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

Bhavesh D. Goswami, Lawrence O. Hall, Dmitry B. Goldgof,
Eugene Fink, and Jeffrey P. Krischer

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

Computer Science and Engineering, University of South Florida, Tampa, FL 33620

Abstract

A clinical trial is a study conducted on a group of patients to evaluate a new treatment procedure. Usually, clinicians manually select patients for a clinical trial; the choice of eligible patients is a labor-intensive process, and clinicians are often unable to identify sufficient number of patients, which delays the evaluation of new treatments. We have developed a web-based system that helps clinicians to determine the eligibility of patients for multiple clinical trials. It uses probabilistic techniques that minimize the amount of manual data entry, by ordering the related data-entry steps. We describe the developed system and give the results of applying it to retrospective data of breast cancer patients at the Moffitt Cancer Center.

1. Introduction

A clinical trial is an experimental evaluation of a new medical procedure. When medical researchers conduct a trial, they specify a list of criteria that determines a patient's eligibility for this trial, and use these criteria to select potential participants among available patients. The selection of patients has traditionally been a manual procedure, and studies have shown that clinicians can miss up to 60% of eligible patients, which often delays the evaluation of new treatments [4, 11].

Computer scientists have developed several artificial-intelligence systems to address this problem. In particular, Musen *et al.* built a rule-based system, called EON, which selected AIDS patients for clinical trials [5]. Ohno-Machado *et al.* developed the AIDS^2 system, which also assigned AIDS patients to clinical trials [7]. Bouaud *et al.* created a decision-tree system, called ONCODOC, which helped to select patients for cancer trials [9, 10]. Papaconstantinou *et al.* built a Bayesian system for assigning patients to breast-cancer trials [8].

The developed probabilistic systems used Bayesian networks, and they inherited the usual drawbacks of Bayesian systems, including complex structure, difficulty of adding new trials, and significant running time. On the other hand, the decision-tree system could check a patient's eligibility for only one trial, and it did not scale to the use of multiple trials. To address this problem, we developed an analytical rule-based system, which efficiently processed multiple clinical trials [1–3]; however, it was unable to estimate the probably of a patient's eligibility for available trials in the absence of complete information. We have then combined the rule-based system with probabilistic techniques, described in this paper.

The developed system consists of knowledge-entry tools [6] and a patient-selection mechanism [3], as shown in Figure 1. The user accesses the system through the web-based

interface, which allows retrieving old patient data and adding new patients. For each new patient, the system presents a list of related questions, and then uses the answers to select matching clinical trials. After each answer, it estimates the probability of the patient's eligibility for each trial, and re-orders the remaining questions to minimize the expected amount of data entry. If the system decides that a patient is ineligible for a trial, it shows the conditions that make the patient ineligible. Furthermore, the system uses the new answers to augment its probabilistic knowledge and to revise the probability estimates.

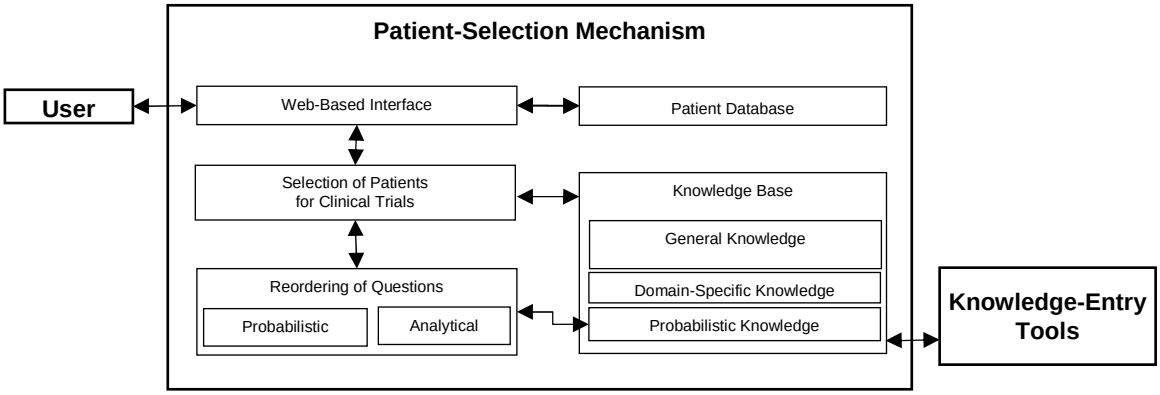

Figure 1: Architecture

2. Ordering of questions

We first explain the use of probabilities for reordering questions to reduce the amount of data entry. The underlying idea is to determine a patient's ineligibility as soon as possible. If a patient is ineligible, the information that is most likely to show her ineligibility should be obtained first. For each question, the system keeps track of how many times it was asked in the past, and how many times a patient was determined ineligible based on this question. It first asks the question that has the highest probability of showing that the patient is ineligible.

These probabilities are also used to estimate the eligibility probability of a patient for a given trial. We assume that all questions have independent probabilities. Although this assumption is not guaranteed to be true, in practice most questions are either completely dependent on each other or completely independent. For example, if a patient's cancer stage is 0 or 1, then the patient has no lymph nodes with cancer cells. Thus, the questions "Does the patient have positive lymph nodes?" and "What is the cancer stage?" are dependent. We can account for such situations by including implication rules into the knowledge base; for example, the system includes the rule "If cancer stage 0 or 1, then the patient has no positive lymph nodes."

We use the Bayes rule to compute eligibility probabilities. We can think of eligibility decisions as a classification problem with two classes, "eligible" and "ineligible." The attributes are the questions, and their values are "favorable" and "unfavorable" for eligibility. For each question and each clinical trial, we can determine the probability that the answer to this question is favorable for the trial. Thus, we have probabilities for the occurrence of each attribute value. To use the Bayes rule, we also need the probabilities of the occurrence of the classification types "Eligible" and "Ineligible." To obtain these probabilities, the system records how many patients were tested for each clinical trial and how many of them were eligible.

For example, suppose that we have a trial T with questions Q_1, Q_2 and Q_3, and that we have tested 100 patients, and found 40 of them eligible. Question Q_1 has been asked 90 times and disqualified patients 10 times, Q_2 has been asked 80 times and disqualified 5 patients, and Q_3

has been asked 70 times and disqualified 15 patients. Then, the probability that a patient is eligible for trial T is $P(T_E) = 40/100 = 0.4$. The probability that question Q_1 is answered favorably is $P(Q_1) = 80/90 = 0.89$; similarly $P(Q_2) = 75/80 = 0.94$, and $P(Q_3) = 55/70 = 0.79$.

Now suppose that we have answers to questions Q_1 and Q_2 for some patient, and both answers are favorable. According to the Bayes rule, the eligibility probability is

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

where $P(Q_1, Q_2 \mid T_E)$ is the probability that answers to Q_1, Q_2 are favorable given that the patient is eligible. If a patient is eligible, then all the questions are answered favorably, which means that $P(Q_1, Q_2 \mid T_E) = 1$. Furthermore, we have assumed that all questions are independent, which implies that $P(Q_1, Q_2) = P(Q_1)\, P(Q_2)$, and therefore

$$P(T_E \mid Q_1, Q_2) = \frac{P(T_E)}{P(Q_1)\, P(Q_2)} = \frac{0.4}{0.89 \cdot 0.94} = 0.48.$$

If the system collects more answers that satisfy the eligibility criteria, the eligibility probability becomes larger. On the other hand, if some answer does not satisfy the eligibility criteria, the system immediately concludes that the patient is ineligible. In general, if the system has collected *n* favorable answers and no unfavorable answers, the eligibility probability is

$$P(T_E \mid Q_1, Q_2, ..., Q_n) = \frac{P(T_E)}{P(Q_1)\, P(Q_2) \ldots P(Q_n)}.$$

3. Experiments

We have tested the developed technique on the retrospective data of six clinical trials and ninety patients at the Moffitt Cancer Center. We have used ten-fold cross validation; that is, we trained the system on the data of eighty-one patients, and then tested it on the other nine patients. We have repeated this test ten times, using different sets of nine test patients. In Table 1(a), we show the number of questions asked by the developed system in each test ("with probs."), and compare it with the number of questions asked by an earlier version of the system, which did not use probabilities ("without probs."). The use of probabilities has reduced the number of questions by 13%, and the *t*-test has shown that this difference is statistically significant with 99.99% confidence.

Table 1(a): Selection of all matching trials for each patient

Test number	Mean number of questions: With probs.	Without probs.	Difference	Percentage difference
1	16.7	20.8	4.1	20%
2	15.2	17.0	1.8	11%
3	15.8	17.6	1.8	10%
4	15.8	18.3	2.5	14%
5	13.8	16.7	2.9	17%
6	15.6	17.8	2.2	12%
7	15.8	18.3	2.5	13%
8	15.5	16.8	1.3	8%
9	16.5	18.5	2.0	11%
10	15.8	19.2	3.4	17%
Mean	15.7	18.2	2.4	13%

Table 1(b): Selection of one matching trial for each patient

Test number	Mean number of questions: With probs.	Without probs.	Difference	Percentage difference
1	20.7	28.7	8.0	28%
2	29.0	34.3	5.3	16%
3	31.7	24.3	−7.4	−30%
4	26.3	33.0	6.7	20%
5	22.3	25.0	2.7	11%
6	18.7	31.7	13.0	41%
7	25.7	33.0	7.3	22%
8	22.7	36.7	14.0	38%
9	19.3	22.7	3.4	15%
10	17.3	24.3	7.0	29%
Mean	23.4	29.4	6.0	20%

Sometimes, clinicans may need to find only one matching trial for a patient, rather than checking the patient's eligibility for all available trials. Thus, we have also experimented with using the system to select one matching trial for each patient. In this experiment, the system uses the probabilistic data to identify the most likely matching trial, and chooses questions relevant to this trial. It continues asking questions until it finds one matching trial or determines that the patient is ineligible for all trials. In Table 1(b), we give the results of this experiment ("with probs."), and compare them with a similar experiment without probabilistic reasoning ("without probs."). The use of probabilities has led to 20% reduction in the number of questions, and the *t*-test has shown that this difference is statistically significant with 95% confidence. The probabilistic system has given better results than the system without probabilities in nine out of ten cases. It has given worse results for test 3, because one of the patients in this test turned out ineligible for clinical trials that initially had a high eligibility probability.

4. Concluding remarks

We have developed a system that helps clinicians to select patients for clinical trials; it reduces the related manual work and helps to avoid human errors, thus increasing the number of selected patients. The system includes a probabilistic mechanism for ordering of the related questions, which helps to minimize the amount of data entry. The web-based interface allows a remote access to the system, and it can potentially enable physicians across the country to access a central repository of clinical trials.

5. References

[1] 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, pages 4541–4545, 2003.

[2] 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, to appear.

[3] 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

[4] 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.

[5] 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.

[6] Savvas Nikiforou. *Selection of clinical trials: Knowledge representation and acquisition.* Master's thesis, Department of Computer Science and Engineering, University of South Florida, 2002.

[7] Lucila Ohno-Machado, Eduardo Parra, Suzanne B. Henry, Samson W. Tu, and Mark A. Musen. *AIDS^2: 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.

[8] 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.

[9] Brigitte Séroussi, 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.

[10] Brigitte Séroussi, 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.

[11] 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.