Finding Appropriate Clinical Trials: Evaluating Encoded Eligibility Criteria with Incomplete Data

Nachman Ash, M.D.,M.S. [1,2], Omolola Ogunyemi, Ph.D. [1], Qing Zeng, Ph.D. [1], Lucila Ohno-Machado, M.D., Ph.D. [1,2]

[1]Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School
[2]Division of Health Sciences and Technology, Harvard Medical School and Massachusetts Institute of Technology, Boston, MA

ABSTRACT

We describe our work on creating a system that selects appropriate clinical trials by automating the evaluation of eligibility criteria. We developed a data model of eligibility for breast cancer clinical trials, upon which the criteria were encoded. Standard vocabularies are utilized to represent concepts used in the system, and retrieve their hierarchical relationships. The system incorporates Bayesian networks to handle missing patient information. Protocols are ranked by the belief that the patient is eligible for each of them. In a preliminary evaluation, we found good agreement (kappa 0.86) between the system and an independent physician in selection of protocols, but poor agreement (kappa 0.24) in protocol ranking. We conclude that our approach is feasible, and potentially useful in assisting both physicians and patients in the task of selecting appropriate trials.

INTRODUCTION

The important role of informatics in all stages of clinical trials is well established, encompassing patient accrual, protocol management and evaluation of results. The National Cancer Institute (NCI) plans to create a web enabled Cancer Informatics Infrastructure (CII) through which all aspects of clinical trials will be accessible[1,2]. Silva describes one of the major aspects of this vision: "...by using their computer, patients and their oncologists can find, for the patient's specific cancer, the best treatments and clinical trials" [1].

While information regarding clinical trials is currently easily accessible via the web[3], the task of finding appropriate clinical trials for a specific patient is tedious, requiring the evaluation of hundreds of eligibility criteria. Physicians often do not have enough time to perform this task, while patients may lack the knowledge and skills required.

Several methodologies were developed for evaluating patients' eligibility for clinical trials[4-8]. All of them aimed at improving the accrual of patients to specific trials. Ohno-Machado et al took a different approach by focusing on the patient. Their system allows the patient or her provider to obtain a ranked list of clinical trials for which the patient is likely to be eligible[9].

In this paper we present our extension to their work. We address the major concerns raised in that study: (1) the authors were able to encode only about 50% of the criteria, ignoring the most complex ones, and (2) they used a deterministic algorithm that did not take into account missing patient data. We designed an object oriented data model, and introduced the use of concepts and relationships from standard medical vocabularies to facilitate the encoding of complex criteria. In addition, our system makes use of Bayesian networks to handle the problem of missing patient data. We also present a preliminary evaluation of the system.

MATERIALS AND METHODS

Source of protocols. The clinical trial protocols were taken from NCI's Physician Data Query (PDQ) database[10]. We focused on phase II and phase III trials for the treatment of metastatic or recurrent breast cancer in women (see [9] for more details). Seventy-nine protocols have been retrieved using these criteria as of February 2001.

Implementation. We redesigned our system based on the following principles (Figure 1):

- Medical knowledge is encapsulated in an object-oriented data model.
- Concepts are represented using standard vocabularies.
- Eligibility criteria are encoded in a logical expression language derived from Arden syntax.
- Encoded eligibility criteria are stored in a database for reuse and future sharing.
- Bayesian networks are incorporated into the system's evaluation process for inferring missing patient data.
- Evaluated protocols are ranked by the likelihood that the patient is eligible for each of them.
- The system has a platform-independent implementation based on Java.

1067—5027/01/$5.00

Knowledge representation. The data model's structure is based on analysis of the breast cancer protocols and the Common Data Elements (CDE) of breast cancer clinical trials developed by NCI[1]. The model captures the data items in these protocols, their temporal aspects, and relationships among them. It is the basis for storing the patient data and checking for allowed values and inconsistencies.

The concepts used in the system are represented using standard vocabularies in the UMLS. We chose to use MeSH and PDQ, which contain the relevant concepts, and capture appropriate hierarchical relationships.

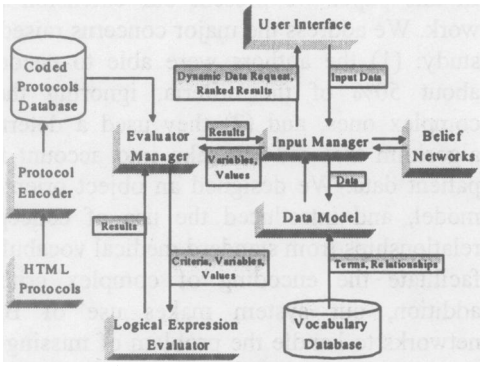

Figure 1: System architecture.

Encoding the protocols. Currently, the first 10 protocols out of the 79 retrieved from the PDQ database have been encoded. The HTML version of each protocol was automatically parsed to extract the textual eligibility criteria. These criteria were encoded manually (by the first author) using a variation of the Guideline Expression Language (GEL)[11]. The language contains the expressions used to retrieve data from the object model (based on pre-defined functions) as well as logical expressions.

We created a special editor for encoding the criteria. It lets the user check the syntax of an expression for correctness, verify the legitimacy of variables' names used in the expression, and assess whether the terms used in the expression map to concepts in the UMLS. When a criterion in a protocol is identical to a previously encoded criterion from a different protocol, its GEL-based encoding is retrieved automatically from the database. The time taken to encode each criterion is measured and saved for analysis.

Inferring missing data. We incorporated Bayesian networks into the new system to infer missing data based on population-based probabilities of patients' characteristics. Some of the probabilities were obtained or calculated from the medical literature and known statistical databases[12]. The first author estimated others based on his medical knowledge. Since the estimated probabilities are not optimal, we plan to augment them by using relevant patient data, as it becomes available, as suggested by Neapolitan[13].

The Bayesian network structure is based on causal and associational relationships identified from the data model and the common data items used in the protocols. Currently, it has 31 nodes and contains 4 separate directed acyclic graphs representing age-related items (Figure 2), liver function tests, white blood cell counts and pulmonary function tests. The software used for creating the network is JavaBayes[14].

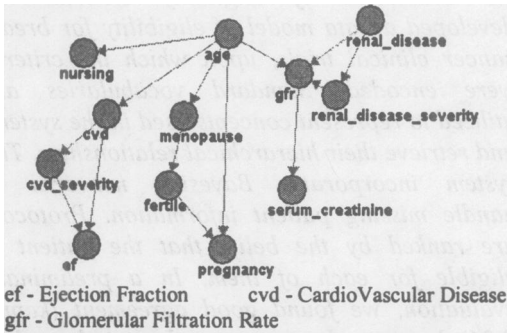

Figure 2: Directional graph of one of the Bayesian networks used in the system.

Evaluating criteria. Encoded criteria are evaluated using a three-valued logic (true, false, unknown) by a parser and interpreter created for GEL.

Ranking the protocols. Protocols for which all eligibility criteria evaluate to "true" given patient's data are ranked highest. Those that contain at least one criterion that evaluates to "false" are filtered out. The remaining protocols, containing at least one criterion that evaluates to "unknown", are ranked according to the belief that the patient is eligible for each of them. The ranking algorithm uses heuristics that take into account the following:

- Number of unknown criteria.
- A discriminatory score of each unknown criterion. An inclusion criterion that is probably true for most patients gets a different score than one that is probably true for only a small subset of patients. For example, "age greater than 18" is more inclusive than "age greater than 65", and therefore if the age of the patient is unknown, there is a greater chance that she meets the first criterion.
- Number of "inferred criteria" (criteria that originally evaluate to "unknown", and later to "true" or "false" based on inferred patient data).
- The evaluation result of the inferred criteria. A protocol containing a criterion that evaluates to false using inferred data is not filtered out, but rather gets a score that will rank it lower.

The final score of a protocol is given on a scale from 1 (definitely inappropriate) to 5 (definitely appropriate).

Criterion Difficulty	Number of Criteria	Average Encoding Time (Min)
Automatic Coding	18	≈ 0
Trivial	8	1.47
Easy	35	3.52
Difficult	9	11.12
Complex	5	28.12
Very Complex	2	36.80

Table 1: Average encoding time of 77 criteria stratified by difficulty of encoding.

Evaluation. Patient data were abstracted from medical records of patients with active metastatic or recurrent breast cancer, who were consecutively hospitalized during 1995 at the Brigham and Women's Hospital, Boston, Massachusetts. Forty-three data items were examined for each patient (items related to patient characteristics, disease characteristics, past treatment, other diseases and test results). The data collection process was separate from the protocol encoding process.

An independent physician (oncologist, but not a breast cancer specialist) evaluated the appropriateness of the protocols for each of the patients, grading the protocols as described above (on a 1-5 scale) and ranking them. The physician was given the patients' data in a short narrative description, and the full abstracts of the protocols as downloaded from NCI's CancerNet web site.

Statistical analysis. The agreement of the system and the physician on selection and ranking of protocols was calculated using the kappa and weighted kappa statistics[15].

RESULTS

Encoding process. We encoded 10 protocols each containing 20 - 41 eligibility criteria (mean 27.2). 228 criteria out of 272 (83.8%) were unique. We were able to encode 269 criteria (98.9%). For two of the three uncoded criteria ("no prisoners" and a request for a specific geographic location), the model could be improved to capture the necessary knowledge. The third ("No other concurrent medical or psychological condition that would preclude study compliance") was difficult to encode for automatic evaluation. A total of 39 other criteria (14.3%) did not represent their text version with 100% accuracy (e.g., "No medical or psychiatric condition that would increase risk" was encoded as "No severe medical or psychiatric condition". Since assessment of risk is subjective, it is difficult to encode for computation).

A significant number (30.3%) of the encoded criteria were lengthy (> 255 characters), suggesting the proportion of more complex criteria.

Table 1 presents the encoding time of 77 criteria from the last 3 protocols. The average encoding time was 5.88 minutes (median 2.1 minute). Therefore, encoding an average sized protocol may take about 3 hours.

Data Item	No. of patients(percent)
Stage:	
Stage IV	5 (25%)
Stage IIIb	5 (25%)
Unknown	10(50%)
Histology:	
Invasive Ductal Ca.	1 (5%)
Unknown	19 (95%)
Confirmed Histology/Cytology	17 (85%)
Measurable/Evaluable Disease	14 (70%)
Menopausal Status	
Postmenopausal	5 (25%)
Premenopausal	8 (40%)
Unknown	7 (35%)
Known Metastases	11 (55%)
Recurrent Disease	3 (15%)
Locally Advanced Disease	8 (40%)
Known Lymph Node Involvement	9 (45%)
Other Diseases	
Hypertension	3 (15%)
NIDDM*	1 (5%)
Asthma	1 (5 %)
Past Treatment	
Chemotherapy	16 (80%)
Radiotherapy	6 (30%)
Biotherapy	8 (40%)
Hormonal therapy	7 (35%)
Surgery	7 (35%)

*Non Insulin Dependent Diabetes Mellitus

Table 2: Patient characteristics.

Preliminary system evaluation. Data from records of 20 patients with metastatic, locally invasive, and recurrent breast cancer were collected. In average, about 25% of the 43 data items collected for each patient had missing values. Age distribution was 25-71 years (mean 44.4). Other patient characteristics are shown in table 2.

The process of protocol selection for these 20 patients involved 5400 evaluations of 272 criteria. Table 3 presents the evaluation results of these criteria.

The system selected 1 - 9 protocols per patient (3.05 protocols on average, overall 61 protocols were selected for 20 patients). None of the protocols evaluated to a score of 5 (definitely eligible) or 4 (probably eligible), 25 were graded 3 (possibly eligible), and 36 were graded 2 (low probability for eligibility).

Evaluation Result	Criteria Number (percent)
TRUE	2287 (42.04%)
FALSE	223 (4.10%)
UNKNOWN	2930 (53.86%)
true (inferred)	543 (9.98%)
false (inferred)	39 (0.72%)
unknown	2348 (43.16%)

Table 3: Results of 5440 evaluations of eligibility criteria.

The system's results were compared to the physician's selection of protocols in two aspects: the agreement on whether the patient is eligible for each protocol (Table 4), and the agreement on protocol ranking for each patient. The kappa statistic for appropriateness of protocols was 0.86 (95% CI 0.72 - 1.00). For 11 out of 20 patients (55%) both the system and the physician ranked the same protocol as first (kappa 0.37). The weighted kappa for ranking the protocols was 0.24.

		Physician Selection		
		Selected	Not Selected	Sum
System	Selected	59	2	61
	Not Selected	10	129	139
	Sum	69	131	200

Table 4: Selection of protocols by the system compared to a physician's selection.

DISCUSSION

Our results show that encoding and automatically evaluating eligibility criteria to find appropriate clinical trials for a specific patient is feasible.

We were able to encode 98.9% of the criteria, as compared to about 50% in the previous version of the system. This is the result of using an elaborated data model and standard vocabularies. Yet, we had difficulty encoding some of the ambiguous criteria that must involve human judgment.

The encoding language requires familiarity with the data model. Nevertheless, we share the vision that authors of clinical trial protocols will encode the criteria by themselves[16], and believe that it will be possible if a library of encoded criteria is provided.

Using terms from standard vocabularies is powerful in many aspects. It enabled us to simplify the data model and make it scalable. Thus, although the system is currently restricted to breast cancer protocols, it may be expandable to other domains.

Different approaches have been used in the past to handle missing data in evaluating eligibility for clinical trials. Tu[17] suggested combining qualitative and probabilistic approaches, while Papaconstantinou[8] used a probabilistic system in which the whole protocol is translated into a Bayesian network.

Our approach is somewhat different from the two mentioned in combining deterministic and probabilistic methods for inferring missing values. Deterministic inference involved, for example, deducing that a patient with metastases has a stage IV disease. Table 2 shows that 11 of the patients were known to have metastasis, while only 5 were known to have stage IV disease. Our system infers that patients with known metastases have stage IV disease (and vice versa). This kind of inference is crucial for appropriate selection of protocols, since we allow filtering out of protocols based on these inferred items.

In addition we modeled several small independent Bayesian networks that capture dependencies among different data items (e.g. liver diseases and liver function tests). Each variable in the network, which has a missing value, due to lack of patient data, has its value inferred by the Bayesian network. Evaluation of criteria that make use of these inferred values produces a qualitative estimate that the patient meets these criteria. Using small networks makes it relatively easy to build and expand them, and it might be simpler to find the needed prior and conditional probabilities that populate them.

The impact of the Bayesian networks was rather small. Although up to 20% of missing variables were inferred, it didn't have a major effect on ranking protocols (the system ranked the protocols the same when used without the Bayesian networks). We believe that this is the consequence of the paucity of patient data (as shown in table 3, more than 50% of criteria were evaluated to unknown). The impact of the Bayesian network will probably be higher if more data are entered into the system.

Our results show fairly good agreement between the system and a physician on protocol selection. It can potentially be a reliable means to select protocols. In this way, it can save practitioners a lot of time since many protocols are filtered out (more than 2/3 in our evaluation). We envision that such a system can be incorporated into the CII project of the NCI.

The agreement on ranking the protocols was much lower. Since the ranking process can be more subjective, these results are not surprising. As we lack a gold standard, we cannot decide which (system's or physician's) ranking is better. We plan to continue investigating this issue.

The study has several limitations that will direct our future work. Independent users will test the coding process, so we can learn about the applicability of the process.

The small number of encoded protocols limited the evaluation of the system. On the other hand, a larger number would probably be less manageable for evaluation by physicians. Since our conclusions are currently based on one physician, we plan to recruit more physicians to evaluate the protocols, some of whom will be domain experts and some general practitioners.

We plan to collect more data items, in particular temporal data, in order to test other aspects of the coded criteria. Finally, we plan to complete the user interface and evaluate the use of the system by practitioners and patients.

ACKNOWLEDGEMENTS

This project was funded by contract 34078PP1024 from the Massachusetts Department of Public Health, and grant DAMD 17-98-1-8093 from the Department of the army. Dr. Ash is also supported by the American Physicians Fellowship Committee (APFC) of the Israel Medical Association, and by Dr. Arthur Holstein's Fund of the American Jewish Joint Distribution Committee. We would like to thank Dr. Ronilda Lacson and Ms. Debra Della Torre for collecting the patient data.

REFERENCES

1. Silva, J and Wittes R, Role of clinical trials informatics in the NCI's cancer informatics infrastructure. Proc AMIA Symp, 1999;p. 950-4.
2. Silva, JS, Fighting cancer in the information age. An architecture for national scale clinical trials. MD Comput, 1999;16:43-4.
3. McCray, AT, Better Access to Information about Clinical Trials. Ann Intern Med, 2000;133:609-614.
4. Butte, AJ, Weinstein DA, and Kohane IS, Enrolling patients into clinical trials faster using real time Recruiting. Proc AMIA Symp, 2000; p. 111-5.
5. Carlson, RW, Tu, SW, Lane, NM, et al., Computer-based screening of patients with HIV/AIDS for clinical-trial eligibility. Online J Curr Clin Trials, 1995;Doc No 179
6. Breitfeld, PP, Weisburd, M, Overhage, JM, Sledge, G, and Tierney, WM, Pilot study of a point-of-use decision support tool for cancer clinical trials eligibility. J Am Med Inform Assoc, 1999;6:466-77.
7. Gennari, JH and Reddy M, Participatory design and an eligibility screening tool. Proc AMIA Symp, 2000:p. 290-4.
8. Papaconstantinou, CG, Theocharous G, and Mahadevan S, An expert system for assigning patients into clinical trials based on Bayesian networks. J Med Syst, 1998;22:189-202.
9. Ohno-Machado, L, Wang, SJ, Mar, P and Boxwala, AA, Decision support for clinical trial eligibility determination in breast cancer. Proc AMIA Symp, 1999:p. 340-4.
10. Physician Data Query (PDQ) - National Cancer Institute comprehensive database that contains cancer information summaries. http://cancernet.nci.nih.gov/trialsrch.shtml, Accessed 2/1/2001.
11. Ogunyemi O, The Guideline Expression Language (GEL) User's guide, Technical Report, DSG-TR-2000-001, 2000, Decision Systems Group, Boston, MA.
12. Ries LAG, Wingo PA., Miller DS, et al., The Annual Report to the Nation on the Status of Cancer, 1973-1997, Cancer, 2000.;88:2398-2424.
13. Neapolitan RE, Probabilistic reasoning in expert systems: Theory and algorithms. Wiley, New York, 1990.
14. Cozman, FG, JavaBayes - a system that handles Bayesian networks implemented in Java. http://www.cs.cmu.edu/~javabayes/Home/, accessed 2/1/2001.
15. Altman, D, Practical Statistics for Medical Research. CRC press, LLC, 1991:p. 403-409.
16. Rubin, DL, Gennari, JH,Srinivas, S. et al., Tool support for authoring eligibility criteria for cancer trials. Proc AMIA Symp, 1999: p. 369-73.
17. Tu, SW, Kemper, CA, Lane, NM, Carlson, RW and Musen, MA, A methodology for determining patients' eligibility for clinical trials. Methods Inf Med, 1993;32:317-25.