7-7-2006

Mining Medical Data in a Clinical Environment

Tim V. Ivanovskiy
University of South Florida

Mining Medical Data in a Clinical Environment

by

Tim V. Ivanovskiy

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Dmitry Goldgof, Ph.D.
Lawrence Hall, Ph.D.
Sudeep Sarkar, Ph.D.

Date of Approval:
July 7, 2006

Keywords: data mining, rule association, medical expert system, apriori, medical implications

DEDICATION

This thesis is dedicated to: my wife Angela, who helped me in many ways during my graduate studies while she herself was completing her undergraduate degree; my parents Diana and Merle, for their constant support during my academic years; and the rest of my family, for working around my busy schedule.

ACKNOWLEDGEMENTS

I would like to thank Dr. Dmitry Goldgof and Dr. Lawrence Hall for giving me the opportunity to work with them and for providing me their input during my work. Thank you Dr. Sudeep Sarkar for being part of the supervisory committee. I would also like to extend my gratitude to Dr. Chris Garrett from Moffitt Cancer Center and his nurse Halina Greenstien for assisting me during the data collection phase of this thesis. Finally, I would like to thank Shibendra Pobi for working with me and all of the nurses and doctors at Moffitt Cancer Center's GI clinic.

I would also like to thank my wife Angela for making sure that my native Russian language did not migrate into the proper English structure of this thesis.

TABLE OF CONTENTS

LIST OF TABLES iii

LIST OF FIGURES v

ABSTRACT vi

CHAPTER 1 INTRODUCTION 1
- 1.1 Overview 1
- 1.2 Statistics 1
- 1.3 Clinical Trials 2

CHAPTER 2 PREVIOUS WORK 3
- 2.1 Similar Systems 3
- 2.2 About the MEANS System 4
- 2.3 MEANS Architecture 5
- 2.4 Knowledge Entry 5
  - 2.4.1 Question Types 7
  - 2.4.2 Protocol Example 7
- 2.5 MEANS Eligibility Algorithm 9
- 2.6 Data Mining 9
  - 2.6.1 Association Rules 11
  - 2.6.2 Mining Medical Data 15

CHAPTER 3 IMPROVEMENTS THROUGH AUTOMATED DATA ANALYSIS 17
- 3.1 Data Collection 17
  - 3.1.1 Implementation in a Clinical Setting 17
  - 3.1.2 GI Clinical Trials 17
  - 3.1.3 Patient Selection 18
- 3.2 Experiments and Results 19
  - 3.2.1 About Patient Data 19
    - 3.2.1.1 Data Extraction 20
  - 3.2.2 Mining Medical Implications 21
    - 3.2.2.1 Subset (Data Set One) 21
    - 3.2.2.2 Superset (Data Set Two) 26
    - 3.2.2.3 Rules of Interest (Subset) 27
    - 3.2.2.4 Rules of Interest (Superset) 30
    - 3.2.2.5 Medical Validation (Subset) 32
    - 3.2.2.6 Medical Validation (Superset) 33
  - 3.2.3 Probability-Based Reordering 35
    - 3.2.3.1 Probability-Guided Agent 38

3.2.3.2 Testing System 42
3.2.3.3 Probability Guided Experiments 44
3.2.4 Data Entry Optimization 48
3.2.5 Interface Changes 49

CHAPTER 4 CONCLUSION 53
4.1 Implication Discoveries 53
4.2 Probability-Based Reordering 54
4.3 Optimization 55
4.4 Future Work 56

REFERENCES 57

LIST OF TABLES

Table 2.1	Types of Questions	7
Table 2.2	Sample Protocol: Criteria	8
Table 2.3	Sample Protocol: Encoding	8
Table 2.4	Sample Protocol: Acceptance and Rejection Expressions	8
Table 2.5	Patient's State(s) During Screening	9
Table 2.6	Apriori Example: Database	12
Table 2.7	Apriori Example: One-item Sets	13
Table 2.8	Apriori Example: Two-item Sets	13
Table 2.9	Apriori Example: Three-item Sets	14
Table 2.10	Apriori Example: Rules	14
Table 3.1	Active Phase II GI Clinical Trials	18
Table 3.2	Patient Type in the Clinic	19
Table 3.3	Implication Example	21
Table 3.4	Subset: Patient Data	24
Table 3.5	Subset: Rule Statistics Per Protocol ($s = 10\%$)	24
Table 3.6	Superset: Patient Data	26
Table 3.7	Superset: Rule Statistics Per Protocol ($s = 5\%$)	27
Table 3.8	Subset: Antecedent-Consequent Matrix ($s = 20\%$)	28
Table 3.9	Subset: Rule Analysis ($s = 20\%$)	28
Table 3.10	Subset: Antecedent-Consequent Matrix ($s = 5\%$)	30
Table 3.11	Superset: Antecedent-Consequent Matrix ($s = 5\%$)	31
Table 3.12	Superset: Discovered Rules	33
Table 3.13	Example: Rule Simplification Based on Medical Knowledge	34

Table 3.14 Analytical Reordering 44

Table 3.15 Probability-Based Reordering (10-Fold Cross Validation) 45

Table 3.16 Probabilistic Thresholding (10-Fold Cross Validation) 46

Table 3.17 Max Number Per Set 47

Table 3.18 Yes/No Question Length 49

LIST OF FIGURES

Figure 2.1	Expert System Architecture	6
Figure 2.2	MEANS Eligibility Algorithm	10
Figure 3.1	Apriori Run (s=70%)	22
Figure 3.2	Decoded Implication	23
Figure 3.3	Subset: All Medically Valid Implications	25
Figure 3.4	Subset: Valid Implications (s = 20%)	29
Figure 3.5	Subset: Rule 206 (Closer Look)	29
Figure 3.6	Superset: All Medically Valid Implications	36
Figure 3.7	Final Rules	37
Figure 3.8	Testing System Algorithm	43
Figure 3.9	Thresholding (10-Fold Cross Validation)	47
Figure 3.10	MEANS Streamlined Initial Questions Page	50
Figure 3.11	Full Text Popup	50
Figure 3.12	Eligibility Color Table	51
Figure 3.13	Protocol Number and Title	51

MINING MEDICAL DATA IN A CLINICAL ENVIRONMENT

Tim V. Ivanovskiy

ABSTRACT

The availability of new treatments for a disease depends on the success of clinical trials. In order for a clinical trial to be successful and approved, medical researchers must first recruit patients with a specific set of conditions in order to test the effectiveness of the proposed treatment. In the past, the accrual process was tedious and time-consuming. Since accruals rely heavily on the ability of physicians and their staff to be familiar with the protocol eligibility criteria, candidates tend to be missed. This can result and has resulted in unsuccessful trials.

A recent project at the University of South Florida aimed to assist research physicians at *H. Lee Moffitt Cancer Center & Research Institute*, Tampa, Florida, with a screening process by utilizing a web-based expert system, *Moffitt Expedited Accrual Network System* (MEANS). This system allows physicians to determine the eligibility of a patient for several clinical trials simultaneously.

We have implemented this web-based expert system at the *H. Lee Moffitt Cancer Center & Research Institute* Gastroenterology (GI) Clinic. Based on our findings and staff feedback, the system has undergone many optimizations. We used data mining techniques to analyze the medical data of current gastrointestinal patients. The use of the Apriori algorithm allowed us to discover new rules (implications) in the patient data. All of the discovered implications were checked for medical validity by a physician, and those that were determined to be valid were entered into the expert system. Additional analysis of the data allowed us to streamline the system and decrease the number of mouse clicks required for screening. We also used a probability-based method to reorder the questions, which decreased the amount of data entry required to determine a patient's ineligibility.

CHAPTER 1

INTRODUCTION

1.1 Overview

The application of Artificial Intelligence to real-world issues has produced promising results. The work presented in this thesis is aimed toward improving the existing medical expert system MEANS–which stands for *Moffitt Expedited Accrual Network System*–to be more physician-friendly. This expert system is designed to be used as a tool for helping physicians screen patients for eligibility for clinical trials. It is our goal to see that the system eventually becomes part of the standard of care at Moffitt Cancer Center, Tampa, Florida, and its affiliates. Various techniques were used to increase user friendliness and acceptability of the expert system, including changes to the interface and system flow. In addition, data mining and statistical analysis techniques were used in order to decrease the amount of data entry needed from the user.

I will begin with a brief introduction to previous work on medical expert systems in Section 2.1 and continue with an explanation of our medical expert system in Section 2.2. Chapter 3 will take us into current work and the changes that were made to the expert system, including work that was done to increase the system's user friendliness and minimize the amount of time required to determine a patient's eligibility for clinical trials. I will conclude with results and future challenges in Chapter 4.

1.2 Statistics

In the United States, cancer is among leading causes of death. The American Cancer Society estimated that the number of new cancer cases in 2005 was 1,372,910 where 570,280 resulted in death. The National Institute of Health estimates that the cost of cancer in the United States was around $189.8 billion in 2004, from which only $64.4 billion was for direct

medical costs. The remaining, $125.4 billion was the cost of lost productivity due to illness or premature death. [4]

1.3 Clinical Trials

Also called clinical studies, clinical trials are defined by National Cancer Institute as: "A type of research study that tests how well new medical approaches work in people. These studies test new methods of screening, prevention, diagnosis, or treatment of a disease" [16]. Clinical trials can also be defined as a way to "evaluate the effectiveness and safety of medications or medical devices by monitoring their effects on large groups of people" [20]. Clinical trials can also be referred as clinical study.

CHAPTER 2

PREVIOUS WORK

2.1 Similar Systems

The use of artificial intelligence in medicine is not a new concept. Researchers realized in as early as 1950's that computers could be used to aid physicians with clinical decision making. Since then, physicians and computer scientists started analyzing medical information in a way that could be used by automated decision aids with the help of artificial intelligence for a certain domain. The term "knowledge engineering" refers to the use of computer-based symbolic reasoning, including knowledge representation, acquisition, explanation, and self modification that comes from self awareness [26]. In Section 2.2, I will show that our system is capable of self modification based on Bayes' learning algorithm.

One of the earliest expert systems, called DENDRAL was developed in early 1965 at Stanford University. Edward Feigenbaum, Joshua Lederberg and Bruce Buchanan were interested in the exploration of the mechanization of scientific reasoning and the formalization of scientific knowledge. Mass spectrometry was emerging as a technology of choice for chemical analysis. Therefore, they decided to apply their idea to the issue of how to properly represent then-existing chemical graph structures and then generate all possible structures in the mass spectral analysis domain [19, 28]. DENDRAL was based on a set of rule-based *if-then* reasoning to deduce the molecular structure of organic chemical compounds from known mass spectrometry data and chemical analyses. The project also created the standards for expert systems by separating internal operations of the system from the explicit rules of the knowledge [28].

The DENDRAL project gave rise to a famous antimicrobial therapy consultation system called MYCIN [25, 27, 28]. Shortliffe *et al.* set out to create a system that will be compatible with the physician's own decision-making process, and MYCIN was the result of their efforts.

It was designed to be used as a tool to assist physicians with the selection of antibiotic treatments for patients with bacterial infections. It was actually the first expert system to incorporate in its design a separate and modifiable knowledge base that consisted of *if-then* or "PREMISE" and an "ACTION" rules as described by Shortliffe *et al.* [25].

The success of DENDRAL and MYCIN not only signifies major achievements in artificial intelligence, but it also helped touch off the development of expert systems. EMYCIN, which actually moved into the commercial software market, was developed from MYCIN. EMYCIN was a generic, domain-independent expert system shell that could be used to build a rule-based expert system in any domain [28]. Janice S. Aikins *et al.* used the EMYCIN framework to build a medical expert system for the interpretation of pulmonary function tests for the patients with the lung disease called PUFF [3]. Another spin-off from MYCIN was a teaching expert system called NEOMYCIN. It was a combination of the knowledge base of MYCIN and a teaching program called GUIDON [7].

2.2 About the MEANS System

The medical expert system MEANS – *Moffitt Expedited Accrual Network System* – was originally developed at the University of South Florida around 1998 by Fletcher *et al.* [5]. The purpose of the system was to automate the selection of patients for cancer clinical trials. Since its conception, MEANS has been improved dramatically with addition of various optimization techniques. The original version, developed by Fletcher *et al.*, was a qualitative rule-based system. Cost-effectiveness of the selection process was addressed by Kokku *et al.* In their work, Kokku *et al.* looked at the cost-optimization problem of ordering related tests and minimizing the pain factor associated with the number of tests needed for answering eligibility criteria questions [18]. Physicians could reduce the overall cost of the screening process by ordering less expensive tests earlier in the process and then use those results to rule out a patient [9]. Savvas Nikiforou created a Knowledge Entry system in 2002 [21]. Until then, protocol expressions had to be coded by a programmer and took a significant amount of time and effort.

Goswami *et al.* experimented with Bayes' probabilistic reordering agent on retrospective data. Their experiments showed that the application of Bayes' probabilistic reordering agent

could reduce data entry by more than 20% [9, 13]. Probability-based reordering was added to the expert system by Goswami *et al.*; however, until now probabilistic knowledge gathering was not properly implemented in MEANS due to conceptual differences of the test system that Goswami *et al.* used to test the probability-based reordering agent. In Section 3.2.3 I will explain the additional modifications which that done to MEANS.

2.3 MEANS Architecture

The MEANS is divided into two main parts: *Patient Assignment* and *Knowledge Entry*. The development of the Patient Assignment subsystem was started by Fletcher *et al.* [5]. Kokku *et al.* continued development of the subsystem and implemented heuristics for ordering medical tests to minimize overall cost [18]. Goswami *et al.* added a probabilistic agent to the system [13]. The Knowledge Entry subsystem was developed by Nikiforou. It has a friendly web-based user interface which allows encoding of clinical protocols for use by the *Patient Assignment* system [21]. To fully complete the system, I refined the implementation of the probabilistic agent, streamlined the system flow and added a reporting subsystem. The detailed system architecture is shown in Figure 2.1.

As seen in Figure 2.1, the knowledge base contains information about clinical trials – stored as general and domain knowledge – together with implications and probabilistic knowledge. The probabilistic knowledge is gathered by the system every time an answer is evaluated by MEANS. Once a sufficient number of patients have been screened, the probabilistic knowledge can be used to reorder the questions. The database contains the information on all previously screened patients which includes the trials for which patients were screened as well as their current eligibility for screened trials and any previously provided answers. At any time, the user has the ability to change or delete an answer via the web-based user interface.

2.4 Knowledge Entry

The medical knowledge acquisition for MEANS is done via the Knowledge Entry subsystem, as seen in Figure 2.1. A protocol must be encoded using only the types of questions, listed in Table 2.1. For a numeric question, a range of numeric values is provided as an answer;

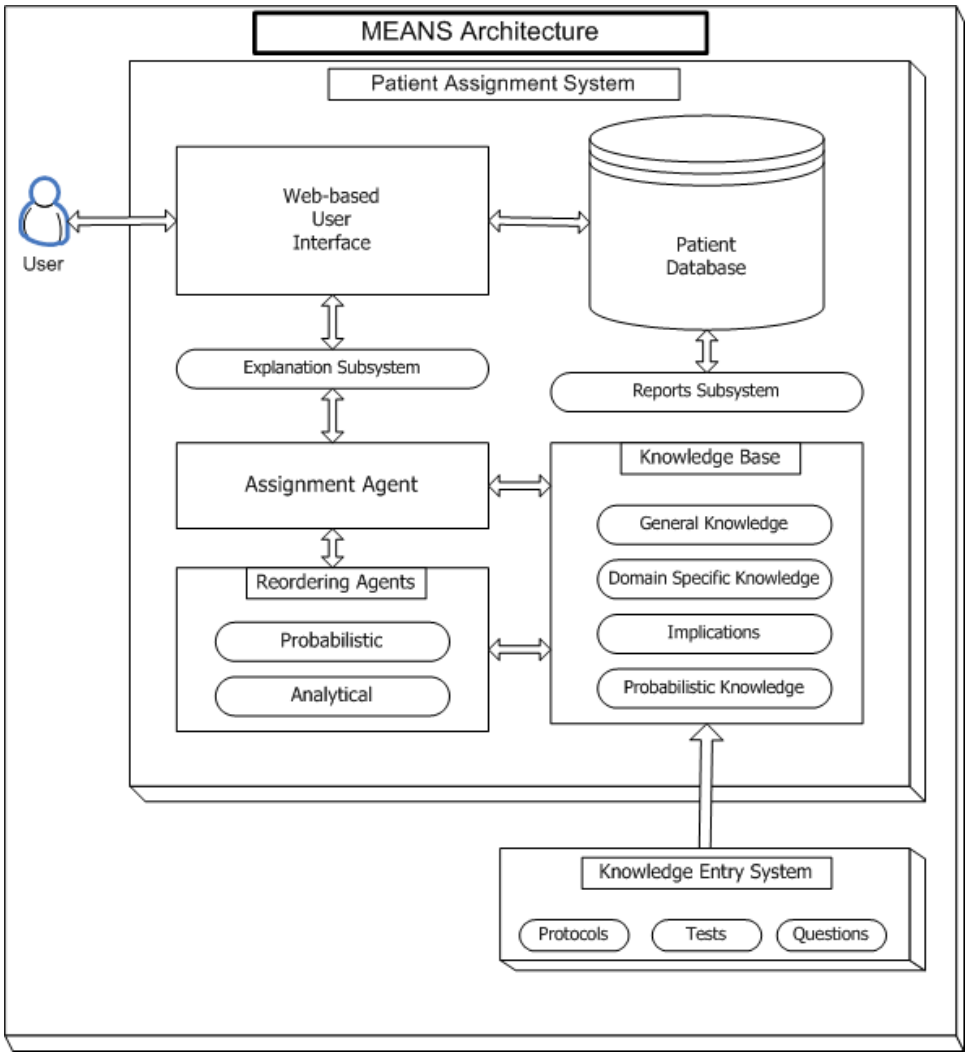


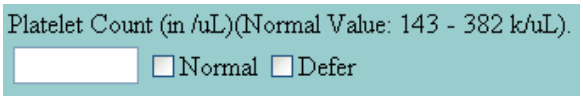
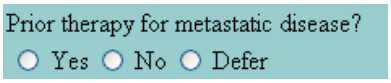
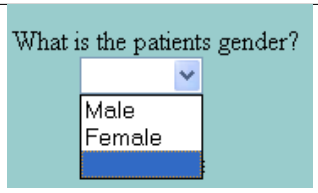
Figure 2.1 Expert System Architecture

for a yes/no question, yes or no may be selected as an eligible answer; for a multiple-choice question, all possible values are provided. The Knowledge Entry subsystem permits the use of logical expressions – such as AND and OR – to create combined questions. Once a protocol is encoded, the Knowledge Entry subsystem generates two expressions per question: 1) the Acceptance Expression, and 2) the Rejection Expression.

2.4.1 Question Types

All of the protocols are encoded utilizing three types of questions: 1) Numeric, 2) Yes/No, and 3) Multiple-choice. A sample of how each question appears in MEANS is shown in Table 2.1.

Table 2.1 Types of Questions

Question Type	Example	Acceptable Answer
Numeric		Numeric Value Normal (Auto Fill) Defer For Later
Yes/No		Yes No Defer For Later
Multiple Choice		Any Value From Dropdown

2.4.2 Protocol Example

This section shows how a simplified protocol can be encoded for use in MEANS. First, the protocol eligibility (inclusion/exclusion) criteria is interpreted. A sample protocol is shown in Table 2.2. Then, the protocol is broken down into separate questions, and the questions that the appropriate answers are entered into the Knowledge Entry system are shown in Table 2.3. Lastly, the Knowledge Entry subsystem converts the protocol into two types of expressions: 1)

the Acceptance Expression and 2) the Rejection Expression. Both expressions for our sample protocol are listed in the Table 2.4. I will discuss the need for both expressions in Section 2.5.

Table 2.2 Sample Protocol: Criteria

Inclusion Criteria	
1.	Patient must be male between the ages of 40 and 55
2.	Have life expectancy of greater than 12 weeks
3.	Able and willing to give written consent
4.	Have pathological diagnosis of adenocarcinoma of the stomach
Exclusion Criteria	
1.	Current tobacco user
2.	Participation in other clinical trials

Table 2.3 Sample Protocol: Encoding

#	Question	Eligibility Answer
1.	Age – – – – – AND – – – – – Sex	≥ 40 and ≤ 55 Male
2.	Life expectancy of > 12 weeks?	Yes
3.	Able and willing to give written consent?	Yes
4.	Pathological diagnosis of adenocarcinoma of the stomach?	Yes
5.	Current tobacco user?	No
6.	Patient on any other trial?	No

Table 2.4 Sample Protocol: Acceptance and Rejection Expressions

Acceptance Expression
{(Age: ≥ 40) *and* (Age: ≤ 55) *and* (Sex = Male)} *AND* (Life expectancy of > 12 weeks = Yes) *AND* (Able and willing to give written consent = Yes) *AND* (Pathological diagnosis of adenocarcinoma of the stomach = Yes) *AND* (Current tobacco user = No) *AND* (On any other trial = No) *AND*
Rejection Expression
{(Age: < 40) *or* (Age: > 55) *or* (Sex = Female)} *OR* (Life expectancy of > 12 weeks = No) *OR* (Able and willing to give written consent = No) *OR* (Pathological diagnosis of adenocarcinoma of the stomach = No) *OR* (Current tobacco user = Yes) *OR* (On any other trial = Yes) *OR*

2.5 MEANS Eligibility Algorithm

As discussed in Section 2.4, the Knowledge Entry subsystem is used to encode a protocol's eligibility criteria into the form that will be interpreted by MEANS. At any given time, a patient's status for a protocol, which is determined by the assignment agent (Figure 2.1), is in one of three states: 1) Eligible, 2) Ineligible, or 3) More Information Needed. If the *acceptance criteria* can be evaluated and is *TRUE*, then the patient is *Eligible* for the protocol. If the *rejection criteria* can be evaluated and is *TRUE*, then the patient status is set to *Ineligible*. If neither the *acceptance* nor *rejection criteria* can be evaluated, then the patient's status is set to *More Information Needed*. The reason for the determination of each state is listed in Table 2.5. The MEANS eligibility algorithm is shown in Figure 2.2. The eligibility algorithm is run for each protocol. Also, I should mention that the user is only presented with 10 questions at a time (line 15). Each time a user provides the system with new answers, the eligibility criteria is evaluated (line 18).

The discovery of new implications will allow questions that have not been asked to be answered by the system, thus decreasing the amount of data entry (Figure 2.2, line 17).

Table 2.5 Patient's State(s) During Screening

State	Reason
Eligible	Acceptance Criteria = TRUE
Ineligible	Rejection Criteria = TRUE
More Information Needed	Acceptance or Rejection can't be determined

2.6 Data Mining

Data mining techniques allow researchers to search through enormous amounts of data in order to discover association rules, emerging patterns and dependency rules. Data mining has been successfully applied to a number of application domains, including telecommunications, commerce, astronomy, geological survey, security, census analysis and text analysis [24, 30]. The identification of sets of items, products, symptoms and characteristics that often occur together in a given database are often seen as basic tasks of data mining.

```
1:  for (each protocol)
2:  {
3:    if (New Patient)
4:    {
5:      Protocol Status = "More Info Needed";
6:    }
7:    while (Protocol Status == "More Info Needed") do
8:    {
9:      Read Unanswered Questions;
10:     if (Unanswered Questions List == 0)
11:     {
12:       Read Deferred Questions;
13:     }
14:     Sort Questions By Importance;
15:     Ask Questions From User;
16:     Read NEW Answers;
17:     Apply Implications;
18:     Evaluate Acceptance Criteria;
19:     if (Acceptance Criteria == TRUE)
20:     {
21:       Generate Reason For Eligibility;
22:       Protocol Status = ELIGIBLE;
23:     }
24:     else
25:     {
26:       Evaluate Rejection Criteria;
27:       if (Rejection Criteria == TRUE)
28:       {
29:         Generate Reason For Ineligibility;
30:         Protocol Status = INELIGIBLE;
31:       }
32:     }
33:     if (Acceptance Criteria ≠ TRUE and Rejection Criteria ≠ TRUE)
34:     {
35:       Protocol Status = More Info Needed;
36:     }
37:   }
38: }
```

Figure 2.2 MEANS Eligibility Algorithm

2.6.1 Association Rules

The problem of *association rule* mining together with an algorithm for its solution was originally introduced by Agrawal *et al.* in 1993 [1]. It was designed for per-transaction base analysis of a supermarket database. The goal was to identify associations between sets of items with some minimal confidence. Agrawal *et al.* proposed a faster algorithm called *Apriori* to solve the *association rule* problem in 1994. [2]

The problem can be described as follows: Let $\mathcal{I} = \{\imath_1, \imath_2 , \ldots ,\imath_m\}$ be a set of all items. Let $\mathcal{D}$ be a database of transactions, where each transaction T consists of a set of items such that $T \subseteq \mathcal{I}$. A transaction T contains X, a set of items in $\mathcal{I}$, if $X \subseteq T$. An *association rule* XY is written in the form of an implication as $X \Rightarrow Y$, where $X \subset \mathcal{I}$, $Y \subset \mathcal{I}$ and $X \cap Y = 0$. An *association rule* must have a confidence and a support. The confidence c for the rule $X \Rightarrow Y$ holds if $c\%$ of transactions in $\mathcal{D}$ that contain X also contain Y. The rule $X \Rightarrow Y$ has a minimum support s in the transaction set $\mathcal{D}$ if $s\%$ of transactions in $\mathcal{D}$ contain $X \cup Y$. [1, 2]

The original motivation for the *Apriori* algorithm came from the need to analyze supermarket analysis data so that customer behavior could be examined in terms of purchased products. How often the products are purchased together are described by a frequent set. An example of this association rule is: *90% of transactions that had butter and eggs also contained milk*. In this example, the confidence c is 0.9 (90%).

Association rules are discovered by looking at *item sets* that have specified minimum support s (coverage). An *item* is a combination of attribute-value pairs. First we generate one-item sets for all attributes that have minimum support greater than a given minimum support value. The next step of the algorithm is the generation of two-item sets (two attribute-value pairs). We must note that we only generate new item sets if their minimum support is greater than a given minimum support value. Item sets in which an attribute takes two separate values will not be generated because it is impossible. For example, it is impossible for the temperature to be hot and cold at the same time; therefore, such an item-set will not be generated. The algorithm will generate new item sets until no more sets can be generated with given minimum support.

Here is an example of how the algorithm works: Suppose we have database $\mathcal{I}$ that consists of five (5) patients (Table 2.6). From this table, we can see that our database consist of 7 attributes and 17 values. Using a minimum support (s) of 3, we can find all possible association rules. I will discuss the confidence (c) level after we find the rules from our database $\mathcal{I}$.

Table 2.6 Apriori Example: Database

Sample Patients							
	Attributes						Goal
#	Metastatic	Biopsy	Complaint	Severity	Race	Age	Treatment
1	No	Yes	Some	High	French	10 to 20	Yes
2	Yes	No	Some	Low	Russian	10 to 20	Yes
3	No	Yes	None	Low	German	40 to 70	No
4	No	Yes	Some	Low	German	20 to 40	Yes
5	No	No	None	High	French	> 70	No

First we must create one-item sets from our database. We will scan our database and locate all possible attribute-value pairs that have occurred. We can see that our database $\mathcal{I}$ contains 17 one-item sets (Table 2.7).

After all one-item sets are found, they are thresholded by a given minimum support. If the support (frequency) of an item set is below a given minimum support, then the item set is excluded from further investigation. From Table 2.7 we can see that only 5 one-item sets have their support above or equal to the given minimum support of 3. For the next step we will only use the item sets that have passed the minimum support test.

In this step we will use one-item sets to compose two-item sets. The resulting two-item sets are shown in Table 2.8. We ended up with 10 two-item sets; however, only 2 out of the 10 pass the minimum support check. Only the first two two-item sets will be used in the next step of generating three-item sets.

Table 2.9 contains the resulting three-item sets. Since our minimum support was set at 3, none of the newly-generated three-item sets pass the minimum support check; hence, we will stop the generation of item sets at this time and will return to our two-item sets.

By looking at the two-item sets (Table 2.8) we will generate association rules. When *Metastatic* is *No* and *Biopsy* is *Yes*, this combination occurs 3 out of 4 times; however, the

Table 2.7 Apriori Example: One-item Sets

One-item Sets				
#	Attribute	Value	Support	Min Sup Check
1	Metastatic	No	4	✓
2	Biopsy	Yes	3	✓
3	Complaint	Some	3	✓
4	Severity	Low	3	✓
5	Treatment	Yes	3	✓
6	Biopsy	No	2	×
7	Complaint	None	2	×
8	Severity	High	2	×
9	Race	French	2	×
10	Race	German	2	×
11	Age	10 to 20	2	×
12	Treatment	No	2	×
13	Metastatic	Yes	1	×
14	Race	Russian	1	×
15	Age	40 to 70	1	×
16	Age	20 to 40	1	×
17	Age	> 70	1	×

Table 2.8 Apriori Example: Two-item Sets

Two-item Sets			
#	Attribute/Value	Support	Min Sup Check
1	Metastatic=No Biopsy=Yes	3	✓
2	Complaint=Some Treatment=Yes	3	✓
3	Metastatic=No Complaint=Some	2	×
4	Metastatic=No Severity=Low	2	×
5	Metastatic=No Treatment=Yes	2	×
6	Biopsy=Yes Complaint=Some	2	×
7	Biopsy=Yes Severity=Low	2	×
8	Biopsy=Yes Treatment=Yes	2	×
9	Complaint=Some Severity=Low	2	×
10	Severity=Low Treatment=Yes	2	×

Table 2.9 Apriori Example: Three-item Sets

Three-item Sets			
#	Attribute/Value	Support	Min Sup Check
1	Metastatic=No Biopsy=Yes Complaint=Some	2	×
2	Metastatic=No Biopsy=Yes Treatment=Yes	2	×
3	Complaint=Some Treatment=Yes Metastatic=No	2	×
4	Complaint=Some Treatment=Yes Biopsy=Yes	2	×

combination when *Biopsy* is *Yes* and *Metastatic* is *No* occurs 3 out of 3 times. These two rules are interpreted as follows: IF Metastatic=No THEN Biopsy=Yes; IF Biopsy=Yes THEN Metastatic=No.

From our sample database (Table 2.6) we were able to discover 4 rules with a minimum support of 3. I listed all of the discovered rules in Table 2.10. I should also point out that the table contains a confidence column. According to the table we can see that rule #1 has occurred 3 out of 4 times in our database. This gives the rule a confidence level of 75% (c=0.75). The other 3 rules have a confidence level of 1 (c=100%).

Table 2.10 Apriori Example: Rules

Sample Rules			
#	Rules	Support	Confidence
1	Metastatic=No ⇒ Biopsy=Yes	3 out of 4	75.00%
2	Biopsy=Yes ⇒ Metastatic=No	3 out of 3	100.00%
3	Complaint=Some ⇒ Treatment=Yes	3 out of 3	100.00%
4	Treatment=Yes ⇒ Complaint=Some	3 out of 3	100.00%

Since we will be mining medical data, we must use a confidence level of 1 (c=1). Based on that, we would only be interested in rules 2, 3, and 4. To complete our process of medical data mining, rules 2, 3, and 4 will need to be validated by a physician before we can declare them to be valid rules.

2.6.2 Mining Medical Data

As seen in Figure 2.1, implications are part of the knowledge base. In Section 3.2.2 I will explore implications in greater detail, but for now a brief overview will be sufficient in order to convey our purpose. To decrease the amount of data entry needed to determine a patient's eligibility, we explored the use of data mining.

Detection of trends and anomalies in populations has produced significant advances. In August of 1854 Dr. John Snow used statistical observations to find the source of Cholera transmission, thus stopping the Cholera outbreak in Britain. Edward Jenner used the observation that milkmaids who suffered from the mild disease of cowpox never contracted the more serious smallpox. After he conducted his famous experiment in 1796, he published his work in 1798, in which he coined the term "vaccine" from the Latin word *vacca*, or "cow" [6].

One of the problems that a data miner is faced with while mining medical data is that unlike other domains, the medical discipline is in itself diverse and complex. Sensible data mining requires significant domain expertise and as a result, an active collaboration between the data miner and the domain specialist [24].

Data availability and accuracy is another issue faced by the data miner. Gathering medical data is a tedious process and any lack of completeness or level of detail may render the data to be useless. Ethical and legal issues also need to be addressed when mining medical data. If, for example, the following rule was discovered:

$$ZipCode(12345), Age(18-25), Gender(Male) \Rightarrow Hepatitis\_B\_Status(Yes) \quad \gamma(20\%)$$

Then such a rule may not only be disturbing but could also be considered offensive should the *Zip Code(12345)* refer to an indigenous community [10, 24].

Researchers at the Pediatric Brain Tumor Research Program at Children's Memorial Hospital in Chicago used data mining when they performed gene expression analysis for pediatric cancers. They were able to isolate pediatric leukemia CD markers (antibodies that bind to proteins on the subraces of white blood cells and leukemic cells) and hope to use that knowledge to improve existing methods of diagnosis and treatment of the disease. [14]

Our goal was to use data mining techniques, specifically the Apriori algorithm, to discover new association rules from current patient data and encode newly-discovered rules into the MEANS implication subsystem [1, 2, 11].

CHAPTER 3

IMPROVEMENTS THROUGH AUTOMATED DATA ANALYSIS

3.1 Data Collection

3.1.1 Implementation in a Clinical Setting

Before the system could be used by physicians, it had to be properly and securely set up for use. Since the system is a web-based tool, no installation was needed on the physicians' computers and only an Internet Explorer shortcut had to be placed on the desktop. Necessary precautions were taken to comply with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule of 1996. HIPAA is the first Federal protection rule for the privacy of personal health information [22].

The MEANS system was launched on February 7, 2006 at Moffitt Cancer Center's Gastrointestinal (GI) clinic. The decision was made to introduce the system to nurses and physicians at the same time. Because the system was part of the study and had a separate study protocol number, all of the patients that were screened by the system had to sign a five-page HIPAA research authorization. Only after the HIPAA authorization was signed by the patient or guardian, could the patient be screened. Since the goal of the system was to screen patients that come to the clinic for the first time, only those who were seeing a gastroenterologist for the first time were considered as potential candidates for screening. As mentioned earlier, only after a patient signed the HIPAA authorization was his or her information entered into the system.

3.1.2 GI Clinical Trials

During April of 2006 the number of all active clinical trials at the *Moffitt Cancer Center* was 140 [15]. Since our study was aimed only at Phase II clinical trials, only *active* and *open*

for enrollment Phase II clinical trials were entered into MEANS. At the time of this study, *Moffitt Cancer Center* had 36 active Phase II clinical trials, five of which were conducted in the GI Clinic. The five active Phase II GI clinical trials, listed in Table 3.1, were encoded into MEANS.

Table 3.1 Active Phase II GI Clinical Trials

	Protocol #	Trial Description
1	13424	A Phase II Study of Capecitabine in Combination with Irinotecan and Oxaliplatin (Eloxatin) in Adult Patients with Advanced Colorectal Cancer
2	13426	A Phase II Study of Cisplatin and Irinotecan Induction Chemotherapy, Followed by ZD 1839 (IRESSA) in Adult Patients with Surgically Unresectable and/or Metastatic Esophageal or Gastric Carcinomas
3	13449	Microarray Analysis of Colon Cancer Outcome-A (MACCO-A) (For advanced or metastatic non-resectable colon cancer)
4	13946	Pharmacogenomic Study of Neoadjuvant Pre-irradiation Docetaxel and Cisplatin, followed by Neoadjuvant Concomitant Docetaxel, Cisplatin and Irradiation, followed by Surgery (DC-DCR-S) in Adult Patients with Operable Adenocarcinomas of the Esophagus or Gastroesophageal Junction
5	14607	Randomized phase II study of AG-013736 and Gemcitabine in Chemo-Naive Pancreatic Cancer Patients

3.1.3 Patient Selection

We collected our data at the GI Clinic of the *Moffitt Cancer Center*. There are three types of patients in the clinic: *New Patient* (NP), *New Established Patient* (NEP) and *Established Patient* (EP). Table 3.2 provides explanations of each patient type. It was our goal to screen every new patient for eligibility for clinical trials in the GI Clinic; hence, our study focused only on NP and NEP patients.

Table 3.2 Patient Type in the Clinic

Type	Explanation
New Patient (NP)	Never seen the doctor at the clinic -First Time Visit
New Established Patient (NEP)	Previously been seen by another doctor at the clinic
Established Patient (EP)	Previously seen this doctor. -Repeat visit

3.2 Experiments and Results

This section describes the data mining experiments that we conducted and our findings.

3.2.1 About Patient Data

We collected data on current NP and NEP patients at the GI Clinic of *Moffitt Cancer Center*. Since patients were going through treatment, not all patient data – such as test results – were available during the initial screening.

MEANS tracks each patient via a patient profile. Each patient profile contains the list of protocols for which he or she was screened together with the list of all answered questions. A patient can be determined ineligible anytime during the screening process. If a patient was ruled out after completion of the initial questions page, then the number of answered questions could be 10 (if all of the questions on the initial page were answered); however, if a patient was ruled out by the last question in the eligibility criteria of each protocol, the number of answered questions in a patient's profile could be 90. Based on that, each patient profile could contain a different number of questions.

Since we were working on a new system, there is no archival data available for analysis. We decided to start the mining patient data after collecting the data for approximately three weeks. Therefore, our data was divided into two sets: *Subset* (Data Set One), which consisted of 161 GI patients whose data was collected from February 7th to March 1st and *Superset* (Data Set Two), which is the extension of the Subset and included GI patient data from February 7th to April 21st.

I must clarify the fact that the data that was used in all of the experiments did not contain any potentially identifiable patient information. An anonymous patient profile was assigned a number and that number was used during the experiments.

3.2.1.1 Data Extraction

Before patient data was mined, some preprocessing was done. We used WEKA, a data mining software toolkit from the University of New Zealand, to discover new association rules [29]. To convert the data from the MEANS format to a format usable by WEKA, a suite of tools was developed by the author of this thesis. These tools permitted necessary pre- and post-processing, such as data extraction, analysis, cleaning, formatting and rule recovery. The suite was written in C and C++. The tools are included on the CD that accompanies this thesis.

The Apriori algorithm does not accept continuous values so the questions that had continuous values as an answer needed to be excluded from the list of questions. The encoding of the GI protocols into MEANS produced a total of 225 questions. Of the 225 questions, 33 of them required continuous values. Default questions answered during the streamlining process, such as *Signed Informed Consent*, would not contribute to the discovery of new association rules since all patients had the same answer; therefore, default questions needed to be excluded from patient profiles.

Using pre-processing tools, patient data was analyzed. Continuous valued questions – as well as default questions – were removed from patient profiles. The maximum number of questions that a patient profile could have was 192. Each patient's information then could be viewed as a vector with a size between 0 (zero) and 192.

Patient data was retrieved from MEANS and stored in a single .arff file so it could be read by the data mining software package. The details about the .arff format are covered in WEKA's documentation [29].

Because we were looking for implications that always hold, we used the confidence c of 1 (100%) for all of our experiments and varied the minimum support s from run to run. The data mining software we used starts each run at $s = 100\%$ and decreases s by 5% until a given minimum support is reached.

3.2.2 Mining Medical Implications

As seen in Figure 2.1, implications are part of the knowledge base. Implications allow us to make inferences based on existing information, therefore decreasing the number of questions needed to be asked for eligibility determination.

Example: Based on the sample questions in Table 3.3, we can construct several implications. If Q1 is answered *Male*, and since it is known that only females can be pregnant, then the following implication can be made: If Q1=*Male* IMPLIES Q2=*No*. This implication is always true (confidence = 100%). An implication that has a confidence $< 100\%$ – for example, Q2=*No* IMPLIES Q1=*Male* – is false in some cases – a female that is not pregnant – can not be considered a valid implication.

Table 3.3 Implication Example

Q1 = Patient's Sex (Male/Female)?
Q2 = Pregnant (Yes/No)?
Q3 = Patient's Age?
Valid Implications (Confidence = 100%):
✓ If Q1=*Male* IMPLIES Q2=*No*
✓ If Q1=*Female* and Age $\geq$ 70 IMPLIES Q2=*No*
Invalid Implication (Confidence $<$ 100%):
$\times$ If Q2=*No* IMPLIES Q1=*Male*

Our goal was to use data mining techniques, specifically the Apriori algorithm, for the generation of association rules. The association rules were mined using WEKA [2, 29]. The encoding of the MEANS data (pre-processing), the decoding of the rules, and discovery of the rules of interest (post-processing) was performed using post-processing tools that where written by the author.

3.2.2.1 Subset (Data Set One)

During the pre-processing phase we found that of the 175 patients, 14 patients contained only default answers in their profile. These patients would not provide any useful information and so were not considered in this experiment. We also found that there were 33 questions that required continuous values for an answer. Since the Apriori algorithm does not deal with

continuous values, these 33 continuous valued questions were excluded from the remaining patient profiles. The result of the pre-processing was a single file that could be viewed as 161 patient vectors with 192 possible dimensions.

For this experiment, 161 records – each with a possible 192 dimensions – were mined. Since we were looking for implications that always hold, a confidence of 100% ($c = 1$) was used. We started with minimum support (s) of 95% and decreased it by 5% until $s = 5\%$.

The first run that produced any result was at a minimum support of 70% (Figure 3.1). With a minimum support of 70%, only 4 rules were generated at $c = 100\%$. After the rules were decoded, the analysis of the rules – shown in Figure 3.1 – sparked some interest. It was found that rules two (2), three (3) and four(4) were permutations of rule one (1):

$$000\_y1y{=}No \Rightarrow 009\_y1y{=}No$$

Rule one (1) had a coverage of 128 which is greater then the rest of the rules generated during this run.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 1000 -T 0 -C 1.0 -D 0.05 -U 1.0 -M 0.7 -S -1.0
Relation:     MEANS.symbolic
Instances:    161
Attributes:   192
              [list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.7
Minimum metric <confidence>: 1
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 7

Best rules found:

  1. 000_y1y=No 128 ==> 009_y1y=No 128    conf:(1)
  2. 000_y3y=Yes 000_y1y=No 120 ==> 009_y1y=No 120    conf:(1)
  3. 000_y1y=No 000_y64y=No 118 ==> 009_y1y=No 118    conf:(1)
  4. 000_y1y=No 000_y125y=No 114 ==> 009_y1y=No 114    conf:(1)
```

Figure 3.1 Apriori Run (s=70%)

The decoded version of rule one (1) is shown in Figure 3.2. This new association rule was reviewed by a physician and deemed medically valid. I should note that [000_y1y] belongs to protocol #13426 and question [009_y1y] belongs to protocol #13946. We found an association rule between two protocols. This was a promising beginning; we found a simple, medically-valid implication. This implication was added to the implication module of MEANS.

1. 000_y1y=No 128 ⇒ 009_y1y=No 128 conf:(1)
Question: [000_y1y]
[Path diagnosis of esophageal SCC or ACA?] = [No]
Occurred [128] Times:
IMPLIES ⇒
Question: [009_y1y]
[Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [No]
Occurred [128] Times:
With conf:(1)

Figure 3.2 Decoded Implication

We continued running experiments by decreasing the minimum support s. As mentioned earlier, each run starts with $s = 100\%$ and decreases s by 5% until a given minimum support is reached; therefore, the rules that were recovered at $s = 70\%$ also showed up in during subsequent runs when minimum support was $< 70\%$. Since 4 rules may not have encompassed everything of interest, I continued to decreased minimum support. When s was decreased to 60%, 14 new rules were generated. Table 3.4 contains the association rule summary for the subset.

We were also curious if there were any implications within a set of questions for each protocol. During data collection, some of the clinicians chose to screen some patients only for a subset of our trials; therefore, not all patients were screened for all five trials. Table 3.5 shows the results of the per protocol mining experiment at s=10% on the subset. Only protocols 13424 and 13946 produced any rules of interest. However, when these rules were examined for medical validity, *none* of the 13 rules for 13424 and 15 rules for 13946 were medically valid.

Since a very large number of association rules could be found, we stopped the analysis of the rules at $s = 5\%$. At minimum support of 5% the last rule had a coverage of 8.

Table 3.4 Subset: Patient Data

Min Support (s) %	Min Coverage (#)	# Rules	# New Rules
70	114	4	4
60	97	18	14
50	83	30	12
40	78	32	2
30	48	103	71
20	32	362	259
10	16	2,537	2,175
5	8	49,564	47,027

Table 3.5 Subset: Rule Statistics Per Protocol ($s = 10\%$)

	13424	13426	13449	13946	14607
Total Patients	158	157	158	154	157
Total Answers	686	590	665	444	502
Rules Recovered	20	0	0	39	0
Rules Of Interest	13	0	0	15	0
Medically Valid Rules	0	0	0	0	0

In the end we discovered 2 true rules and 5 conditional rules. The final list of medically valid rules from the subset is listed in Figure 3.3 (Subset: All Medically Valid Implications). An explanation of this figure will be discussed in Section 3.2.2.3.

Subset: All Medically Valid Implications
··· 2 True Rules ···
#1: [Path diagnosis of esophageal SCC or ACA?] = [No] IMPLIES ⇒ [Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [No]
#5471: [Metastatic colon or rectal cancer (tissue proven) and not suitable for surgery?] = [Yes] IMPLIES ⇒ [Histologically confirmed metastatic colorectal cancer?] = [Yes]
··· 5 Conditional Rules ···
#3393: [Tissue proven pancreas adenocarcinoma?] = [Yes] IMPLIES (if same tissue sample) ⇒ [Metastatic colon or rectal cancer (tissue proven) and not suitable for surgery?] = [No]
#3394: [Tissue proven pancreas adenocarcinoma?] = [Yes] IMPLIES (if same tissue sample) ⇒ [Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [No]
#3395: [Tissue proven pancreas adenocarcinoma?] = [Yes] IMPLIES (if same tissue sample) ⇒ [Histologically confirmed metastatic colorectal cancer?] = [No]
#3397: [Tissue proven pancreas adenocarcinoma?] = [Yes] IMPLIES (if same tissue sample) ⇒ [Path diagnosis of esophageal SCC or ACA?] = [No]
#206(modified) Rule has 1 antecedent and 1 consequent: [001_y1y] [Measurable disease (RECIST)?] = [No] IMPLIES (if same tissue sample) ⇒ [000_y138y] [Tissue proven pancreas adenocarcinoma?] = [No]
TOTAL: 7 implications

Figure 3.3 Subset: All Medically Valid Implications

3.2.2.2 Superset (Data Set Two)

Our superset was the extension of the subset. As more and more patients were screened, the data available for mining increased. Our superset had to go through some pre-processing before the mining could begin.

We started with 393 patient records; however, after pre-processing, we discovered that 26 of the patient records contained only default questions and/or a combination of default questions and continuous values. We ended up with only 367 patient records, which meant that our superset contained 206 more patient records than the subset. Since the pool of the questions had not changed, each patient record could contain up to 192 possible dimensions. The superset contained a total of 5055 answers from 367 patients.

Like before, we started data mining with a minimum support of 95% and decreased it by 5% until it reached 5%, keeping the confidence at 100%. In comparison to the subset, we did not get any rules at s=70%. The first result for the superset was produced at s=30%. As shown in Table 3.6, at s=30% we found 27 rules. The minimum coverage for the last rule at s=30% was 111 instances. We continued to decrease the minimum support until 5%. We stopped at s=5% with the discovery of 7079 rules.

Table 3.6 Superset: Patient Data

Min Support (s) %	Min Coverage (#)	# Rules	# New Rules
30	111	27	27
20	73	71	44
10	37	343	272
5	18	7,079	6,736

We also took a look at the data on a per protocol basis. The result of this experiment on the superset at s=5% is shown in Table 3.7. As mentioned earlier, clinicians have a choice of selecting any combination of protocols for which to screen a patient. As a result, the number of patients screened differed between protocols. However, protocols 13424, 13449 and 14607 had the same number of patients.

Even though we a discovered significant number of rules, they needed to be analyzed before anything could be said about them. We still needed to validate them to make sure they made

Table 3.7 Superset: Rule Statistics Per Protocol ($s = 5\%$)

	13424	13426	13449	13946	14607
Total Patients	362	361	362	355	362
Total Answers	1658	1426	1588	907	1303
Rules Recovered	2	0	0	2	0
Rules Of Interest	1	0	0	1	0
Medically Valid Rules	0	0	0	0	0

sense and to confirm their medical validity with a physician. By comparing the subset rules (Table 3.5) and the superset rules (Table 3.7), we can see that the subset had a greater number of rules that were discovered; however, the number of medically valid rules for both were the same – zero.

3.2.2.3 Rules of Interest (Subset)

Before we could claim that we found new implications, we needed to analyze the rules for permutations and medical validity. I will first discuss our findings on the subset data. Later, I will demonstrate how these discoveries helped us come up with a filtering heuristic to narrow our search during the superset analysis. The rule numbers referred to from now on refer to the discovery sequence of the rules and not the rule sequence of valid rules.

The run at $s = 20\%$ on the subset produced a total of 362 new association rules. The rule set was analyzed in terms of the number of *antecedents* and *consequents*. The matrix in Table 3.8 shows the breakdown. The rows represent the number of antecedents and columns represent the number of consequents in a rule. We can see that our rule set contains rules with up to 7 antecedents and 3 consequents.

The rules from the subset run at $s = 20\%$ were analyzed for medical validity. We found that of 362 rules, 201 rules were medically invalid; however, 161 rules were medially valid and were of interest. Further analysis of the rules revealed that of 161 rules of interest, 158 were permutations of valid rules. Only 3 rules (1, 206, and 213) out of 362 were found to be medically valid. The rule summary table for the subset rule analysis at $s = 20\%$ is shown in Table 3.9.

Table 3.8 Subset: Antecedent-Consequent Matrix ($s = 20\%$)

Ant/Cons Matrix			
	1	2	3
1	1	0	0
2	26	9	2
3	78	25	4
4	102	20	0
5	66	5	0
6	20	1	0
7	3	0	0

Table 3.9 Subset: Rule Analysis ($s = 20\%$)

Rule Type	#
Invalid Rules	201
Permutations	158
Valid Rules	3

A closer look at the 3 valid rules (Figure 3.4) revealed that rule one (1), which was actually discovered at s=70%, was the only simple rule. Rules two (2) and three (3) both had two antecedents, one of which they shared.

The results of our experiments on the subset data indicated to us that we should have concentrated on the rules that contained a maximum of two antecedents and only one consequent. For example:

$$\mathrm{A} \Rightarrow \mathrm{C}$$

$$\mathrm{A} \wedge \mathrm{B} \Rightarrow \mathrm{C}$$

As mentioned earlier, rules #206 and #213 had the same first antecedent (Figure 3.4).

We first examined rule #206, which was the second rule on our list (Figure 3.4). After taking a closer look at the subset data, we found that of 160 patients, 2 patients did not have the answers for two questions, [000_y64y] and [000_y138y]. We removed these two patients from our data and ran the experiment again. We found that aside from the existing rule #1 (Figure 3.4), there were two additional simple rules, 1 antecedent and 1 consequent (Figure 3.5). So with the additional analysis of our data, we were able to come up with the simple rule.

Subset: Valid Implications ($s = 20\%$)
1) Rule #1 has 1 antecedent and 1 consequent: [000_y1y] [Path diagnosis of esophageal SCC or ACA?] = [No] IMPLIES ⇒ [009_y1y] [Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [No]
2) Rule #206 has 2 antecedents and 1 consequent: [001_y1y] [Measurable disease (RECIST)?] = [No] – AND – [000_y64y] [Any histopathologically proven diagnosis of malignant GIST that is not amenable to standard therapy with curative intent?] = [No] IMPLIES ⇒ [000_y138y] [Tissue proven pancreas adenocarcinoma?] = [No]
3) Rule #213 has 2 antecedents and 1 consequent: [001_y1y] [Measurable disease (RECIST)?] = [No] – AND – [000_y125y] [Metastatic colon or rectal cancer (tissue proven) not suitable for surgery?] = [No] IMPLIES ⇒ [000_y138y] [Tissue proven pancreas adenocarcinoma?] = [No]

Figure 3.4 Subset: Valid Implications ($s = 20\%$)

Subset: Rule 206 (Closer Look)
1) Rule has 1 antecedent and 1 consequent: [001_y1y] [Measurable disease (RECIST)?] = [No] IMPLIES ⇒ [000_y138y] [Tissue proven pancreas adenocarcinoma?] = [No]
2) Rule has 1 antecedent and 1 consequent: [001_y1y] [Measurable disease (RECIST)?] = [No] IMPLIES ⇒ [000_y64y] [Any histopathologically proven diagnosis of malignant GIST that is not amenable to standard therapy with curative intent?] = [No]

Figure 3.5 Subset: Rule 206 (Closer Look)

After the simple rules (Figure 3.5) were reviewed a physician, it was discovered that rule #2 was not valid; however, rule #1 was validated, but only under certain assumptions. I will discuss medical validation in detail in Section 3.2.2.5.

When the minimum the support was dropped to 5%, mining of the subset produced 49,564 rules (Table 3.4). After analysis of the rules, we discovered that 100 of them were simple rules. Table 3.10 shows us that this set of rules contained only 100 simple rules. We will restrict our analysis to these rules.

Table 3.10 Subset: Antecedent-Consequent Matrix ($s = 5\%$)

Antecedents(row)/Consequents (col) Matrix								
	1	2	3	4	5	6	7	8
1	100	175	241	245	177	85	24	3
2	997	1,566	1,719	1,307	662	200	27	0
3	3,205	4,274	3,669	2,035	672	100	0	0
4	4,858	5,308	3,429	1,274	213	0	0	0
5	3,938	3,364	1,506	2,94	0	0	0	0
6	1,825	1,108	269	0	0	0	0	0
7	477	160	0	0	0	0	0	0
8	58	0	0	0	0	0	0	0

When we look at the subset data at a minimum support of 5% we find that of 100 simple rules, only 6 were valid. Rule #1, was previously seen at s=20%. Rules 5471, 3393, 3394, 3395, and 3397 were newly discovered. I should mention that rule 5471 is a *true rule*, which means it refers to the same tissue type (colorectal).

3.2.2.4 Rules of Interest (Superset)

Let's take a look at the rules discovered in our superset. All rule numbers from now on refer to the superset rules unless otherwise noted. Table 3.11 is a snapshot of the discovered rules from the superset. Again, our rules of interest only include rules with a maximum of two antecedents and exactly one consequent. In the table we can see that there are 13 one-antecedent, one-consequent rules and 241 two-antecedents, one-consequent rules, a total of 254 rules of interest.

Table 3.11 Superset: Antecedent-Consequent Matrix ($s = 5\%$)

	1	2	3	4	5	6
1	13	7	4	1	0	0
2	2,41	128	82	40	13	2
3	903	475	226	82	14	0
4	1,303	716	303	65	1	0
5	1,034	509	134	5	0	0
6	461	159	12	0	0	0
7	116	17	0	0	0	0
8	13	0	0	0	0	0
Antecedents(rows)/Consequents (cols)						

After examining the 254 rules of interest of in our superset, we found that 73 rules had some potential; however, closer examination was needed. Since our 254-rule pool consisted of rules with a maximum of two antecedents and only one consequent, it is interesting to note that only 4 of the 73 rules had one antecedent and one consequent pair. The rest had two antecedents. Out of the 4 rules with one antecedent, only one rule (rule #998) was the true rule (not dependent on any assumptions). The other three (#1355, #1356, #1357) could only be validated under the assumption that the antecedent and consequent of the rule referred to the same tissue sample.

Now let's take a look at the rules with two antecedents and once consequent. We can see that there are five (5) rules that looked promising (rules #1, #343, #1347, #7036, #7076). It was interesting to discover that the rule #1 from Figure 3.2, which was found during analysis of the subset, acquired an additional antecedent (Sex = Female). A similar case occurred with rules #343 and #7036. Superset rules #343 and #7036 were the same as the subset rule #5471 except that #343 acquired an additional antecedent (Take Medications By Mouth = YES) while #7036 acquired a different antecedent (Uncontrolled Medical Conditions = NO). I will discuss these findings in greater detail during the Medical Validation discussion (section 3.2.2.5).

The final count of rules that was discovered from the superset was 9. Table 3.12 lists the newly discovered rules from the superset together with the number of antecedents and consequents per rule and medical comments. I will expand on the medical comments in the next section.

3.2.2.5 Medical Validation (Subset)

Because we were mining medical data and not supermarket data, we were presented with an additional challenge during the rule-interpretation phase–that of rule validation. The domain knowledge needed for medical validation of the rules was provided by the physicians at the *Moffitt Cancer Center*, Dr. Chris Garrett and Dr. Amit Pathak. All of our findings were checked by these physicians for medical validity. Rule interpretation from a medical standpoint can, at times, be subjective. I spent a lot of time with Dr. Amit Pathak discussing how the discovered rules should be interpreted for our purpose.

Here are some of the issues we were faced with. Figure 3.3 lists all of the 7 valid implications that were discovered from our subset data. We can see that only 2 (#1 and # 5471) of the 7 implications were specific enough because the questions mention the same region of the body in the antecedent and consequent. As seen in Figure 3.3 in implication #5471, both the antecedent and consequent refer to colorectal type cancer; therefore, this rule is tissue-specific and did not require any additional assumptions.

In order for the other 5 implications to be valid, we had to make the assumption that both the antecedent and consequent refer to the same tissue sample. Let's take a closer look at the subset rule #3395 from Figure 3.3:

[Tissue proven pancreas adenocarcinoma?] = [Yes]

IMPLIES (With assumption that the same tissue was examined) $\Rightarrow$

[Histologically confirmed metastatic colorectal cancer?] = [No]

Explanation: The logic behind this is that, if a physician is looking at the pancreatic tissue and it is proven that this tissue is cancerous, then it is also known that this tissue is not from the colorectal region and therefore will not contain a colorectal cancer.

Since the rest of the questions from the subset were somewhat general, the medical decision about them was based on the assumption that the antecedent and consequent refers to the same tissue sample.

3.2.2.6 Medical Validation (Superset)

Now let's look at the superset data (all of the subsequent rule numbers will refer to superset rule numbers). After analyzing the superset data, we found 9 rules of interest (Table 3.12). We see that of the 9 rules there was only 1 true rule (#998) since both the antecedent and consequent of the rule referred to the cardiac region. Rules #1347, #1355, #1356, #1357, and #7076 are tissue dependent rules. Two of these 5 rules contained 2 antecedents; however, upon medical evaluation it was determined that one of the antecedents was not necessary in order for the implication to remain valid (1 ANT not needed). I will give an example of such a case later in this section. Superset rule #1 was a variation of simple rule #1 discovered in our subset. Rules #343 and #7036 had two antecedents and were variations of the subset's simple rule #5471.

Table 3.12 Superset: Discovered Rules

Superset Discovered Rules				
Rule #	Type	#ANT	#CON	Medical Comments
1	Variation of Subset #1	2	1	1 ANT not needed
343	Variation of Subset #5471	2	1	1 ANT not needed
998	True	1	1	Superset
1347	Tissue Dependent	2	1	Specific to General 1 ANT not Needed
1355	Tissue Dependent	1	1	Must Be Same Tissue
1356	Tissue Dependent	1	1	Must Be Same Tissue
1357	Tissue Dependent	1	1	Must Be Same Tissue
7036	Variation of Subset #5471	2	1	1 ANT not needed
7076	Tissue Dependent	2	1	1 ANT not needed

If rule $A \wedge B \Rightarrow C$ was discovered during data mining, it would be incorrect to say that it could be broken down into two rules ($A \Rightarrow C$ and $B \Rightarrow C$) and that these new rules would have the same confidence level as $A \wedge B \Rightarrow C$. If the data could support it, then these two simple rules would have shown up prior to the combination rule $A \wedge B \Rightarrow C$ but only if they had an equal confidence level to $A \wedge B \Rightarrow C$ (we require 1.0). However, in our case, we were able to use medical knowledge to simplify such rules. The example in Table 3.13 demonstrates our reasoning behind doing simplification based on medical knowledge. If a rule A $\wedge$ B $\Rightarrow$ C

was discovered, and after medical evaluation, it was determined that the first antecedent (A) of the rule was not needed, then based on medical knowledge the rule could be reduced to B $\Rightarrow$ C and still maintain the confidence level of 100%. I must emphasize again that such a rule reduction would not be correct if based only on data.

Table 3.13 Example: Rule Simplification Based on Medical Knowledge

RULE: $A \wedge B \Rightarrow C$ A) [Age Over 18]† = [Yes] B) [Sex] = [Male] IMPLIES $\Rightarrow$ C) [Pregnant] = [No]
MEDICAL KNOWLEDGE: Only female can be pregnant. REASONING: Since male can not be pregnant, question A [Age Over 18] is not necessary for the above rule to be valid.
SIMPLIFIED RULE: $B \Rightarrow C$ B) [Sex] = [Male] IMPLIES $\Rightarrow$ C) [Pregnant] = [No]
† - Antecedent Not Necessary

Let's take a closer at look why the simple rule #1 from Figure 3.3 that was discovered during the mining of the subset ended up with 2 antecedents in our superset. Since we used 100% confidence, we only got rules that were correct 100% of the time based on our data. So when the Apriori algorithm was run on the superset with the confidence of 100%, we got our subset rule #1 with an additional antecedent (Sex=Female). However, when the confidence level was dropped to 90%, we got a one antecedent, one consequent superset rule that matched the subset rule #1 exactly. A closer look at the patient data revealed that one of the patients contain the following answers:

A) [Sex] = [MALE]

B) [Path diagnosis of esophageal SCC or ACA?] = [No]

C) [Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [Yes]

Medical examination revealed that if B *[Path diagnosis of esophageal SCC or ACA]* is NO, then C *[Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction]* must also be NO. Since the answer for B was Yes, the antecedent A [Sex=Female] surfaced for the rule. I've already stated that to be medically valid C must be NO, and since it was Yes this case revealed an *error in the data entry.*

We see that by using data mining we could find rules of interest. Because we were dealing with the medical field, we needed to recruit the services of a physician in order to validate our rules of interest. By relying on his or her medical knowledge, the physician is able to trim down a rule by dropping an antecedent if the presence of the antecedent is not medically necessary for the rule to be valid (Table 3.13). Figure 3.6 contains the final result of our data mining from the superset.

Even though we discovered 7 rules in the subset and 9 rules in the superset (two of which were previously discovered in the subset) we can only use the rules that do not rely on any assumptions. We call such rules *true rules.* By listing only the true rules, which are marked with ($\diamond$) and shown in Figure 3.6, we end up with our final list of five (5) implications. Our final list of implications is seen in Figure 3.7. These implication have been entered into the implication module of MEANS.

3.2.3 Probability-Based Reordering

MEANS uses analytical and probabilistic agents to reorder the list of unanswered questions (questions to ask). The Analytic heuristic is based on the cost of medical tests together with the structure of the acceptance and rejection expressions. The probability-based reordering is done with the help of the probabilistic agent. The probabilistic agent uses data that was accumulated over time to calculate the probability for a question. In this section, I will discuss the probability-based heuristic in detail.

Bayesian Networks seem to be preferred for probabilistic expert systems; however, there are some drawbacks to using them. They can be very complex and computationally expensive.

Superset: All Medically Valid Implications
#1:◇ ‡ [Sex]† = [Female] AND [Path diagnosis of esophageal SCC or ACA?] = [No] IMPLIES ⇒ [Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [No]
#998:◇ [Prior cardiac condition in the last 6 months?] = [No] IMPLIES ⇒ [Unstable angina?] = [No]
#1355: [Tissue proven pancreas adenocarcinoma?] = [Yes] IMPLIES (if same tissue sample) ⇒ [Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [No]
#1356: [Tissue proven pancreas adenocarcinoma?] = [Yes] IMPLIES (if same tissue sample) ⇒ [Histologically confirmed metastatic colorectal cancer?] = [No]
#1357: [Tissue proven pancreas adenocarcinoma?] = [Yes] IMPLIES (if same tissue sample) ⇒ [Path diagnosis of esophageal SCC or ACA?] = [No]
#343: [Take medications by mouth?]† = [Yes] AND [Histologically confirmed metastatic colorectal cancer?] = [No] IMPLIES (if same tissue sample) ⇒ [Metastatic colon or rectal cancer (tissue proven) AND not suitable for surgery?] = [No]
#1347:◇ [Histologically confirmed metastatic colorectal cancer?]† = [No] AND [Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [Yes] IMPLIES ⇒ [Path diagnosis of esophageal SCC or ACA?] = [Yes]
#7036:◇ ‡ [Metastatic colon or rectal cancer (tissue proven) and not suitable for surgery?] = [Yes] AND [Uncontrolled medical conditions?]† = [No] IMPLIES ⇒ [Histologically confirmed metastatic colorectal cancer?] = [Yes]
#7076:◇ [Histologically confirmed metastatic colorectal cancer?] = [Yes] AND [Prior cardiac condition in the last 6 months?]† = [No] IMPLIES ⇒ [Measurable disease (RECIST)?] = [Yes]
† Antecedent NOT Necessary ‡ Rule Previously Found in Subset ◇ True rule
TOTAL: 9 implications

Figure 3.6 Superset: All Medically Valid Implications

Final List of Discovered Association Rules	
True Rules:	
1	[Path diagnosis of esophageal SCC or ACA?] = [No] IMPLIES ⇒ [Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [No] Superset Rule #1 Subset Rule #1
2	[Metastatic colon or rectal cancer (tissue proven) and not suitable for surgery?] = [Yes] IMPLIES ⇒ [Histologically confirmed metastatic colorectal cancer?] = [Yes] Superset Rule #7036 Subset Rule #5471
3	[Prior cardiac condition in the last 6 months?] = [No] IMPLIES ⇒ [Unstable angina?] = [No] Superset Rule #998
4	[Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction] = [Yes] IMPLIES ⇒ [Path diagnosis of esophageal SCC or ACA?] = [Yes] Superset Rule #1347
5	[Histologically confirmed metastatic colorectal cancer?] = [Yes] IMPLIES ⇒ [Measurable disease (RECIST)?] = [Yes] Superset Rule #7076

Figure 3.7 Final Rules

Introduction of new evidence requires recalculation of corresponding numerical probabilities and may require modification to an existing network [8, 23]. However, one of the greatest advantages of Bayesian Networks is that they are able to predict an outcome even with absence of some information. This ability gives Bayesian Networks an advantage over rule based systems. Due to this feature, the naïve Bayes approach was adapted for the probabilistic agent.

Every time a new set of questions and answers is submitted, the answer to each question is examined and MEANS makes a determination of the patient's eligibility on a per protocol basis. If, after the evaluation of all newly-submitted answers the eligibility of a patient can not be determined, MEANS compiles a list of questions to ask and presents the user with the top 10 questions from that list. This cycle continues until the eligibility of the patient is determined for all selected protocols.

We wanted to make the screening process less time-consuming. If a patient is eligible for a trial, then all of the questions in the acceptance expression need to be answered; however, only a small set of questions is required for a rejection expression to be TRUE, thus making a patient ineligible. One way of decreasing the number of questions that needed to be answered was to try to make a patient ineligible as soon as possible during the screening process. We needed to present the user with the questions that would most likely to make a patient ineligible. This was possible with the use of a probability-based heuristic.

By using a probability-based heuristic we could reorder questions such that the questions with a higher probability-influenced ranking value of determining a patient's ineligibility for a protocol would be closer to the top of the list of questions to ask. Since a question with a higher ranking of making the patient ineligible for a protocol would be asked first, such an ordering had the potential to reduce the number of questions that a user needed to answer.

3.2.3.1 Probability-Guided Agent

The original version of the probability-guided agent was developed by B. Goswami [12]. We modified the original implementation of data collection and introduced a threshholding feature during probability calculation.

Even though it is not entirely true, in terms of our research, we use class conditional independence (an assumption that all of the questions are independent of each other). We eliminated known dependencies by using implications; therefore, the dependent (implied) questions were not in the list of questions to ask. The screening process could be viewed as a 2-class problem: *Eligible* and *Ineligible*. We treated the questions of the protocols as our attributes and the values of the questions as an eligibility status: either *Favorable to be Eligible* or *Favorable to be Ruled Out*. To use naïve Bayes, we needed know the probabilities of occurrences of each class type. This was accomplished every time a question was evaluated by the system.

During the question evaluation phase, we used the following heuristic to keep track of asked questions and the outcome. Each question had two numbers attached to it together with a protocol number. The first number kept track of how many times a question ruled out a patient (made a patient ineligible for a protocol); the second number was the number of times a question made a patient eligible for a protocol (or did not rule out a patient). Every time an answer was evaluated for a protocol, one of the numbers would increase. If, the eligibility of a patient could not be determined even after processing all of the newly submitted answers, a new list of questions to ask was compiled.

When a patient was screened, the user had a choice of either answering a question, leaving it blank or deferring the question for later. At this time the system does not capture which question was answered first on the page or the order of submitted data on a per-page basis; therefore, we were unable to determine if the question that was answered was the first question on the list of presented questions. This could be implemented in the future versions of the system by adding a JavaScript to track the order in which the questions were answered on a page and capture the data during processing.

We were interested in calculating the likelihood probability that the question would make a patient ineligible (*ruled out*) for a protocol. The calculation of the probability was based on the Bayes' rule of conditional probability (Equation 3.1). Pr[A] is the probability of the event A and Pr[A|B] is the probability of the event A conditional to another event B. The evidence

that consists of a particular combination of attribute values is denoted by E [29]:

$$Pr[H|E] = \frac{Pr[E|H] \quad Pr[H]}{Pr[E]} \tag{3.1}$$

In our case, the probability that a question would rule out a patient from a protocol was calculated as follows:

$$Pr[Ineligible|E] = \frac{Pr[Ineligible]}{Pr[E]}$$

Example: When we collected the statistics for a question *Q1*, the line that contained the question was read as *Q1 13424 10 80*. This could be interpreted as: Question *Q1* for protocol *13424*, which was asked 90 times, ruled out a patient (made ineligible) *10* times, and did not rule out a patient (left eligible) *80* times. If we denote A for the number of times a patient was ruled out and B for the number of times a patient was eligible, then we have:

$$Pr[rule \quad out|E] = \frac{\frac{A}{(A+B)}}{\frac{A}{(A+B)} + \frac{B}{(A+B)}}$$

which in our case is simplified to:

$$Pr[rule \quad out|E] = \frac{A}{(A+B)} \tag{3.2}$$

By substituting the numbers from our example into Equation 3.2, to calculate the probability that the question *Q1* is likely to rule out a patient from the protocol *13424* would be:

$$Pr[rule \quad out|E] = \frac{A}{(A+B)} = \frac{10}{(10+80)} = 0.11$$

It is not uncommon for some protocols to share the same questions, like age and sex. When a patient is being tested for more than one protocol, a question may appear in the acceptance criteria in each of the protocols. To take this into consideration, we have chosen to use a *probability-influenced ranking value* for our final ordering of the questions. For a

question that appears in more than one protocol, the probability-influenced ranking value is calculated by taking an average of individual probabilities and multiplying it by the number of trials in which the question appears.

If we have question Q1 in three protocols and each individual probability for question Q1 has been already calculated and is:

$$P(Q1_{13424}) = .11 \quad P(Q1_{13426}) = .72 \quad P(Q1_{13946}) = .51$$

then the probability-influenced ranking value will be calculated as follows:

$$Ranking \quad Value = \frac{(Q1_{13424}) + (Q1_{13426}) + (Q1_{13946})}{3} \times 3 =$$

$$\frac{(.11) + (.72) + (.51)}{3} \times 3 = 1.34$$

The resulting probability-influenced ranking value will be between zero and the number of trials in which a question appears. If a question appears only in one trial, then the probability-influenced ranking value is equal to the real probability of the question. However, if a question appears in more than one trial, the probability-influenced ranking value will give an advantage to such a question because it the has potential of ruling out a patient from multiple trials.

Other ways of calculating a probability-influenced ranking value would be to use a weighted average of the individual probabilities or only use the maximum probability of a question. If we use the maximum probability of a question for a trial, that will give the rule-out advantage to the trial from which the probability was used; however, that may not give the overall advantage that we are seeking. Such heuristics still need to be explored.

When the probability-based heuristic is used, the list of questions to ask is ordered with the highest rule-out question ranking at the top.

In order to have a meaningful probability for a question, the number of times a question should have been asked must be sufficient, that way the questions presented first will have a greater impact. Let's say that a question *Q2* contains some rare condition and was asked a total of 9 times, and ruled out a patient only twice. The probability for this question is

0.22. If we compared the probabilities for *Q1* and *Q2*, the rule-out ranking for *Q2* is greater; therefore, it will be closer to the top of the top of the list of questions to ask. If *Q2* did not rule out any more patients, after some time the rule-out ranking would decrease and *Q1* would move closer to the top of the list; however, this requires additional data entry.

Before using equation 3.2 to calculate the probability on a per-protocol basis, we used a threshold to evaluate the total number of times a question had been asked. If that number was above the given threshold, then we calculated the probability for a question. Otherwise, we assigned the question a probability of zero and did not use the it to reorder.

3.2.3.2 Testing System

To test the probability-guided method we developed a testing system. The testing system, which was written in C and is on the enclosed CD, was able to screen the patient by submitting existing answers from patient's data one question at a time. After an answer was evaluated and the questions-to-ask list compiled, the testing system attempted to find an answer for the question from the list of existing answers of a patient. If the answer could not be found, the system would increase the number of asked questions, and would try to find an answer for the next question on the list. When an answer for a question was found, the number of answered questions increased and the answer was submitted into the system for evaluation. If a patient's eligibility for a protocol was determined eligible or ineligible, no other questions from that protocol would appear in the list. If the eligibility for all protocols was determined, the questions-to-ask list would then be blank and the testing program terminated. The testing system algorithm is shown in Figure 3.8.

We have created two versions of the testing system: web-based and command prompt-based. The web-based test system permits screening a patient without logging into the server. We can test one or two patients and view the results via the browser. However, the command prompt version is more robust if screening a list of patients.

The command-prompt version takes in one parameter, which is the MEANS ID of the existing patient. During our experiments, the command-prompt testing system was called via shell scripts.

```
1:  for (Each testing patient)
2:  {
3:     Load default answers
4:     Load first page answers
5:     Compile questions-to-ask list by using ranking to reorder
6:     while (there are questions to ask)
7:     {
8:           Take top question from questions-to-ask list (ordered by ranking value)
9:           Try to find the answer in patient's profile
10:          if (answer not found)
11:          {
12:                while (answer not found AND there are questions to ask)
13:                {
14:                      Increment number of asked questions
15:                      Take next question from questions to ask list
16:                      Try to find the answer in patient's profile
17:                }
18:          }
19:          Increment number of answered questions
20:          if (answer found)
21:          {
22:                Evaluate patient's eligibility
23:                Compile questions to ask list
24:          }
25:    }
26: Record Patient's statistics
27: }
```

Figure 3.8 Testing System Algorithm

3.2.3.3 Probability Guided Experiments

For our experiment, we chose 100 ineligible patients. These were the ineligible patients from the latter part of the data-collection phase. First, each of the 100 test patients were evaluated by the testing system and a number of asked questions and answered questions were tracked together with the statistics for each question. Next, the same 100 patients were evaluated by testing system again, except this time the questions-to-ask list was sorted by probability-influenced ranking value with the highest rule-out ranking value at the top of the list.

I should reiterate that the testing system screens a patient one question at a time. A question is considered *presented* or *asked* when a question is at the top of the questions-to-ask list and the testing system is done searching for the answer in a test patient's profile.

The results of the analytical reordering experiment are shown in Table 3.14. From the *Questions Answered* column we can see that, on average, the system required 7.89 questions in order to determine a patient's eligibility. The *Less Needed* column shows that, compared to the test profile, the number of questions that were answered during data collection but were not needed for eligibility determination decreased by an average of 4.46 questions per patient. The *Questions Asked* column refers to the number of questions that were presented (searched-for answers in a test patient) before eligibility was determined. The average number of questions that was presented was 86.7.

Table 3.14 Analytical Reordering

	Questions Asked	Questions Answered	Questions in Test Profile	Questions in Original Profile	Less Needed
Total ($\sum$)	8,670	789	1,787	2,233	446
Mean ($\bar{x}$)	86.7	7.89	17.87	22.33	4.46
Median (Md)	106	8	18	19	2.5
Mode (Mo)	106	8	18	19	1
St. Dev (σ)	52.17	5.19	5.23	8.75	5.73

To compare how the analytical reordering stacked up against the probability-based heuristic, we used the 10-fold cross validation method. We divided our 100 patients into 10 equal sets of 10 patients per set. We gathered the probability of the 9 sets and then used that prob-

ability on the 10th set that was left out during the probability gathering. This was repeated 10 times, once per set.

Table 3.15 Probability-Based Reordering (10-Fold Cross Validation)

	Questions Asked	Questions Answered	Questions in Test Profile	Questions in Original Profile	Less Needed
Total ($\sum$)	1,887	511	1,509	2,233	724
Mean ($\bar{x}$)	18.87	5.11	15.09	22.33	7.24
Median (Md)	12	4	14	19	5
Mode (Mo)	11	4	14	19	5
St. Dev (σ)	23.13	3.65	3.69	8.75	6.84

The comparison between Table 3.14 and Table 3.15 indicates that using the probability-based heuristic to reorder the questions-to-ask list decreased the number of questions that were needed to rule out a patient from clinical trials. By looking at the mean ($\bar{x}$) we can see that the average total questions asked for eligibility determination (*Questions Asked*) decreased from 86.7 to 18.87. That is a drop of 67.83, which is quite a significant number. The average number of questions that were answered (*Questions Answered*) was 5.11, a decrease of 2.78.

I must note that during the initial screening of the patients, clinicians had a choice of answering any number of the 10 questions on the page in any order before submitting the answers into the system. Since the system compiled the questions-to-ask list only after evaluation of all of the newly-submitted answers, the number of asked questions will differ when multiple answers were submitted at a time instead of one answer at a time.

As I mentioned before, we were also interested in finding out what happens if we threshold the number of times a question was asked before calculating probability for the question. For our next experiment, we used the same 100 ineligible patients and ran the 10-fold cross validation while varying the threshold value before calculating the probability for each question.

First we ran the 10-fold cross validation with a threshold (T) of zero. Then we ran the 10-fold cross validation for each threshold value starting at $T = 10$ until $T = 120$ while varying the threshold by 5. As seen in Table 3.16, the number of questions asked when T increased from 0 to 10 decreased by an average of 8.03. The lowest average of questions asked was 8.60,

which occurred at $T = 55$. This can be seen better in the graph in Figure 3.9. We were interested in the curve with the diamond-shaped data points.

Table 3.16 Probabilistic Thresholding (10-Fold Cross Validation)

Threshold Value	Questions Asked	Questions Answered	Questions in Test Profile	Questions in Original Profile	Less Needed
0	17.11	5.02	15.00	22.01	7.01
10	9.08	5.54	15.52	22.01	6.49
15	9.09	5.77	15.75	22.01	6.26
20	9.02	5.74	15.72	22.01	6.29
25	8.87	5.71	15.69	22.01	6.32
30	8.87	5.71	15.69	22.01	6.32
35	8.86	5.71	15.69	22.01	6.32
40	8.77	5.71	15.69	22.01	6.32
45	8.62	5.71	15.69	22.01	6.32
50	8.61	5.71	15.69	22.01	6.32
55	8.60	5.71	15.69	22.01	6.32
60	8.68	5.64	15.62	22.01	6.39
65	8.68	5.64	15.62	22.01	6.39
70	8.68	5.64	15.62	22.01	6.39
75	8.68	5.64	15.62	22.01	6.39
80	8.68	5.64	15.62	22.01	6.39
85	14.24	5.80	15.78	22.01	6.23
90	23.30	6.44	16.42	22.01	5.59
95	23.30	6.44	16.42	22.01	5.59
100	23.28	6.44	16.42	22.01	5.59
105	23.61	6.46	16.44	22.01	5.57
110	23.98	6.60	16.58	22.01	5.43
115	24.36	6.55	16.53	22.01	5.48
120	24.36	6.55	16.53	22.01	5.48

In Figure 3.9 we can see that as T increased from zero, the average number of questions that needed to be asked in order to rule out a patient decreased. However, when T approached 85, the curve started going up, with a noticeable jump at T=90. This can be explained by looking at Table 3.17, which lists for each set the maximum number of times a question was asked for a protocol. We can see that the average maximum number of times a question was asked

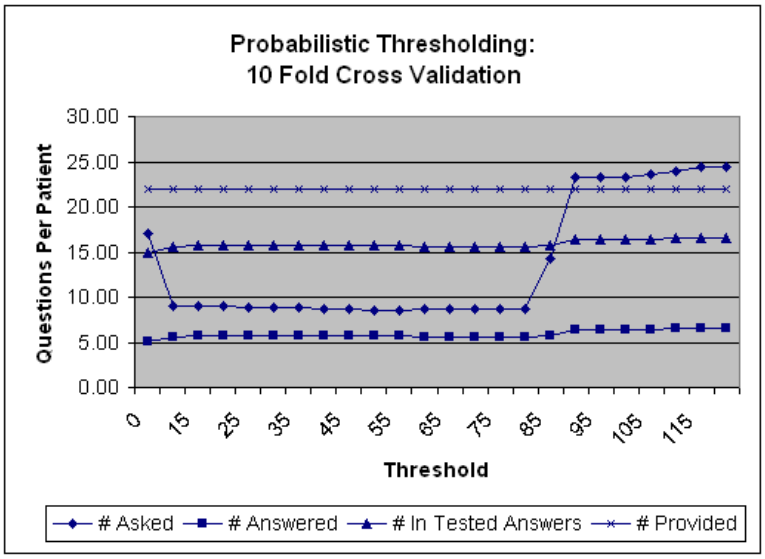


Figure 3.9 Thresholding (10-Fold Cross Validation)

for a particular protocol was 88. So, when a threshold was set too high, the probability-based reordering was not in effect and the questions were presented at random.

Table 3.17 Max Number Per Set

Max Number of Times a Question for a Protocol Was Asked									
Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10
87	87	88	88	88	88	88	88	88	90
Average: 88									

When sorting by probability-influenced ranking value, it is useful to know when to pick a starting point in order to insure that the probability calculation for a question will be significant enough. Counting the number of patients that have been screened is not sufficient. Since we were reordering the questions, we needed to use information about each individual question. We showed that by thresholding the number of times a question was asked before computing the probability for a question, it decreased the number of questions asked, compared to non-thresholding. By decreasing the number of questions needed to rule out a patient we decreased the amount of data entry required and possibly decreased the amount of time spent on the screening process.

3.2.4 Data Entry Optimization

During the patient data collection phase, we received continuous feedback from physicians and nurses at the *Moffitt Cancer Center* GI Clinic. To make MEANS more user-friendly, we implemented the clinicians' suggestions.

One of the concerns was the amount of mouse clicks that clinicians needed to make during the selection of protocols. Based on that, we implemented a protocol bypass option. That option, which is preselected, automatically selects all of the protocols for which a patient will be screened. A user has the option of deselecting the automatic selection of protocols if he or she would like to screen a patient only for a subset of active protocols. If a user decides to deselect it, then the user may manually pick which protocols to screen a patient for. At least one of the protocols must be selected in order for the system to proceed.

Another optimization measure, which was implemented based on the suggestions from the physicians, was the entry of normal lab values. Physicians are only interested if a patient has abnormal lab values. If the lab values of a patient are all within the normal range, then the actual lab values are not of particular interest. The same can be said about MEANS' screening process. As long as the lab values are within the normal range, a patient will be considered eligible. Based on that, we implemented a *normal* checkbox for each question that requires a lab value. Another related option – *All Labs Normal* – was also requested by the physicians. Now, when a physician checks a patient's labs and is able to determine that all of the lab values fall within the normal range, a physician can check the *All Labs Normal* checkbox and the system will automatically enter all of the predefined lab values into that patient's profile. This can be seen in Figure 3.10. Such measures significantly reduces the number of mouse clicks that a physician must make during the screening process. Implementation of the above-mentioned changes also had impact on the system flow.

We also found that there was a significant difference between the verbiage used by the physicians and the legal language in which the protocol inclusion/exclusion criteria was written. The legal version can sometimes be much longer, as seen in Table 3.18. To conserve a physician's time, only the shorter version of a question was displayed on the screen. The full

version of the question was viewable by placing the mouse cursor over the question or clicking on the *Full Text* link.

Table 3.18 Yes/No Question Length

Full Text
Does the patient have histological or cytological confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction? (This refers to tumors at the junction of the esophagus and the stomach, where > 50% of the tumor mass is above the diaphragm)
Short Version
Tissue confirmed esophageal adenocarcinoma or adenocarcinoma of the gastroesophageal junction?

Automatic selection of all encoded clinical trials and the *All Labs Normal* option were introduced based on feedback from the physicians. These changes are reflected in Figure 3.10.

3.2.5 Interface Changes

To improve usability of the system, a variety of changes were made to the interface of the system. The initial questions page is shown in Figure 3.10. We can see that beside each question there are three choices. If a question is a numeric question, a physician can enter the actual value, select *Normal*, and have the system enter a value, or select *Defer* to delay answering the question for later.

By clicking on the *Full Text* link a physician can see the full version of any question. An example of a full-text pop-up can be seen in Figure 3.11.

To help physicians see when a patient becomes eligible, we added a table that lists the status of all of the trials during the screening process. When a patient becomes eligible for a trial, the color of the eligibility slot of the table changes to green. As seen in Figure 3.12, the first cell of the table is highlighted green, signifying that the patient with ID 105 is eligible for a trial.

Most of the physicians do not know the clinical trials by their protocol number because they refer to them by the name of the study. Because of this, we implemented a pop-up that includes both the protocol number and the full title of the clinical trial. By clicking on the

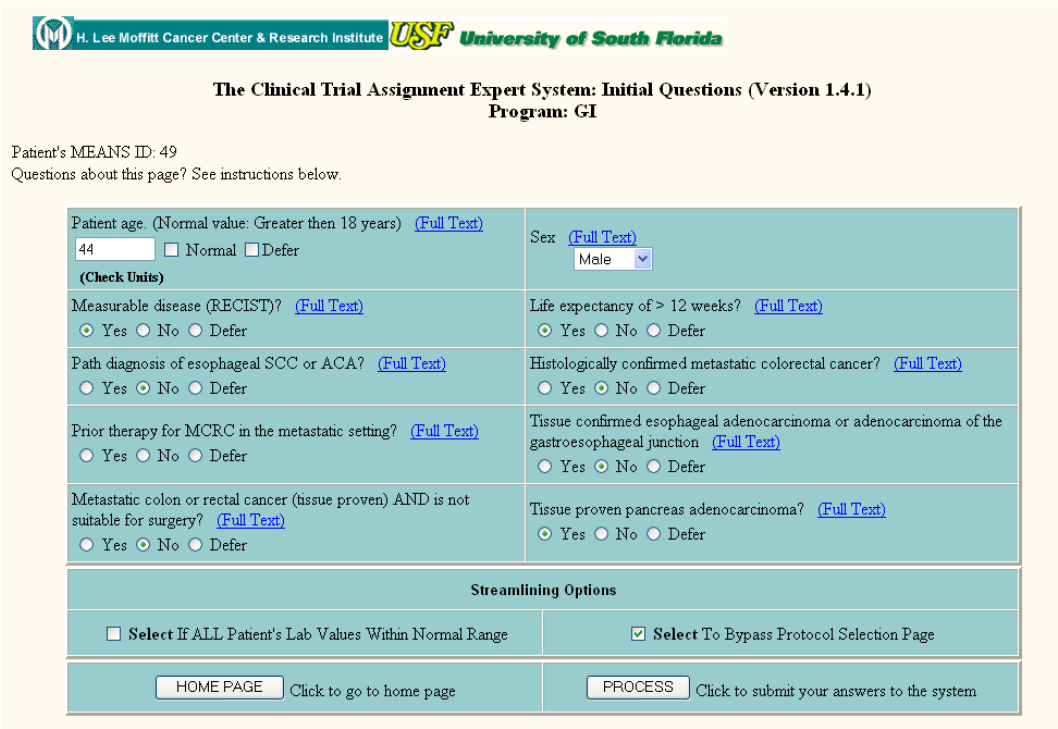


Figure 3.10 MEANS Streamlined Initial Questions Page

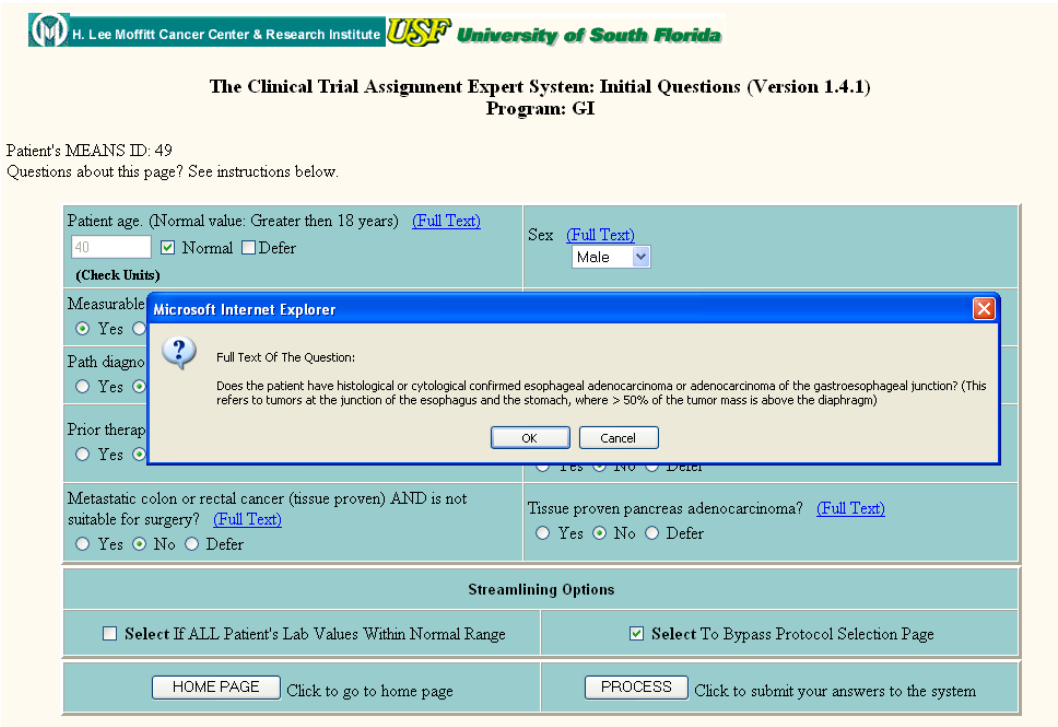


Figure 3.11 Full Text Popup

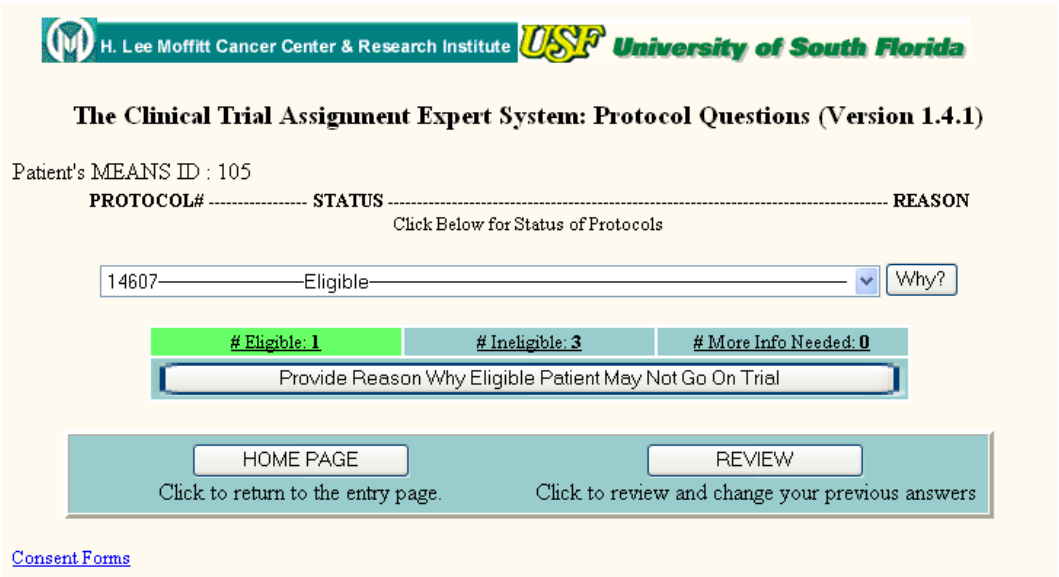


Figure 3.12 Eligibility Color Table

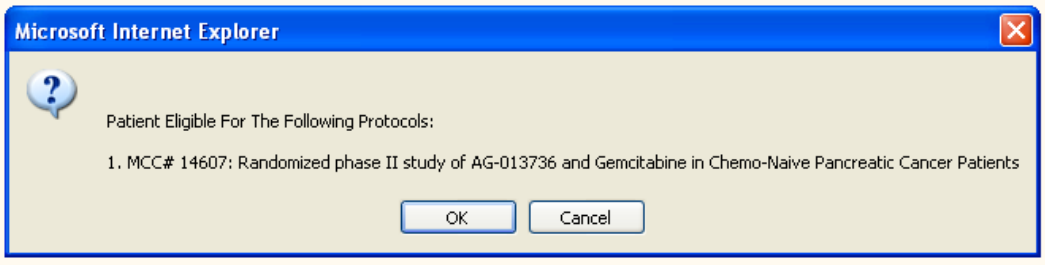


Figure 3.13 Protocol Number and Title

#Eligible link, a pop-up with the explanation will appear (Firgure 3.13). We received positive feedback on the changes that we have implemented.

CHAPTER 4

CONCLUSION

4.1 Implication Discoveries

We used data mining on acquired medical data to flag implications for further physician evaluation. As a result of our efforts, 14 rules were discovered; however, only 5 of these rules were true rules and could be entered into MEANS. These 5 implications were added to MEANS' implication module.

I should note that one reason we did not find more was because of the way the initial questions in MEANS were set up. Because we wanted to identify ineligible patients as early as possible during the screening process, the initial questions page in MEANS contained a set of questions that were most-likely to make a patient ineligible for a protocol. The initial page usually contained two key questions from each protocol, and each question was specifically chosen by a physician to most-likely rule out a patient from a protocol. The majority of the patients were ruled out after the questions on the first page were answered. The patients that were not ruled out after the first page were ruled out soon after, unless they were eligible for a protocol. Since the questions on the first page were not medically related, the number of rules that could be medically valid decreased tremendously.

We found that by using Apriori we could find many rules. We could isolate our rules of interest from the rest of the rules by looking at the number of antecedents and consequents. Our experiments showed that when we were using a confidence level of 100%, it was beneficial to not only look at one antecedent, one consequent rules but also include two antecedents, one consequent rules in our rules-of-interest pool. With the help of physicians, we could determine whether a rule with two antecedents could be simplified by dropping one of the antecedents. Such a simplification was only possible with the support of medical knowledge.

I must also note that some of the discovered rules were medically validated based on certain assumptions. Because some of the questions did not specify the region of the body to which they were referring, we assumed that the reference was either to the same section of the body or both the antecedent and consequent of the rule were referring to the same tissue. The data supported our assumption; however, since the rules could only be validated based upon the assumption, it was not a good idea to include such rules in our implication subsystem. Due to this, the rules that required any assumptions, were not included in our final list.

Many decisions in medicine can be subjective; therefore, two physicians may evaluate the same condition differently. This presents another challenge for medical validation of newly-discovered rules.

4.2 Probability-Based Reordering

We were interested in minimizing the amount of data entry needed to determine a patient's eligibility for a clinical trial (protocol). If a patient is eligible, it is obvious that, all of the questions must be answered. However, if a patient is ineligible as a result of the screening, it would be beneficial to ask the question that made the patient ineligible at the beginning of the screening process.

We were interested in calculating the likelihood probability that a question would make a patient ineligible (*ruled out*) for a protocol. We based our calculation of probability on the Bayes' rule of conditional probability (Equation 3.1). By calculating probability for each question and presenting the user with a question that has a higher rule-out probability-influenced ranking value, we showed that the average number of questions that were presented (*Questions Asked*) by the test system decreased from 86.7 to 18.87 and that the average number of questions that needed to be answered before eligibility was determined decreased from 7.89 to 5.11.

We also showed that by thresholding the number of times a question was asked, it was possible to decrease the amount of questions asked during the screening process even more. By thresholding during the probability calculation, the average number of questions asked by the system was the lowest when the threshold was between 45 and 80; however, the

optimal threshold was at 55. At $T = 55$ the average number of questions asked by the test system to determine eligibility was 8.60 and the average number of answers required for eligibility determination was 5.71. Compared to analytical and non-thresholding probability-based heuristics, this is a significant reduction in the number of questions that were asked (presented) and answered in order to determine eligibility. At this time the probabilistic thresholding is only being implemented in test systems.

Since the probability was gathered every time a question was answered, we could set the threshold value and, as soon as the threshold value was reached, the probability-based reordering for that question would come into play. If the question was below the threshold value, probability-based reordering would have no effect on the question ranking and would be placed in the lower part of the list of questions to ask after all of the questions with a rule-out ranking value. As time goes by, our system "learns" from submitted answers and becomes "smarter" over time.

4.3 Optimization

Physicians are very reluctant to use new software, which makes it difficult to introduce systems to physicians. By using the method participatory design, we used the feedback that was given to us by the clinicians during the patient screening phase to streamline our system [17]. The feedback that was received during the fielding of the system was invaluable to the overall success of MEANS. If the system is not accepted by the users, no matter how well it functions or how great it is, no one will use it and, as a result, it will fail.

We found that the success of the acceptance of MEANS in a clinical environment depends heavily on the amount of time that a physician will spend screening a patient. Physicians are very busy and if the system takes a long time to learn and use, it is less likely that it will be utilized. We used medical knowledge to select the questions for display on the initial page because most of the rule-out questions were based on cancer site.

Based on the feedback, we modified the system to decrease the amount of time a physician spends reading the questions. We also decreased number of mouse clicks required. With one mouse click a physician can mark all of the lab values as *Normal*. Since we automated the

selection of the protocol for physician, no mouse clicks are required to select all protocols for screening. During the screening process, optimization occurs prior to the probability-based reordering.

4.4 Future Work

Because of HIPAA regulations, we were unable to interface MEANS with the the electronic medical records system at *H. Lee Moffitt Cancer Center & Research Institute*. Therefore, every time a new patient was screened, all of the patient's information had to be entered manually. Because of this, the amount of information available for data mining was very limited. It would be a great advantage to have some kind of interface where information from medical records can be imported into MEANS. This would significantly increase the amount of data available for data mining. It would also decrease the amount of data entry required for screening patients.

We did not explore this option, but it is possible to automate the entire process of data extraction from MEANS, running the Apriori algorithm on the data, rule recovery, and rule filtering based on the number of antecedents in the rule. It would be interesting to see, with the addition of new data, what rules could be recovered on a weekly basis.

Mining medical data presents a challenge all its own because of the nature of the medical domain. Recovering rules based on medical data has proved to be a difficult task. Many questions in the protocols are not specific enough, which in turn presents a gray area when trying to use medical knowledge to validate a rule. It would be of an advantage to have clearer wording in the inclusion/exclusion criteria of the protocols. This would help alleviate confusion if questions are referring to the same tissue.

We only explored the Bayes' rule of conditional probability with the thresholding option. The other probabilistic methods should also be explored to see if they may be more effective in reordering questions for rule-out probability. We used a 10-fold cross validation of 100 patients to conduct our experiments. It would be interesting to see how the system behaves with a larger number of patients. If the number of patients increases, it may be possible that a different threshold value would perform better.

REFERENCES

[1] R. Agrawal, T. Imielin'ski, and A. Swami. Mining association rules between sets of items in large databases. pages 207–216. Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. Proceedings of the 20th Very Large Databases Conference Santiago, Chilie, Sept. 1994.

[3] J. S. Aikins, J. C. Kunz, E. H. Shortliffe, and R. J. Fallat. Puff: An expert system for interpretation of pulmonary function data. *Computers and Biomedical Research*, 16:199–208, July 1982.

[4] American Cancer Society. Cancer facts and figures 2005. www.cancer.org, 2005.

[5] S. Bhanja, L. M. Fletcher-Heath, L. O. Hall, D. B. Goldgof, and J. P. Krischer. A qualitative expert system for clinical trial assignment. pages 84–88. Proceedings of the Eleventh International Florida Artificial Intelligence Research Society Conference, 1998.

[6] British Broadcasting Corporation. Historic figures, http://www.bbc.co.uk/history/historic_figures, 2006.

[7] W. J. Clancey and R. Letsinger. Neomycin: Reconfiguring a rule-based expert system for application to teaching. In *IJCAI*, pages 829–836, 1981.

[8] F. J. Díez, J. Mira, E. Iturralde, and S. Zubillaga. DIAVAL, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*, 10:59–73, 1997.

[9] E. Fink, L. O. Hall, D. B. Goldgof, J. P. Krisher, B. Goswami, and M. Boonstra. Experiments on the automated selection of patients for clinical trials. volume 5, pages 4541 – 4545. IEEE, Systems, Man and Cybernetics, 2003. IEEE International Conference on, Oct 2003.

[10] P. Fule and J. F. Roddick. Detecting privacy and ethical sensitivity in data mining result. volume 26. Proceedings of the Twenty-Seventh Autralian Computer Science Conference (ACSC2004), Australian Computer Society, Inc., January 2004.

[11] B. Goethals. *In The Data Mining and Knowledge Discovery Handbook*, chapter 17, pages 377–397. Springer, 2005.

[12] B. D. Goswami. Optimizing cost and data entry for assignment of patients to clinical trial using analytical and probabilistic web-based agents. Master's thesis, University of South Florida, 4202 E Fowler Avenue, Tampa, FL 33620, Nov 2003.

[13] B. D. Goswami, L. O. Hall, D. B. Goldgof, E. Fink, and J. P. Krisher. Using probabilistic methods to optimize data entry in accrual of patients to clinical trials. pages 434 – 438. Computer-Based Medical Systems, 2004. CBMS 2004. Proceedings. 17th IEEE Symposium on, June 2004.

[14] M. Hagland. Stronger computer tools allow deeper analysis of medical research, patient care and insurance data. *Healthcare Informatics*, page 33, April 2004.

[15] H. Lee Moffitt Cancer Center & Research Institute. Prevention and treatment: Clinical trials., April 2006.

[16] National Cancer Institute. Dictionary of cancer terms - clinical study (http://www.cancer.gov), May 2006.

[17] M. R. John H. Gennari. Participatory design and an eligibility screening tool. In *In Proceedings of the American Medical Informatics Association Annual Fall Symposium*, pages 290–294, 2000.

[18] P. K. Kokku, L. O. Hall, D. B. Goldgof, E. Fink, and J. P. Krischer. A cost-effective agent for clinical trial assignment. volume 1, pages 60–65. IEEE, Oct 2002.

[19] J. Lederberg. Proceedings of the acm conference on history of medical informatics, bethesda, maryland, usa, november 5-6, 1987. In B. I. Blum, editor, *History of Medical Informatics*. ACM, 1987.

[20] MedicineNet.com Diagnosis information about clinical trials. www.MedicineNet.com, 1996-2001.

[21] S. Nikiforou. Selection of clinical trials: Knowldege representation and acqusition. Master's thesis, University of South Florida, 4202 E Fowler Avenue, Tampa, FL 33620, May 2002.

[22] National Institute of Health. Hipaa privacy rule and its impact on research (http://privacyruleandresearch.nih.gov/), May 2006.

[23] C. Papaconstantinou, G. Theocharous, and S. Mahadevan. An expert system for assigning patients into clinical trials based on bayesian. *Journal of Medical Systems*, 22(3):189–202, 1998.

[24] J. F. Roddick, P. Fule, and W. J. Graco. Exploratory medical knowledge discovery: experiences and issues. *ACM SIGKDD Explorations Newsletter*, 5(1):94–99, July 2003.

[25] E. H. Shortliffe, S. G. Axline, B. G. Buchanan, T. C. Merigan, and S. N. Cohen. An artificial intelligence program to advise physicians regarding antimicrobial therapy. *Computers and Biomedical Research*, 6:544–560, 1973.

[26] E. H. Shortliffe, B. G. Buchanon, and E. A. Feigenbaum. Knowledge engineering for medical decision making: A review of computer-based clinical decision aids. In *Proceedings of the IEEE*, volume 67, pages 1207 – 1224. IEEE, September 1979.

[27] T. Shortliffe and R. Davis. Some considerations for the implementation of knowledge-based expert systems: The mycin project. In *AIM Workshop*, pages 9–12. SIGART Newsletter, December 1975.

[28] National Research Council (U.S.). *Funding a Revolution: Government Support for Computing Research*. National Academy Press, Washington, D.C., 1999.

[29] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition edition, 2005.

[30] H. Yaho and H. J. Hamilton. Mining itemset utilities from transaction databases. Data and Knowledge Engineering, Accepted October 2005, 2005.