Estimating Frequency of Disease Findings from Combined Hospital Databases: A UMLS Project

Lee-Shing Fu[1], Stan Huff[1], Omar Bouhaddou[1,3], Bruce Bray[2], Homer Warner[1]

[1] Department of Medical Informatics,
[2] Division of Cardiology,
University of Utah, School of Medicine, and
[3] Applied Informatics, Inc.
Salt Lake City, Utah

Abstract

Merging data from the Salt Lake VA hospital database and the LDS hospital HELP system into a UMLS sponsored unified patient database has demonstrated that distribution of variables within a disease is hospital independent. Although disease prevalence is clearly not the same among hospitals, analysis of data within a disease group across hospitals can be done using such a merged database. This unified patient database would allow study of unusual diseases not possible using data from a single institution.

Introduction

Obtaining useable statistical estimates about uncommon medical events is often not possible using data from a single institution. The development of a system for combining data from multiple sources could be a means for solving this problem if indeed such merging of data is valid. In previous papers [1, 2], we presented a model for a Unified Patient Database (UPD). It is the purpose of this project to present a test of the homogeneity of value distributions of selected medical variables within the UPD.

As part of its UMLS project [3, 4], the National Library of Medicine is developing tools such as the metathesaurus, a semantic net, and an information sources map to facilitate access to medical knowledge. The primary focus has been on the published medical literature, but it is also recognized that data recorded in the patient medical record is an important source of information that must be indexed using these same tools. The principle effort of the Utah component of the UMLS project has been to develop a prototype UPD. The prototype UPD developed to date has demonstrated that it is possible to represent clinical events from different patient databases using a unified frame structure or Event Definition [2]. We now present results obtained with the UPD prototype to demonstrate that data from electronic patient records from two or more institutions can be combined to obtain valid estimates of the frequency of findings in a disease.

First, we briefly describe the components of our Unified Patient Database. A detailed description was made at last year's SCAMC conference [2]. The prototype UPD was populated from subsets of patient cases from LDS hospital and VA hospital for five diseases and 35 medical variables. Using graphical displays and statistical tests (i.e., analysis of variance and t-test), we then show that the distribution of each variable is similar for the two hospitals within the disease groups. Finally, we discuss the problems and potential usefulness of implementing a working system, that is a system that contains enough data to evaluate and demonstrate the usefulness of a pooled patient database for various applications in clinical medicine and health care systems research such as providing statistical data for knowledge engineering.

Sources of Data: LDS and VA Hospitals

Two hospital clinical databases were sampled to populate the prototype UPD and test the hypothesis that clinical data can be combined: LDS and VA hospital in Salt Lake City. Data obtained from the LDS hospital are in the HELP clinical database [5]. The electronic record of most patients do not have history and physical exam data, but all have laboratory results and discharge diagnoses. In HELP, all computerized data is coded using a large clinical vocabulary [6], much of which has been incorporated into the metathesaurus (Meta-1) of the UMLS. The VA hospital information system (DHCP) has become the standard for almost all VA hospitals. Utah is a regional data center for development of tools for integration of data among VA hospitals in our area. Potentially 170 VA hospitals can supply patient cases to the UPD.

Methods

The building components of the prototype unified patient database are 1) the event definitions, 2) the master object index (MOI) file, and 3) the event definition instantiations built for each dictionary source (i.e., HELP, DHCP). First, event definitions are descriptions or templates of the

0195-4210/91/$5.00

structure of clinical data as it resides in a database. Examples of two event definitions for laboratory results and diagnoses are shown in Figure 1 and Figure 2 respectively. Second, the MOI file contains the list of all vocabulary terms compiled from the sources and are used to fill the slots in the event definitions. The UMLS metathesaurus will ultimately serve as the MOI of the UPD. Third, instances of event definitions are built automatically to represent patient parameters transferred from each source to the UPD. For example, serum creatinine kinase (CK) and ferritin will be instances of the laboratory event definition while acute MI and iron deficiency anemia are instantiations of the diagnosis event definition.

Description: Used for recording diagnoses.
Event Identifier: 5
Body:
- (A) Name, Diagnosis, Syndrome, Problem (CHF, alcohol, pneumonia, ICD-9 terms)
- (B) Etiology (M. tuberculosis, pneumococcus)
- (C) Anatomy (body parts)
- (D) Severity (mild, moderate, final stage)
- (E) Chronicity (recent, remote, chronic)
- (F) Probability (definite, support, probable)
- (G) Source/Method (attending MD, pathologist, consultant, ILIAD, QMR)
- (H) Type of Source (discharge, admitting, surgical, autopsy, other)
- (I) Primary Dx (yes/no flag)

Figure 1: The event definition for diagnosis events.

Description: Used for recording laboratory results data.
Event Identifier: 3
Body:
- (A) Test Name (SMAC7)
- (B) Time Recorded (13:04,11/03/87)
- (C) Time Collected (10:00,11/01/87)
- (D) Specimen Type (serum, blood, sputum, urine, bone marrow)
  - (D.1) Specimen Type Modifier (random, peak, trough, 24 hour collection)
- (E) Coded Comment (specimen hemolyzed)
- (F) Result Name (Na, K, Ca, rbc,wbc, plt)
  - (F.1) Numerical Value (numerical, 10000, 13.2)
    - (F.1.1) Comparator (greater than, < , is)
    - (F.1.2) Units (mg/ml, count)
  - (F.2) Coded Value (positive, negative, 2+, abnormal)
  - (F.3) Titer Value (1:20)
    - (F.3.1) Comparator (greater than, < , is)
  - (F.4) Hi/Low Flag (hi, low, normal, panic hi, panic low, abnormal)
  - (F.5) Delta Check Flag (lower, higher)
  - (F.6) Coded Comment (additional dilution required)
  - (F.7) High Reference (mean + 2*SD)
  - (F.8) Low Reference (mean - 2*SD)
  - (F.9) Normal Control
    - (F.9.1) Comparator (greater than, < , is)
    - (F.9.2) Units (mg/ml, count)
  - (F.10) Trend (increased)
- (G) Probability (shows evidence of, present)

Figure 2: The event definition for laboratory result events.

Combining Data from the Sources

The sources used in the current prototype were two hospital information systems: HELP and the DHCP. Two internal medicine diagnostic systems: Iliad [7] and QMR [8] were primarily used to test the database model on history and physical examination data since these two classes of data are not available in HELP and the DHCP system at present. However, LDS and VA hospitals could provide a large number of cases to make the estimates of distribution for other classes of variables. The subset of patient information transferred in each case is comprised of:

- Patient demographic information
- Complete blood count (CBC) data
- Chemistry panel results (e.g. CHEM20)
- Blood gases (pH, pO_2, pCO_2)

To provide a manageable data set to test our hypothesis that data from distinct institutions could be merged to obtain valid statistical estimates, we decided to study a subset of the diseases related to these parameters. The subset of diseases included are listed below, along with the number of patient cases obtained from each hospital (Table 1):

Table 1: List of diseases included in the current UPD prototype, with the count of patient cases from each participating hospital.

	LDS	VA	Total
Acute MI	345	61	406
Urinary Tract Infection	218	162	380
Pulmonary Embolism	54	15	69
Iron Deficiency Anemia	24	89	113
Cirrhosis of Liver	23	36	59
Total	664	363	1027

For this study we developed event definitions for laboratory results and diagnoses which are shown in Figures 1 and 2 above. Each event definition has been tested to insure correct representation of the subset of terms obtained from the two sources considered. The transfer of patient records from the different sources is done via two utility programs: the first program is used to automatically instantiate (build an instance of) an event definition to represent each new medical parameter encountered in the two vocabulary sources. Then, a second utility (data loader) converts all data to be transferred from the host system into a text file and stores

a pointer to the appropriate instantiated event definition in the UPD record for that patient along with a time stamp and its value. The 5 diseases and 35 medical parameters (e.g., CK, LDH, SGOT, WBC) were instantiated automatically.

The implementation of the current prototype uses 4th Dimension, a relational DBMS, on a Macintosh SE/30 with 4 Mbytes of RAM and 80 Mbytes of hard disk space. About 16 Mbytes of disk space is required to store the current 1027 patient records and associated UPD files.

Statistical Testing of Data Homogeneity between the LDS and VA Hospital Databases

Each patient included in the unified database has at least one of the five diseases and a value for each of the 35 variables which is the initial value obtained after admission. Also, for each patient we record the admitting diagnosis, the age and sex and the hospital of origin. Thus, the total number of observations is (35+1+2+1) x 1,027= 40,053 in the unified data set. In our analysis, we tried to identify the relationship of each variable to the diseases and the hospitals. In order to merge data from two hospitals and get from this merged database (UPD) a true estimate of the distribution of values or frequency of findings within a given disease, we must first demonstrate that these distributions are similar in the two institutions. We have chosen 5 common diseases for which we have enough relevant data from each institution to test this hypothesis. In the case of rare diseases, an estimation of the distribution of variables may only be obtainable from the proposed UPD source.

Results

Figure 3 illustrates an analysis of one variable and one disease across both hospitals. In this case, the serum CK on admission is shown in patients with (Dz) and without (~Dz) acute myocardial infarction in each of the two hospitals. Although the distribution of the values from the disease and non-disease groups are quite different, the difference between hospitals is not significant in the disease group, but although small , is significant in the non-disease group (Table 2 and 3). The non-disease group includes patients with one of the other 4 diseases beside acute MI, and the prevalence of these diseases in our prototype differs in the two hospital samples.

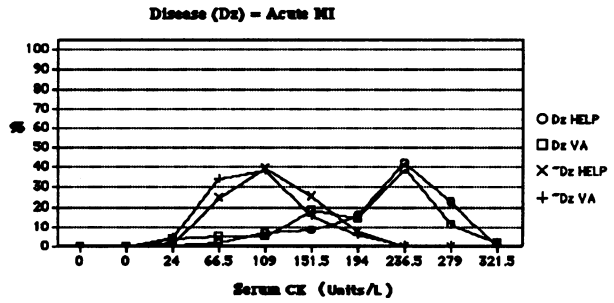

Figure 3: This figure illustrates how the distributions of initial CK enzyme values obtained from the two hospitals overlap and help discriminate between the disease (acute MI) and the non-disease (non-MI) group.

Table 2: Statistics about the distribution of the serum CK in patients with acute MI from each of the two sources.

Sub-population	Dz UPD	Dz HELP	Dz VA	non-Dz HELP	non-Dz VA
# of patients	399	345	54	317	249
Mean	206.6	217.3	195.9	115.8	104.6
Standard Deviation	74.4	80.8	68.0	43.2	45.8

Table 3: Statistics about the distribution of the serum CK in patients without acute MI from each of the two sources.

	HELP Dz vs non-Dz	VA Dz vs non-Dz	HELP Dz vs VA Dz	HELP non-Dz vs VA non-Dz
T-tests	25.98	12.05	1.84	2.97
F-tests	not tested	not tested	1.22	1.06

Referring to Figure 3, clearly, initial serum CK values are useful in separating acute MI patients from patients without acute MI. The analytical program that generates this graph and statistical data can be used to explore the effects of varying threshold values on estimates of true and false positive rates. Thus, the program can generate a ROC curve and can calculate positive and negative likelihood ratios and allows a user to change threshold values interactively. Similar analyses were done for all 35 variables across all 5 diseases. In the case where a variable is not related to a disease, the distribution of the parameter in patients with and without the disease across the two hospitals, all 4 curves overlap. Figure 4 illustrates such a case for serum glucose measurements in patient with and without acute MI.

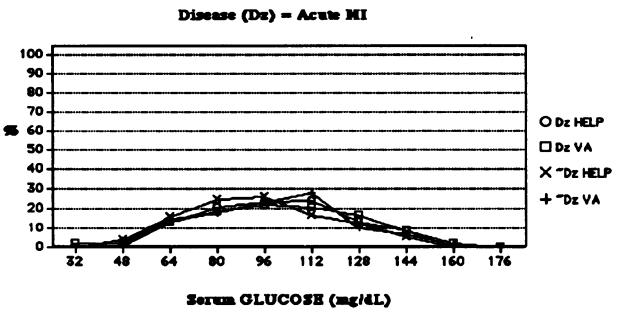

Figure 4: Distribution of serum glucose distribution in patients with and without acute MI across the two hospitals.

To further illustrate that the distribution of values of a variable depends on disease rather than hospital source, Figure 5 shows the average of the initial arterial pO_2 values on patients with each of the 5 diseases in both hospital groups. Notice that, there is no difference between hospitals but a striking lower mean value in pulmonary embolus patients in whom hypoxia is an important finding.

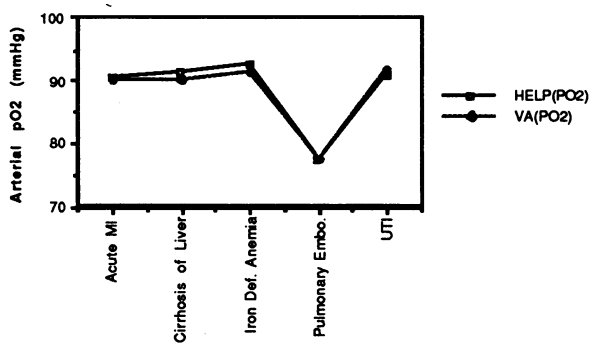

Figure 5: Profile of mean initial arterial pO_2 values across the 5 diseases in the two hospitals.

Within the same disease, 15% of the variables on average showed a significant difference at the 5% level in their distribution based on values obtained from the two hospitals. On the other hand, if the comparisons of the distributions were made without regard to the diseases, then 45% of these would be rejected at this level. At the 1% level of confidence the numbers given above become 5% and 37% respectively.

Discussion

The HELP and VA are two prominent examples of computerized patient databases. They constitute two real world instances of patient data resources potentially useful in answering questions important to health care providers, knowledge engineers, and administrators. We have already demonstrated [2] that a unified representation scheme can be developed to describe clinical events from such diverse systems. In this work, we described experiments which evaluate the homogeneity of data across these sources, since we think this is a key question that must be answered if one is to judge the potential usefulness of a Unified Patient Database.

The current results are sufficiently encouraging that we are planning a much larger implementation with enough data to answer meaningful clinical questions. Although limited conclusions can be drawn from our current implementation, the experience we have gained has made us aware not only of the potential for such a public domain UPD, but also of some of the challenges to be faced in building a full working implementation. The database needed to study unusual diseases and clinical events must be very large to be useful (we estimate at least 200,000 patients are needed) and must include much more details such as time course of a variable within a given patient. The challenge of implementing such a system in a way that would make access convenient and acceptable has yet to be met as well. Our data to date does not allow us to extrapolate with complete confidence to a broader base of patients with the full range of diseases and disease manifestations. The data studied are quantitative laboratory results, which make the comparisons of distribution between different hospital population relatively accessible, however, we have experimented with a general methodology to compare qualitative medical parameters described in two different terminologies [9]. A scoring system was developed to reflect the "degree of match" between two similar qualitative terms from two different vocabularies. It uses a distance metric which captures the degree of similarity between two terms. For instance, "knife-like pain" and "sharp and stabbing pain" might be related as 75% similar, where as "dull pain" sharp or stabbing pain" might be related as 100% dissimilar (opposite).

Our initial interest in using such a merged database is for knowledge engineering. We think that the experience reported here makes it likely that a fully implemented UPD would supply statistics such as finding sensitivity and specificity, and disease prevalence. In addition, the study of the time course of clinical variables in hospitalized patients with a disease and the influence of treatment on these variables would be facilitated if investigators have access to a UPD. Other uses by health care administrators would no doubt benefit as well. Finally, the clinical usage pattern of medical terms in a

patient database could contribute considerably to expand the UMLS metathesaurus and perhaps guide many ongoing efforts at standardization.

Conclusion

In summary, with the current UPD, we have demonstrated that a unified database structure and vocabulary component can be used to represent patient information coming from two or more hospital databases. Using two independently developed hospital databases, we showed that the distribution of a given variable in a disease group was independent of the hospital source. We conclude that the database structures developed for the UPD prototype are adequate to represent merged data in a much larger implementation. This, together with the observed similarity of disease manifestations among hospital sources, justifies our proceeding with a working implementation based on a large enough data set to be useful in the public domain.

Acknowledgments

This work is supported in part by contract number N01-LM-8-3515 from the National Library of Medicine. Gordon Moreshead and Robert Andrews supplied all the VA DHCP patient cases.

References

[1] Huff SM, Craig RB, Gould BL, Castagno DL and Smilan RE; *Medical Data Dictionary for Decision Support Applications*; Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care (11th SCAMC), p. 310, 1987.

[2] Fu LS, Bouhaddou O, Huff SM, Sorenson DK, Warner HR; *Toward a public domain patient database..* Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care (14th SCAMC), p. 170, 1990.

[4] Tuttle M, Sherertz D, Erlbaum M, Olson N and Nelson S; *Implementing Meta-1: The First Version on the UMLS Metathesaurus*; Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care (13th SCAMC), p. 483, 1989.

[5] Pryor TA, Gardner RM, Clayton PD and Warner HR; *The HELP system*; Journal of Medical Systems, Vol. 7, No. 2, p. 87, 1983.

[6] Huff SM, Warner HR; *A comparison of Meta-1 and HELP terms: Implications for clinical data.* Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care (14th SCAMC), p. 166, 1990.

[7] Warner HR, Haug PJ, Bouhaddou O, Lincoln M, *et. al.*; *Iliad As An Expert Consultant to Teach Differential Diagnosis*; Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care (12th SCAMC), p. 371, 1988.

[8] Miller RA, Masarie FE and Myers JD; *Quick Medical Reference (QMR) for Diagnostic Assistance*; MD Computing, Vol. 3, No. 5, p.34, 1986.

[9] Masarie FE, Miller RA, Bouhaddou O, Guise NB, and Warner HR; *An Interlingua for Electronic Interchange of Medical Information: Using Frames to Map between Clinical Vocabularies*; Computers and Biomedical Research, Vol. 24, No. 4, p. 379, 1991.