss of elasticity of the supporting structures due to increase in age.
3. Shortening of the esophagus due to fibrosis caused by esophagitis.
4. Previous surgery.

Although hiatal hernia is one of the most common abnormalities of the gastrointestinal (GI) tract in Western World [1, 2], its true prevalence is still unknown due to the lack of commonly agreed-upon diagnostic standards and the large number of asymptomatic patients [45]. Despite this uncertainty, it is well established that this condition is more prevalent in the Western World as compared to Asia or Africa and, also, that it is more common in the elderly and in women [1].

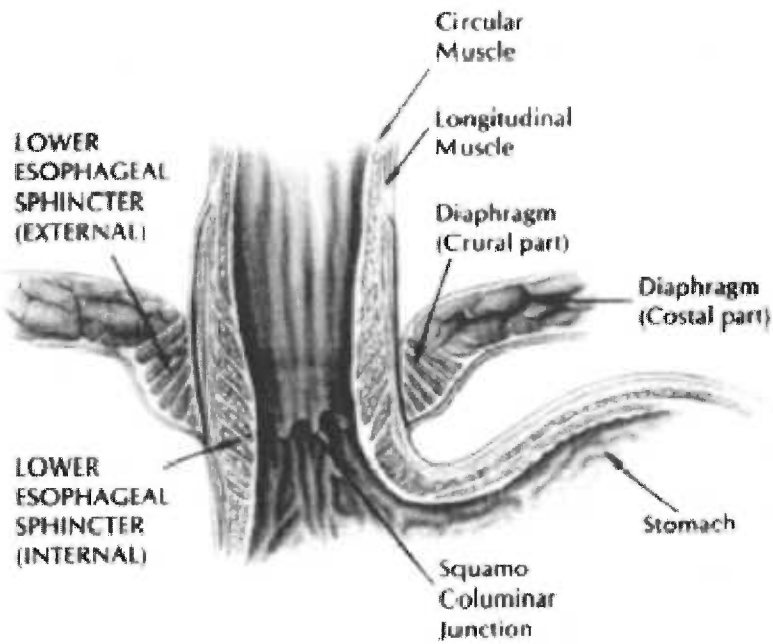


Figure 1. Normal anatomic structure at the esophagogastric junction (also called the squamo-columnar junction). Note that the esophagogastric junction is lined up with the diaphragm (adapted from Mittal (1997) [1]).

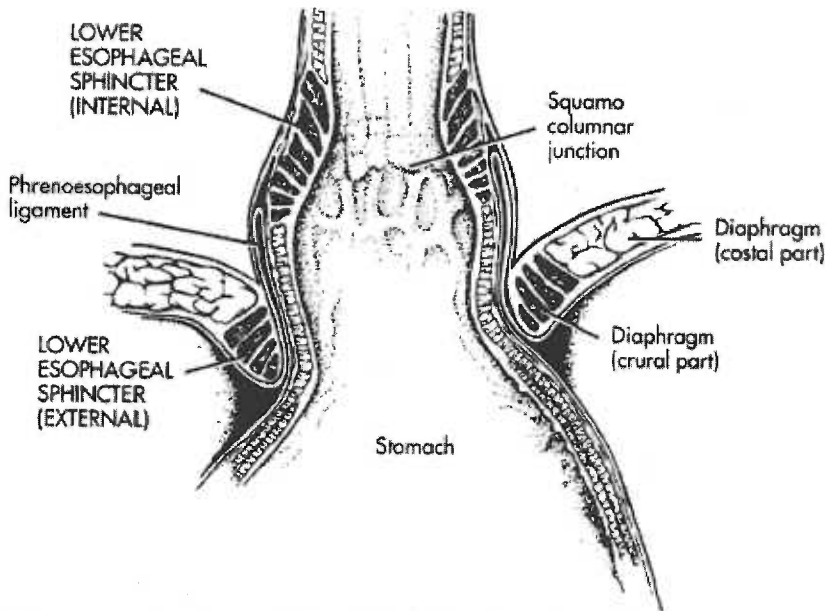


Figure 2. Type I or sliding hiatal hernia where the esophagogastric junction (squamo-columnar junction) is above the diaphragm (adapted from Mittal (1997) [1]).

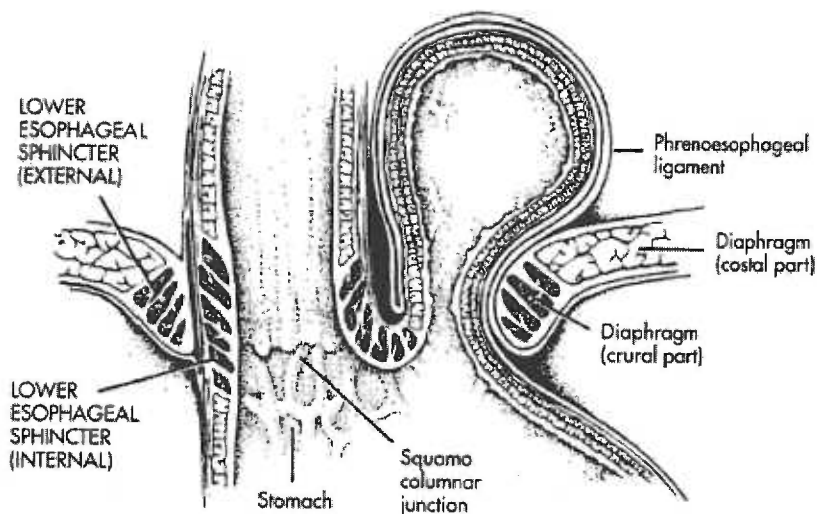


Figure 3. Type II or paraesophageal hernia. The esophagogastric junction (squamocolumnar junction) is in the normal location but the gastric fundus protrudes into the thoracic cavity through the widened diaphragmatic hiatus (adapted from Mittal (1997) [1]).

Hiatal hernias may be classified into three types. The first type is the "sliding" or type I hiatal hernia, where the esophagogastric junction (EG junction) is above the diaphragmatic hiatus. Figure 2 shows the upward dislocation of the EG junction and may be compared with the normal anatomic structures as shown in Figure 1. This is by far the most common type of hiatal hernia, with seven times higher incidence rate than the other two types [45]. The second type is the "paraesophageal" or type II hiatal hernia, where the EG junction is at the normal position but the widened hiatus allows the fundus of the stomach to protrude into the chest (Figure 3). This type is very rare. A little more common, but still rare, is the mixed hiatal hernia [1, 2]. Mixed hiatal hernias, or type III hernias are characterized by both an upward dislocation of the GE junction and the protrusion of the gastric fundus. Although type II and III hernias are both rare, they are far more dangerous than type I: 25% of patients [45] with these two types of hernia develop incarceration and strangulation of the hernia, which may lead to perforation,

excessive bleeding or volvulus. If these emergency conditions occur, the mortality rate is high [3]. Because of these catastrophic life-threatening complications, more attention has been paid to type II and III hiatal hernias than to the relatively benign sliding hiatal hernias. For the same reason, both type II and III hernias were the target of this study and, for convenience, will for the remainder of this document both be called paraesophageal hernias (PEH).

1.2 Collagens and Related Diseases

Collagens are a family of extracellular matrix proteins. They are organized into insoluble fibers of great tensile strength. Due to its stress-bearing structural characteristics, collagen has a dominant role in maintaining the structural integrity of various organs and tissues, such as bone, teeth, cartilage, tendon, ligament and the fibrous matrices of skin and blood vessels.

At least 23 members of the collagen superfamily have been described. In addition, there are 15 more proteins with collagen-like domains. Collagen is formed by three parallel polypeptide chains that wind around each other with a right-handed, ropelike twist to form a triple-helical structure (Figure 4). Each helix is composed of the Gly-X-Y pattern where every third amino acid in the chain is a glycine moiety. More than a thousand mutations in 22 genes have been characterized which lead to collagen disorders [4]. These diseases are generally categorized into two classes, diseases caused by mutations in genes for collagens and diseases caused by mutations in genes for collagen-

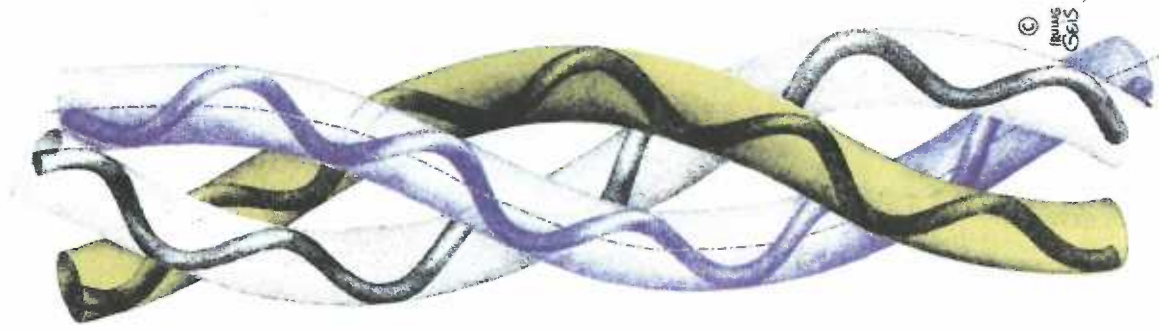


Figure 4 Artistic illustration of the triple helices of collagen fiber (adapted from Geis (1990) [5])

processing enzymes. They include osteogenesis imperfecta, many chondrodysplasias, several subtypes of the Ehlers-Danlos syndrome, Bethlem myopathy, Alport syndrome, some subtypes of epidermolysis bullosa, and Knobloch syndrome. It is also postulated that some cases of osteoporosis, arterial aneurysms, osteoarthrosis, and intervertebral disc disease are caused by collagen defects [4]. Table 1 lists the diseases caused by mutations in genes for collagens or collagen-processing enzymes.

1.3 Previous Work

Since collagen is the major stress-bearing component of connective tissue [4], it would be reasonable to suspect that collagen defects are responsible for the connective tissue weakness that leads to some hiatal hernia formation. To date, there are no direct studies addressing this issue. There is some evidence in the literature, however, pointing in this direction. In a 1997 paper, Horvath and associates reported that cardiac investigation of 23 pediatric patients with sliding hiatal hernia revealed mitral valve prolapse in 18 patients, suggesting a high association rate of 78.2% [6]. Mitral valve prolapse is a congenital anomaly and is often associated with other signs of connective tissue

Gene or enzyme	Disease
COL1A1; COL1A2	Osteogenesis imperfecta Ehlers-Danlos syndrome type I, II, VIIA, VIIB Osteoporosis
COL2A1	Several chondrodysplasias Osteoarthritis
COL3A1	Ehlers-Danlos syndrome type IV Arterial aneurysms
COL4A3; COL4A4; COL4A5	Alport syndrome
COL4A5 and COL4A6	Alport syndrome with diffuse oesophageal leiomyomatosis
COL5A1; COL5A2	Ehlers-Danlos syndrome types I and II
COL6A1; COL6A2; COL6A3	Bethlem myopathy
COL7A1	Ehlers-Danlos syndrome, dystrophic forms, epidermolysis bullosa
COL9A1; COL9A2; COL9A3	Multiple epiphyseal dysplasia Intervertebral disc disease Osteoarthritis
COL10A1	Schmid metaphyseal Chondrodysplasia
COL11A1; COL11A2	Several mild chondrodysplasias Non-syndromic hearing loss Osteoarthritis
COL17A1	Generalized atrophic benign Ehlers-Danlos syndrome, epidermolysis bullosa
COL18A1	Knobloch syndrome
Lysyl hydroxylase 1	Ehlers-Danlos syndrome type VI
Procollagen N-proteinase	Ehlers-Danlos syndrome type VIIC

Table 1. Diseases caused by mutations in genes for collagens or collagen-processing enzymes (adapted from Myllyharju (2001) [4]).

weakness, such as Ehlers-Danlos syndrome and Marfan's syndrome, both of which are collagen disorders. In addition, an increased risk for primary inguinal hernias in Ehlers-Danlos patients has been reported in the literature [43]. Furthermore, Giroto and associates conducted a retrospective chart review of patients with recurrent ventral herniation and found that 10% of these patients had Ehler-Danlos syndrome, a rate considerably higher than expected [7]. In a 1999 paper, a group of researchers led by Pans made a connection between collagen matrix defects and groin hernias. Using

histochemical and biomechanical techniques, the authors demonstrated the pathological characteristics of connective tissue in the herniated fasciae, revealing a malformed collagen framework with areas of disorganization and an increased number of isolated fibers [8]. Recently, Zheng and associates used RT-PCR and Northern Blot techniques to investigate the expression of procollagen type I to type III ratio and matrix metalloproteinase (MMP) 1 and 13 (enzymes that break down collagens) mRNAs in the skin fibroblast of inguinal hernia patients, as compared to healthy controls. The results strongly suggest that inguinal hernia is a disease of the collagen matrix due to a decreased ratio of collagen type I to type III, and increased MMP-1 and MMP-13 expressions [9, 11, 13]. These researchers also used the same techniques to demonstrate the connection between collagen matrix defects and incisional hernias [10, 12].

1.4 Goal of This Study

This study addressed the question of association between PEH and collagen diseases using exploratory data mining techniques. The goal of this study was to elucidate the probable connections between PEH and other conditions, such as collagen disorders. Unlike traditional hypothesis-oriented statistical analysis, however, which looks at associations between two specific conditions, this exploratory approach is able to ask and answer open-ended questions and, therefore, potentially detect connections that have not yet been considered.

The information obtained from this study might be used to formulate hypothesis-oriented clinical epidemiological studies. Experimental research using animal models might also

be undertaken in hopes of revealing the mechanism of detected associations and obtaining a clearer understanding of the underlying pathophysiology. In addition, the information from this study could be important in preventive care and screening of high-risk patients to avoid the emergency condition of PEH perforation. It could also be useful for supporting therapeutic strategies in hernia surgery, such as the decision to use surgical meshes in hernia repair.

1.5 Data Mining

With computerization of virtually every sector of the world, information collected in databases is growing at a phenomenal rate. Traditional ad hoc analysis of data using statistical techniques and data management tools is no longer sufficient for handling this large volume of data. It is estimated that only 5-10% of commercial databases have ever been analyzed [42]. This situation led to the emergence and rapid growth of data mining techniques in the 1990's. Data mining, also known as Knowledge Discovery in Databases (KDD), is defined as the extraction of implicit, previously unknown, and potentially useful patterns from data.

Data mining is closely related to statistics in that it uses statistical algorithms in addition to other methods and technologies. KDD focuses on data-oriented exploratory analysis, in the sense that patterns and hypotheses are automatically extracted from the data. Traditional statistics, on the other hand, is often hypothesis-oriented. Once a hypothesis is formed, it is then validated against the data. Another way of saying this is that data mining tends to be data-driven and generates hypotheses whereas statistics is human-

driven and formally tests hypotheses. Furthermore, data mining draws upon methods, algorithms and technologies from a variety of diverse fields in data analysis, including artificial intelligence, data warehousing, pattern recognition and computer visualization.

Some of the common tasks in data mining, as defined by Zhang and Zhang (2002), include:

- **Association rule mining.** Association rule mining discovers the association relationships among a set of items, i.e. detects the items that frequently occur together. The dataset for analysis consists of a set of transactions, each of which contains a set of data items. The number of items in each transaction in the dataset can be different. A classic example of association rule mining (also known as known as market basket analysis) is the analysis of supermarket transaction data. In this example, the goal of association rule mining is to find out which items are commonly purchased at the same time. An association rule found in the mining process might look like this: $\text{milk} \Rightarrow \text{bread}$ [cover = 80%, support = 56%, confidence = 70%]. This says that 80% of the customers buy milk (the cover); 56% of the customers buy milk and bread together (the support) and those who buy milk buy bread 70% of the time (the confidence). The following are the definitions of the parameters that measure the strength of a rule: in the association rule $A \Rightarrow B$, *cover* of the rule is defined as the percentage of transactions that contain A; *support* is the percentage of transactions that contain both A and B, and *confidence* is $\text{support}/\text{cover}$, i.e., ($\% \text{ transactions that contain both A and B} / \% \text{ transactions that contain A}$).

- **Classification.** Classification assigns unknown data patterns to a set of established classes. For example, one of the tasks of this study is to assign patients into classes of having a PEH and not having a PEH (non-PEH) based on a set of variables including other diagnoses and demographics. A very popular method for classification is decision tree-based classification, in which the classification result can be presented in an easy-to-understand graphical tree structure. In the above PEH example, various factors (age, race, gender, and other medical conditions) can be fit into a decision tree to help investigators predict to which class a patient is likely to belong. In decision tree-based classification, a training process, commonly referred to as model fitting, uses part of the dataset to create the decision tree, which is followed by testing the model for its validity using another part of the dataset. There are parameters that describe the characteristics and the performance of the decision trees built in the mining process: *prediction success rate* is the percentage a tree can predict correctly for each class; *complexity* describes how complex a tree is and is usually represented by the number of terminal nodes; *cost* of a tree is a numerical value that is calculated from a matrix specifying the relative weight of classifying cases into the wrong classes. The lower the cost of the tree, the better it is. The matrix used to calculate the cost can be assigned by the investigator to emphasize the importance of correct prediction of certain classes. For example, if it is more important to correctly predict PEH cases than non-PEH cases, misclassification of PEH cases into non-PEH can be assigned to carry a 50% higher matrix value than misclassification of non-PEH cases into PEH.

- **Clustering.** Clustering is the unsupervised grouping of patterns (observations, data points, or attribute values) into classes. Unlike classification, the final classes are not predefined. Given a set of data points, the aim of clustering is to find a schema to group the data into classes such that each data point is similar in some way to its class members but different from members of the other classes. Research has focused largely on Euclidean distance-based clustering analysis. Bayesian automatic classification systems (such as is used in the program AutoClass, developed by the Bayes group at Ames Research Center of NASA) and self-organizing Kohonen neural networks are other examples of clustering applications.
- **Time-series analysis.** Time-series analysis is a data mining technique involving periods of time. For example, time-series analysis can be used to predict the trend of the Dow-Jones Industrial Average using past history, current market conditions, and current economic and political situations of the country.

In general, the process of knowledge discovery in databases consists of the following steps [36]:

- **Problem definition:** the goals of the knowledge discovery are identified.
- **Data preprocessing.** Data preprocessing is the most time consuming and laborious step. The success of the project depends heavily on this step. It consists of the following five processes:
 - **Data collection:** obtain necessary data from various internal and external sources.

- Data cleaning: resolve data conflicts, missing data and ambiguity.
 - Data selection: select relevant data for the data mining problem.
 - Data transformation: transform data into the form appropriate for the mining program.
- Data mining: intelligent methods and algorithms are applied to extract data patterns.
 - Post-data mining: model comparison and evaluation for relevancy.

Data mining has been used successfully as a decision support tool in a number of business and industrial sectors, such as marketing, banking and telecommunications. Now these techniques are beginning to be used in biomedical research. For example, cluster analysis is used with micro-array data. In this study, data mining was used instead of traditional statistical methods because of data mining's exploratory characteristics. By using a data mining approach, this study also demonstrated the usefulness of this analytical technique for biomedical data.

1.6 Data Source

The data source used for this study was the Nationwide Inpatient Sample (NIS) database, which is part of the Healthcare Cost and Utilization Project (HCUP). HCUP [34] is the result of a federal-state-industry partnership to build a standardized, multi-state health data system. HCUP is maintained by the Agency for Healthcare Research and Quality (AHRQ). It is the largest collection of all-payer, encounter level hospital care data that is publicly available in the United States. There are four individual datasets in HCUP: the

Nationwide Inpatient Sample, State Inpatient Databases, State Ambulatory Surgery Databases, and Kids' Inpatient Databases.

The Nationwide Inpatient Sample [35] contains the patient-level clinical and resource use information included in a typical discharge abstract. It is designed to approximate a 20-percent stratified sample of U.S. community hospitals. Strata are defined using the hospital characteristics of ownership, bed size, teaching status, urban/rural location, and region of the United States. All discharges from sampled hospitals are included in the NIS database. There is one dataset for each year from 1988 to 2001, each containing from five to eight million stays. The datasets have the following information for each encounter:

- Primary and secondary diagnoses (coded both in ICD-9-CM and Clinical Classification System codes)
- Primary and secondary procedures (coded both in ICD-9-CM and Clinical Classification System codes)
- Admission and discharge status
- Patient demographics (e.g., gender, age, race, median income for the ZIP Code in which the patient lives)
- Expected payment source
- Total charge
- Length of stay
- Hospital characteristics (e.g., ownership, size, teaching status)

2. Methodology

The four steps in the knowledge discovery/data mining process described above were followed in the study. They are problem definition, data preprocessing, data mining, and post-data mining analysis.

2.1 Problem Definition

A hypothesis-driven approach would ask a question such as: Are there associations between collagen disorders and PEH? Although an interest in collagen disorders still underlies this research, for the data mining process this question must be phrased differently, in a non-hypothesis-driven fashion. This rephrasing leads to the following two research questions:

1. What other diseases or factors are associated with PEH?
2. Can we predict which patients have a PEH given some other medical conditions and demographic status, e.g. collagen diseases, other hernias, age range, *etc.*?

Special steps can still be taken to try to detect the association of PEH with collagen disorders, even though not specifically stated in these questions.

2.2 Data Preprocessing

SAS® for Windows version 8 (The SAS Institute) was used in data preprocessing because it allows intensive programmed data processing. This capability is especially important for dealing with a large dataset such as the NIS.

Data collection. Since the NIS is an existing database, the data collection step was already done. The 1999 dataset was used.

Data cleaning. The dataset appeared to be clean enough for this study due to the considerable effort put into the consolidation and maintenance of the data by AHRQ.

Data selection. Irrelevant data elements, such as date, charge amount, physician ID, discharge status, etc., were removed from the dataset (see Appendix A for a complete list of the data elements in the 1999 NIS dataset). The remaining variables are:

1. Key: a unique record number
2. Age: age in years at admission
3. Dx1 through DX15: principle and up to 14 secondary diagnoses using ICD-9-CM codes
4. Female: gender of patient -- (0) male, or (1) female
5. Hospst: state postal code for the hospital
6. Pay1: expected primary payer, uniform -- (1) Medicare, (2) Medicaid, (3) private including HMO, (4) self-pay, (5) no charge, or (6) other
7. Pr1 through PR15: principal and up to 14 secondary procedures using ICD-9-CM codes
8. Race: (1) white, (2) black, (3) Hispanic, (4) Asian or Pacific Islander, (5) Native American, (6) other
9. Zipinc: Median household income for patient's ZIP code -- (1) \$1-\$24,999, (2) \$25,000-\$34,999, (3) \$35,000-\$44,999, or (4) \$45,000 and above

This study used a case-control design nested in a cross-sectional design. The nature of the NIS dataset (i.e. snapshot-like discharge abstract data from sampled hospitals)

defined it as a cross-sectional study. A case-control design was used in order to study less common conditions, such as PEH.

Choosing cases. After consultation with GI surgeons and coding experts at OHSU, it was determined that the correct ICD-9-CM code for PEH is the procedure code 53.7, which corresponds to the more commonly used CPT code of 39502 (repair, paraesophageal hiatus hernia, transabdominal, with or without fundoplasty, vagotomy). Other procedural codes were felt to be less specific and no diagnostic codes are available that clearly identify PEH patients. Since PEH is almost always treated with surgery, it is unlikely that we have excluded a significant number of cases with this criterion.

We did not wish to include congenital hiatal hernias, which are likely to be an etiologically different disease from adult hiatal hernia. In order to filter out the congenital cases, only patients aged 18 years or older were selected. This age criterion was also used for all controls. This process found 1633 cases of PEH (approximately 0.02% of the 1999 NIS dataset), which will be referred to as 'PEH cases' in the following sections.

Choosing controls. Given the nature of this data, i.e. that all patients had some type of morbidity requiring hospitalization, no "normal" controls could be used. The choice of appropriate controls from this hospital data is difficult. By choosing patients with diseases where there is unlikely to be a relationship to PEH, the major predictors for PEH should not be masked. However, any results are likely to include factors that are major

predictors of the control disease, i.e., which predict non-PEH rather than predicting PEH.

For these reasons, four types of controls were selected.

1. Cases with coronary artery diseases: using ICD-9-CM codes 410 to 414 (Ischemic heart diseases)
2. Cases with malignant neoplasms excluding those of the GI tract: using ICD-9-CM codes 140-149 (malignant neoplasm of lip, oral cavity, and pharynx) and 160 -199 (malignant neoplasm of respiratory and intrathoracic organs, bone, connective tissue, skin, and breast, genitourinary organs, other and unspecified sites)
3. Cases with esophageal reflux excluding those with hiatal hernia: using ICD-9-CM code 530.81 (gastroesophageal reflux) excluding codes for diagnoses or procedures indicating hiatal hernia: 551.3, 552.3 and 553.3 (diaphragmatic hernia) and 53.7 and 53.80 (repair of diaphragmatic hernia abdominal and thoracic approach)
4. Cases with type I hiatal hernia: using ICD-9-CM procedure codes 44.65 (esophagogastroplasty, including Belsey operation) and 44.66 (other procedure for creation of esophagogastric sphincteric competence including Nissen's fundoplication), which captures surgically-treated patients only.

Five random controls of each type were selected for each case, matching on age (within a five-year range) and gender for the first two types of controls. The last two types of controls were matched only on gender, thus allowing age as a potential predictor in the classification.

Definition of PEH by diagnosis codes. Cases with type I hiatal hernias and PEH could not clearly be identified using ICD-9-CM diagnosis codes. The available diagnostic codes are:

- 551.3 (Diaphragmatic hernia with gangrene)
- 552.3 (Diaphragmatic hernia with obstruction)
- 553.3 (Diaphragmatic hernia)

A comparison of the use of these diagnostic codes in cases identified as type I hiatal hernia or PEH by procedure codes is shown in Table 2.

Data transformation. In the ICD-9-CM coding system, there is no single code which can be used to identify cases with collagen diseases. In order to use the diagnosis codes in the dataset to study the relationship between collagen diseases and PEH, a variable was created to replace all potential collagen disease diagnoses with a single code. Deciding

Procedure codes for repair of hiatal hernias	Diagnostic codes for Diaphragmatic Hernia				Total
	551.3	552.3	553.3	No diagnosis code	
53.7 (PEH)	13	339	1206	75	1633
44.65 or 44.66 (type I)	6	233	3535	3427	7201
No procedure code	17	349	102409		
Total	36	921	107150		

Table 2. A comparison of diagnostic and procedures codes for hiatal hernia and hiatal hernia repair in the 1999 NIS dataset showing the number of cases. Cases aged < 18 years were excluded.

what diagnoses to include as collagen diseases was not a trivial process. Some disorders, such as Ehlers-Danlos Syndrome and osteogenesis imperfecta, clearly belong to the category. A number of other disorders are not so straightforward. For example, collagen defects are believed to be one of the underlying causes for subsets of many common disorders [4], such as osteoporosis, osteoarthritis, atherosclerosis, intervertebral disc disease, and others, but one should not conclude that all cases with these diseases are caused by collagen defects. Hence, including these diseases in the collagen diseases category is not possible. Therefore, only diseases with more definitive proof as being caused by collagen gene defects were included into the category [4]. Early onset abdominal aortic aneurysm (i.e. in patients under age 70) is likely to be caused by genetic factors, and was therefore included [14-17]. The cutoff value on age is empirical only. One known collagen disease, Knobloch Syndrome, was not included because no ICD-9-CM code clearly identified it. Epidermolysis bullosa is not included because it is one of several disorders identified by ICD-9-CM code 757.39 (Other specified anomalies of skin). Table 3 shows the collagen diseases included in this study.

After the first data mining step, further data transformation was required. This is described in later sections.

2.3 Data Mining

Two well-known algorithms were used in this two-step data mining process. In the first step, association rule mining was performed using the program Classification Based on Associations (CBA) developed at School of Computing, National University of

Collagen diseases	ICD-9-CM code	ICD-9-CM name
Osteogenesis imperfecta	756.51	Osteogenesis imperfecta
Multiple epiphyseal dysplasia	756.56	Multiple epiphyseal dysplasia
Ehler-Danlos syndrome	756.83	Ehler-Danlos syndrome
Chondrodysplasia Skeletal dysplasia Dyschondrosteosis	756.4	Chondrodystrophy
Alport's syndrome	759.89	Alport Syndrome Hereditary nephritis
Bethlem myopathy related to limb-girdle muscular dystrophy	359.1	Hereditary progressive muscular dystrophy
Marfan syndrome	759.82	Marfan syndrome
Aortic aneurysm, < 70 years old	441	AORTIC ANEURYSM AND DISSECTION

Table 3. Collagen diseases included in the study with corresponding ICD-9-CM codes and names.

Singapore [19, 20]. CBA implements the Apriori algorithm.

In this step, only the 1633 PEH cases were used. The set of diagnostic codes (DX1-DX15) from each PEH cases acted as the set of transaction items. This required transformation of this dataset into a comma delimited file.

A minimum support of 0.6% and minimum confidence of 50% were chosen. Since this is the hypothesis-generating step, the low 0.6% support was deliberately chosen in hopes of finding associated diseases even though they have a low prevalence. Only those resulting rules which contained PEH on the right hand side were retained (for example, Obesity=>PEH). As a result, therefore, the left side of the remaining rules contained

diseases found to have an association with PEH. These associated diseases were analyzed to remove acute conditions associated with surgery (since all PEH cases had surgery); for instance, 'pulmonary collapse' was removed because it is very likely to be related to the surgery rather than specifically to PEH. The remaining conditions were grouped into diseases groups. The grouping was done empirically by judging a number of parameters, including the diseases' similarity, probability that the codes represented the same disease, codes representing diseases having the same cause, etc. The 40 disease groups with highest support generated in this process were used as variables in the next data mining step and are listed in Appendix C.

The second data mining task was classification using the program CART® (Salford Systems). CART® is a commercial decision tree-based classification tool that can automatically sift large, complex databases, searching for and isolating significant patterns and relationships. The discovered knowledge can then be used to generate reliable, easy-to-grasp predictive models [40]. It has been used successfully in business applications such as for profiling customers for targeting direct mailings and for managing credit risk, as well as in academic research analyses [22- 27]. CART® uses a binary recursive partitioning algorithm in the classification process. Unlike the classical regression model, which requires a clear idea of the model to be estimated, including the predictors and the probable interactions, CART® can do data exploration even if there is no prior knowledge of the possible model. This study took advantage of this characteristic of CART® to test on a range of variables and build a decision tree to classify patients into PEH, or non-PEH.

The 1633 PEH cases were combined with each of the four control groups resulting in four separate datasets for analysis. The 40 associated diagnostic groups discovered in the first step of this project were used in a data transformation step. Forty binary variables were created to indicate presence or absence of these 40 associated diagnostic groups respectively. To our disappointment, collagen diseases were not in the 40 diagnostic groups found in the first step. Since we still wished to try and detect the association of PEH with collagen disorders, one more binary variable was created to represent the presence or absence of collagen diseases (see Table 3). Some original variables of the dataset, such as race and zipinc (median household income for patient's ZIP code) were retained. Appendix D lists all variables assigned to CART®.

The four datasets were then processed using CART®, one at a time. The program was directed to randomly select two thirds of the dataset to build the classification rules (the model fitting process) and use the remaining one third for validation (the model testing process). The cost matrix value for misclassifying PEH into non-PEH was set at 1.5 (50% higher than misclassifying non-PEH into PEH, which is set at 1.0). The program defaults were used for the rest of the parameters.

2.4 Post-Data Mining Analysis

The output from CART® was a set of classification trees for each of the datasets. Although the program designated the tree with the minimum cost as the "best" tree, detailed evaluation of all trees was performed manually. This process included

examining each tree's prediction success rate, cost, variable importance (a numeric value calculated by CART® to measure a variable's importance in the classification process; the higher the value the more important the variable is) and its complexity as represented by number of terminal nodes. A plot similar to a Receiver Operating Characteristic (ROC) curve of sensitivity vs. (1-specificity), including only trees with ≤ 20 terminal nodes, was prepared for each of the four tree sets (Figure 5). Trees with more than 20 nodes were not plotted because they were believed empirically to be too complex to be the best tree. A best tree was selected from each set using these plots and by keeping the following criteria in mind:

1. All major predictors (with importance $> 10\%$) should be retained.
2. Cost for a tree should be within one standard error of the minimum cost in the tree set.
3. Prediction success rate should be no less than 90% of the highest rate in the set.
4. The tree should have the least possible number of terminal nodes.
5. The most optimal tree is the one closest to the upper left corner of the ROC-like plot.

The best trees from the four datasets contained a number of "predictors", i.e. the disease groups which appeared in the classification nodes. The common major predictors from the four datasets were selected as the important predictors for PEH. Further search of the medical literature for supporting evidence was then done to evaluate the model's and the predicting variables' clinical relevancy.

3. Results

3.1 Descriptive Statistics

The 1999 NIS dataset contains 1633 cases identified as having PEH using the ICD-9-CM procedure code 53.7. Of these, 1043 (63.9%) are female; 91.1% (of the 1056 where race was indicated) are white. Mean age is 60.5 years with a standard deviation of 17.2 years. Over 94% had zip codes where average household income was higher than \$25,000/year. These descriptive statistics are shown graphically in Appendix B.

The comparison of PEH cases with a diagnosis of collagen diseases with the prevalence of that diagnosis in the whole NIS 1999 dataset is shown in Table 7. There were very few patients with this type of disorder. This was true for both the PEH cases and for the whole inpatient dataset. Only 0.3% of the PEH cases had an identifiable collagen disease, which was comparable to the percentage in the whole 1999 dataset.

Diseases	1999NIS dataset, n=7,198,929		PEH cases, n=1633	
	Number of cases	Percentage of dataset	Number of cases	Percentage of dataset
Osteogenesis imperfecta	678	0.01	0	0.00
Multiple epiphyseal dysplasia	17	0.00	0	0.00
Ehler-Danlos Syndrome	356	0.00	0	0.00
Chondrodysplasia	707	0.01	0	0.00
Alport's Syndrome	2676	0.04	0	0.00
Bethlem myopathy	2080	0.03	1	0.06
Marfan Syndrome	769	0.01	0	0.00
Aortic aneurysms, age < 70	10801	0.15	3	0.18
Total	51408	0.71	10	0.61

Table 7. Breakdown of collagen diseases in the whole 1999 NIS dataset as compared to the PEH cases.

3.2 Association Rule Mining

The result of association rule mining is a list of disease groups, members of which have met a minimum support of 0.6% as the left side of association rules where PEH is on the right hand side. The disease with the highest cover (38.9%) was esophageal reflux disease (ICD-9-CM 530.81), which is not surprising since the connection between this and PEH has been well documented [1, 30-32]. Interesting disease groups include hypertension, obesity, peritoneal adhesions, and gall bladder/bile duct diseases, all of which had group supports of 8%-28%. Appendix C shows the complete list of the diseases and the results of grouping after removing acute conditions related to surgery. A total of 40 groups of diagnoses were used. Collagen diseases as a group did not appear in the list.

Transformation of the datasets for classification was performed based on these 40 groups and collagen diseases as discussed in the methodology section. Each group was represented as a dichotomous variable indicating presence or absence as described previously. The complete list of variables used can be found in Appendix D.

3.3 Classification

Four sets of classification trees were built from the four datasets. Tree sets for CAD, cancer, GI and type I hernia control datasets have numbers of terminal nodes ranging from 2-251, 2-313, 2-547 and 3-576 respectively. ROC-like plots of sensitivity versus (1- specificity) for the four sets of trees are presented in Figure 5. In these graphs, a tree closer to the upper left corner of the graph is a "better" tree. The data points representing

trees close to the upper left corner were connected with lines to show the "frontier" of the set. Trees on the frontier line are candidates for the best tree. Using the graphs and the criteria described in the methodology section, a best tree was selected for each set.

Following are the detailed rationales for the selection of the best trees.

In the coronary artery diseases control set, trees with 12, 15, 17, 9 and 2 terminal nodes were candidates for best tree as illustrated by the frontier line in Figure 5-A. The tree with nine nodes was selected because it is less complex than the trees with 12, 15 and 17 nodes and has similar prediction success rate. The tree with two nodes gives 2% higher success rate for PEH, but has a 4% lower rate for non-PEH, and therefore was not chosen. The graphic presentation of the "best" classification tree for this dataset is in Figure 6-A.

Similarly, in the malignant neoplasm control set, the six-node tree was selected over the four-node tree (Figure 5-B) because the four-node tree gives up 4% success rate in non-PEH but only gains 2% in PEH. The decision tree for this control set is in Figure 6-B.

In the esophageal reflux disease control set, the nine-node tree was selected over the six-node tree. By the same token, the nine-node tree was chosen the best tree over the ones with three, five, six and seven-node (as depicted in Figure 5-D) in the type I hiatal hernia control set. The graphic presentation of these two trees is in Figure 6-C and D.

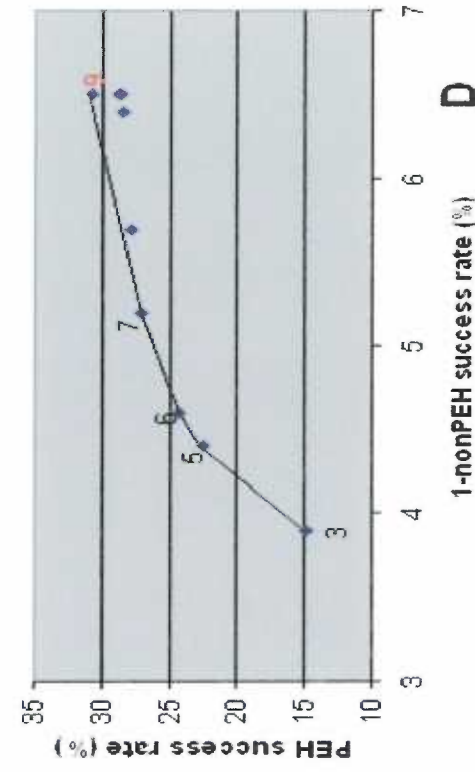
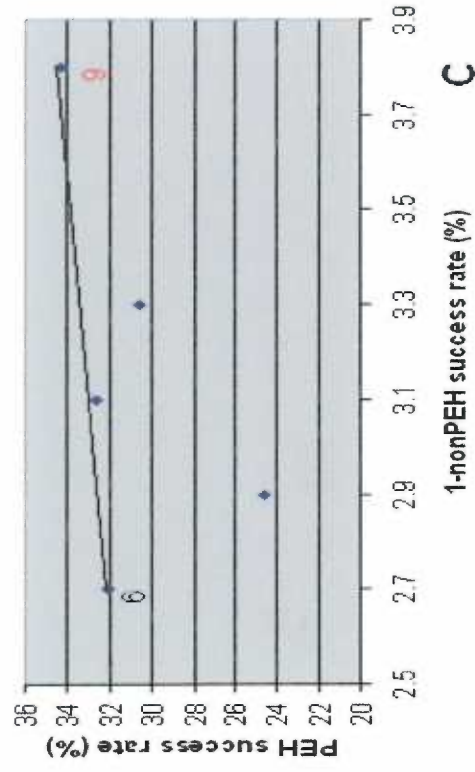
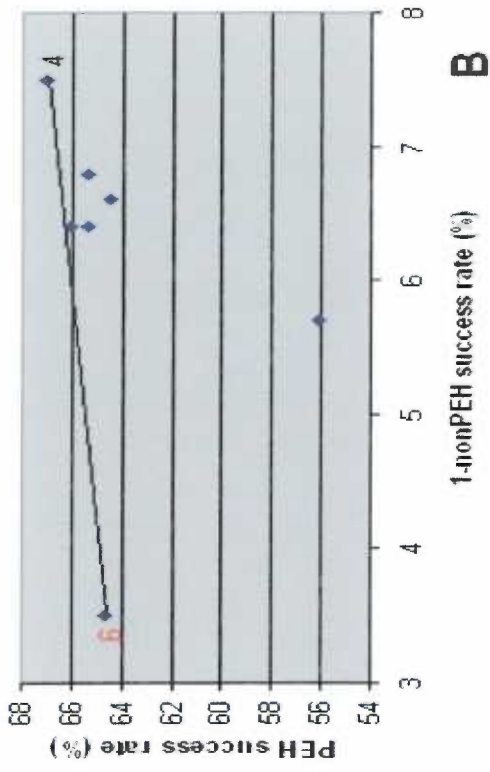
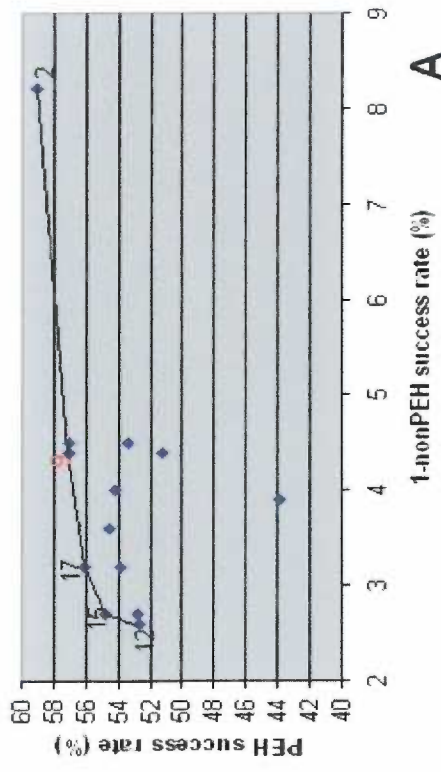


Figure 5. Rate of successfully predicting paraesophageal hernia (PEH) patients versus 1 – (rate of successfully predicting non-PEH patients) for four tree sets built in the CART® classification process. Only trees with 20 or fewer terminal nodes were plotted. The points closest to the upper left corner were connected with lines to show the frontier. Numbers next to the points indicate the number of terminal nodes, with the best tree's number in red. See text for the rationale for selecting the best tree. A: coronary artery diseases control set, B: malignant neoplasm control set, C: reflux disease control set, and D: type I hiatal hernia control set.

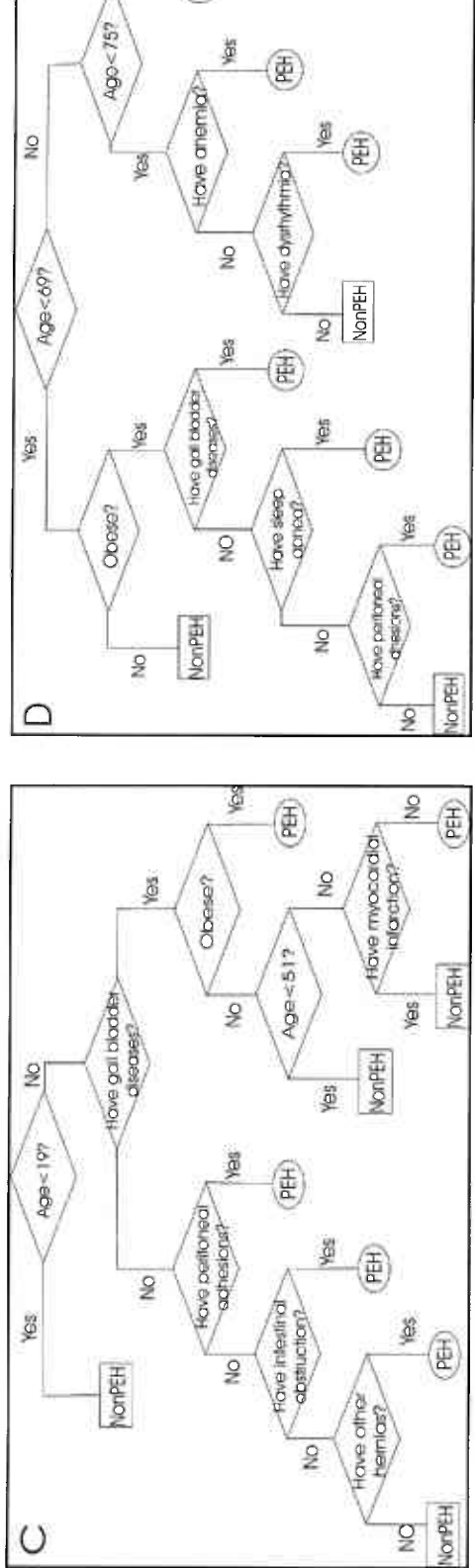
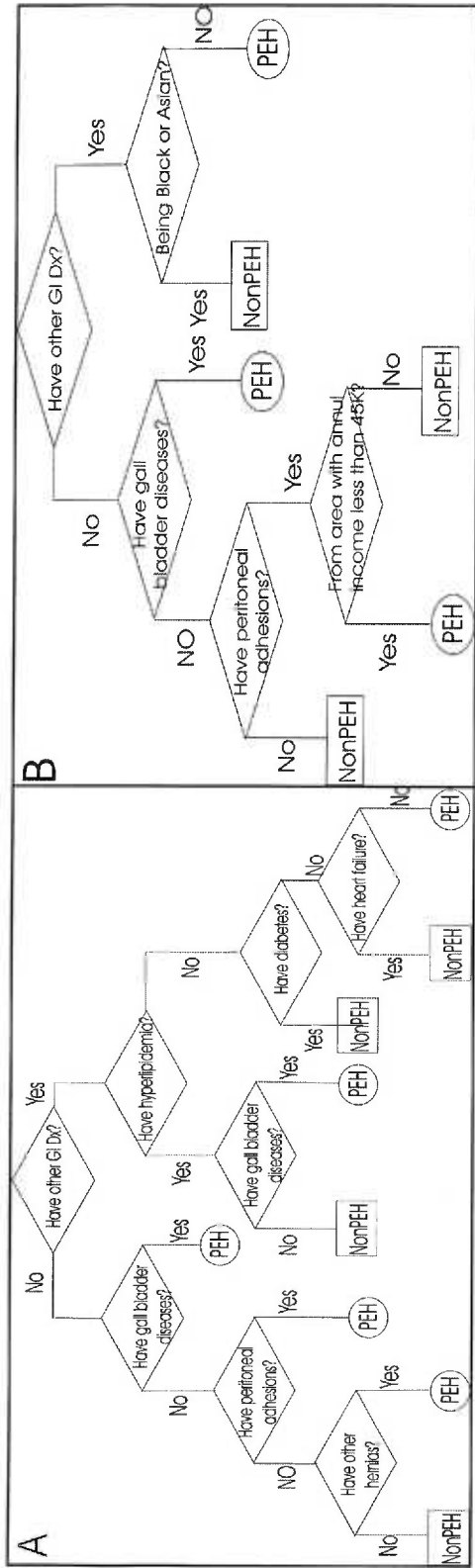


Figure 6. Decision trees for the four datasets. The best trees selected from the four tree sets built in the classification process are presented in an easy-to-understand binary tree structure. The prediction success rates for each tree are summarized in Table 8. A: coronary artery disease control, B: malignant neoplasm control, C: esophageal reflux disease control and D: type I hiatal hernia control.

Table 8 summarizes the parameters of the best trees. Appendix E shows the detailed trees from the CART® output. Except for the tree for the cancer control set, which has six terminal nodes, all other trees have nine terminal nodes. The percentages of correct prediction are similar in the fitting and the testing processes of all four datasets, which gives validity to the models built by classification. The correct prediction rates in the later two sets with GI controls are lower than those with non-GI controls. Considering that PEH cases represent approximately 16.7% of each dataset, all results show improvement over chance.

Success rate		Controls			
		Coronary Artery Diseases	Cancer	Esophageal Reflux	Type I hiatal hernia
Fitting	Non PEH	95.7%	94.0%	97.5%	94.0%
	PEH	56.8%	67.1%	31.4%	34.0%
Testing	Non PEH	95.6%	93.5%	97.3%	93.5%
	PEH	57.1%	64.7%	32.1%	30.7%
Complexity		9 terminal nodes	6 terminal nodes	9 terminal nodes	9 terminal nodes

Table 8. Complexity and prediction success rates in the fitting and testing processes of the four best classification trees, by control group. PEH: paraesophageal hernia cases, nonPEH: control cases.

As shown in Table 9, the common important predictors in the four case-control datasets are gall bladder/bile duct diseases and peritoneal adhesions. Age, obesity, other GI diagnoses, other hernia types, intestinal obstruction and dysrhythmia appeared in two of the four classification trees, while hypertension, race, income level, myocardial infarction, sleep apnea and anemia appeared in only one of the four classification trees.

Important variables ¹	Controls			
	Coronary Artery Diseases	Cancer	Esophageal reflux	Type I hiatal hernia
Other GI diagnoses	X	X		
Gall bladder/bile duct diseases	X	X	X	X
Peritoneal adhesions	X	X	X	X
Other hernia	X		X	
Intestinal obstruction	X		X	
Hypertension	X			
Dysrhythmia	X			X
Race		X		
Income level		X		
Age			X	X
Obesity			X	X
Myocardial infarction			X	
Sleep apnea				X
Anemia				X

¹ The variable names are the same as the disease groups identified in the association rule mining process and their definitions and ICD-9-CM codes used to identify them can be found in Appendix D

Table 9. Variables used as predictors in the classification process in the four datasets. An X indicates the presence of the predictor in the classification tree of that control dataset. Gall bladder/bile duct diseases and peritoneal adhesions are found in all of the four trees. The rest only appeared in one or two of the datasets. .

4. Discussion

4.1 Major Findings

Diseases that result from defects in collagen are difficult to recognize using ICD-9-CM diagnostic codes. Some collagen diseases like Knobloch syndrome do not have ICD-9 codes and some are grouped into ‘other anomalies’ with other unrelated diseases, e.g. epidermolysis bullosa. In addition, the collagen diseases defined in this study have a very low prevalence. In the 1999 NIS dataset, 0.25% could be identified as having collagen defect-related disease, a prevalence almost identical to that in the subpopulation of patients found to have PEH. It is not surprising, based on equal prevalence alone, that no association between collagen defect-related diseases and PEH could be demonstrated using these methodologies.

Two explanations for this lack of demonstrable association can be proposed. One explanation is that there truly is no association between PEH and collagen diseases, either because collagen diseases are not the causes of PEH or because the association is not manifest at the clinical or phenotype level, on which this study tested. It is worth noting that a recent study by Pans and associates attempting to demonstrate the connection between collagen defect and groin hernia at the clinical level also failed [33]. The other explanation is that the association was not detected due to the limitations of the NIS dataset and/or the study design. Limitations will be discussed later in this section.

From a positive viewpoint, however, this approach uncovered a number of possible important predictors which may deserve future hypothesis-oriented studies. It is easy to

understand the associations in some cases, for example, the association with other GI diagnoses. The relationship to PEH of GI diseases like esophageal reflux has been the interest of studies recently [1]. It is also not surprising that obesity is associated with PEH, since there are reports in the literature that obesity is associated with gastroesophageal reflux disease [28, 29], which again is associated with hiatal hernias. The finding of these known associated factors assures that this algorithm is working correctly. Other important but less easily explained associations found in this study are those with peritoneal adhesions, intestinal obstruction, gall bladder/bile duct diseases, and dysrhythmias. Some of these associations might be worth further study.

4.2 The Advantage of Using Data Mining Technology

In this study a large dataset was analyzed using exploratory data analysis technology. The two-step data mining method is a suitable preliminary study methodology for exploring the data and finding possible predictors where no prior experience or scientific theories are available to aid the analysis.

Some of the disease associations discovered in this process may be confounders rather than true associations. For instance, hypertension may be a confounder of the elderly age of the PEH cases. We saw that in the second step of the study, hypertension presented as an important predictor only in the dataset using cases with coronary artery disease as a control, where there is a well-established association with the control disease. Therefore, it is reasonable to assume that hypertension is a confounding variable. Other variables that appeared in only one of the control groups (e.g. race and income) are likely to be

confounders based on the same reasoning. This study design, using the results of association rule mining to inform the classification process appears to be a good method for both exploring for possible associated diseases and for eliminating possible confounders.

4.3 Results from the Classification Process

In the second step, a classification tree was built using type I hiatal hernia cases as a control for the PEH cases. It is only mildly successful with a 30.7% correct rate, but that is twice the number of correctly predicted cases as by chance alone. This decision tree may have potential use in clinical practice after further verification in real-life settings. Compared with type I hiatal hernia patients, patients with PEH tend to be older, obese, and have gall bladder and bile duct diseases. By keeping these major predictors in mind, physicians might be able to prescribe diagnostic tests for PEH to patients who meet the criteria and are more likely to have PEH.

Gall bladder/bile duct diseases and peritoneal adhesions were predictors in all four datasets in the classification step. A search of the medical literature has not yielded any support for these associations. While these associations might be due to coding bias, as discussed later in this section, it is likely that gall bladder/bile duct diseases and peritoneal adhesions were truly associated with PEH. The appearance of these predictors in the type I hiatal hernia control set supports this allocation, as these controls should be subject to the same coding bias, being surgical patients like the PEH cases. Further

research on these two factors, using hypothesis-oriented methods such as traditional case control studies, is needed to elucidate the nature of these associations.

Other possible associated factors that are found in at least two of the four datasets include age, obesity, other hernia types, intestinal obstruction and dysrhythmia. Age and obesity are predictors in the two GI disease control groups which suggest that, despite the similarity between these diseases and PEH, this severe form of hiatal hernia tends to occur in older and more obese patients. One might even postulate that esophageal reflux disease and type I hiatal hernia progress to PEH. The connection with other hernia types tends to suggest that the formation of a hernia might be caused by the weakness of the supporting connective tissue [1, 8-10], again pointing to collagen defects as an underlying cause. As for intestinal obstruction and dysrhythmia, like peritoneal adhesions and gall bladder diseases, further studies need to be done to elucidate the mechanism of the association. Finally, other factors that only appeared once in the classification trees, including MI, sleep apnea, anemia, race and income level, most likely are present because of their independent association with a control set.

4.4 Limitations of the Study

The first and most limiting factor in this study is the data itself. As with all studies on retrospective datasets, the results can only be as reliable as the data. The data used in this study were collected from the discharge abstracts of the sampled hospitals mainly for use in hospital facility utilization studies. Using data collected for other purpose to answer questions of our interest may lead to some problems. For example, in the NIS dataset,

the diagnosis codes may be incomplete and even erroneous. It is likely that only diagnoses related to the current hospital stay are reported. This leads to the probability that many of the diagnoses unrelated to the current hospitalization (such as, possibly, collagen diseases) were under-represented and that diagnoses related to the procedure done during that hospital stay were more emphasized and fully coded. Since this data comes from the discharge record, there is no means for tracking of patients longitudinally, which might partially alleviate this problem. In addition, codes might be chosen to maximize the billed amount rather than to reflect accurately what happened during the hospital stay. It also has been shown that there are significant variations in coding errors between different types of hospitals and geographic areas in the NIS dataset [41].

Secondly, the classification principles of ICD-9-CM codes are not aligned with the needs of this study. For example, we were unable to discern different types of hiatal hernia using the diagnosis codes in the dataset, which made it difficult to have clear-cut cases and control samples. We have substituted procedure codes, but this means that patients with the diagnosis who have not had surgery on that hospitalization will not have been detected.

The third limitation of this study involves the detection of collagen diseases. Many diseases, such as osteoporosis, may have variants that have a true association to collagen defects. ICD-9-CM codes, however, do not differentiate these subtle variants. The approach used in this study was to include only diagnoses with a well-established relationship to collagen diseases, making this category very specific but not sensitive. A

more sensitive means of identifying these diseases might have resulted in the finding of an association with PEH.

Finally, choosing controls in this hospital population was difficult. Using controls with diseases totally unrelated to GI diseases made the existence of GI diseases a very important predictor, which is not an especially useful piece of information for purposes of this study. Furthermore, some of the nodes of the decision trees contain factors which may be characteristics of the controls rather than of the PEH cases. For example, myocardial infarction was found to be the most significant predictor in the CAD control set, which almost certainly is a factor identifying CAD patients. In other words, this predictor selected for controls, not for PEH and can probably be eliminated from further consideration.

On the other hand, if diseases close to PEH are chosen as controls, such as was done here with esophageal reflux and type I hiatal hernia, other GI diseases and diseases that are closely related to the controls are no longer found to be predictors, and the important predictors are more likely to be the features that can differentiate PEH. However, because of the close relationship and similar characteristics between the PEH cases and the controls, the prediction success rates were relatively low in these two controls.

4.5 Future Work

One major problem encountered in this study is the fuzziness in the definition of collagen diseases. Even though some inheritable disorders such as osteogenesis imperfecta and

Elhers-Danlos Syndrome can definitely be defined as collagen diseases -- as they are caused by mutations of the collagen genes or genes for post-translational enzymes of collagen synthesis -- collagen mutations have also been found in certain common diseases, for example osteoporosis, osteoarthritis and aortic aneurysms, and it is now evident that subsets of patients with these diseases have defects in types I, II or III collagen, as a predisposing factor[44]. In addition, inherited collagen diseases are rare in the population and some are under-diagnosed [7], so that their existence might be overlooked and underestimated in most clinical records. Therefore, retrospective data analysis may not be the best analytical choice for this type of disease. Instead, traditional case-control studies are recommended. Cases could be recruited from patients undergoing PEH repair procedures with controls from the general population without the disease, matching on possible confounders, such as gender and age. Surveys, clinical examinations and genetic tests regarding collagen diseases could then be conducted to collect data on the true prevalence of collagen diseases in the case and control groups. This design overcomes the limitations mentioned in previous section.

Although the NIS dataset has its limitations, the method used in this study can be applied to the NIS or other datasets to answer other types of questions. Because clear-cut diagnoses and procedures are less likely to be missed and miscoded, this method could be used to answer questions about possible risk factors, for example, for Cesarean section versus natural birth. This dataset and methodology might also be well suited for analysis of demographical and regional indications and economical consequences for certain diseases. One might ask, for instance, what factors influence asthma patients' length of

hospital stay and total charges in addition to the severity of the patients' medical condition.

5. Conclusion

In this study, data mining algorithms were used to explore the possible associations between PEH and other diseases or conditions. No association could be demonstrated between PEH and collagen diseases, the hypothesized association of interest. We recognize the limitations of both the dataset (1999 NIS) and our definition of collagen diseases, as well as those of the methodology used, which might contribute to this lack of detected association. Instead, other diseases were found to have an association with PEH, including gall bladder/bile duct diseases and peritoneal adhesions. Further hypothesis-oriented research might be able to elucidate the possible mechanism of these associations. The classification tree built using type I hiatal hernia cases as a control showed that PEH patients tend to be older, more obese and have gall bladder or bile duct diseases as compared to patients with type I sliding hernias. With further testing and validation in real-life clinical settings, this type of classification tree and the major predictors discovered in the process may be useful as a starting-point tool for the early diagnosis of this potentially severe type of hiatal hernia.

6. References

1. Mittal, R.K., Hiatal Hernia: Myth or Reality? *Am J Med.* 1997 Nov 24;103(5A):33S-39S.
2. Mittal, R.K., The spectrum of diaphragmatic hernia. *Hosp Pract (Off Ed).* 1998 Nov 15;33(11):65-6, 69-70, 73-5 passim.
3. Part Eleven: Disorders Of The Gastrointestinal System, Harrison's principles of internal medicine. <http://harrisons.accessmedicine.com/> On May 28, 2003
4. Myllyharju, J., Kivirikko, K. Collagens and collagen-related diseases. *Ann Med* 2001; 33:7-21
5. Geis, I. Three-dimensional structures of proteins. In: Stiefel, J. Editor, *Biochemistry.* John Wiley & Sons, Inc. 1990.
6. Horvath, M. Association of hiatal hernia with mitral valve prolapse. *Eur J Pediatr* 1997; 156:35-36
7. Giroto JA, Malaisrie SC, Bulkely G, Manson PN. Recurrent ventral herniation in Ehlers-Danlos syndrome. *Plast Reconstr Surg* 2000 Dec;106(7):1520-6
8. Pans, A. New prospects in the etiology of groin hernias. *Chirurgie* 1999; 124(3):288-97
9. Zheng H, Si Z, Kasperk R, Bhardwaj RS, Schumpelick V, Klinge U, Klosterhalfen B. Recurrent inguinal hernia: disease of the collagen matrix? *World Journal of Surgery* 2002; 26(4):401-8
10. Si Z, Bhardwaj R, Rosch R, Mertens PR, Klosterhalfen B, Klinge U, Rhanjit B, Rene PM. Impaired balance of type I and type III procollagen mRNA in cultured fibroblasts of patients with incisional hernia. *Surgery* 2002 Mar;131(3):324-31
11. Rosch R, Klinge U, Si Z, Junge K, Klosterhalfen B, Schumpelick V. A role for the collagen I/III and MMP-1/-13 genes in primary inguinal hernia? *BMC Med Genet* 2002;3(1):2
12. Klinge U, Si ZY, Zheng H, Schumpelick V, Bhardwaj RS, Klosterhalfen B. Collagen I/III and matrix metalloproteinases (MMP) 1 and 13 in the fascia of patients with incisional hernias. *J Invest Surg* 2001 Jan-Feb;14(1):47-54
13. Rodrigues Junior AJ, Rodrigues CJ, Cunha AC, Jin Y. Quantitative analysis of collagen and elastic fibers in the transversalis fascia in direct and indirect inguinal hernia. *Rev Hosp Clin Fac Med Sao Paulo* 2002 Nov-Dec;57(6):265-70
14. Plenz GA, Deng MC, Robenek H, Volker W. Vascular collagens: spotlight on the role of type VIII collagen in atherogenesis. *Atherosclerosis* 2003 Jan;166(1):1-11
15. van Vlijmen-van Keulen CJ, Pals G, Rauwerda JA. Familial abdominal aortic aneurysm: a systematic review of a genetic background. *Eur J Vasc Endovasc Surg* 2002 Aug;24(2):105-16
16. Carmo M, Colombo L, Bruno A, Corsi FR, Roncoroni L, Cuttin MS, Radice F, Mussini E, Settembrini PG. Alteration of elastin, collagen and their cross-links in abdominal aortic aneurysms. *Eur J Vasc Endovasc Surg* 2002 Jun;23(6):543-9

17. Treska V, Topolcan O. Plasma and tissue levels of collagen types I and III markers in patients with abdominal aortic aneurysms. *Int Angiol* 2000 Mar;19(1):64-8
18. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P.Smyth, and R. Uthurusamy, Editors, *Advances in Knowledge Discovery and Data Mining*, pages 307-328, Menlo Park, California, 1996. AAAI Press/MIT Press.
19. Bing Liu, Wynne Hsu, Yiming Ma, Shu Chen, "Discovering Interesting Knowledge using DM-II" to appear in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99)*, Industrial Track, August 15-18, 1999, San Diego, CA, USA.
20. Bing Liu, Wynne Hsu, Yiming Ma, "Pruning and Summarizing the Discovered Associations" to appear in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99)*, August 15-18, 1999, San Diego, CA, USA.
21. Bohanec, M. & Bratko, I. (1994). Trading accuracy for simplicity in decision trees. *Machine Learning*, 15, 223-250
22. Altman, D. G. & Goodman, S. N. (1994). Transfer of technology from statistical journals to the biomedical literature: Past trends and future predictions. *JAMA: Journal of the American Medical Association*, 272, 129-132
23. Alarcon, G. S., Willkens, R. F., Ward, J. R., Clegg, D. O., Morgan, J. G., Ma, K. N., Singer, J. Z., Steen, V. D., Paulus, H. E., Luggen, M. E., Polisson, R. P., Ziminski, C. M., Yarboro, C. & Williams, H. J. (1996). Early undifferentiated connective-tissue disease. *Arthritis and Rheumatism*, 39, 403-414
24. Albert, R., Muller, J. G., Kristen, P., Schindewolf, T., Kneitz, S. & Harms, H. (1996). New method of nuclear grading of tissue sections by means of digital image-analysis with prognostic significance for node-negative breast-cancer patients. *Cytometry*, 24, 140-150
25. Badgett, R. G., Tanaka, D. J., Hunt, D. K., Jelley, M. J., Feinberg, L. E., Steiner, J. F., Petty, T. L. (1994). The clinical evaluation for diagnosing obstructive airways disease in high-risk patients, *Chest*, 106,1427-1431
26. Barrett, R. J., Harlan, L. C., Wesley, M. N., Hill, H. A., Chen, V. W., Clayton, L. A., Kotz, H. L., Eley, W., Robboy, S. J. & Edwards, B. K. (1995). Endometrial cancer: Stage at diagnosis and associated factors in black and white patients. *American Journal of Obstetrics and Gynecology*, 173, 414-423
27. Barriga, K. J., Hamman, R. F., Hoag, S., Marshall, J. A. & Shetterly, S. M. (1996). Population screening for glucose intolerant subjects using decision tree analyses. *Diabetes Research and Clinical Practice*, 34, S17-S29
28. Gomez Escudero O, Herrera Hernandez MF, Valdovinos Diaz MA. Obesity and gastroesophageal reflux disease *Rev Invest Clin* 2002 Jul-Aug;54(4):320-7
29. Barak N, Ehrenpreis ED, Harrison JR, Sitrin MD. Gastro-oesophageal reflux disease in obesity: pathophysiological and therapeutic considerations. *Obes Rev* 2002 Feb;3(1):9-15

30. Tougas G, Banemai M. Gastroesophageal reflux disease pathophysiology. *Chest Surg Clin N Am* 2001 Aug;11(3):485-94
31. Orlando RC. Overview of the mechanisms of gastroesophageal reflux. *Am J Med* 2001 Dec 3;111 Suppl 8A:174S-177S
32. Kahrilas PJ. Supraesophageal complications of reflux disease and hiatal hernia. *Am J Med* 2001 Dec 3;111 Suppl 8A:51S-55S
33. Pans A, Albert A. Joint mobility in adult patients with groin hernias. *Hernia* 2003 Mar;7(1):21-4
34. <http://www.ahcpr.gov/data/hcup/> Accessed on July 28, 2003
35. <http://www.ahcpr.gov/data/hcup/hcupnis.htm> Accessed on July 28, 2003
36. C. Zhang, S. Zhang. Chapter 1, Introduction In: *Association rule mining: models and algorithms*, Berlin; New York: Springer, c2002.
37. U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, From data mining to knowledge discovery: an overview. In: *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT Press, 1996: 1-36.
38. W. Frawley, G. Piatetsky-Shapiro, and C. Matheus, Knowledge discovery in databases: an overview. *AI Magazine*, 13(3) (1992): 57-70.
39. C. Zhang, S. Zhang. Chapter 2. Association rules. In *Association rule mining: models and algorithms*, Berlin; New York: Springer, c2002.
40. <http://www.salford-systems.com/> Accessed on May 28, 2003
41. Benthelsen, C.L. Evaluation of coding data quality of the HCUP National Inpatient Sample. *Topics in Health Information Management* 2000; 21(2): 10-23
42. U. M. Fayyad and E. Simoudis, Data mining and knowledge discovery. In: *Proceedings of 1st International Conf. Prac. App. KDD&Data Mining*, 1997: 3-16.
43. Liem, M.S.L., van der Graaf, Y., Beemer, F.A., and van Vroonhoven, T.J. Increased risk for inguinal hernia in patients with Ehlers-Danlos syndrome. *Surgery* 122: 114, 1997.
44. Kivirikko KI. Collagens and their abnormalities in a wide spectrum of diseases. *Ann Med*. 1993 Apr;25(2):113-26.
45. DeMeester T. R. Esophagus and Diaphragmatic Hernia. In: Schwartz, S.I. editor. *Principles of Surgery* (7th Ed.). New York: McGRAW-HILL, 1999.

7. Appendices

Appendix A. NIS 1999 dataset detail description

Appendix B. Statistics of the paraesophageal hernia cases selected by procedure code '53.7'

Appendix C. Cover values and ICD-9-CM codes for the possible associated diseases and the 40 diseases groups identified in the association rule mining process

Appendix D. Variables used in CART® classification.

Appendix A. NIS 1999 dataset detail description

1	KEY	1	14	Num	HCUP record identifier
2	AGE	15	17	Num	Age in years at admission
3	AGEDAY	18	20	Num	Age in days (when age < 1 year)
4	AMONTH	21	22	Num	Admission month
5	ASOURCE	23	24	Num	Admission source (uniform)
6	ASOURCE_X	25	27	Char	Admission source (as received from source)
7	ATYPE	28	29	Num	Admission type
8	AWEEKEND	30	31	Num	Admission day is a weekend
9	DIED	32	33	Num	Died during hospitalization
10	DISCWT	34	43	Num	Weight to discharges in AHA Universe
11	DISPUB92	44	45	Num	Disposition of patient (UB-92 standard coding)
12	DISPUNIFORM	46	47	Num	Disposition of patient (uniform)
13	DQTR	48	49	Num	Discharge quarter
14	DRG	50	52	Num	DRG in effect on discharge date
15	DRG10	53	55	Num	DRG, version 10
16	DRG18	56	58	Num	DRG, version 18
17	DRGVER	59	60	Num	DRG grouper version used on discharge date
18	DSHOSPID	61	73	Char	Data source hospital identifier
19	DX1	74	78	Char	Principal diagnosis
20	DX2	79	83	Char	Diagnosis 2
21	DX3	84	88	Char	Diagnosis 3
22	DX4	89	93	Char	Diagnosis 4
23	DX5	94	98	Char	Diagnosis 5
24	DX6	99	103	Char	Diagnosis 6
25	DX7	104	108	Char	Diagnosis 7
26	DX8	109	113	Char	Diagnosis 8
27	DX9	114	118	Char	Diagnosis 9
28	DX10	119	123	Char	Diagnosis 10
29	DX11	124	128	Char	Diagnosis 11
30	DX12	129	133	Char	Diagnosis 12
31	DX13	134	138	Char	Diagnosis 13
32	DX14	139	143	Char	Diagnosis 14
33	DX15	144	148	Char	Diagnosis 15
34	DXCCS1	149	152	Num	CCS: principal diagnosis
35	DXCCS2	153	156	Num	CCS: diagnosis 2
36	DXCCS3	157	160	Num	CCS: diagnosis 3
37	DXCCS4	161	164	Num	CCS: diagnosis 4
38	DXCCS5	165	168	Num	CCS: diagnosis 5
39	DXCCS6	169	172	Num	CCS: diagnosis 6
40	DXCCS7	173	176	Num	CCS: diagnosis 7
41	DXCCS8	177	180	Num	CCS: diagnosis 8
42	DXCCS9	181	184	Num	CCS: diagnosis 9
43	DXCCS10	185	188	Num	CCS: diagnosis 10
44	DXCCS11	189	192	Num	CCS: diagnosis 11

45	DXCCS12	193	196	Num	CCS: diagnosis 12
46	DXCCS13	197	200	Num	CCS: diagnosis 13
47	DXCCS14	201	204	Num	CCS: diagnosis 14
48	DXCCS15	205	208	Num	CCS: diagnosis 15
49	FEMALE	209	210	Num	Indicator of sex
50	HOSPID	211	215	Num	HCUP hospital identification number
51	HOSPST	216	217	Char	Hospital state postal code
52	HOSPSTCO	218	222	Num	Hospital modified FIPS state/county code
53	LOS	223	227	Num	Length of stay (cleaned)
54	LOS_X	228	233	Num	Length of stay (uncleaned)
55	MDC	234	235	Num	MDC in effect on discharge date
56	MDC10	236	237	Num	MDC, version 10
57	MDC18	238	239	Num	MDC, version 18
58	MDID_S	240	255	Char	Attending physician number (synthetic)
59	NDX	256	257	Num	Number of diagnoses on this record
60	NEOMAT	258	259	Num	Neonatal and/or maternal DX and/or PR
61	NPR	260	261	Num	Number of procedures on this record
62	PAY1	262	263	Num	Primary expected payer (uniform)
63	PAY1_X	264	273	Char	Primary expected payer (as received from source)
64	PAY2	274	275	Num	Secondary expected payer (uniform)
65	PAY2_X	276	285	Char	Secondary expected payer (as received from source)
66	PR1	286	289	Char	Principal procedure
67	PR2	290	293	Char	Procedure 2
68	PR3	294	297	Char	Procedure 3
69	PR4	298	301	Char	Procedure 4
70	PR5	302	305	Char	Procedure 5
71	PR6	306	309	Char	Procedure 6
72	PR7	310	313	Char	Procedure 7
73	PR8	314	317	Char	Procedure 8
74	PR9	318	321	Char	Procedure 9
75	PR10	322	325	Char	Procedure 10
76	PR11	326	329	Char	Procedure 11
77	PR12	330	333	Char	Procedure 12
78	PR13	334	337	Char	Procedure 13
79	PR14	338	341	Char	Procedure 14
80	PR15	342	345	Char	Procedure 15
81	PRCCS1	346	348	Num	CCS: principal procedure
82	PRCCS2	349	351	Num	CCS: procedure 2
83	PRCCS3	352	354	Num	CCS: procedure 3
84	PRCCS4	355	357	Num	CCS: procedure 4
85	PRCCS5	358	360	Num	CCS: procedure 5
86	PRCCS6	361	363	Num	CCS: procedure 6
87	PRCCS7	364	366	Num	CCS: procedure 7
88	PRCCS8	367	369	Num	CCS: procedure 8

89 PRCCS9	370	372	Num	CCS: procedure 9
90 PRCCS10	373	375	Num	CCS: procedure 10
91 PRCCS11	376	378	Num	CCS: procedure 11
92 PRCCS12	379	381	Num	CCS: procedure 12
93 PRCCS13	382	384	Num	CCS: procedure 13
94 PRCCS14	385	387	Num	CCS: procedure 14
95 PRCCS15	388	390	Num	CCS: procedure 15
96 PRDAY1	391	393	Num	Number of days from admission to PR1
97 PRDAY2	394	396	Num	Number of days from admission to PR2
98 PRDAY3	397	399	Num	Number of days from admission to PR3
99 PRDAY4	400	402	Num	Number of days from admission to PR4
100PRDAY5	403	405	Num	Number of days from admission to PR5
101PRDAY6	406	408	Num	Number of days from admission to PR6
102PRDAY7	409	411	Num	Number of days from admission to PR7
103PRDAY8	412	414	Num	Number of days from admission to PR8
104PRDAY9	415	417	Num	Number of days from admission to PR9
105PRDAY10	418	420	Num	Number of days from admission to PR10
106PRDAY11	421	423	Num	Number of days from admission to PR11
107PRDAY12	424	426	Num	Number of days from admission to PR12
108PRDAY13	427	429	Num	Number of days from admission to PR13
109PRDAY14	430	432	Num	Number of days from admission to PR14
110PRDAY15	433	435	Num	Number of days from admission to PR15
111 RACE	436	437	Num	Race (uniform)
112 SURGID_S	438	453	Char	Primary surgeon number (synthetic)
113 TOTCHG	454	463	Num	Total charges (cleaned)
114 TOTCHG_X	464	478	Num	Total charges (as received from source)
115 YEAR	479	482	Num	Calendar year
116 ZIPINC	483	484	Num	Median household income category for patient's zip code

Appendix B. Statistics of the paraesophageal hernia cases selected by procedure code '53.7'

Figure 1: **Distribution of genders.**

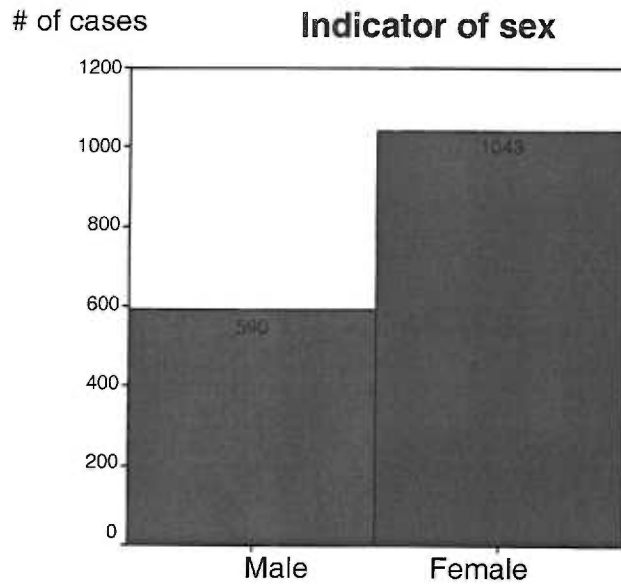


Figure 2: Distribution of race. Note that there were 577(35.3%) missing values.

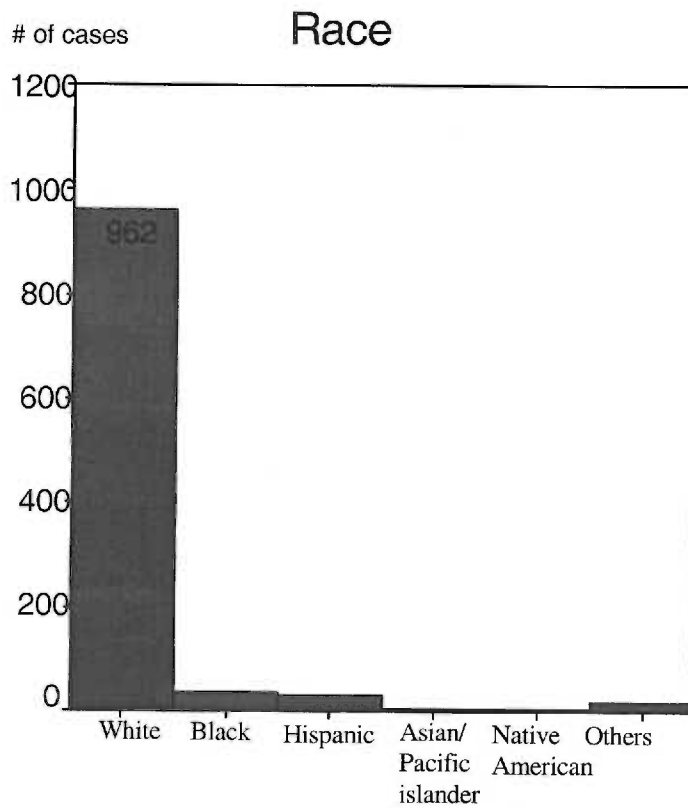
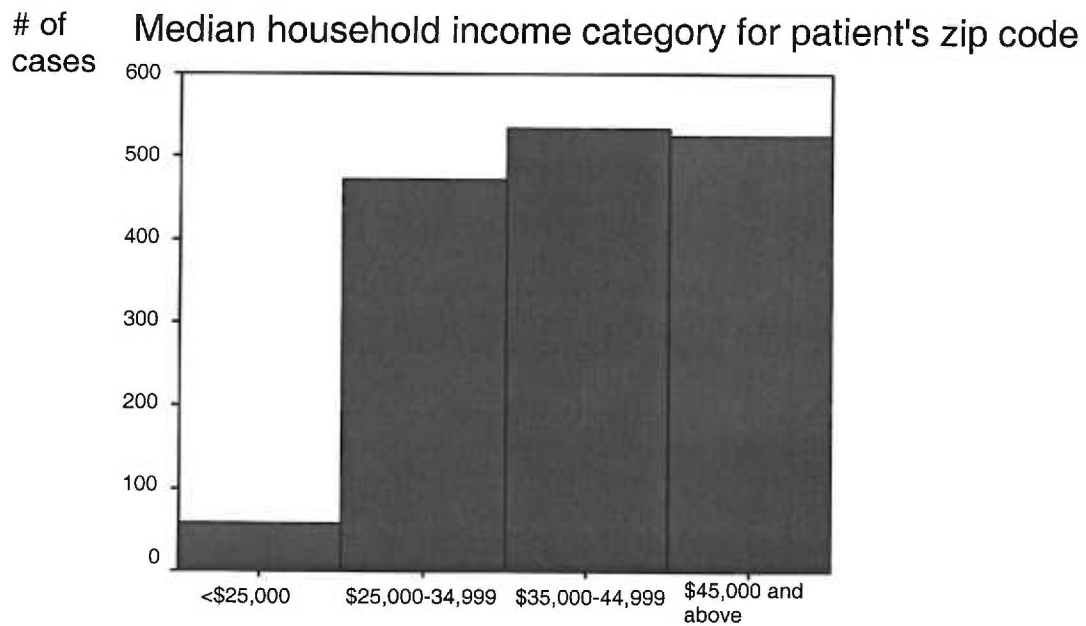


Figure 3: Distribution of age in years at admission.



Figure 4: Distribution of median household income category for patient's zip code.



Appendix C. Cover values and ICD-9-CM codes for the possible associated diseases and the 40 diseases groups identified in the association rule mining process.

ICD-9	Description	Cover	Group ID	Definition
53081	ESOPHAGEAL REFLUX	38.9	Gidx	Other GI diagnoses
5306	ACQ ESOPHAG DIVERTICULUM	0.8		
5368	STOMACH FUNCTION DIS NEC	0.8		
5300	ACHALASIA & CARDIOSPASM	.9		
53011	REFLUX ESOPHAGITIS	7.2		
5302	ULCER OF ESOPHAGUS	4.6		
53789	GASTRODUODENAL DIS NEC	4.1		
5303	ESOPHAGEAL STRICTURE	2.6		
53010	ESOPHAGITIS, UNSPECIFIED	2.0		
7872	DYSPHAGIA	1.7		
53019	OTHER ESOPHAGITIS	1.3		
53190	STOMACH ULCER NOS	1.1		
53550	GSTR/DDNTS NOS W/O HMRHG	1.0		
53390	PEPTIC ULCER NOS	0.7		
53140	CHR STOMACH ULC W HEM	0.7		
53540	OTH SPF GSTRT W/O HMRHG	0.7		
5370	ACQ PYLORIC STENOSIS	2.1		
4019	HYPERTENSION NOS	28.4	Hprtnsn	Hypertension
4011	BENIGN HYPERTENSION	2.1		
27801	MORBID OBESITY	13.3	Obsty	Obesity
27800	OBESITY NOS	2.3		
5680	PERITONEAL ADHESIONS	8.6	Adhesion	Peritoneal adhesions
57410	CHOLELITH W CHOLECYS NEC	6.9	Choles	Gall bladder/bile duct diseases
57511	CHRONIC CHOLECYSTITIS	4.5		
5756	GB CHOLESTEROLOSIS	1.7		
57400	CHOLELITH W AC CHOLECYST	0.7		
42731	ATRIAL FIBRILLATION	6.2	Dysrhytm	Dysrhythmia
42789	CARDIAC DYSRHYTHMIAS NEC	1.4		
V4589	Presence of neuropacemaker or other electronic device	0.7		
42769	PREMATURE BEATS NEC	0.7		
V4501	Cardiac pacemaker	0.7		
2449	HYPOTHYROIDISM NOS	6.0	Hypothy	Hypothyroidism
496	CHR AIRWAY OBSTRUCT NEC	5.9	Resp	Respiratory diseases
4928	EMPHYSEMA NEC	1.2		
78609	RESPIRATORY ABNORM NEC	1.2		
49320	CH OB ASTH W/O STAT ASTH	1.1		
49390	ASTHMA W/O STATUS ASTHM	5.8		

2720	PURE HYPERCHOLESTEROLEM	5.0	Hyperlipid	Hyperlipidemia
2724	HYPERLIPIDEMIA NEC/NOS	1.4		
25000	DMII WO CMP NT ST UNCNTR	5	DM	Diabetes
25001	DMI WO CMP NT ST UNCNTRL	0.9		
2851	SIDEROBLASTIC ANEMIA	4.8	Anemia	Anemia
2859	ANEMIA NOS	3.5		
2800	CHR BLOOD LOSS ANEMIA	1.8		
2809	IRON DEFIC ANEMIA NOS	1.3		
41401	CRNRY ATHRSCL NATVE VSSL	4.7	MI	Myocardial infarction
V4581	Aortocoronary bypass status	2.6		
41400	COR ATH UNSP VSL NTV/GFT	2.4		
412	OLD MYOCARDIAL INFARCT	1.7		
4139	ANGINA PECTORIS NEC/NOS	1.2		
V4582	Percutaneous transluminal coronary angioplasty status	1.0		
4280	CONGESTIVE HEART FAILURE	4.3	HtFail	Heart failure
78057	OTH UNSPCF SLEEP APNEA	4.1	SlpApnea	Sleep apnea
311	DEPRESSIVE DISORDER NEC	3.0	Depress	Depression
2639	PROTEIN-CAL MALNUTR NOS	2.6	Malnutr	Malnutrition
3051	TOBACCO USE DISORDER	2.4	Tobacco	Tobacco use
V1582	History of tobacco use	2.4		
5531	UMBILICAL HERNIA	2.4	Ohernia	Other hernia
55321	INCISIONAL HERNIA	1.3		
71590	OSTEOARTHROS NOS-UNSPEC	2.3	Arthros	Osteoarthritis
V4365	Organ or tissue replaced by other means, knee	0.7		
71690	ARTHROPATHY NOS-UNSPEC	1.4		
4240	MITRAL VALVE DISORDER	2.2	Valve	Valve disorder
4241	AORTIC VALVE DISORDER	0.7		
73300	OSTEOPOROSIS NOS	2.0	Osteo	Osteoporosis
56081	INTESTINAL ADHES W OBSTR	1.7	Obstr	Obstruction
5602	VOLVULUS OF INTESTINE	0.7		
5780	HEMATEMESIS	0.9	Hematemesis	Hematemesis
5789	GASTROINTEST HEMORR NOS	1.7		
78039	Other convulsions	1.7	Convul	Convulsions
5718	CHRONIC LIVER DIS NEC	1.6	Liver	Chronic liver disorder
7242	LUMBAGO	1.4	Lumbago	Lumbago
600	HYPERPLASIA OF PROSTATE	1.3	Prostate	Hyperplasia of prostate
56210	DVRTCLO COLON W/O HMRHG	1.3	Colon	Diverticula of intestine
56211	DVRTCLI COLON W/O HMRHG	1.1		
V103	Personal history of malignant neoplasm, Breast	1.2	Cancer	History of cancer

V1005	Personal history of malignant neoplasm, Large intestine	0.9		
V1046	Personal history of malignant neoplasm, Prostate	0.8		
4571	OTHER LYMPHEDEMA	1.2	Lymphedema	Lymphedema
7891	HEPATOMEGALY	1.2	Hepato	Enlargement of liver
6256	FEM STRESS INCONTINENCE	1.2	Incont	Incontinence
72939	PANNICULITIS, SITE NEC	1.1	Pncfts	Panniculitis
5570	AC VASC INSUFF INTESTINE	1.0	Vasc	Vascular insufficiency of intestine
5939	RENAL & URETERAL DIS NOS	0.9	Renal	Renal disorder
6111	HYPERTROPHY OF BREAST	0.9	Hprbr	Hypertrophy of breast
30000	ANXIETY STATE NOS	0.8	Anxiety	Anxiety
2948	ORGANIC BRAIN SYND NEC	0.8	Brain	Organic psychotic conditions
5770	ACUTE PANCREATITIS	0.8	Pncrtts	Diseases of pancreas
3320	PARALYSIS AGITANS	0.7	Paralysis	Paralysis

Appendix D. Variables used in CART® classification. The 40 disease groups from the association rule mining process are used as binary variables in CART®. The variable names for these 40 groups of diseases are the same as the group ID presented in appendix B. In addition, a variable was created for collagen diseases.

Variable Name	Definition	ICD-9-CM codes	Type
Case	PEH patient		Categorical (1/0)
Age	Age		Continuous
Female	Gender		Categorical (1/0)
Race	Race		Categorical
ZipInc	Income level		Continuous
Hospst	State postal code for hospital		Categorical
Pay1	Expected primary payer		Categorical
GIdx	Other GI diagnoses	530-537 7871 7872	Categorical (1/0)
Hprtnsn	Hypertension	401	Categorical (1/0)
Obsty	Obesity	278	Categorical (1/0)
Adhesion	Peritoneal adhesions	5680	Categorical (1/0)
Obstr	Obstruction	560	Categorical (1/0)
Choles	Gall bladder/bile duct diseases	574-576	Categorical (1/0)
Dysrhytm	Dysrhythmia	427 V450 V4589	Categorical (1/0)
Hypothy	Hypothyroidism	244	Categorical (1/0)
Resp	Respiratory diseases	490-496 78609	Categorical (1/0)
Hyperlipid	Hyperlipidemia	272	Categorical (1/0)
DM	Diabetes	250	Categorical (1/0)
Anemia	Anemia	280-285	Categorical (1/0)
MI	Myocardial infarction	410-414 V4581 V4582	Categorical (1/0)
HtFail	Heart failure	4280	Categorical (1/0)
SlpApnea	Sleep apnea	78051,53,57	Categorical (1/0)
Depress	Depression	311	Categorical (1/0)
Malnutr	Malnutrition	263	Categorical (1/0)
Tobacco	Tobacco use	3051 V1582	Categorical (1/0)
Ohernia	Other hernia	550-553 less	Categorical (1/0)

		5513, 23, 33	
Arthros	Osteoarthritis	715-716 V4365	Categorical (1/0)
Valve	Valve disorder	424	Categorical (1/0)
Osteo	Osteoporosis	73300	Categorical (1/0)
Hematemesis	Hematemesis	578	Categorical (1/0)
Convul	Convulsions	78039	Categorical (1/0)
Liver	Chronic liver disorder	571	Categorical (1/0)
Lumbago	Lumbago	7242 7245	Categorical (1/0)
Prostate	Hyperplasia of prostate	600	Categorical (1/0)
Colon	Diverticula of intestine	562	Categorical (1/0)
Cancer	History of cancer	V10	Categorical (1/0)
Lymphedema	Lymphedema	4571	Categorical (1/0)
Hepato	Enlargement of liver	7891	Categorical (1/0)
Incont	Incontinence	6265	Categorical (1/0)
Pnclts	Panniculitis	7293 7236 7248	Categorical (1/0)
Vasc	Vascular insufficiency of intestine	557	Categorical (1/0)
Renal	Renal disorder	5939	Categorical (1/0)
Hprbr	Hypertrophy of breast	6111	Categorical (1/0)
Anxiety	Anxiety	3000	Categorical (1/0)
Brain	Organic psychotic conditions	294	Categorical (1/0)
Pncrtts	Diseases of pancreas	577	Categorical (1/0)
Paralysis	Paralysis	3320	Categorical (1/0)
CD	Collagen diseases	In Table 2	Categorical (1/0)

Appendix E. Detailed illustrations from the CART® output of the four best trees. Class 0 = no paraesophageal hernia; Class 1 = paraesophageal hernia

Figure 1. Best tree selected from the set built by CART® using coronary artery diseases

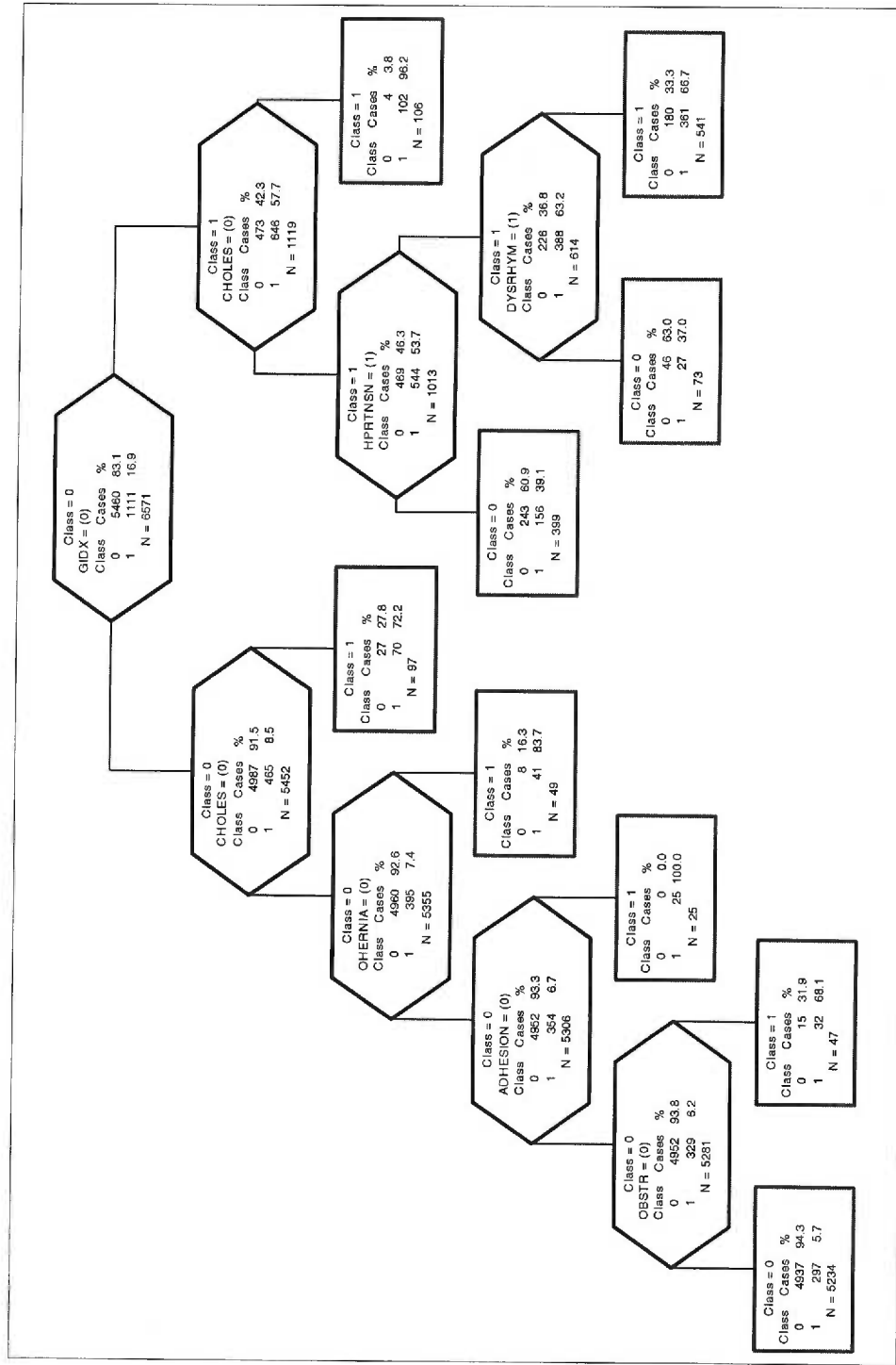


Figure 2. Best tree selected from the set built by CART® using malignant neoplasm as control:

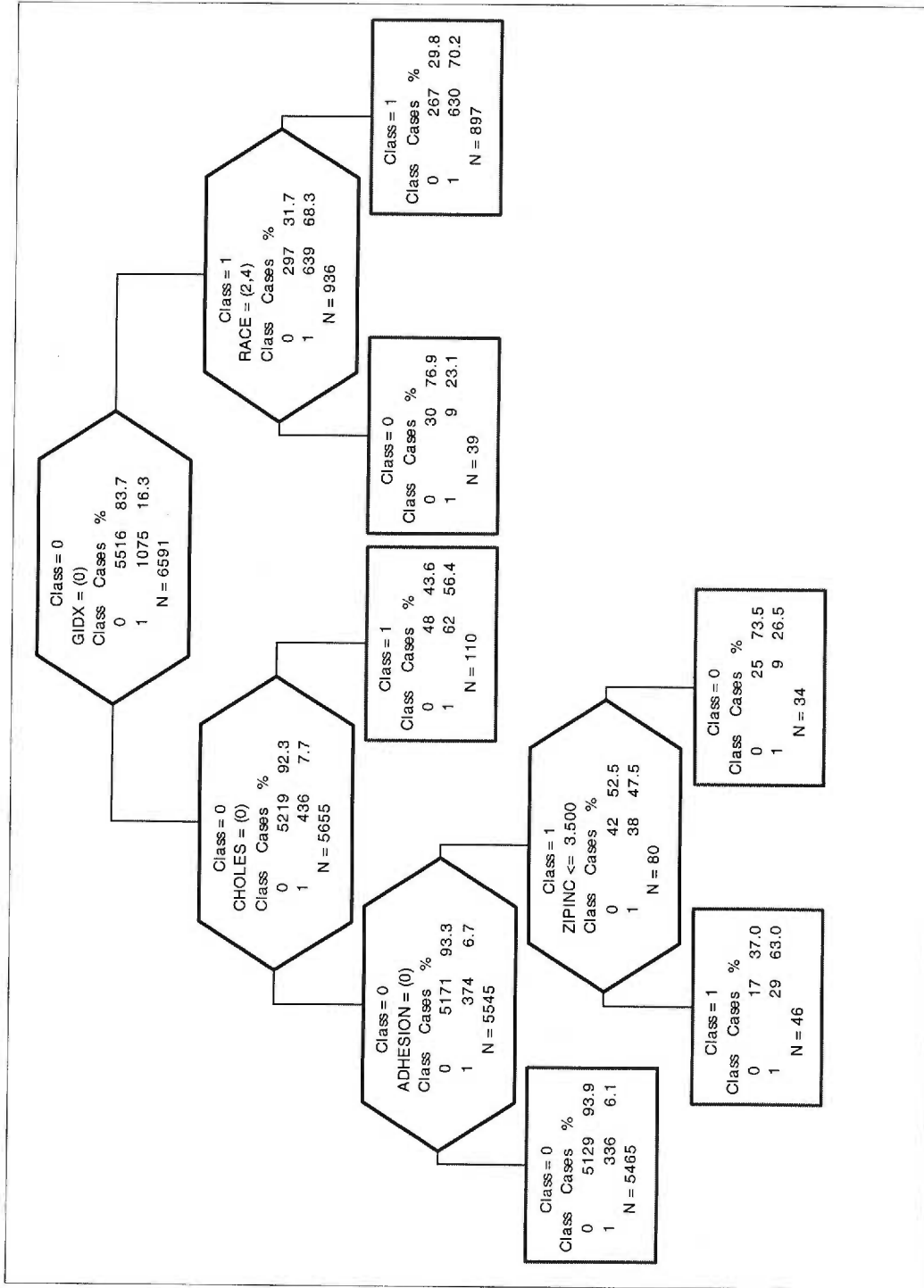


Figure 3. Best tree selected from the set built by CART® using reflux disease as control:

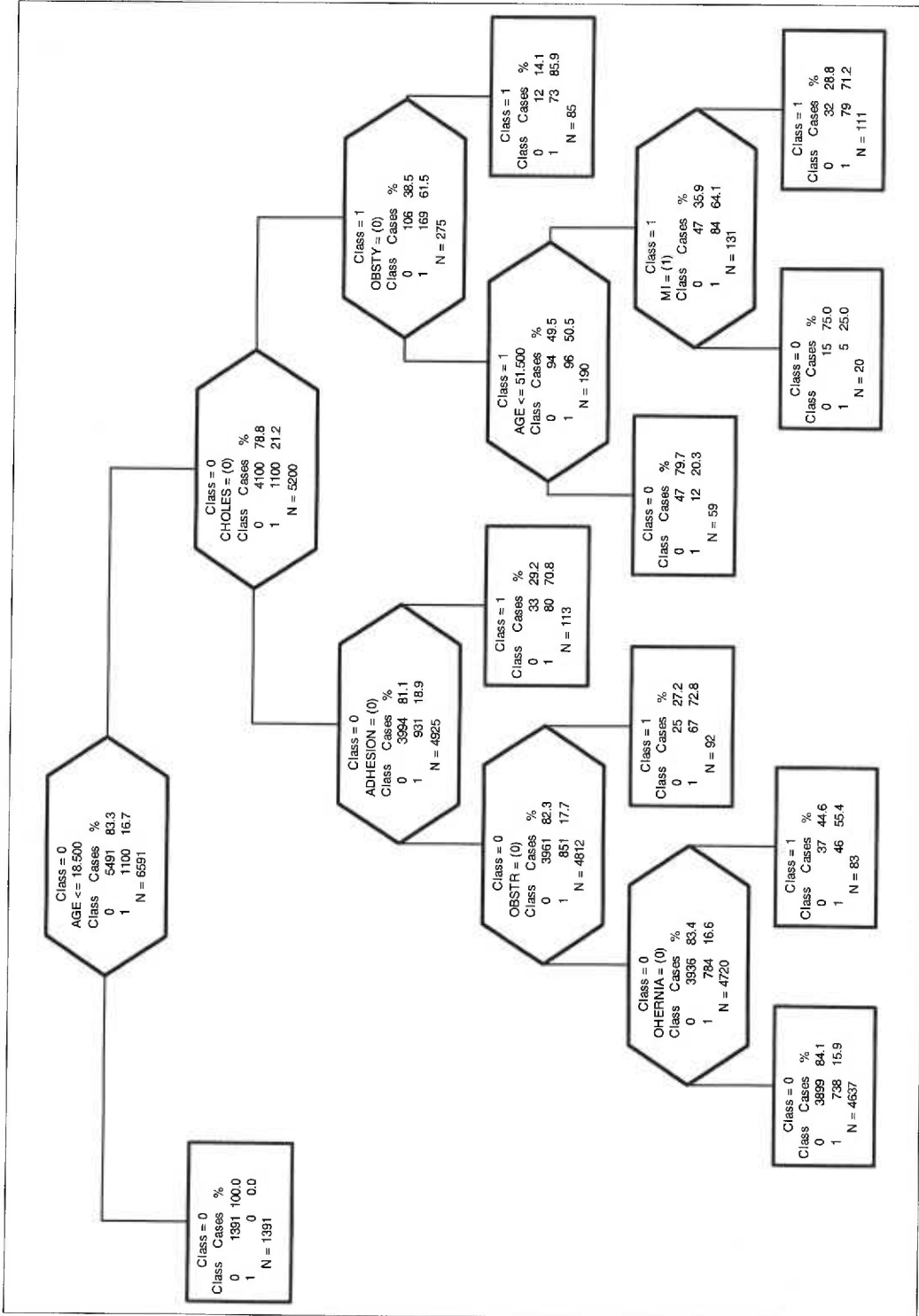


Figure 4. Best tree selected from the set built by CART® using type I hiatal hernia as control:

