Using Computer Modeling to Help Identify Patient Subgroups in Clinical Data Repositories*

Gregory F. Cooper[1], M.D., Ph.D., Bruce G. Buchanan[2], Ph.D.,
Mehmet Kayaalp[1], M.D., M.S., Melissa Saul[3], John K. Vries[4], M.D.
[1]Center for Biomedical Informatics, University of Pittsburgh
[2]Department of Computer Science, University of Pittsburgh
[3]University of Pittsburgh Medical Center Health System
[4]University of Pittsburgh School of Medicine

OBJECTIVE: The ability to accurately and efficiently identify patient cases of interest in a hospital information system has many important clinical, research, educational and administrative uses. The identification of cases of interest sometimes can be difficult. This paper describes a two-stage method for searching for cases of interest.
DESIGN: First, a Boolean search is performed using coded database variables. The user classifies the retrieved cases as being of interest or not. Second, based on the user-classified cases, a computer model of the patient cases of interest is constructed. The model is then used to help locate additional cases. These cases provide an augmented training set for constructing a new computer model of the cases of interest. This cycle of modeling and user classification continues until halted by the user.
MEASUREMENTS: This paper describes a pilot study in which this method is used to identify the records of patients who have venous thrombosis.
RESULTS: The results indicate that computer modeling enhances the identification of patient cases of interest.

INTRODUCTION

A hospital information system (HIS) can greatly facilitate retrospective studies of patient cases. With such a system, researchers also can identify potential participants in prospective clinical studies. Both uses of a HIS require the identification of patient subgroups of interest. The more accurately and efficiently such subgroups of patients can be identified, the better.

The accurate and efficient identification of patient subgroups of interest extend beyond research studies to include clinical, educational and administrative uses. For example, one educational application would be the identification of interesting patient cases for use in teaching medical students.

If a patient subgroup can be defined by a relatively small and simple combination of coded data fields in a HIS (e.g., ICD-9 codes, laboratory values, and patient demographic variables), then finding the relevant hospital records is relatively straightforward. The task becomes considerably more challenging, however, when a simple Boolean combination of data fields does not readily locate all the patient cases of interest. In such situations, it may be necessary to perform more complex searches that involve coded information and possibly unstructured information (e.g., free-text history and physical examinations, radiology reports, and discharge summaries). This paper describes a semi-automated technique that is intended to assist users in performing these more complex searches.

BACKGROUND

We are using the MARS (Medical ARchival System) hospital information system at the University of Pittsburgh Medical Center [1], which contains a rich store of both coded and free-text clinical information. This section describes a current method that is commonly used in identifying patient subgroups in MARS.

Patient subgroups are often identified in MARS as a team effort by MARS staff and users. We characterize here a typical collaboration. The user (e.g., a researcher) first describes the target population to the staff member, and through a discussion, they refine the description to construct an initial MARS query. Typically the staff and user first identify a set of patient *reference cases*, which are patient records in MARS that match the study criteria. These cases often are obtained using a Boolean search of MARS. The reference cases are reviewed by the staff member to get an idea of the search terms to use in finding additional relevant records, which we call *record matches*.

After developing a refined search query in consultation with the user, the staff member performs the search and reviews the results with the user. If there are too many records to review exhaustively, then a sample of the records is reviewed. Through this process, a set of record matches is assembled. A review of these records often provides additional hints about how to further refine the search query, which is run, and the cycle repeats.

The search and review cycle terminates when the user is satisfied with the set of record matches that have been found. The union of the reference cases and all record matches constitute the set of records that define the patient cases of interest to be used in further analyses.

* To appear in the Proceedings of the 1998 AMIA Fall Symposium (formerly SCAMC).

There are two labor-intensive aspects to the process of constructing a patient subgroup of interest: (a) constructing complex search criteria beyond the initial Boolean search, and (b) reviewing the retrieved results of searches to identify record matches and refine the queries further. We propose a method to facilitate these two steps for the staff and user team, which for brevity we generically call *the user*. The current paper focuses primarily on the first step.

COMPUTER-ASSISTED IDENTIFICATION OF PATIENT SUBGROUPS

In this section, we provide an overview of our basic approach to using computer modeling to help identify patient subgroups in MARS. Figure 1 provides a schematic summary of the approach. We currently are in the process of implementing and integrating all the steps in this approach.

The reference cases hold important clues to the patient features that distinguish the patient subgroup of interest from the other patients with records in MARS. We are investigating the use of computer modeling methods to characterize explicitly those distinguishing clues. The reference cases are used as positive instances in a training set. Negative instances are either provided directly by the user or are randomly selected from the entire set of MARS records if it is highly unlikely that a random hospital record satisfies the study criteria. The training set containing positive and negative instances is used to construct a computer model. The variables in the model can represent both coded and free-text patient data, although initially we are using only coded data.

The computer model contains a set of patient features (clues) that characterize approximately the patient subgroup of interest. Conceptually, each patient record in MARS is compared to the model to determine a numeric score that indicates how closely the record and the model match (in practice, the implementation can be more efficient computationally). The most closely matching records are presented to the user, sorted in descending order of their scores. As outlined in Figure 1, the user reviews records in the order they appear in the sorted list, and in the process indicates record matches, until the yield of matches becomes so low that he or she stops the review. To help maintain confidentiality, these records contain no explicit patient or clinician identifiers (e.g., name, address, social security number); currently we perform this de-identification process manually, but we plan to investigate methods for automating it. We hypothesize that (a) this selective review of the records will be significantly faster than the current more manual process of reviewing records, and (b) more record matches will be found using the system than are found using current methods. The current paper describes a preliminary study that provides a step toward addressing these hypotheses.

The records that are marked by the user as matches or non-matches become part of a growing training set of patient records. The augmented training set is used to learn a new, refined computer model, and the cycle of search and review continues. Ultimately, the user will decide to terminate the search and review process. At that point the union of the reference cases and all record matches become the study set that is exported for use.

RELATED RESEARCH

Conceptually, the basic methodology presented in the current paper is closely related to relevance feedback research in information retrieval [2]. In relevance feedback, a user performs an initial search and classifies a sample of the retrieved documents regarding whether or not they are of interest. This feedback is used to re-weight and possibly expand the search terms that are then used in subsequent retrievals. Other related research is being pursued by machine-learning researchers [3].

The line of research we are pursuing is distinguished along two basic ways from most of the prior work on relevance feedback. First, we are interested in using a highly heterogeneous collection of information in the electronic medical record, including free text of several types (e.g., dictated history and physical examinations, surgical pathology reports, and discharge summaries) as well as a variety of types of coded information (e.g., demographic information, patient charge codes (see below), and laboratory results). The current paper focuses on learning models using patient charge codes; it serves as a baseline study for future extensions that involve modeling with additional information in the electronic medical record. Second, we are interested in applying a wide variety of statistical and machine-learning methods for modeling. Some of these methods (e.g., neural networks and Bayesian networks) are able to model multivariate interactions among features. By using a simple Bayes classification model, the current paper serves as a baseline for planned investigations that use more sophisticated modeling techniques.

EXPERIMENTAL METHODS

This section describes a pilot study that involves the identification of those patients in the intensive care unit with a venous thrombosis (*VT*). The primary goal of the original study was to eventually locate just those patients with a deep venous thrombosis.

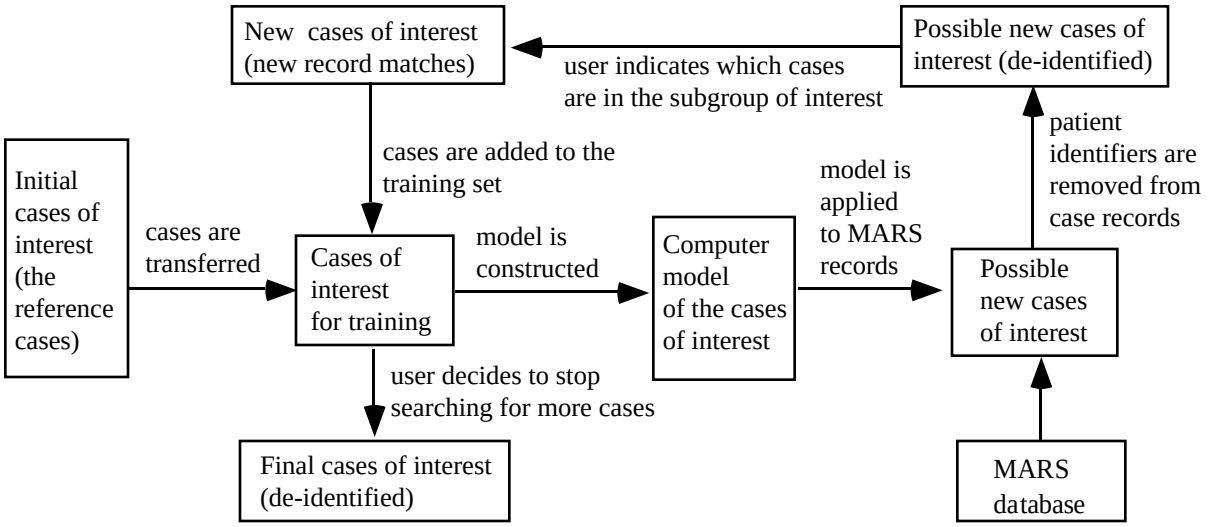

Figure 1. An outline of a method for identifying patient subgroups that uses computer modeling.

We investigated an approach in which initial cases of *VT* are obtained by a simple search of coded features of the total study population. The question is whether there are other cases of *VT* that computer modeling can help a user efficiently locate.

The total study population consists of all patients admitted to two medical ICUs at the University of Pittsburgh Medical Center Health System (UPMC-HS) between January 1, 1993 and December 31, 1995. There were a total of 3020 patients. Of these, 124 patients were identified as having *VT*, based on being assigned one or more of the following ICD-9 codes at discharge: other venous thrombosis (ICD-9 code 453), Budd-Chiari syndrome (453.0), thrombophlebitis migrans (453.1), vena cava syndrome (453.2), renal vein thrombosis (453.3), venous thrombosis nec (453.8), and venous thrombosis nos (453.9). An ICU specialist reviewed the MARS patient records and verified that the records support a *VT* diagnosis.

In the current experiment, we randomly divided the 124 *VT* cases into two parts. A total of 74 cases were placed in a training dataset. For the purpose of our study, we imagine that this dataset contains all the cases found by a Boolean search of the ICD-9 codes; suppose, for example, that the ICD-9 coding of records had been less complete, and only 74 cases were coded as *VT*. We use the remaining 50 cases as a test dataset. We investigate to what extent computer modeling can help efficiently locate some or all of those 50 cases without using ICD-9 codes; so, for example, if the ICD-9 coding had been incomplete, we are investigating the extent to which the uncoded 50 cases could be located by means other than the ICD-9 codes.

Of the 3020 ICU patient cases, we randomly selected 1802 cases for use in preliminary experiments that investigated model construction methods. To avoid testing bias, in performing the final experiment reported here, we used only the remaining 1218 cases (3020 - 1802). These 1218 cases contained 50 *VT* cases, as mentioned previously; none of these 50 cases appeared in the 1802 cases used in preliminary experiments. The fraction of *VT* cases in the entire set of 3020 is approximately 4 percent. Likewise, the fraction of *VT* cases in the set of 1218 cases is approximately 4 percent; thus, our evaluation dataset of 1218 cases provides an unbiased sample of the original patient population of 3020 cases.

We used patient charge codes as the features (variables) with which to construct computer models of *VT* cases. A charge code is a single item charged to a patient during his or her hospitalization. For example, the code 33001371 is used to represent a charge for "warfarin sodium 5 mg tablet". There are approximately 45,000 distinct items that can be charged at UPMC-HS. Among the 3020 cases in our study database, only 7035 distinct charge codes appeared. Since charge codes can be very specific, and thus lead to small sample sizes, we augmented these 7035 charge code features with 1081 abstract features. In the current pilot study, an abstract feature was created for each unique first word within the charge code descriptions; as an example, the abstract charge code *heparin* corresponds to the disjunction of all charge codes whose descriptions begin with the word "heparin". For the current study, we choose to use charge codes because (a) they represent many of the actions taken in caring for patients, and thus, are a rich source of clinical information, and (b) they are sufficiently voluminous and complex in their organization that using them in a structured Boolean search would be difficult for most users. We represented a given charge code as a binary variable. The value of a charge code is set to *true* if the given

charge was made at any time during a patient's hospitalization; otherwise, the value was set to *false*.

We used a simple Bayes system as a statistical model, because it is simple to code, efficient to learn and apply, and it often performs well in practice. A simple Bayes system assumes that features (e.g., charge codes) are independent conditioned on a diagnosis (e.g., *VT*). Although this assumption usually is not valid, simple Bayes models have been shown to be robust to violations of the assumption when the models are used for classification [4]. That is not to say that improvements cannot be made by using more sophisticated models, which we intend to do in future research.

Suppose that d represents a diagnosis or prediction (e.g., *VT*). Let d' denote that d is bound to some particular value (e.g., $VT = present$). Let f_i denote a feature (e.g., a charge for a 5 mg warfarin tablet) that is bound to some particular value (e.g., either bound to the value *true* or *false*). Let F denote a set of n features, each of which is bound to some value. A simple Bayes model computes the probability of d' given F as follows:

$$P(d' \mid F) = \frac{P(d')\prod_{i=1}^{n} P(f_i \mid d')}{\sum_{d}(P(d)\prod_{i=1}^{n} P(f_i \mid d))} \quad (1)$$

where n is the number of features in the model and the sum is taken over all the values of d (e.g., d = *present* and d = *absent*). For any specific values of f_i and d, we estimated the term $P(f_i \mid d)$ as $(freq(f_i, d) + 1)/(freq(d) + 2)$, where $freq(f_i, d)$ is the number of times that f_i and d co-occur in the training database. Similarly, we estimated $P(d)$ as $(freq(d) + 1)/(N + 2)$, where N is the total number of cases in the training database. When the sample size is small, these estimates are less extreme than are maximum likelihood estimates based on the simple ratios $freq(f_i, d)/freq(d)$ and $freq(d)/N$.

When there are a large number of features, simple Bayes models often have better predictive performance if feature selection is applied. Thus, rather than use many thousands of charge codes in Equation 1, a search method is applied to select a relatively small number of features that are highly predictive of the diagnosis. We used a simple feature selection method: in Equation 1 we only used those features with the value *true* in at least one case in the training set.

Initially (cycle 0) we started with a training set containing 74 cases of patients known to have had *VT* and 74 cases known not to have had *VT*. These constitute the *reference cases* denoted in Figure 1. Following the figure, a simple Bayes model was constructed from those 148 cases. This model was applied to the 1218 cases in the test set and a probability of *VT* was computed for each case using Equation 1. The probabilities were sorted in descending order. We assumed that a user would be willing to review the 40 most probable *VT* cases and classify each case as *VT* or not. In actual practice, such a review would involve the user examining the MARS record of each of the 40 patients to assess whether the patient had experienced *VT*. Such an examination could involve looking at structured information (e.g., laboratory values) or free text (e.g., discharge summaries). In our experiments we simulated the assessments that a user would give. We were able to do so because we assume that among the 3020 cases the only ones with *VT* were those 124 cases described above.

We repeated this cycle of 40 patients for 10 cycles. We stopped after 10 cycles because it seems unlikely that a user would be willing to review more than 400 cases.

RESULTS

Figure 2 shows the key results. The *x* axis represents the number of cycles taken in Figure 1. In a single cycle, the 40 most probable *VT* cases (outside the training set) are presented to the user for classification. The *y* axis shows the total number of *VT* cases found, which is equal to the 74 reference cases plus the new record matches as verified by the user.

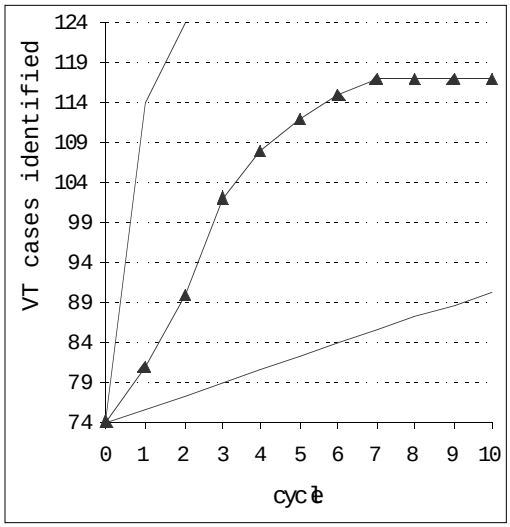

Figure 2. A plot of the total number of *VT* cases located as a function of the modeling cycle.

The upper line in Figure 2 is the maximum performance that could be achieved; since only 40 cases are selected in each cycle, the maximum number of cases that can be identified just after the first cycle is 40 + 74 = 114. The line containing triangles

shows the performance of applying the method described in the previous section. The lower line is the performance that would be expected when randomly selecting 40 patient cases (from outside the training set) during each cycle; this performance corresponds to randomly searching through the ICU patients in search of *VT* cases beyond the original 74.

DISCUSSION

The results in Figure 2 indicate that computer modeling is helping locate *VT* cases that are not included in the original set of 74 reference cases. In the first cycle, 7 new *VT* cases are included among the 40 cases presented to the user. Based on *VT* having a prior probability of 4 percent in this dataset, we would expect to find only 1 or 2 *VT* cases in a random sample of 40 cases. The performance of computer modeling continues at about the same rate, until the fourth cycle. A plausible interpretation is that as more and more *VT* cases are discovered, the remaining cases become relatively fewer and are therefore more difficult to locate; thus, the slope begins to flatten at about the fourth cycle. Also, the most apparent *VT* cases are likely to be located in the early cycles, leaving the more difficult cases for later cycles. Another factor is that we presently are using only charge codes as model features. We intend to expand the features modeled to include demographic information, ICD-9 codes, and DRG codes; we then plan to investigate the extent to which these additional features change the system's performance.

In the present study, we have assumed that all 2896 (i.e., 3020 - 124) cases that are not known to be *VT* are in fact not *VT*. As a check on that assumption, we plan to have an ICU specialist examine cases of *VT* that are predicted by the system and that are not among the 124 cases already established to be *VT*. We are interested in the fraction of those cases that the specialist designates as being *VT* cases. The larger the fraction the better. As a supplementary check, we also plan to select a feasible-sized random sample of the 2896 cases and present these to the specialist for evaluation. From the fraction designated by the specialist as *VT* cases we can estimate the total number of *VT* cases within the entire dataset of 3020 cases.

In the near term, we also plan to expand our investigation to the identification of other patient subgroups. In the longer term, our plans include the representation of features that are extracted from free text in the medical record (e.g., the discharge summary). Our emphasis in representing free text will be to explore relatively simple methods that will be computationally tractable on a database the size of MARS. For example, one basic approach is to map free-text phrases into UMLS Metathesaurus concepts using available mapping tools. Each identified concept represents one variable that characterizes a part of the free-text report. We also will explore a wide variety of statistical and machine-learning models, including neural networks, Bayesian networks, and rule-based systems. We plan to provide a web interface that will allow a user to perform an initial set of Boolean queries on coded features, then extend the search for patient cases using computer modeling. The interface will permit the user to examine and directly alter the model, if desired. An audit trail will be maintained of all user decisions regarding classification of cases in order to document for each case the user's justification for it being of interest or not.

When the overall search system is sufficiently mature, we plan to perform a randomized, control experiment with a wide variety of users and search tasks in order to compare the computer modeling approach to the current approach. The primary metrics we anticipate using are the total patient cases of interest that are found using each approach (case capture), the total amount of user and staff time expended in obtaining those cases, and the users' assessments of the utility of the case capture given the total expended time.

Acknowledgments

We thank Dr. Michael Donahoe for his help in defining the dataset that we used in the pilot study reported here. We thank Dr. Charles Friedman for helpful comments on the design and development of this project. This research was supported in part by IAIMS grant 1-G08-LM06625 from the National Library of Medicine to the University of Pittsburgh.

References

1. Yount RJ, Vries JK, Councill CD. The Medical Archival System: An information retrieval system based on distributed parallel processing. Information Processing Management 1991;27:379.

2. Harman D. Relevance feedback and other query modification techniques. In: Frakes WB, Baeza-Yates R, eds. Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice Hall, 1992: 241-263.

3. Nigam K, McCallum A, Thrum S, Mitchell T. Learning to classify text from labeled and unlabeled documents. National Conference on Artificial Intelligence (AAAI), 1998.

4. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 1997;29:103-130.