Metadata Repository for Population Health Research

by

Mingxiu Li
B00338297
mingxiu@cs.dal.ca

Performed at

Population Health Research Unit
Department of Community Health & Epidemiology
Faculty of Medicine

5790 University Ave.
Dalhousie University
Halifax, NS B3H 1V7

In partial fulfillment of the requirements of the Master of Health Informatics Program,
Dalhousie University

Report of Internship for the period January 25 – April 22, 2005

Date Submitted: May 23, 2005

Acknowledgment

This report has been written by Mingxiu Li and has not received any previous academic credit at Dalhousie University or any other institution.

I would like to thank Mr. Mark Smith, Acting Director of the Population Health Research Unit, for his effort of making this internship possible and providing me with such a great opportunity to work and learn at PHRU.

I would like to thank Leslie Anne Campbell, Research Coordinator of the Health Outcomes Research Unit, for providing me with the opportunity to participate in the Health Outcomes Data Warehouse project.

My special thanks go to Victor Maddalena, course instructor, for his help in making this internship possible and Kevin Druhan, Data Manager at the Population Health Research Unit, for his great guidance and support of my work during this internship.

I would like to thank Grace Paterson for her help in the grant proposal application and all the PHRU staff for their support of my work.

I would also like to thank Dr. Zitner, Dr. Shepherd, Dr. Abidi, and Deirdre Harvey for their support of this internship.

(signature)

Mingxiu Li

Executive Summary

The internship was performed at the Population Health Research Unit (PHRU) and the Health Outcomes Research Unit (HORU). The objectives were to work with the health data management group and practice health informatics knowledge in population health information managing. The author participated in three major activities:

1) Generate the online documentations for the PHRU database and assist in web development.
2) Participate in the initiative of developing a metadata repository for Population Health Research and the writing of a grant proposal.
3) Coordinate the initial planning for the Health Outcomes Data Warehouse project in Capital Health.

The online documentation provides users with specific information about the health datasets at PHRU. The author had the opportunity to learn more about the population health data and research. The process of generating the online documentation required manual processing of each dataset, which was time consuming and would be difficult to maintain over time. In addition, the online documentation consists of static data files that only allow users to navigate files and variables. It is considered a temporary solution for making database information available online.

The Metadata Repository project was a proposed solution to automatically produce searchable metadata for each dataset and allow other institutes to add their metadata to the repository. It will also provide a dynamic searching tool that allows users to perform their own database querying and database exploration. The Metadata Repository for Population Health database would centralize and automate the management of health information, as well as provide a new opportunity for easier human-computer communication. The author found it a valuable experience to work from a manual solution to planning an automated solution for managing health information.

The Health Outcomes Data Warehouse project allowed the author to have a closer look at the Capital Health information systems. The author learned that because of different information systems and confidentiality and security policies, there are still many barriers preventing the sharing of information automatically within the hospital. A Data Warehouse is a repository that will gather all the health data from different systems in the hospital and link them together. The raw data would be pre-processed and ready for decision-making, research or other purposes.

A major objective of Health Informatics is to use existing or emerging information technologies to facilitate sharing and management of health information. The author found that this internship achieved the goal of learning and applying Health Informatics knowledge in practice on centralizing health data management, increasing human-computer communication and providing high-quality information to improve population health and health outcomes research.

Table of Contents

1. Introduction

The objective of this internship was to work with the health data management group at the Population Health Research Unit (PHRU) and investigate the potential of health informatics theory to improve the storage and usage of the population health data housed there. The internship was involved in three major tasks. The three tasks are tightly connected to the management of health care data and they all have a close relation to Health Informatics.

The first task was generating the online documentations for the PHRU database and assisting in web development. Through this activity, the author became familiar with the content of all datasets housed at PHRU, and gained an improved understanding of the privacy and confidentiality concerns and their implications for population health research. The author also learned to program in SAS, which is a powerful statistical and data management software package. This task gave the author an 'inside look' at the population health database to learn what data has been captured in different fields, (e.g. pharmacare, medicare etc); where they come from (e.g. CIHI, DOH etc), and how they are stored, (e.g. flat file format vs relational database).

The second task was participating in the initiative of developing a metadata repository for Population Health Research and prepared for the writing of a grant proposal. The author was required to understand the concepts of metadata repository, metadata standards, concept mapping and semantic web as well as conducting background research for the project.

The third task was coordinating the initial planning for the Health Outcomes Data Warehouse project in Capital Health. The internship required the understanding of the data warehouse concepts and the system requirements. The author communicated with the data holders and researchers in Capital Health. During this task, the author had the opportunity to learn about the current systems in Capital Health, and began to recognize the importance of a centralized data warehouse to integrate, clean and format disparate sources of health care data for improving health outcomes research.

2. Description of the organizations

2.1 Population Health Research Unit

The Population Health Research Unit (PHRU) is a university-based research and support group that conducts systematic research in population health, health services and explores their inter-relationships. It is a sub-unit of the Department of Community Health and Epidemiology at Dalhousie University. PHRU houses and manages provincial administrative health data obtained from the Department of Health of Nova Scotia and provide expertise in data analysis to support population health research. PHRU's mission is to "*advance the level of knowledge and develop innovative research methods for the betterment of the health of the general population, in a cost effective manner.*" [1]

PHRU currently houses over 240 million provincial administrative healthcare data records as well as linked national census and survey data. Figure 1 [1] provides an overview of the databases held at PHRU. All the data has been cleaned and is ready to be used for research.

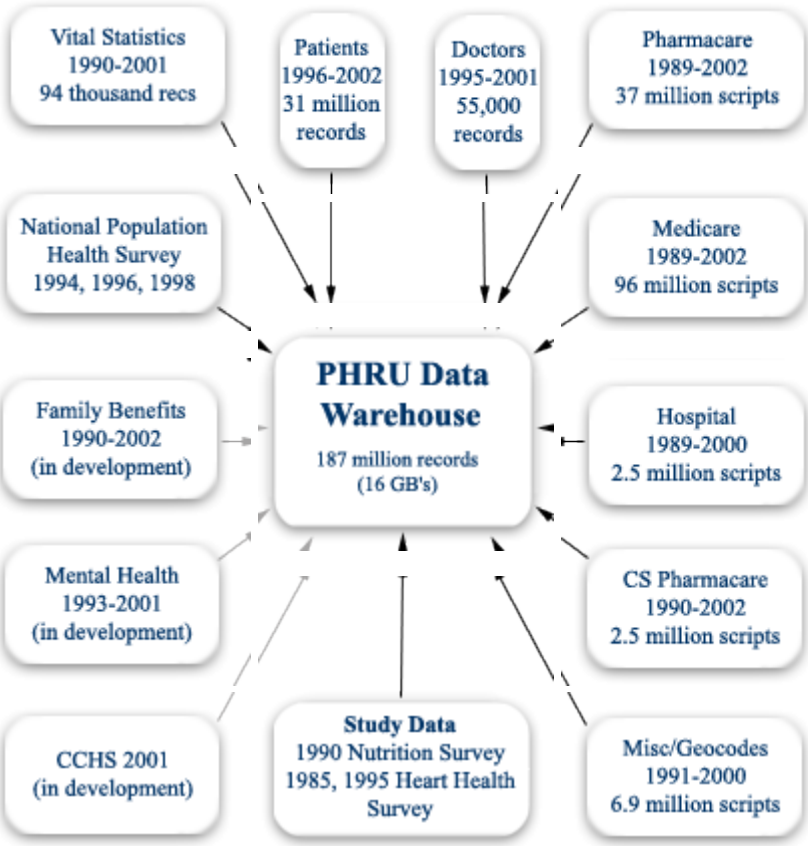

Figure 1

2.2 Health Outcomes Research Unit

The Health Outcomes Research Unit (HORU) was created in September 2002 to strengthen and promote health outcomes research within Capital Health in Nova Scotia. The HORU provides support and guidance for health outcomes research initiatives at Capital Health, coordinates collaborative projects between a variety of researchers, clinicians, and policy planners, and conducts health outcomes research. HORU has a close working relationship with the Population Health Research Unit and departments of psychiatry and community health and epidemiology at Dalhousie University. The Health Outcomes Data Warehouse is one of the current research projects at HORU.

3. The work performed and lessons learned

The internship involved in three major tasks. 1) Generate the online documentations for the PHRU database and assisted in the web development. 2) Participate in the initiation to develop a Metadata repository for the Population Health Research and prepared for the writing of a grant proposal. 3) Coordinate the initial planning for the Health Outcomes Data Warehouse in Capital Health. The three tasks are tightly connected to the management of health care database and they all have a close relation to Health Informatics.

3.1 Online documentation

3.1.1 Internship Role

Although PHRU had been working towards developing an online repository of documentation for their administrative population health database, information about most of the PHRU datasets was not yet available online. In order for the researchers to search information online without having to consult the data manager or senior analyst, there was a need to put the documentations online. All of the database documentation at PHRU, however, was contained either in the SAS data files themselves, or a paper format with any patient identification information removed. The

author's role was to inspect each dataset and select the variables to be published while making sure to avoid any privacy and confidentiality concerns. The author also modified numerous SAS programs to generate HTML documents for all of the datasets and linked them to the PHRU website.

3.1.2 Lessons Learned

The task gave the author the opportunity to learn the following things:

- Understand the process of population health data collection and learn what kind of data and information were collected in different health sectors.
- Gain familiarity with how raw data is cleaned, encrypted and stored.
- Understand the concept of privacy and confidentiality of patient or doctor's information.
- Understand how population health data can be used and analyzed for research.
- Learned SAS programming in a secure VMS operating system environment.

3.1.3 Discussion

The process of generating the online documents is time consuming and would be very difficult to maintain in the future because any change in the data necessarily requires a re-run of the HTML generating programs. Although SAS programming was used to produce the documents, each file had to be processed individually due to differing variables in each dataset. Furthermore, the online documentations are static files. The user needs to navigate through all the files to look for their specific information. If the researchers need to navigate through hundreds of files to get to the information that meets their needs, they might prefer to just call the data manager for help. The value of the online documentation is then reduced. In addition, one dataset is documented within one HTML file. As a result, some HTML files are more than ten pages long because some datasets contain large numbers of variables. For example, the hospital discharge abstract data from CIHI (Canadian Institute of Health Information) can contain more than six hundred variables. Because it can take a long time to go through all the variables individually, some kind of search tool is preferable. A potential solution to this and other problems with the static HTML data documentation became the next task of the internship, which was to propose a metadata repository solution to enhance the online searching for quality information for population health research.

3.2 Metadata Repository for Population Health Research

3.2.1 Background

The Population Health Research Unit at Dalhousie University manages and maintains large volumes of Nova Scotian health data to provide a broad level of data support for population health research. This data can be used for many types of studies including primary care needs assessment, cost-benefit and cost-effectiveness analyses, compliance studies, determinants of health studies, patterns of care and practice studies, and service and technology utilization studies. Every year, dozens of research projects use PHRU data for population health research. Researchers encounter barriers when designing their projects because health data housed at PHRU is in coded format and has restricted access due to privacy and confidentiality concerns. Researchers have to consult the PHRU staff (typically the senior analyst or database manager) to be able to obtain sufficient information to decide whether the dataset is suitable for their research. This one-to-one process takes a lot of time both for the researcher and the data analyst at PHRU. In addition, PHRU currently doesn't have an efficient system that can make the information of the huge amount of datasets available online. As a result, many datasets collected in Nova Scotia are not known to the general research community. Data providers such as PHRU spend a lot of effort providing data resources that are comprehensive and user-friendly; but different data providers use different technologies for data storage and access. This means that a researcher might have to search ten or more different databases to find all the information pertaining to a particular research project. Since 'data exploration' is so difficult and time-consuming, many researchers have indicated that they tend to use data already known to them instead of seeking out new sources. If they were doing these kinds of searches on a regular basis, a centralized data information system would save a lot of time. Maintaining up-to-date and fully functioning versions of all those databases and the tools to search them, however, is an expensive and complex task.

3.2.2 Proposed solutions

Data access barriers such as these are a severe obstacle to building international, cross-provincial and inter-institutional research capacity and facilitating collaborative research in population

health. To promote data exploration and build capacity for population health researchers to expand the scope of their research, PHRU began to explore the concept of a metadata repository and tools for rapidly locating, browsing, and searching all available datasets. Their proposed solution was:

- Develop a pilot healthcare metadata repository containing information about all population health data resources available in Nova Scotia including a description of each database, variable lists and descriptions, and contact information for obtaining access to the data. A framework would be established for automatically producing searchable metadata for each dataset, allowing other institutions to easily add their own datasets to the repository. If successful, the repository would stimulate new population health research by allowing researchers to perform their own data exploration and assess the utility of a particular dataset quickly and easily. Metadata standards such as DDI (Data Documentation Initiative) or Dublin Core metadata standard would be used to develop the metadata architecture.
- Develop and evaluate online analytical search tools allowing researchers to quickly locate data related to a particular research topic, investigate the data at the variable level to assess the suitability of the data for a specific project, and contact the holding institution directly to inquire about data access procedures. Researchers would be able to browse all available datasets and perform searches of the metadata using plain language queries. A significant innovation to be developed would be the capacity to search for data at the diagnosis, procedure, or drug code level, allowing researchers to rapidly make a high-level analysis of a particular dataset without requiring the assistance of programming and analytical staff.

An example of usage is a researcher interested in brain trauma. After typing in "brain injury", the tool would generate a list of Nova Scotia datasets containing data with ICD9/10 codes related to brain injury. The metadata of each dataset can then be viewed in more detail including a description, variables, number of observations, and contact information for obtaining access to the dataset. Figure 2 illustrates how the data navigation part of the project might look:

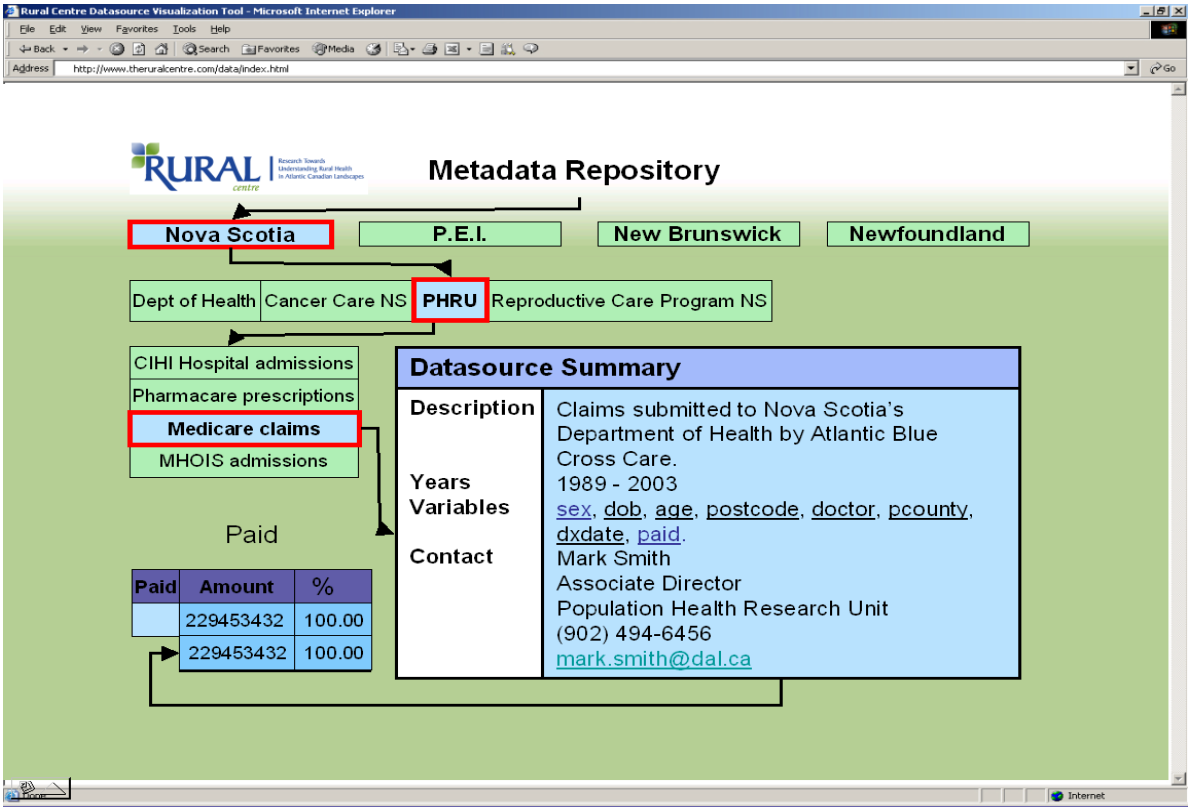

Figure 2

3.2.3 Proposed methodologies

The design of the system involves two innovative functions.

1. Use natural language 'free text' queries to search the *coded* health data.

2. Build a metadata repository to allow in-depth health data information exploration.

Although some research have been done in employing natural language processing with metadata standards [2-4], not much integrating work has been done to serve the fundamental needs of the population health researcher. It is a technical challenge, as well as a lot of work to develop a metadata standard suitable for presenting the massive volume of health care data in a succinct yet usable manner. The two components to be used in this project would be the Unified Medical Language System (UMLS) medical term translation system and the Metadata Repository itself.

The **Unified Medical Language System** (UMLS) is a long-term project began in 1986 by the National Library of Medicine (NLM). Its purpose is to overcome two barriers when retrieving information from machine-readable databases [5]:

1) the variety of expressions for the same concept among different machine-readable sources and by different people
2) the distribution of useful information among disparate databases and systems.

UMLS is a knowledge source that collects medical concepts, terms and relationships from different classifications and systems. Table 1 explains the principle of the UMLS concepts. The synonymous terms from different sources are clustered into concepts (e.g. Essential Hypertension). When a medical term is queried, the term is semantically mapped into related concepts and within the concepts the source is identified. It is as if the machine "understands" the meaning of the medical term.

Terms	**Source**	**ID**
Essential hypertension	SNOMED-CT	59621000
Essential (primary) hypertension	ICD-10	I10
hypertension	LOINC	LP32933
Idiopathic hypertension	Read Codes	XE0Uc
Essential Hypertension	**UMLS**	**C0085580**

Table 1

The UMLS knowledge source has been widely used "by system developers in building or enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information, as well as in informatics research." [6] PubMed is one example of the NLM's applications that uses the UMLS [7] concepts and semantic network in indexing and locating the information sources. Other research projects that used UMLS knowledge source include the project ARIANE which used UMLS to express the user request to match with and connect to the pre-generated information sources [8].

In this project, the UMLS Knowledge Source Server will be used as the mediator between the user and the metadata repository. User's free text query will be semantically mapped into a related coding scheme in order to match with the coded data.

Metadata Repository - Metadata is simply data about data. A metadata repository is the collection of metadata for various databases for the purpose of centralized information searching. Metadata constitutes the information that enables the effective, efficient, and accurate use of health care datasets and increases the interoperability among different organizations and their information systems. There are three type of metadata: descriptive, structural and administrative metadata [9].

Descriptive metadata describes the content of a resource of identification, searching and retrieval. Examples are bibliographic information and coded diagnoses for a patient. The coding schemes include ICD-10, SNOMED etc.
Structural metadata describes the architecture and relationships of the different sections of resources for navigation. Examples are table of contents and index of the type of drugs.
Administrative metadata describes technical aspects of an information resource for processing and management. Examples are publishing information, privacy and confidentiality rules of the data usages.

In order for metadata to be used and communicated in an interoperable way among different systems, a standardized structure must be used. [9] The standard involves character encoding, language, controlled vocabulary, message structure and formatting. Many metadata standards have been developed, including the Data Documentation Initiative (DDI), Dublin Core (DC) and HL7. But there is no existing metadata standard that is designed for health metadata. The challenge for us is that the breadth of detailed information in the database makes it infeasible to use only one of the metadata standards to complete the task. In addition, in order to secure buy-in to the centralized health data information repository from other health organizations, a specific health metadata standard will need to be developed to support interoperability between different database systems.

Dublin Core Metadata was developed in 1995 based on the discussion in a workshop of "how a core set of semantics for web-based resources would be extremely useful for categorizing the web for easier search and retrieval" [10] in Dublin, Ohio. "Dublin Core Metadata" was named based on the location of the workshop. Dublin Core has been adopted by national libraries in

Australia, Denmark, and Finland [10]. The Dublin Core Metadata Element Set (DCMES) has 15 descriptive elements that provide vocabularies for describing the "core" information properties, such as "Description", "Creator", and "Date". It represents a core set of elements likely to be useful across a broad range of vertical industries and disciplines of study. Dublin Core Metadata has been used in health information resources indexing and searching [3]. However, there are limitations using Dublin Core for health care database information due to the complexity and diversity of health data. In order to more completely describe the health data, more elements need to be added to the current elements sets [4].

Data Documentation Initiative (DDI) is “an effort to establish an international criterion and methodology for the content, presentation, transport, and preservation of metadata about datasets in the social and behavioral sciences.” [10] It is a documentation standard for social science database that is created by the North American and European survey research and data archive organizations. DDI is continuously supported by Health Canada [11]. Before DDI was developed, systems that searched for social science datasets could only search on limited fields (i.e. name, author, study number, and abstract). The codebooks, which contained all the metadata about the datasets, had to be used manually in order to find out detailed information about the study (e.g. variables, methodology, and structure of the data). With the achievements of the DDI, codebooks can now be created in a uniform, highly structured format that is easily and precisely searchable on the Web. Further, this specification may have far-reaching implications for improvement of the entire process of data collection, dissemination, and analysis. eXtensible Markup Language (XML) is the language used in DDI.

Although DDI was created primarily for social and behavioral sciences data, it has a similar structure to the nature of health metadata. Unlike Dublin Core, DDI has a diversity of tags that may allow it to be adapted for a health metadata repository. Further research needs to be done to determine the potential of using DDI for health care metadata. Before determining what metadata standard to be used, an evaluation of all the exiting metadata standards is also essential to come up with a better solution.

3.2.4 Discussion

The objective of the project is to provide a tool for facilitating the sharing of health information and supporting automatic management of data exploration by researchers. The project is still in the planning stage. If developed, the tool would reduce the significant overhead costs associated with exploratory data requests. The project would also involve an evaluation component to identify potential areas for improvement including small focus group sessions with researchers, administration of user surveys, review of submitted user queries with follow-up interviews, and monitoring of usage patterns.

3.3 Health Outcomes Data Warehouse

3.3.1 Background

Currently there are numerous clinical databases housed within Capital Health. Some of the databases are created by individual researchers or clinicians for their specific research. Some of them are generated within the hospital information systems. However, they are stored in different formats and have limited degrees of access. There is no linkage between them. In addition, some of the individual datasets are stored in personal laptops, which is under low security guard. All the clinical databases are valuable resources for individual or collaborative health outcomes research. There is a need for a secure and readily accessible data source that can meet the needs of clinicians, administrators, and researchers, using methods that require data validity, integrity, and confidentiality. From consultation with the Health Outcomes Strategic Group and Information Technology Services at Capital Health, it has become evident that an inventory of existing clinician databases is needed. Such an inventory would provide the basis of the Health Outcomes Data Warehouse.

3.3.2 Objective

The objective of the Health Outcomes Data Warehouse is to develop a structured data repository that will be used to centrally manage and link the health outcomes datasets, facilitate population based research, provide quality assurance with feedback to data providers, and serve as a coordinated, single access point for population level health data at Capital Health.

The data warehouse will enable the combination of different types of data dynamically, while leaving the source data unaffected. Data from a variety of sources, including clinical research programs, research projects, and laboratory, would be collected and updated quarterly. Data transformations will likely be automatic or semi-automatic where applicable. All data will be reviewed and signed off before being entered into the data warehouse. At this point, discrepancies or anomalies in the data will be identified and provided back to the source, serving as a quality assurance feedback loop. Data updates may be either complete data refreshment (certainly for the smaller "data marts" or project-specific databases), or changed data capture, dependant upon the nature of the data.

3.3.3 Internship Role

The internship objective was to coordinate the initial planning of this data warehouse project. The author's tasks included collection of data inventory, user analysis and system analysis.

After developing the objective and vision of the project, the first step was to locate the data owners for the preparation of the data inventory collection. The information was collected by meeting with the Health Outcomes Committee and individual group members. An inventory form was created previously to collect all of the database information including the research area, the data that was collected, number of records, the period covered, and the database format, as well as the data holder's concerns, comments, etc. After gathered all of the contact information, the author sent out more than fifty inventory forms to the researchers. The author followed up by email and telephone to provide more information about the project and answer questions. Meanwhile, the author interviewed several database managers to discuss the data warehouse project and collect more detailed information about their databases. From the feedback of the inventory request, the author collected all the responses for the user analysis and system requirement analysis.

3.3.4 Interviews

The interviews with database managers allowed the author to learn more about the different information systems in Capital Health. The first interview was with Bryan Crocker, District Manager of the Laboratory Information System. The interview gave the author a better

understanding of how the “Cerner Classic PathNet” lab system works and how it collects and stores the lab information. The Cerner Classic PathNet is a laboratory system that is used for the whole Capital Health Region. It tracks all the patient lab record within CDHA. Currently most of the lab result are collected automatically through the test instruments and stored in the data files. Because the patient record is still in paper format, the lab system is not connected to the patient clinical records. However, it connects to the STAR (administrative system) so there is no duplicated entry of patient demographic information. All lab reports are sent paper-based by fax or mail or over the phone to the clinic even though the lab data are computerized. It takes time and human effort to delivery the lab result and it is easy for it to get misdirected or lost. Only after the Electronic Health Record is in place will this problem be solved. In the mean time, there are frequent requests for lab data for research. It takes the lab data manager extra time to extract the data from the huge file-based database and copy the data to CD for researcher use. It is also time consuming for the researcher to first clean the data before being able to analyze it. Most importantly, when the data is away from the secured system, the data protection is hard to control. The data warehouse eliminates both of these concerns. First of all, the data will be cleaned and encrypted in the data warehouse that is ready for analysis. Secondly, there will be a secured network that only allows the authorized users to access to the database. No portable hardware is necessary.

3.3.5 Analysis and Discussion

There were about 40 responses from the data inventory request. The databases covered various clinical research areas. 98% of the databases were created for research purposes. Among these, 45% were also used for administrative purposes. In addition, only 44% of all of the databases have been completed. This implies that more than half of the databases would be continuously updating sources and would require ongoing update and maintenance. The various identifiers for the data collections included patient Medical Services Insurance number (MSI), Hospital Unit Number (HUN) etc. It would be complicated to link all the datasets without a unique identifier for each patient. However, there may not be many connections between the datasets. Based on the characteristics of the individual research for which the data was collected, each dataset only contains observations specifically around the research question. The chances of the same patient occurring within several different research datasets are rare. According to the response from the

researchers, some of them expressed strong interest in linking to other databases for the purpose of further collaborative research. PHRU houses all health administrative databases in Nova Scotia, and it is possible and potentially beneficial to link all of the health outcomes datasets with PHRU data. This would meet many of the needs for databases linkage for research purposes. The database formats are mostly Access. Some of them are in excel sheet, dBase or Dos. The variety of formats would be synchronized into a standard database format before put into the data warehouse.

Based on the research on the existing health data warehouse, most of data warehouses are housing regional or hospital data. The amount of data is huge and requires a lot of system space. A relational database is the most common choice for the database structure. A relational database creates a table for each object (e.g. patient, diagnosis, procedures) and relates the tables by using primary and secondary keys. It is easy to manage and saves space from duplicate observations. To determine the system requirements, the project team had consulted with the IT department within Capital Health. Due to the complexity and variety of the clinical research datasets, the cost of a relational database system and its implementation is huge. PHRU has a great deal of experience with the storage and linkage of sizable health services databases, having demonstrated success with the cleaning, management, and linkage of the various administrative databases held by the province. PHRU has automated the process of documenting data sets, generating data dictionaries, and creating metadata, which is critical to the management of large data sets such as data warehouse. As such, they are an invaluable and unique resource for data management. Further, to utilize their services would be cost-effective, as they currently run a similar system.

4. Conclusions

Health care database management is one of the most important aspects of Health Informatics. A good database management system decides how health information is created and used to improve people’s health. The metadata repository combined with the UMLS natural language search tool is an innovative step for health information sharing and searching, as well as reducing the gap between human-computer interactions. The tool makes information about each database

in a distributed health database network available online for easy searching and navigation. It will be an important tool towards increasing capacity in population health research.

The development of an Electronic Health Record (EHR) is in progress in many countries, and is considered to be the solution that will make all health information available electronically to health professionals and patients. A data warehouse, on the other hand, will serve as a data and information repository that integrates, cleans and formats all the health care data from different information systems and sources [12]. The cleaned and aggregated data will then be ready for administrative or clinical decision-making and health research.

5. Recommendations

The Population Health Research Unit is an experienced provider of health data and information resources for population health research. A better data management system would increase the usage of health data and information for research and other decision-making in Nova Scotia. Currently all the data stored at PHRU are in flat file format. This means one table contains all the variables for one dataset, e.g. all the patient visits were input into one table. Although simpler conceptually than a relational database, it creates a lot of duplication and occupies a lot of space. There is approximately 20% duplicated data at PHRU. A relational database will eliminate the duplicate data and save a lot of space. However, it might be inefficient for the analyst because a relational database will slow down the processing speed. This is still somewhat of a controversial issue that deserves future investigation.

An optimized information system will reduce data management costs and better support the use of health data. The process of optimizing the PHRU information system would benefit from cooperation with Health Informatics. The Metadata Repository and Data Warehouse Projects mentioned earlier are typical health informatics projects. The collaboration between PHRU and Health Informatics will be beneficial for both sides. Based on her experiences in the Health Informatics program, the author found the following areas of possible collaborations between Health Informatics and PHRU.

1. **Make good use of real health data in Health Informatics' teaching**

   Statistics for Health Informatics and *Data Mining for Health Informatics* both require real health data to fulfill the course requirements. It will be beneficial for the students to manage and analyze real health data to understand the practical use of health informatics. In addition, using real health data might facilitate research projects. However, some Health Informatics students found it hard to search for real health data.

   PHRU houses 240 million health care records in Nova Scotia and it could be a valuable resource for students. At the same time, data mining is a new technique in health care, and the collaboration between PHRU and data mining will facilitate new methods and techniques in population health.

2. **Internship opportunities**

   PHRU is an organization that deals with health information. It is a perfect place for Health Informatics students to practice their knowledge and get on-the-job training and for PHRU to get knowledgeable people for short-term projects.

3. **Research Cooperation**

   There have been ongoing research collaborations between Medical Informatics and PHRU such as the Metadata Repository project. The author believes there will be more and more opportunities of collaboration in the coming future.

Practicing is the best way for learning. The author feels that there is a need to build cooperation between PHRU and Health Informatics in order to bring new knowledge into practice to improve the management of health information.

Reference

[1] Population Health Research Unit, "PHRU Data," [http://www.phru.dal.ca/data/index.htm], May 2005.

[2] G. Singh, S. Bharathi, A. Chervenak, E. Deelman, C. Kesselman, M. Manohar, S. Patil and L. Pearlman, "A Metadata Catalog Service for Data Intensive Applications," *Proceedings of the 2003 ACM/IEEE conference on Supercomputing*, 2003.

[3] M.N. Boulos, A.V. Roudsari and E.R. Carson, "Towards a semantic medical Web: HealthCyberMap's tool for building an RDF metadata base of health information resources based on the Qualified Dublin Core Metadata Set," *Med.Sci.Monit.*, vol. 8, pp. MT124-36, Jul. 2002.

[4] B. Thirion, G. Loosli, M. Douyere and S.J. Darmoni, "Metadata element set in a quality-controlled subject gateway: a step to a health semantic Web," *Stud.Health Technol.Inform.*, vol. 95, pp. 707-712, 2003.

[5] O. Bodenreider, J. Willis and W. Hole, "The Unified Medical Language System: What is it and how to use it?" *MEDINDFO*, vol. 2005, pp. 189, 2004.

[6] U.S. National Library of Medicine, " About the UMLS Resources," [http://www.nlm.nih.gov/research/umls/about_umls.html], April 2005.

[7] U.S. National Library of Medicine, "UMLS Applications," [http://www.nlm.nih.gov/research/umls/umlsapps.html], April 2005

[8] M. Joubert, M. Fieschi, J.J. Robert, F. Volot and D. Fieschi, "UMLS-based conceptual queries to biomedical information databases: an overview of the project ARIANE. Unified Medical Language System," *J.Am.Med.Inform.Assoc.*, vol. 5, pp. 52-61, Jan-Feb. 1998.

[9] G. Kim, "Metadata in Medicine," [http://meld.medbiq.org/primers/metadata_in_medicine_kim.htm], April 2005.

[10] "Metadata Reference Guide," [http://libraries.mit.edu/guides/subjects/metadata/standards/ddimag.html], May 2005

[11]"Data Documentation Initialtive," [http://www.icpsr.umich.edu/DDI/], April 2005,

[12] R. Verma and J. Harper, "Life cycle of a data warehousing project in healthcare," *J.Healthc.Inf.Manag.*, vol. 15, pp. 107-117, Summer. 2001.