Clinical Warehouse Enhancement: A Methodology for Linking Heterogeneous Databases

Sharon Kromhout-Schiro, University of North Carolina, Chapel Hill, NC
Rose M. Reedy, Eisenhower Army Medical Center, Fort Gordon, GA
Rob Lenderman, The Bradford Groupe, Ltd., Raleigh, North Carolina,
James Zadinsky, Office of Telemedicine, University of North Carolina, Chapel Hill, NC
Vicki Harp, Science Applications International Corporation, Fort Gordon, GA

Abstract

Health care information is crucial to making decisions that will help to contain the costs of health care. Within the military community, the effort to improve medical care, to meet multi-level management information needs, and to support the readiness initiative is hampered by the storage of Department of Defense (DoD) medical information in large legacy databases in a format that is difficult for physicians and administrators to use for analysis and reporting.

The military medical database system is fragmented over multiple physical sites, and includes several legacy systems (Zadinsky, 1997). Portions of a patient's medical record may be stored in multiple databases of differing structures at multiple medical treatment facilities (MTFs). Access to the medical information requires the linking of multiple databases, both military and civilian. Due to the heterogenity of the databases, the linking of the databases necessitates the translation of the data and data structures so that the data in the composite database have the same meaning.

This paper describes the principles behind the linking of multiple databases through the development of a platform of common data definitions. This platform of common data definitions is henceforth called the Foundation Library (Kalet, 1989). The use of the Foundation Library reduces the effort required to develop code for the translation of data. Only one translation needs to be developed for each database: the translation to the Foundation Library. The coupling of a military medical database, the Composite Health Care System (CHCS) and the North Carolina Trauma Registry (NCTR) database is described as a demonstration of this technique.

Introduction

In the Department of Defense, electronic medical record information may be fragmented across many files or data sources. Data may be stored in different types of structures, such as text files or databases, and have different field definitions and field value options. These differences result from differences in clinical practice and computer facilities. This variation in structure makes the aggregation of data for medical care or administrative decision making more complex. However, integration of the various pieces of the medical record system is crucial for improving the quality of medical care and reducing the cost of that care.

In order to accomplish the integration of the Department of Defense (DOD) medical systems, an architecture must be developed that permits the linking and consolidation of these data sources. This merging and conversion of the data must be transparent to the report-generating user. To combine the data from these various sources, a common platform of communication, the Foundation Library, was defined through which all data processing activities will occur. This Foundation Library consists of definitions of all objects that are fundamental to patient care and medical administration. The structures of the databases that were linked together were

mapped to the structure of the Foundation Library through translation code. Analysis software was written so that it depended only on the data structures defined in the Foundation Library. The basing of the analysis software on the Foundation Library allows other databases to be linked and analyzed without changes in the analysis software. This methodology will enhance access to all of the DoD medical system data without requiring changes to the structure of any of the individual databases.

The steps involved in the development and use of the Foundation Library for this demonstration project were:

1) identification of a subset of variables of interest for trauma research using a military database and a civilian database;
2) development of common data definitions for the variables of interest;
3) development and testing of SAS® code for the translation of data from each of the two databases to the structures and values defined in the Foundation Library;
4) development of analysis code based on the data structure of the Foundation Library.

Datasets

The two databases chosen for this project have significantly different structures, field definitions, and field values. The CHCS is an information system sponsored by the U.S. Army Information Systems Selection and Acquisition Activity and Defense Medical Information Management. It is designed to support the administration and delivery of health care in Department of Defense Medical Treatment Facilities. The North Carolina Trauma Registry (NCTR) is a cooperative undertaking of the designated trauma center hospitals in North Carolina and the North Carolina Office of Emergency Medical Services. The NCTR collects data on all trauma patients admitted to the hospital for at least one day, as well as all patients declared dead in the Emergency Department.

Twenty variables were chosen for the demonstration project that represent a cross section of demographic and clinical variables. These variables were patient name (last, first, middle, birth), sex, attending physician name, attending physician specialty, admission and discharge dates, mode of transportation to the Emergency Department (ED), admission condition, chief complaint, temperature, pulse, respiration, blood pressure, diagnoses, disposition, units of blood given, and length of stay.

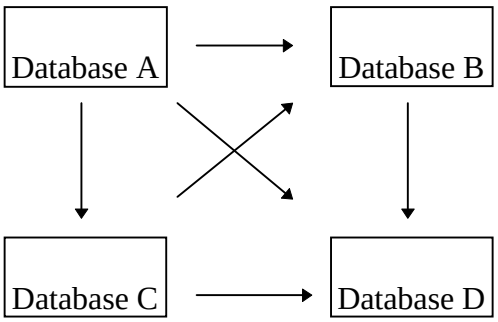
Foundation Library

The Foundation Library is the platform of common definitions through which all other databases communicate. For this demonstration project, the Foundation Library was developed for a subset of twenty variables from the CHCS and NCTR databases. The definitions consisted of a field name, description of the field, data type (numeric, text, or date), field length, and definitions of all allowed values for the field (including units of measure). This information was developed as a text file to be used as a reference when writing the translation code.

Translation Code

The Foundation Library platform reduces the programming necessary to connect heterogeneous databases by requiring only one translation program for each database, since each database would only have to communicate with the Foundation Library. For example,

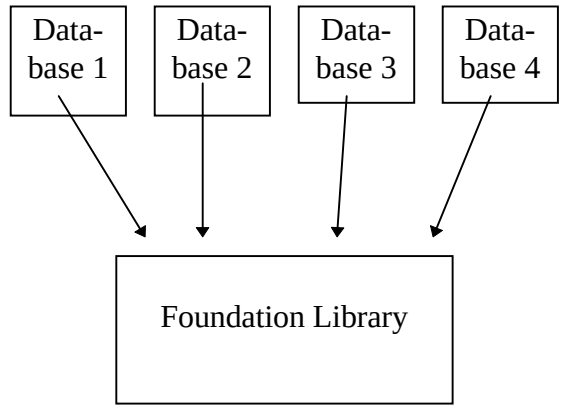
Figure 1:Translations required for communication between databases without Foundation Library.



without the Foundation Library, the linking of four heterogeneous databases to each other in pairs would require the development of six versions of translation code (Figure 1).

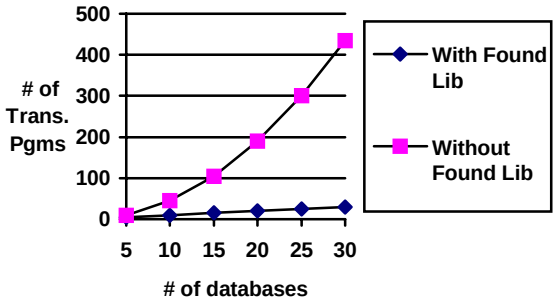
The connection of these databases through a Foundation Library necessitates only four versions of the translation code (one from each of the four databases to the Foundation Library) (Figure 2).

Figure 2: Translations required for communication between databases using Foundation Library



As the need arises to link additional datasets of heterogeneous structures to those already linked, the only programming requirement is the development of the code to translate the new structures to the Foundation Library structure. The impact of the Foundation Library on the programming effort required to link databases increases dramatically with the number of databases to be linked (Figure 3). The number of translation programs required to link *n* databases with a Foundation Library is *n*; the number of translation programs required without the Foundation Library is (0.5 * *n*) * (*n*-1).

Figure 3: Comparison of number of translation programs required for linking databases with and without a Foundation Library.



Discussion

The development of the Foundation library definitions was relatively straightforward for some of the variables such as name and sex. In the Foundation Library, as in NCTR, the patient's name was stored as last name, first name, middle name, and birth name. In CHCS, the patient's name is stored in one field (last name, first name, minit), and birth name is stored in a separate field. The translation code for the CHCS database involved parsing the name field. The code for parsing the name field is dependent on consistency in the way that text data are entered into this field. Thus, data entry error and its impact on the translation code is an issue that constantly needs to be addressed.

The definition of other variables required significant discussion with the trauma physicians and nurses at the sites that generated the participating databases. For example, admission condition was a difficult variable for which to arrive at a common definition. The CHCS definition combined values for the patient's state of alertness (comatose, lethargic) with values that indicated the patient's ability to ambulate (stretcher, wheelchair). The NCTR definition of admission condition limited the values to the patient's level of alertness (alert, responds to verbal commands, responds to painful stimuli, unresponsive). The decision for this demonstration Foundation Library was to create two variables, one to record ability to ambulate, and one to record level of alertness. The translation code for the CHCS admission

condition variable to the two Foundation Library variables was unclear at the time of publication, and would require further discussion with trauma care experts.

Testing of the translation programs was done for the variables that were defined in the Foundation Library, and to which the CHCS and NCTR data structures were mapped. This testing involved the generation of sample data sets in both the CHCS and NCTR structures, and the running of several simple analyses demonstrate the correct conversion of the data.

Conclusions

The long term goals of this project were threefold: (1) to provide data for medical research, including disease management and outcome studies, (2) to provide data for medical administrative decision making, e.g., cost and length of stay analysis and utilization review for the management and delivery of health care in the Department of Defense (DoD) managed care environment, (3) to improve the timeliness and ease of access to data and provide new functionality to support the Tri-Service Care (TRICARE).

This project demonstrates the utility and complexity of developing a library of common data definitions to allow heterogeneous data bases to communicate. Future work will involve the development of common definitions for more fields for trauma research, as well as the development of translation programs to connect other databases to foundation library.

Correspondence:
Sharon Kromhout-Schiro
University of North Carolina at Chapel Hill
Campus Box 7227
Chapel Hill, NC 27599-7227
Telephone: 919-966-6263
FAX: 919-966-6433
E-mail: eileen@med.unc.edu

References

(Zadinsky, 1997): Zadinsky, J. (1997) Data Warehouse and Telemedicine. Southeast SAS Users Group. Cary, NC: SAS Institute, Inc. p 453.