

# Using Probabilistic Methods to Optimize Data Entry in Accrual of Patients to Clinical Trials

Bhavesh D. Goswami<sup>1</sup>, Lawrence O. Hall<sup>1</sup>, Dmitry B. Goldgof<sup>1</sup>, Eugene Fink<sup>2</sup>,  
Jeffrey P. Krischer<sup>1</sup>

*bgoswami@csee.usf.edu, hall@csee.usf.edu, goldgof@csee.usf.edu, e.fink@cs.cmu.edu,  
jpkrischer@moffitt.usf.edu*

<sup>1</sup>*Computer Science and Engineering, University of South Florida, Tampa, FL 33620*

<sup>2</sup>*Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213*

## **Abstract**

*A clinical trial is defined as a study conducted on a group of patients to determine the effect of a treatment. Assignment of patients to clinical trials is a data and labor intensive task. Usually, medical personnel manually check the eligibility of a patient for a clinical trial based on the patient's medical history and current medical condition. According to studies, most clinical trials are under-enrolled which negatively affects evaluation of their efficacy. We have developed web-based agents that can test the eligibility of patients for many clinical trials at once. An agent that uses analytical methods to reorder questions for optimizing cost and data entry has been developed. This paper introduces a probabilistic version of the system. Agents with analytical and probabilistic heuristics were then tested on retrospective data of relatively current breast cancer patients at the Moffitt Cancer Center. It is shown that less data entry is required when probabilistic agents are used to reorder questions.*

## **1. Introduction**

A clinical trial is an experimental research study that evaluates a specific new treatment for a specific population of patients. The trial protocol is like a rulebook which clearly identifies the criteria for a patient to be eligible for the trial. The criteria are based on the medical history and the present medical condition of the patient. Some criteria are general information such as the age and sex of the patient while others requires specific tests be done to determine if the patient matches them. As each eligibility protocol may have many criteria, it is a labor intensive task to check the eligibility of a patient for many clinical trials at once. Studies have found that clinicians miss up to 60% of eligible patients and many clinical trials fail due to under-recruiting [1, 2].

Several researchers have used artificial intelligence techniques to address this problem. A system called AIDS [3] was developed to assign patients to HIV clinical trials using Bayesian belief networks. A rule based system called EON [4] was also developed for selecting participants for AIDS clinical trials. ONCODOC [5, 6] was developed for accrual of patients in breast cancer clinical trials using decision trees. Papaconstantinou and colleagues [7] developed

---

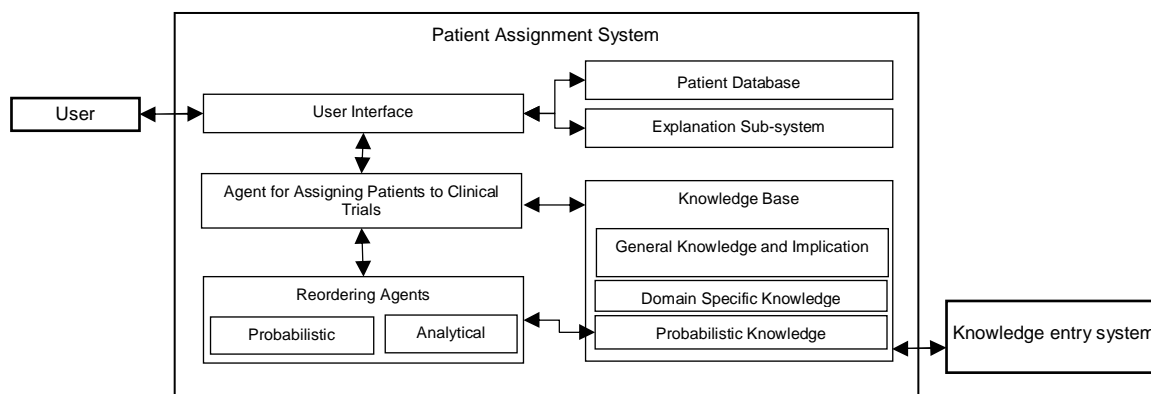
This work has been partially supported by the Breast Cancer Research Program of the U.S. Army Medical Research and Materiel Command under contract DAMD17-00-1-0244.

a similar system for assignment of patients to breast cancer clinical trials using Bayesian networks.

Most probabilistic patient assignment systems used Bayesian networks. Thus, they had problems inherent to Bayesian networks like complex structures, difficulty in addition of new trials and slow execution time. When decision trees were used, the system could check the eligibility of the patient for only one clinical trial at a time. This would make it time inefficient to check the eligibility of a patient for many clinical trials. Thus the authors first developed an analytical rule based system [8, 9]. The system was user friendly, new clinical trials could easily be added into the system and when tested on retrospective data of relatively current patients at the Moffitt Cancer Center in Tampa, FL, it resulted in identification of 160 assignments of patients to clinical trials potentially missed by medical personnel [8, 9]. The system had some drawbacks inherent to rule based systems. The most prominent one was the inability of the system to estimate the probability of a patient being eligible for a given clinical trial in the absence of complete information. Consider a clinical trial protocol which has 20 eligibility rules, and 19 of them are met, which seemingly makes it likely for a patient to be eligible. A rule based system will still not be able to predict anything about the probability of the patient's eligibility while a probabilistic system can estimate the eligibility of the patient using its prior probabilities. To attempt to exploit the advantages of the probabilistic systems and to avoid its drawbacks, we used the probabilistic methods discussed in this paper.

## 2. System design

The system is divided into 1.) A knowledge entry system and 2.) A patient assignment agent. Nikiforou [10] implemented the knowledge entry system. It is a web-based system that has a user-friendly interface for encoding the clinical trials into a form that is understood by the agents. Fletcher, Kokku [8] and colleagues have built the initial agent for matching of patients to clinical trials. We added the probabilistic agents to the system and conducted experiments to compare the data entry needed by each system to decide the eligibility of a patient for a set of clinical trials. Figure 1 represents the basic structure of the system.



**Figure 1: System architecture**

As shown in Figure 1, the user interacts with the system through the web-based interface. A user can access old patient data, add new patients and enter data for existing patients. The system interacts with the user by presenting a list of questions. If the user has the information required to answer the questions, she does so, or else additional test needs to be done on the patient to obtain the information needed to answer the questions. Each time a question is answered, new information is obtained about the patient and her eligibility is rechecked for the

protocols selected. At any time the patient is either eligible, ineligible or more information is required to decide on eligibility for a particular protocol. The system augments its probabilistic knowledge base by monitoring the information entered for the current patients and its effect on the eligibility of the patient. After checking eligibility, if the patient has protocols for which eligibility is not yet decided upon, the reordering agents reorder the relevant questions using analytical and probabilistic agents. At all times, the system has an explanation subsystem which can provide an explanation for the system's decisions. More protocols can be added to the system using the knowledge entry system.

### **3. Probabilistic reordering agent**

As previously discussed, estimation of eligibility probability becomes important in certain cases, especially when you have many trials. The Moffitt Cancer Center at USF has about 15 active breast cancer trials at once. It is time-consuming and expensive to test the eligibility of patients for all trials. Thus rather than testing the patients for all trials, we can just test her eligibility for trials which show high initial eligibility probability. We can also use probabilistic knowledge accumulated over time by the system to reorder the questions.

First we discuss using the probabilistic knowledge base to reorder questions to reduce data entry. The basic idea is to try to classify a patient as ineligible as soon as possible. If a patient is ineligible, the information that is most likely to determine her ineligible should be obtained first. This would optimize the data entry needed to decide upon a patient's eligibility. The system gathers this probabilistic knowledge over time. For each question in the system, it keeps a log of how many times a question is asked and how many times a patient is ruled out for a particular clinical trial after that question was answered. This gives us the probability of that question, when asked, ruling out a patient for a particular trial. So the approach would be to ask the question which has the highest probability of ruling out a patient first. To test the effectiveness of this method, we did 10-fold cross-validation experiments on retrospective patient data from Moffitt Cancer Center. The details of the experiment are presented in the experiments section and the results are presented in Table 1 also in the experiments section.

These probabilities are also used to estimate the eligibility probability of a patient for a clinical trial. In this case we make an important assumption that all questions have independent probabilities. Although this assumption is not entirely true, it is practical and close enough. Most questions are either completely dependent on each other or are not at all dependent. For example if a patient has no positive lymph nodes, which means that the cancer stage is either 0 or 1. Thus the two questions "Does the patient have positive lymph nodes?" and "What is the cancer stage?" are dependent on each other. We can take care of such situations by an implication sub-system. We can add implication rules to the system such as "If cancer stage 0 or 1 then it implies that patient has no positive lymph nodes". When the system has information that the cancer stage is either 0 or 1, the implication subsystem automatically generates information that no lymph nodes are positive and the system does not ask for that information. Thus, all such completely dependent questions are taken care of by the implication subsystem. There are very few questions which have a partial implication, like "If a patient is ER positive then there is an 80% chance of positive lymph nodes". We ignore such conditional probabilities among questions and treat all questions as either completely dependent or independent.

To compute eligibility probabilities, Bayes rule appeared to fit well. We can think of the patient enrolment procedure as a classification problem. The classification classes will be "Eligible" or "Ineligible". The attributes will be the questions and the values for the question are "Favorable for eligibility" or "Unfavorable for eligibility" for each clinical trial. We have a

probability for each question to be favorable and unfavorable for each clinical trial. Thus we have a classification problem where we have probabilities for the occurrence of each attribute value. To use Bayes rule we also needed probabilities of the occurrence of each classification type, which is “Eligible” or “Ineligible” in our case. The system recorded how many patients were tested for each clinical trial and how many patients were decided to be eligible and ineligible. Thus, we could now use Bayes rule to calculate the eligibility probability of a patient with partial information.

Let us assume that we have a clinical trial T with three questions  $Q_1, Q_2$  and  $Q_3$ . Out of 100 patients that were tested for the clinical trial, we found 40 to be eligible and 60 to be ineligible. Question  $Q_1$  was asked 90 times and it disqualified patients 10 times,  $Q_2$  was asked 80 times and disqualified patients 5 times, and  $Q_3$  was asked 70 times and disqualified patients 15 times.

Thus  $P(T_E)$ , the probability of patient being eligible for protocol T is  $\frac{40}{100}$ .  $P(Q_1)$ , the

probability that question  $Q_1$  is answered favorably for clinical trial T is  $\frac{80}{90}$ . Similarly  $P(Q_2) =$

$$\frac{75}{80} \text{ and } P(Q_3) = \frac{55}{70}.$$

When we don't have answers for any questions of clinical trial T, the probability of a given patient being eligible for it is  $\frac{40}{100}$ . Now assume that we have answers to questions  $Q_1$  and  $Q_2$

and the answers are favorable for eligibility. Thus the new eligibility probability will be  $P(T_E | Q_1, Q_2)$ . According to Bayes rule:

$$P(T_E | Q_1, Q_2) = \frac{P(T_E) P(Q_1, Q_2 | T_E)}{P(Q_1, Q_2)}$$

$P(Q_1, Q_2 | T_E)$  will be the probability that questions  $Q_1, Q_2$  are favorable for eligibility given that the patient is eligible for protocol T. Now, if a patient is eligible for a clinical trial, all the questions must be answered favorably for eligibility. Thus,  $P(Q_1, Q_2 | T_E) = 1$ . Also we have assumed that all questions are independent and thus

$$P(Q_1, Q_2) = P(Q_1) P(Q_2)$$

Substituting all the values, we get the new equation:

$$P(T_E | Q_1, Q_2) = \frac{P(T_E)}{P(Q_1) P(Q_2)}$$

Substituting the values of  $P(T_E)$ ,  $P(Q_1)$  and  $P(Q_2)$  we get the value 0.48. Thus after answering two questions  $Q_1$  and  $Q_2$ , the eligibility probability of a patient for clinical trial T is 48%. As more questions answered fit the eligibility criteria, the eligibility probability increases. When a question's answer does not fit the eligibility profile, the patient becomes ineligible and the eligibility probability becomes zero. When the patient becomes eligible the eligibility probability is 1. The eligibility probability varies between 0 and 1 when the eligibility of a patient is undecided.

The generic equations for eligibility probability for a clinical trial when we have favorable answers for eligibility for  $n$  questions will be

$$P(T_E | Q_1, Q_2, \dots, Q_n) = \frac{P(T_E)}{P(Q_1) P(Q_2) \dots P(Q_n)}$$

#### 4. Probabilistic experiments

To test the effectiveness of the technique that use probabilistic knowledge base in reordering questions we did a ten-fold cross validation experiments on retrospective data of 90 relatively current patient at Moffitt Cancer Center at USF. We randomly selected 90% of the patients and their data was used to generate a probabilistic knowledge base for the system. The remaining 10% of the patients were tested using the system which used this probabilistic knowledge base to reorder the questions. This process was repeated ten times with each 10% of the testing patients being unique. Six clinical trials were used in the experiments. These six clinical trials were selected out of about 15 possible clinical trials, as these trials were in “open” status for a long enough duration during the experiments to have an adequate number of patients that were tested. In the probabilistic experiments it was necessary to train the system on an adequate number of patients before it can be used to reorder questions for other patients. We selected 90 patients at random from our list of patients and used their data. As mentioned above, a ten-fold cross validation was carried out, so that the system was trained on 81 patients and the remaining 9 patients were tested using the system. Table 1(a) shows the results of each fold of the cross validation.

**Table 1(a): Results using probabilistic system to reorder questions**

Ten-fold cross validation				
Test Num	Average number of questions			Diff %
	System A	System B	Diff	
1	16.67	20.75	4.08	19.68
2	15.17	17.00	1.83	10.78
3	15.83	17.58	1.75	9.95
4	15.75	18.25	2.50	13.70
5	13.83	16.67	2.83	17.00
6	15.58	17.75	2.17	12.21
7	15.83	18.25	2.42	13.24
8	15.50	16.83	1.33	7.92
9	16.50	18.50	2.00	10.81
10	15.83	19.17	3.33	17.39
Average	15.65	18.08	2.43	13.42

**Table 1(b): Results using eligibility probability to reorder questions**

Ten-fold cross validation				
Test Num	Average number of questions			Diff %
	System A	System B	Diff	
1	20.67	28.67	8.00	27.91
2	29.00	34.33	5.33	15.53
3	31.67	24.33	-7.33	-30.14
4	26.33	33.00	6.67	20.20
5	22.33	25.00	2.67	10.67
6	18.67	31.67	13.00	41.05
7	25.67	33.00	7.33	22.22
8	22.67	36.67	14.00	38.18
9	19.33	22.67	3.33	14.71
10	17.33	24.33	7.00	28.77
Average	23.37	29.37	6.00	20.43

System A is the probabilistic system and system B is the analytical system. As seen in Table 1 the probabilistic system reduces data entry by 13.42% on average compared to the analytical system. The average number of questions asked by the system is reduced by 2.43. One important observation is that the probabilistic system **always** asks fewer questions than the analytical system for all the patients tested. Using the t-test, the probabilistic system is statistically significantly better at the 99.99% confidence interval in the number of questions asked.

The eligibility probability of the patients was derived as explained before. An important use of the eligibility probability can be to try to quickly assign a patient to a clinical trial. This can be achieved by generating an initial eligibility probability for all the available trials for that patient. We then check the eligibility of the patient for the trial which has the highest eligibility probability. After every piece of information is obtained, the probabilities are regenerated and the system asks for more information about the trial with the highest eligibility probability until the patient is found eligible for a trial or is determined ineligible for all the trials. The system stops seeking further information after the patient is found eligible for a clinical trial as the purpose of these experiments is to find a single matching protocol with least the number of questions being answered. To test the effectiveness of this approach we did a ten-fold cross validation on the available patients. We used the same six clinical trials. We compare the results

with the analytical system in which we stop answering questions when it finds a clinical trial for which the patient is eligible. Results are shown in Table 1(b).

The results show that the Bayes method of computing eligibility probability and using it in reordering the questions reduces the data entry needed by 20.43 % on an average. Also, the probabilities generated can be used to give feedback to the user. Using the t-test, the probabilistic system is statistically significantly better at minimizing questions at the 95% confidence interval.

## 5. Conclusions

Recruiting patients to clinical trials is very time and labor intensive work. Many clinical trials can't be fully evaluated due to under recruitment. Reducing the data entry needed to determine eligibility of a patient can result in a patient being tested for more clinical trials in less time. Thus the system can play a critical part in the success of a clinical trial. Also the web-based interface of the system makes it possible to have a central system which can be accessed by clinical personnel from any medical institute around the country. All large research centers have clinical trials of their own and it is very hard to exchange the trial information between them as the same trials can be interpreted in a different way by different clinicians. Having an electronic version of the clinical trials encoded using our knowledge entry system can make sharing of clinical trials between different hospitals very convenient and effective. This in turn can increase accrual for clinical trials as the pool of potential participants increases. The system also effectively reorders the tests and reduces the cost incurred in determining eligibility. Data entry was successfully optimized by as much as 20% using the probabilistic reordering agent.

## 6. References

- [1] Cyrus Kotwall, Leo J. Mahoney, Robert E. Myers, and Linda Decoste. *Reasons for non entry in randomized clinical trials for breast cancer: A single institutional study*. Journal of Surgical Oncology, 50:125-129, 1992.
- [2] 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.
- [3] Lucila Ohno-Machado, Eduardo Parra, Suzanne B. Henry, Samson W. Tu, and Mark A. Musen. *AIDS: A decision-support tool for decreasing physicians' uncertainty regarding patient eligibility for HIV treatment protocols*. In Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care, pages 429-433, 1993.
- [4] Mark A. Musen, Samson W. Tu, Amar K. Das, and Yuval Shahar. *EON: A component based approach to automation of protocol-directed therapy*. Journal of the American Medical Informatics Association, 3(6):367-388, 1996.
- [5] Brigitte S'erotoussi, Jacques Bouaud, and Eric-Charles Antoine. *Users' evaluation of ONCODOC, a breast cancer therapeutic guideline delivered at the point of care*. Journal of the American Medical Informatics Association, 6(5):384-389, 1999.
- [6] Brigitte S'erotoussi, Jacques Bouaud, and Eric-Charles 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.
- [7] Constantinos Papaconstantinou, Georgios Theocharous, and Sridhar Mahadevan. *An expert system for assigning patients into clinical trials based on Bayesian networks*. Journal of Medical Systems, 22(3):189-202, 1998.
- [8] Princeton K. Kokku, Lawrence O. Hall, Dmitry B. Goldgof, Eugene Fink, and Jeffrey P. Krischer. *A cost-effective agent for clinical trial assignment*. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 2002.
- [9] Eugene Fink, Lawrence O. Hall, Dmitry B. Goldgof, Bhavesh D. Goswami, Matthew Boonstra, and Jeffrey P. Krischer. *Experiments on the automated selection of patients for clinical trials*. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 2003.
- [10] Savvas Nikiforou. *Selection of clinical trials: Knowledge representaton and acquisition*. Master's thesis, Department of Computer Science and Engineering, University of South Florida, 2002.