Decision Support for Clinical Trial Eligibility Determination in Breast Cancer

Lucila Ohno-Machado, MD, PhD, Samuel J. Wang, MD, PhD,
Perry Mar, Aziz A. Boxwala, MBBS, PhD
Decision Systems Group and Division of Health Sciences and Technology
Harvard Medical School and Massachusetts Institute of Technology, Boston, MA

ABSTRACT

We have developed a system for clinical trial eligibility determination where patients or primary care providers can enter clinical information about a patient and obtain a ranked list of clinical trials for which the patient is likely to be eligible. We used clinical trial eligibility information from the National Cancer Institute's Physician Data Query (PDQ) database. We translated each free-text eligibility criterion into a machine executable statement using a derivation of the Arden Syntax. Clinical trial protocols were then structured as collections of these eligibility criteria using XML. The application compares the entered patient information against each of the eligibility criteria and returns a numerical score. Results are displayed in order of likelihood of match. We have tested our system using all phase II and III clinical trials for treatment of metastatic breast cancer found in the PDQ database. Preliminary results are encouraging.

INTRODUCTION

Historically, accrual of patients for clinical trials has not been very successful, particularly for certain clinical domains. Studies demonstrate that just a small percentage of eligible patients (3 to 10%) are actually enrolled in such trials [1,2]. The low accrual rates are attributed to: (1) physician factors such as lack of knowledge about clinical trials, (2) patient factors such as lack of patient-oriented information regarding trials, (3) organizational barriers, and (4) health care system obstacles. If clinical trial information can be made more accessible to patients and their primary care providers (PCPs), we believe that clinical trial accrual rates can improve.

The increasing participation of patients in decisions regarding their own health has created a demand for health information resources oriented towards the patient and PCP, rather than the specialist [5]. A few systems have been previously designed to help with the determination of clinical trial eligibility. Tu et al. developed systems for this purpose, described in [6]. Ohno-Machado et al. previously developed a system that could reason under conditions of uncertainty [7]. However, these systems have focused on helping investigators identify eligible patients for a specific clinical trial. In contrast to these systems, the purpose of our system is to enable PCPs and patients to identify the best trials for a specific patient.

MATERIALS AND METHODS

Data. We used the National Cancer Institute's Physician Data Query (PDQ) database [8] as the source of information for clinical trials. The clinical trial summaries in the PDQ database contain free-text lists of eligibility criteria organized by patient characteristics (e.g., age, menopausal status); disease characteristics (e.g., histology, metastases); and prior and concurrent therapy. For the preliminary phase of this study, we selected from the PDQ database all Phase II and Phase III trials for the treatment of metastatic or recurrent breast cancer. Breast cancer was chosen because this is the oncology domain that contains the largest number of clinical trials. We chose advanced stage cancer because we hypothesized that these patients would be more interested in seeking participation in clinical trials after exhausting traditional treatment venues. We decided to limit our initial set to Phase II and Phase III trials since these studies are further developed, and typically involve several patients. We found a total of 85 clinical trials in the PDQ database (as of July 1998) that fit these parameters.

Clinical Trial Eligibility Database. Each clinical trial summary was encoded into a structured format. The encoded summary was stored in an XML document (Figure 1). This document contains elements describing identifying information about the clinical trial (name of trial, protocol number) and a collection of criteria elements. Each criterion element contains the original narrative text description from PDQ and the criterion encoded in a computable expression. The criterion is encoded in a modified version of the grammar used for specifying logic statements in the Arden Syntax [9]. Modifications had to be made to the Arden Syntax specification in order to accommodate a data model that contains hierarchical term relationships and compound data-types. (Details and discussion of our modifications to the Arden Syntax are presented elsewhere [10].) The resulting extended syntax for conditional expressions is also being incorporated into proposed extensions to GLIF, a clinical guideline interchange format developed by The InterMed Collaboratory [11].

1091—8280/99/$5.00

```
<PROTOCOL ID="09251">
 <NAME> Phase II Randomized Study of
Cyclophosphamide/Methotrexate/Fluorouracil (CMF)
vs Mitoxantrone in Elderly Patients with Advanced Breast
Cancer
 </NAME>

<!--Disease Characteristics-->
<CRITERION>
  No CNS metastases
  <SPEC>
    (metastases_locations where it is a "CNS") == []
  </SPEC>
</CRITERION>

<!--Patient Characteristics-->
<CRITERION>
 Over 70
  <SPEC>
    age &gt; 70
  </SPEC>
</CRITERION>

<CRITERION>
 Postmenopausal
   <SPEC>
     menopausal_status == "postmenopausal"
   </SPEC>
</CRITERION>

<!--Hematopoietic:-->
<CRITERION>
 WBC at least 3,000
   <SPEC>
    WBC &gt;= 3000
   </SPEC>
 </CRITERION>
```

Figure 1. Excerpt of clinical trial protocol structured in XML format.

The translation of the original free-text criterion descriptions from PDQ into a machine-interpretable representation was largely a manual process performed by informatics fellows and faculty in our laboratory. We used text parsing tools such as Perl scripts to automate portions of this process. We established a uniform basis for encoding criteria. For example, a certain clinical trial summary may have specified "estrogen receptor negative," and another may have specified "ER (-)." These refer to the same eligibility criterion and are encoded using the same expression ("estrogen_receptor == negative").

```
<!-- Patient Characteristics -->

<VARIABLE NAME='age' TYPE='number' CUI='C0001779'>
</VARIABLE>

<VARIABLE NAME='birthdate' TYPE='date' CUI='C0421451'>
</VARIABLE>

<VARIABLE NAME='gender' TYPE='enum' CUI='C0079399'>
Gender of patient
  <VALUE CUI='C0024554'>male</VALUE>
  <VALUE CUI='C0015780'>female</VALUE>
</VARIABLE>

<VARIABLE NAME='menopausal_status' TYPE='enum'
CUI='C0025320'>
Menopausal status of patient.
  <VALUE CUI='C0279752'>premenopausal</VALUE>
  <VALUE CUI='C0279753'>postmenopausal</VALUE>
</VARIABLE>
```

Figure 2. Excerpts from data dictionary containing definitions of clinical concepts used in the eligibility criteria.

In order to adequately model eligibility criteria, we found it necessary to create a data model that was sophisticated enough to accommodate hierarchical relationships among clinical concepts, sub-attributes of concepts, and temporal relationships among concepts. The concepts used in the eligibility criteria were defined in a data dictionary (also an XML document) (Figure 2), and mapped to concepts in the UMLS Metathesaurus [12]. We analyzed all the encoded criteria to assess which concepts occurred most frequently and were also relatively easy for the patient or PCP to obtain. This information was taken into consideration to construct web-based entry forms, shown in Figure 3.

Clinical Trial Ranking. Upon entry of patient data, the application produces a ranked list of clinical trials that the patient is eligible for. The ranking algorithm is tolerant of missing data. All criteria are considered as having equal weight (importance) when used in protocol ranking. The algorithm sequentially processes all the criteria in all the clinical trials. The algorithm first rules out all clinical trials for which at least one eligibility criterion was not met. For the remaining clinical trials, the ones that have fewest unknown criteria are placed higher on the list. Resulting trials are displayed with links to the original PDQ clinical trial summaries (Figure 4). The search can be refined with data entered in dynamically created forms (Figure 5). For each clinical trial, we also provide a summary of which criteria have been met and which still need to be evaluated (Figure 6).

Application. We are developing two versions of the application: one for the primary care provider and one for the patient. The version for the patient will provide a simplified user interface and will only request data that a patient would be expected to know. The application runs on the Microsoft Windows platform. HTML pages are dynamically generated on the server using Microsoft's Active Server Pages (ASP). The application logic was written in Visual C++ and wrapped as an ActiveX object that is invoked by ASP.

RESULTS

A total of 2188 criteria in the set of 85 clinical trials were chosen for this study. In this set, the least, most, and median number of criteria in a protocol were 6, 45, and 25 respectively. To date, we have encoded about 50% of the criteria in these clinical trials. We are first encoding frequently occurring criteria and those that are readily accommodated by the criteria representation syntax. (See [10] for details on difficulties encountered in encoding the eligibility criteria.) Figures 3 to 6 show an example of the PCP version of the application for a sample breast cancer patient: a premenopausal, 55 year-old woman with stage IV breast cancer with metastases to liver and bone, previous mastectomy, chemotherapy and

radiotherapy. This patient also suffers from coronary artery disease and diabetes mellitus. Figure 3 shows the initial data input form in which the PCP has entered some clinical information about the patient. Using this information, the program returns a preliminary list of trials. This list is ranked, with the most likely matches at the top (Figure 4).

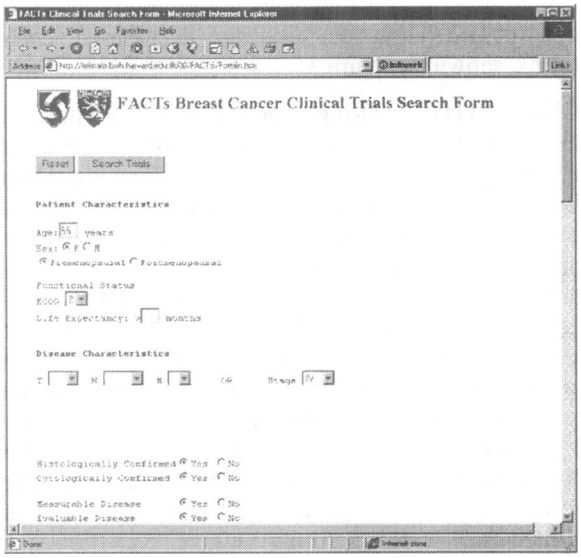

Figure 3. The initial entry form requests items that are most frequent and easiest to obtain.

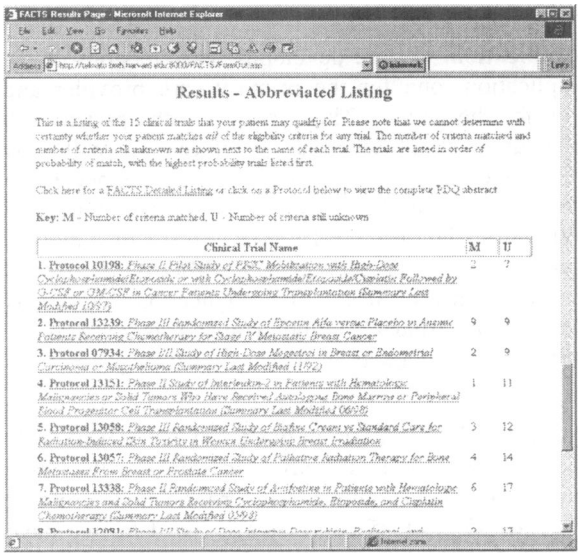

Figure 4. Results page showing a ranked list of clinical trials.

If the list is long, the application offers the PCP an opportunity to fill in additional patient information to narrow the search. The program dynamically constructs the secondary input form to request the information that would be more likely to narrow the number of clinical trials (Figure 5). Again, the PCP fills in as much additional information as he or she can. This process can be repeated as many times as desired until either the resulting list is short enough, or there is no additional information required or available.

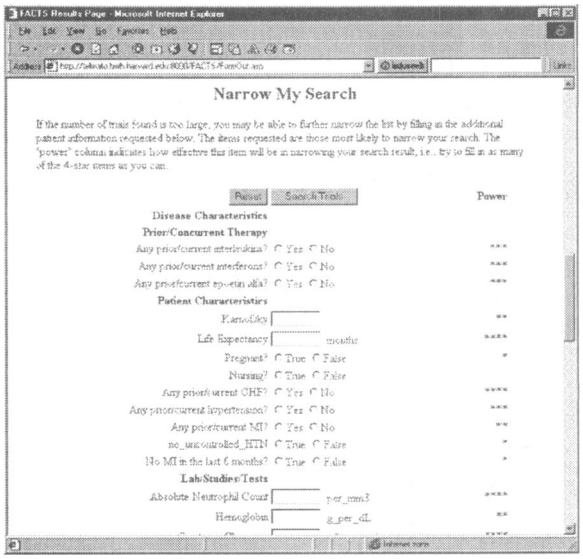

Figure 5. Secondary entry forms are created dynamically and request information that will be most useful in narrowing the search.

The final list is presented in order of likelihood of match. In this example, the system narrowed the list to 15 trials that the patient is potentially eligible for. A summary of all the entered information is provided. Detailed information about these clinical trials (Figure 6) can be displayed, along with a list of the criteria still to be checked.

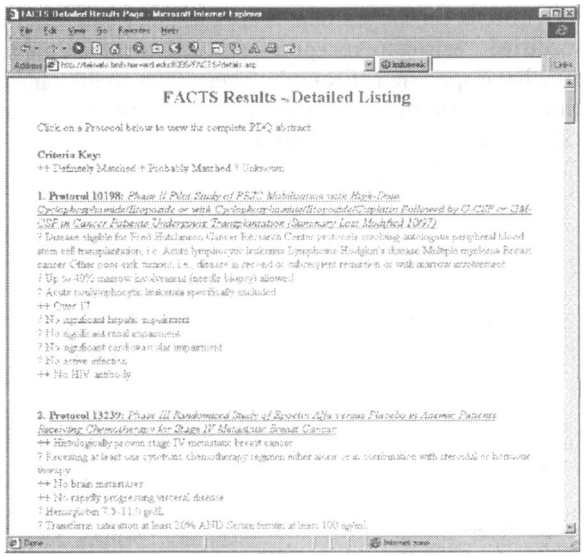

Figure 6. Detailed information about remaining trials is displayed.

Discussion and Future Directions

The current ranking algorithm makes two simplifying assumptions: (1) all criteria have equal importance and equal probability of being met if their values are

unknown, and (2) all criteria are independent. Regarding the first assumption, a more accurate approach would be to assign a weight to each criterion or data item, and then use these weights to compute the ranking. We may be able to obtain these weights by asking domain experts, from the literature, or by analysis of large patient data sets. Tu [6] has proposed that some criteria variables are mutable over time (e.g., age) or controllable (e.g., stop current chemotherapy), and therefore might bear less weight in ruling-out or ranking one clinical trial against others. We have not decomposed criteria into "atomic" parts, each containing just one variable, hence this approach has not been yet tested.

The other simplifying assumption, criteria (and data item) independence, also introduces inaccuracies in ranking. For example, a clinical trial may specify two separate data items for the liver function tests, AST and ALT: "AST < 2 times normal" and "ALT < 2 times normal." These criteria are currently considered independent, when in fact a better approximation would be to consider them just *conditionally* independent given a certain liver disease. For example, if AST is high, there is an increased probability that ALT is high because the disease that causes the former to increase is also likely to cause the latter to do so. The independence assumption causes some criteria to be unfairly "counted twice." A more accurate approach would be to identify dependencies among the data items and adjust the scoring accordingly. In this version of the application, we considered all criteria to be Boolean (i.e., "true" or "false"), and have not further characterized their nature.

The current clinical trial selection algorithm is deterministic. We have not attempted to deal with uncertainty using probabilities in this prototype. A global model to infer the value of missing values for common criteria and specification of criteria dependencies will be built using expert knowledge. This model will be based on a belief network, the structure and probabilities of which will be extracted by interviews with specialists, analysis of literature, or "learned" from clinical databases. A future version of this system will take into account "proxies" for certain criteria (e.g., known renal disease as a proxy for laboratory values that measure renal function, or "severity of cancer" as a proxy for staging). The probabilities of eligibility will be determined by inferencing values for required data from the proxies.

Other prototype applications have been built with the assumption that certain medical domains may require very few eligibility criteria to reasonably eliminate a large percentage of the candidate trials for a given patient [13]. In contrast, our approach has been to attempt to encode as many criteria as we reasonably can in an attempt to arrive at a more accurate list of potentially matching clinical trials. However, it is difficult to algorithmically determine eligibility with 100% accuracy because of the clinical judgement that is necessary for evaluating several of these criteria. Our objective is to narrow and rank the list of matching trials, as much as possible, before turning the list over to a specialist for final determination of eligibility. Encoding complex criteria is a time-consuming effort. Although we have developed some automated parsing tools to facilitate this task, it remains a largely manual process. We predict that our application will perform better as we encode more criteria. However, an open question that deserves further study is how much encoding is "enough," i.e., at what point is it not cost-beneficial to encode more complex criteria. Since software applications cannot determine clinical trial eligibility with 100% certainty, it may not be worth the extra effort to encode very complex criteria.

The criteria encoded for this study were taken from clinical trial summaries from PDQ. These summaries are abstracted from the original protocol documents and may lose some fidelity in the process. Our encoding is only as good as the translated text descriptions. For improving accuracy, an alternative approach would be to go directly to the original full research clinical trial descriptions to obtain the eligibility criteria. The future development and routine use of computer-based protocol authoring tools may reduce these problems.

Currently, we have not taken into account patient preferences in ranking the clinical trials, such as modality of treatment, potential toxicity, potential for cure, and geographic constraints. The system currently ranks trials solely based on the likelihood that the patient will satisfy the eligibility requirements. It is a very different question to ask what types of trials a patient may prefer. While eligibility criteria are obviously a firm prerequisite to enrollment, in cases with incomplete information, there may be some benefit to introducing patient preferences even before eligibility has been completely determined. This could help narrow the list more quickly so as not to waste the patient's or clinician's time in reviewing eligibility requirements for trials that the patient would never consider enrolling in.

We plan to automatically retrieve some of the required patient data from the clinical information system at our institution in order to ease the data entry burden on the user. The user will only need to provide information not available in the clinical system. For the institutional version, we will link the

eligibility component to other tools that automate the enrollment process, such as display of informed consent forms, and detailed explanation of the clinical trials. A more general version of the application will be available on the WWW. In addition to UMLS, we also plan to map the concepts used in our system to the Common Data Elements (CDE) that are being developed under the supervision of the informatics group at the National Cancer Institute [14]. Mapping to the CDE will make the system more robust for national scale use. The open architecture and facility to add customized dictionaries will also make it easy to adapt the system for integration to electronic medical record systems of different institutions.

This initial version of the application has been designed for use by PCPs. For the patient version, we intend to customize the user interface according to different levels of user sophistication. The user interface will be designed in consultation with patient advocacy groups, health educators, and PCPs. Reduction and simplification of data items to be entered is necessary. We will utilize a decision analytic approach to determine the data items needed.

CONCLUSIONS

We have developed a WWW-based decision support system to help patients and providers determine the patient's eligibility for certain clinical trials. The system currently contains all Phase II and III treatment clinical trials for metastatic breast cancer from the NCI's PDQ database. It rules out trials that the patients are not eligible for and ranks the remaining trials according to how many criteria still need to be checked to determine eligibility. This initial prototype system has helped us identify relevant issues in machine-readable criteria representation, user interface design, and clinical trial ranking under uncertainty. Preliminary testing of the system with a few clinical cases has been promising. A formal evaluation of usability and reliability is underway. Future versions of this application will include a belief network that will allow the system to impute missing data values and reason under conditions of uncertainty.

ACKNOWLEDGMENTS

This project was funded by contract 34078PP1024 from the Massachusetts Department of Public Health and grant DAMD17-98-1-8093 from the Department of the Army. The contents of this article do not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. Dr. Wang was funded by NLM grant 2 T15 LM07092. We would like to thank Jeremy Theal, Dr. Jeff Huhn and Dr. Ross Martin for helping to encode the eligibility criteria. We also thank Dr. Ursula Matulonis, Dr. Craig Bunnel, and Dr. Darrel Smith for sharing their expertise in breast cancer. Mr. John Ehresman implemented parts of this work. Prof. Robert Greenes provided valuable suggestions to this project.

REFERENCES

[1] Mansour EG. Barriers to clinical trials. Part III: Knowledge and attitudes of health care providers. *Cancer*, 1994, 74(9 Suppl):2672-5.

[2] Winn RJ. Obstacles to the accrual of patients to clinical trials in the community setting *Seminars in Oncology*, 1994, 21(4 Suppl 7):112-7.

[3] Taylor KM; Margolese RG; Soskolne CL. Physicians' reasons for not entering eligible patients in a randomized clinical trial of surgery for breast cancer. *New England Journal of Medicine*, 1984, 310(21):1363-7.

[4] Klabunde C; Kaluzny A; Ford L. Community Clinical Oncology Program participation in the Breast Cancer Prevention Trial: factors affecting accrual. *Cancer Epidemiology, Biomarkers and Prevention*, 1995, 4(7):783-9.

[5] Sweeney MA; Skiba D. Combining telecommunications and interactive multimedia health information on the electronic superhighway. *Medinfo*, 1995, 8 Pt 2:1524-7.

[6] Tu SW; Kemper CA; Lane NM; Carlson RW; Musen MA. A methodology for determining patients' eligibility for clinical trials. *Methods of Information in Medicine*, 1993, 32(4):317-25.

[7] Ohno-Machado L, Parra E, Henry SB, Tu SW, Musen MA. AIDS2: A decision-support tool for decreasing physicians' uncertainty regarding patients eligibility for HIV treatment protocols. *Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care,* 429-33, 1993.

[8] NCI. http://cancernet.nci.nih.gov/.

[9] American Society for Testing and Materials. E 1460 Standard Specification for Defining And Sharing Modular Health Knowledge Bases (Arden Syntax for Medical Logic Modules). ASTM Standards, v 14.01. Philadelphia: ASTM, 1992; 539-87.

[10] Wang SJ, Ohno-Machado L, Mar P. Representing Criteria in Guidelines and Clinical Trial Protocols: Common Needs and Solutions. *Decision Systems Group, Brigham and Women's Hospital, Technical Report* 1999-04.

[11] Ohno-Machado L, Gennari JH, Murphy SN, et al. The guideline interchange format: a model for representing guidelines. *J Am Med Inform Assoc.* 1998;5:357-72.

[12] Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. Bull Med Libr Assoc 1993;81(2):170-7.

[13] Gennari J. personal communication, 1998.

[14] NCI Cancer Informatics Infrastructure Home Page. http://hiip-wkstn.hpc.org/.