# Mining for Implications in Medical Data

Cindy L. Bethel and Lawrence O. Hall and Dmitry Goldgof
Department of Computer Science and Eng. ENB118
University of South Florida
Tampa, FL 33620
USA
{hall,goldgof, clbethel}@csee.usf.edu

## Abstract

*Accruing patients for clinical trials has been a tedious and time consuming task for clinicians. It requires extensive knowledge of the specific criteria for all available clinical trials. Through interviews with clinicians, implications were discovered which reduced the number of required questions/answers to determine eligibility. After gathering and recording data on past breast cancer patients, the answers to the questions asked by an expert system were extracted. An association rule learner, was used to generate implication rules such as: male => not pregnant. It was determined that all current implication rules could be recovered with 100% confidence. Further searching for additional rules resulted in the discovery of several which provided an improvement in the clinical ease of use of the web-based Clinical Trial Assignment Expert System.*

## 1 Introduction

Cancer is a disease that can affect people from all walks of life. The focus of this research was on identifying patients eligible for breast cancer clinical trials. In 2004 it was estimated that there will be 215,990 new female and 1,450 new male cases of invasive breast cancer diagnosed [2]. Additionally, it is estimated that there will be another 59,390 cases of *in situ* breast cancer [2]. In 2004, it is expected that 40,110 women and 470 men will die from this disease [2]. Therefore, it is important to match patients to clinical trials so that new treatments can be adequately evaluated in a timely manner.

Clinical trials provide a method for evaluating the effectiveness and safety of new cancer treatments on human subjects. A continual problem is obtaining adequate participation to sufficiently test these treatments. Enrollment in clinical trials is a time consuming and tedious process for clinicians. It requires them to have extensive knowledge of the protocol treatment(s) and the inclusion/exclusion criteria for participation. The clinician must match this information with their current patient base to determine if there are any matches. Currently, the matching process is performed manually by clinicians. Studies have indicated that up to 60% of eligible patients do not participate in clinical trials, which in turn leads to significant delays in the evaluation and possible approval of new beneficial treatments for diseases [5, 6, 7, 8].

We have developed a web-based expert system to match eligible patients with open clinical trials, or if the patient was eligible and not put onto the trial, categorize the reasons [4]. It has been used on historical data. The system allows for testing against all available clinical trials or a subset of them at a time. It is necessary to precisely follow the wording of inclusion/exclusion criteria when creating questions. This results in multiple, slightly different, questions which ask for the same type of information. Further, a physician will immediately know the answers to some questions based on answers to others. That is, there is a clear implication that always holds. However, the people entering the protocols may or may not know this. A busy clinical practitioner will not be willing to answer redundant questions and hence will not use a system that asks them. So, it is of paramount importance to develop implication rules which assert all answers implied by any given answer. For example, once we know the sex is male no question about pregnancy should appear on a displayed page of questions. Once a question about biopsy results is answered, there should be no question about whether they have ever had any type of surgery.

There have been other systems developed to help increase accrual to clinical trials. These and our own research results have indicated that a significant increase in accrual is possible [3, 4, 9]. However, to get a system fielded in the clinic it must be extremely easy and quick to use. It should request no extraneous information. This paper discusses improving usability of such a system over time by applying association rule mining to data that is acquired

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

through the initial use of our clinical trial assignment expert system at the Moffitt Cancer Center at the University of South Florida. We have used a tool based on the Apriori algorithm [1] to show that it is possible to recover what we call implication rules that were obtained from experts, as well as some unknown (to us) implication rules. These implication rules allow facts to be asserted when a related question is answered and allow decisions to be made with less data entry which is extremely important to enable clinical use.

Figure 1 shows the task flow for this work.

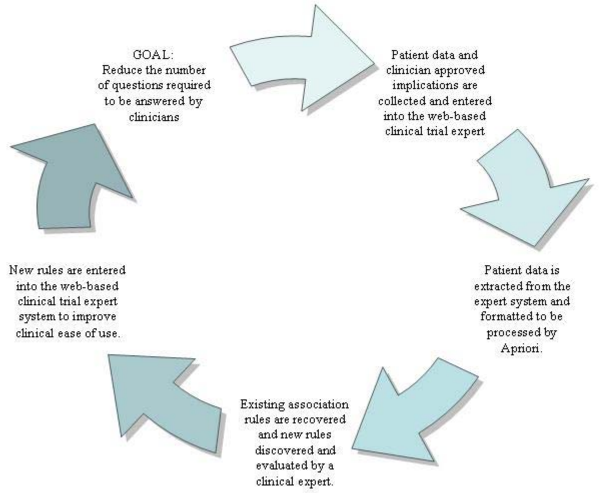

GOAL:
Reduce the number of questions required to be answered by clinicians
Patient data and clinician approved implications are collected and entered into the web-based clinical trial expert
Patient data is extracted from the expert system and formatted to be processed by Apriori.
Existing association rules are recovered and new rules discovered and evaluated by a clinical expert.
New rules are entered into the web-based clinical trial expert system to improve clinical ease of use.

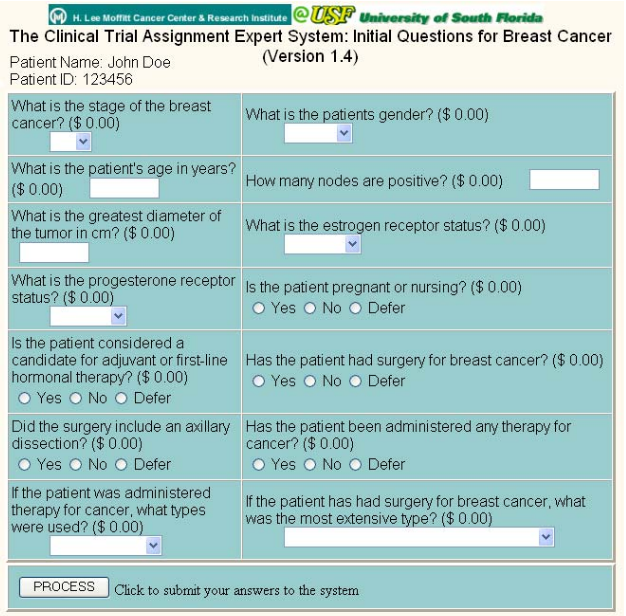
**Figure 1. Flow of research for this project.**

## 2 Patient Data Collection

Data was collected for 135 past breast cancer patients treated at the Moffitt Cancer Center. Data mining techniques were applied to a database of 135 patients in one set of experiments and 100 patients in a second set of experiments. The use of the reduced database for the second set of experiments was due to version changes of the expert system and changes in the open clinical trial protocols at the time of these experiments. The initial page of questions is shown in Figure 2.

### 2.1 Data Extraction

Patient data for the 135 patients entered into the system was extracted from the Clinical Trial Assignment expert system database and consolidated into a single file. For each patient a different number of questions could be answered. Maybe 1 question would rule them out or the answers to 30 would show eligibility, for example. There were 299 questions but 46 required continuous values and were ignored because our association rule mining tool, like most, did not accept continuous values. So, each patient feature vector consisted of answers to between 1 and 253 questions.


H. Lee Moffitt Cancer Center & Research Institute @USF University of South Florida

The Clinical Trial Assignment Expert System: Initial Questions for Breast Cancer
(Version 1.4)

Patient Name: John Doe
Patient ID: 123456

| | |
|---|---|
| What is the stage of the breast cancer? ($ 0.00) | What is the patients gender? ($ 0.00) |
| What is the patient's age in years? ($ 0.00) | How many nodes are positive? ($ 0.00) |
| What is the greatest diameter of the tumor in cm? ($ 0.00) | What is the estrogen receptor status? ($ 0.00) |
| What is the progesterone receptor status? ($ 0.00) | Is the patient pregnant or nursing? ($ 0.00)<br>○ Yes ○ No ○ Defer |
| Is the patient considered a candidate for adjuvant or first-line hormonal therapy? ($ 0.00)<br>○ Yes ○ No ○ Defer | Has the patient had surgery for breast cancer? ($ 0.00)<br>○ Yes ○ No ○ Defer |
| Did the surgery include an axillary dissection? ($ 0.00)<br>○ Yes ○ No ○ Defer | Has the patient been administered any therapy for cancer? ($ 0.00)<br>○ Yes ○ No ○ Defer |
| If the patient was administered therapy for cancer, what types were used? ($ 0.00) | If the patient has had surgery for breast cancer, what was the most extensive type? ($ 0.00) |

Click to submit your answers to the system

**Figure 2. Screenshot of Initial Questions page for the Clinical Trial Assignment System.**

## 3 Data Mining Experiments

The experiments were done with the Apriori and Apriori Rules software (http://www.adrem.ua.ac.be/~goethals/software/). We used two experimental data sets.

For completeness we provide some definitions. For a given set of transactions, where each transaction is a set of items, an association rule is an expression XY, where X and Y are subsets of items and $I$ is the set of all items. Usually expressed as $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$ [1].

An association rule has a minimum support. Rule $X \Rightarrow Y$ has *support s* in a transaction set *D* if *s*% of transactions in *D* contain $X \cup Y$ [1]. It also has a confidence. The rule $X \Rightarrow Y$ holds in the transaction set *D* with *confidence c* if *c*% of transactions in *D* that contain X also contain Y [1].
The goal is to generate all the association rules found within the dataset that are greater than or equal to the minimum support and confidence values.

### 3.1 Data Set One

In the first experiment, 135 patient records each with a possible 253 dimensions were processed through the Apriori software. Since the implications we search for **always hold**, a confidence value of 1 is required. The lower the minimum support value the more rules to sift. In our case, we were searching for known implications which we hypothesized would occur frequently (high support).

For the purpose of this experiment a minimum support value of 88 and a confidence value of 1 which is equivalent to 100% confidence was used. This produced 15 association rules. It was determined that this was not sufficient for the purposes of this investigation because none of the original implication rules were recovered and of the 15 association rules found, none were considered meaningful.

The minimum support value was lowered to 85 resulting in 114 new association rules. From this experiment two meaningful rules were discovered; however none of the original implication rules were recovered. Next, a minimum support of 70 and a confidence level of 100% was used. This resulted in 1812 new association rules. Most of the rules generated were permutations on the earlier rules discovered; however from the results we were able to locate 83 rules of interest. Of these 83 association rules discovered, 16 of these rules were of particular interest and required further investigation. Of the 16 new rules discovered, it was determined that four of these rules should be potentially added back into the expert system as *new* implications, a nice result. However, we still did not find the existing clinician generated implication rules. The newly discovered rules were:

1. If the patient has had surgery for breast cancer, what was the most extensive type? - Breast_sparing_procedure_or_lumpectomy ⇒ Has the patient had surgery for breast cancer? - Yes
   (This combination occurred 79 times out of 135 records)

2. What is the estrogen receptor status? - Positive ⇒ Has the patient had surgery for breast cancer? - Yes
   (This combination occurred 80 times out of 135 records)

3. What is the progesterone receptor status? - Positive ⇒ Has the patient had surgery for breast cancer? - Yes
   (This combination occurred 70 times out of 135 records)

4. Did the surgery include an axillary dissection? - Yes ⇒ Has the patient had surgery for breast cancer? - Yes
   (This combination occurred 116 times out of 135 patient records)

It was apparent that *Rules* one (1) and four (4) should be included as additions to the implication set in the current knowledge base. It was not clear to us whether *Rules* two (2) and three (3) were valid. A clinical expert was consulted and determined these implications were correct. When a patient is determined to have either a positive estrogen and/or progesterone status then breast cancer surgery would have occurred.

### 3.2 Data Set Two

We noticed that there had been some significant changes in the clinical trial protocols that were open during data collection. So, we chose the most recent 100 patients' data and focused just upon the two protocols for which we had 4 clinician generated implication rules. We used just a subset of patients that were ineligible for each of the two protocols for two reasons. One, it was expected that implication rules that existed would be used in these cases to quickly rule out the patients and we believed that it might be possible to find some new implications to quickly determine patients were not eligible.

Each protocol was processed independently using *Apriori* with a minimum support value of ten and a confidence level of 100%.

First, all 100 ineligible patients for the first of the two protocols analyzed were used. There were 831 rules that were discovered which included three of the four current system implications. From this set of rules, there were ten additional rules of interest that required further investigation to verify their medical validity, but just one was found valid by our clinical experts. The current system implications recovered from the first run of this experiment were as follows:

1.) Did the surgery include an axillary dissection? No
=> Was an axillary dissection performed? No
(This combination occurred 14 times out of 100 records.)
2.) Was an axillary dissection performed? Yes
=> Did the surgery include an axillary dissection? Yes
(This combination occurred 71 times out of 100 records. The reverse implication also held.)

New rule: If the patient was administered therapy for cancer, what types were used? Chemotherapy
=> Has the patient signed an informed consent? Yes
(This combination occurred 36 times out of 100 records.)

In the second experiment, only the 85 patients determined ineligible for the second of the two protocols analyzed were used. There were 487 rules discovered which included one of the current system implications. There were an additional 13 new rules of interest that would require further investigation to establish medical validity. The last current implication rule was recovered:

Has the patient been administered any therapy for can-

cer? No
=> Has the patient received a total dose of doxorubicin or daunorubicin of greater than 450 mg/mˆ2? No
(This combination occurred 53 times out of 100 records.)

Three of the thirteen additional rules discovered were verified to be valid (by the clinicians in our acknowledgments) in all situations and should be considered for addition to the Clinical Trial Assignment expert system. An example is: If the patient was administered therapy for cancer, what types were used? Chemotherapy
=> Has the patient been administered any therapy for cancer? Yes
(This combination occurred 49 times out of 100 records.)

From these results, it appears that in future experiments; current implications can be recovered using the same methods. Additionally, this experiment clearly verifies that we received good information from interviews with clinicians.

## 4 Conclusions

Research has shown that clinicians are reluctant to use automated clinical trial selection systems due to time constraints; therefore it was the goal of this research to attempt to discover new implications within the knowledge base that would reduce the number of questions required to be answered by clinicians. By reducing the number of questions it is our hope that the system will become easier to use and will encourage clinicians to actively utilize the system as a tool for matching patients to currently available clinical trials. The system described was focused on breast cancer trials, but has been adapted for different cancer types.

The result of this research is that eight (8) additional association rules or implications were discovered. These rules have been verified as correct by a clinical expert. The newly discovered rules have been added to the Clinical Trial Assignment Expert System to improve clinical ease of use, system efficiency, and accuracy.

Another goal of this research was to verify that all the current implications are valid with 100% confidence. This was done by looking at ineligible cases. All system implications being used in the currently open clinical trial protocols were recovered, which provides support to the use of data mining techniques to determine new association rules. All new implications should be verified with a clinical expert to determine the medical validity of the implication as was done here. As shown in the Experiments section, there are many associations that can be derived from the use of the association rule learning tool, *Apriori*. However, very few of those implications are medically verifiable and accurate. Even with a confidence of one there are many potential rules to sift, which is an issue. There will always be a need to discuss potential new rules with a medical expert before adding these implications into the expert system.

**Acknowledgements:** This work has been partially supported by the Breast Cancer Research Program of the U.S. Army Medical Research and Materiel Command under contract DAMD17-00-1-0244, the National Institutes of Health under grant number 5R21CA108653-2 and by the Moffitt Cancer Center. Thanks to Margaret E. Gross-King, RN, MS, and Jamie Bolen, RN, of the H. Lee Moffitt Cancer Center & Research Institute for their clinical expertise. Thanks to Matthew Boonstra, Bhavesh Goswami, and Prodip Hore, for their assistance and training and to Tim Ivanovskiy for a readable screenshot.

## References

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. *Proc. of the 20th Int'l Conference on Very Large Databases*, Santiago, Chile, Sept. 1994

[2] Cancer facts and figures 2004. American Cancer Society, http://www.cancer.org/docroot/STT/stt_0_2004.asp?sitearea =STT&level=1, 2004.

[3] Robert W. Carlson, Samson W. Tu, Nancy M. Lane, Tze L. Lai, Carol A. Kemper, Mark A. Musen, and Edward H. Shortliffe. Computer-based screening of patients with HIV/AIDS for clinical trial eligibility. *Online Journal of Current Clinical Trials*, 4(179), 1995.

[4] Eugene Fink, Princeton K. Kokku, Savvas Nikiforou, Lawrence O. Hall, Dmitry B. Goldgof, and Jeffrey P. Krischer. Selection of Patients for Clinical Trials: An Interactive Web-Based System, Artificial Intelligence in Medicine, 31(3), 241-254, July 2004.

[5] John H. Gennari and Madhu Reddy. Participatory design and an eligibility screening tool. In Proceedings of the American Medical Informatics Association Annual Fall Symposium, pages 290 294, 2000.

[6] Carolyn Cook Gotay. Accrual to cancer clinical trials: Directions from the research literature. Social Science and Medicine, 33(5):569 577, 1991.

[7] B. Seroussi, J. Bouaud, and E-C. Antoine. Enhancing clinical practice guideline compliance by involving physicians in the decision process. In Werner Horn, Yuval Shahar, Greger Lindberg, Steen Andreassen, and Jeremy C. Wyatt, editors, Artificial Intelligence in Medicine, pages 76 85. Springer-Verlag, Berlin, Germany, 1999.

[8] Samson W. Tu, Carol A. Kemper, Nancy M. Lane, Robert W. Carlson, and Mark A. Musen. A methodology for determining patients eligibility for clinical trials. Journal of Methods of Information in Medicine, 32(4):317 325, 1993.

[9] B. Seroussi, J. Bouaud, and E.-C. Antoine. ONCODOC: A successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. Artificial Intelligence in Medicine, 22(1):43 64, 2001.