Assessing Privacy Implications and Data Quality Issues for the Implementation of a Provincial Client Registry

B00421725
lypowy@dal.ca

Performed at

Nova Scotia Department of Health
1690 Hollis Street
Halifax, Nova Scotia
B3J 2R8

HINF 7000

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

Report of Internship for the period May 5 – August 31, 2005

Date Submitted: *September 26, 2005*

Acknowledgements

I am grateful to Steven Parker and Carole Arsenault of The Barrington Consulting Group for providing the opportunity for this rich internship. I would also like to thank Lyn Kilroy, Business Solutions for approving the addition of one more resource to the project team, and for providing all required resources within the Nova Scotia Department of Health. Michelle Gignac, Information Access & Privacy Unit, has my gratitude for being so generous in her availability and guidance throughout this internship. Many thanks go to Dr. Sunny Marche, Masters of Health Informatics Executive Committee for his help in clarifying the scope of portions of this report, and to Dr. John McHugh, Director of the Privacy and Security Laboratory, Faculty of Computer Science for his guidance on sources for the study of Intrusion Detection Systems.

The following project team members and authorities at the Nova Scotia Department of Health have reviewed this report and approved of its content for release into the public domain:

- Carole Arsenault, Associate Partner, The Barrington Consulting Group
- Lyn Kilroy, Manager, Business Solutions, Nova Scotia Department of Health
- Michelle Gignac, Acting Manager, Information Access and Privacy Unit, Nova Scotia Department of Health

My wife Christine humbles me with her support and inspiration. She also provided the sweetest distraction an internship could ever have – the birth of our son Aleksander.

Endorsement

This report has been written by me and has not received any previous academic credit at this or any other institution.

_____

Executive Summary

Once implemented, a Client Registry is considered to be the authoritative source of patient demographic information throughout a jurisdiction. This makes it an important foundational element of the Electronic Health Record and a key element for study in any Health Informatics program. This past summer the author spent his internship at the Nova Scotia Department of Health working with a team which, in partnership with Canada Health Infoway, performed detailed planning for the implementation of a prvince wide Client Registry. Acting as the Privacy Lead for the team, the author conducted a Privacy Impact Assessment for the Client Registry which identified two key implementation issues. The author also contributed research on best practices in Data Quality Assessment and Data Cleansing, offered an approach to 'seeding' the Client Registry with data from stakeholder systems, and aided with the Use Case analysis of the processes underlying the use of the Client Registry. All documents generated by the author were submitted to Canada Health Infoway for their approval, as part of the requirements of the project.

While conducting the Privacy Impact Assessment, the author identified an opportunity for improvement in the handling of log files generated by Client Registry stakeholder systems. A high-level solution architecture is offered, based upon an established Intrusion Detection System model. The author proposes using Data Mining techniques to analyze system log files looking for abnormal user behaviour (anomaly detection) as an indication of a breach of privacy.

Table of Contents

1. Introduction

This document describes the author's internship experience while working on the Client Registry Phase 1 Detailed Planning Team at the Nova Scotia Department of Health (NSDoH). For the period of May 5 - August 31, 2005 the author was part of a team working in partnership with Canada Health Infoway (CHI) to develop requirements, high-level design, planning deliverables, and a budget to enable the NSDoH to both procure and implement a provincial Client Registry.

For the duration of the internship the author was officially retained as a sub-contractor to the Barrington Consulting Group, the company contracted by NSDoH to complete the Phase 1 Detailed Planning project. All work done by the author was performed under the authority of the NSDoH's Business Solutions department.

With a mission to, "…promote, maintain and improve the health status of Nova Scotians at a cost that is sustainable for Nova Scotia…" [1], the Nova Scotia Department of Health is the funding and oversight body for healthcare in Nova Scotia. The Business Solutions department has a province-wide mandate to deliver information systems and technology solutions to meet the information management needs of the NSDoH. The author also worked closely with the Information Access and Privacy Unit of the NSDoH. This department leads the development of policy, standards and processes for the responsible management of health information throughout the province, with a focus on access and privacy.

2. Author's Role and Responsibilities

Working as the Privacy Lead for the project team, the author's primary responsibility was to plan and conduct a Privacy Impact Assessment of the implementation of the Client Registry. The output from this assessment, the PIA Report, was one of the required deliverables to be submitted to CHI for their approval. As with any assessment process, there were a series of reviews which had to take place before

the final PIA report was accepted as complete. While awaiting responses during these reviews, the author was given the following additional responsibilities:

- Conduct research into Best Practices for Data Quality Assessment and Data Cleansing;
- Provide an approach for loading the Client Registry with data from other systems; and,
- Aid in documenting Client Registry-related business processes through Use Case analysis.

As stated in the application for this internship, the author was slated to be involved in the creation of a Knowledge Management Strategy for the project. Unfortunately, due to the timing of the start of the internship this portion of the project was near completion when the internship began. Therefore the author did not participate in that portion of the project other than to share his knowledge of the topic of Knowledge Management with the team's Project Manager, and to review the requirements (as mandated by CHI) for the Knowledge Management Strategy.

3. Client Registry in Nova Scotia

A brief description of the proposed provincial Client Registry is offered to help provide context for the discussion of the work performed by the author.

A Client Registry provides a single data warehouse, or "White Pages", of demographic information to ensure the comprehensive and unambiguous identification of clients of health services (i.e. patients). Once implemented, it will be considered the authoritative source of patient demographic information throughout the province. A high-level view of the patient data flow with the Client Registry is depicted in Figure 1.

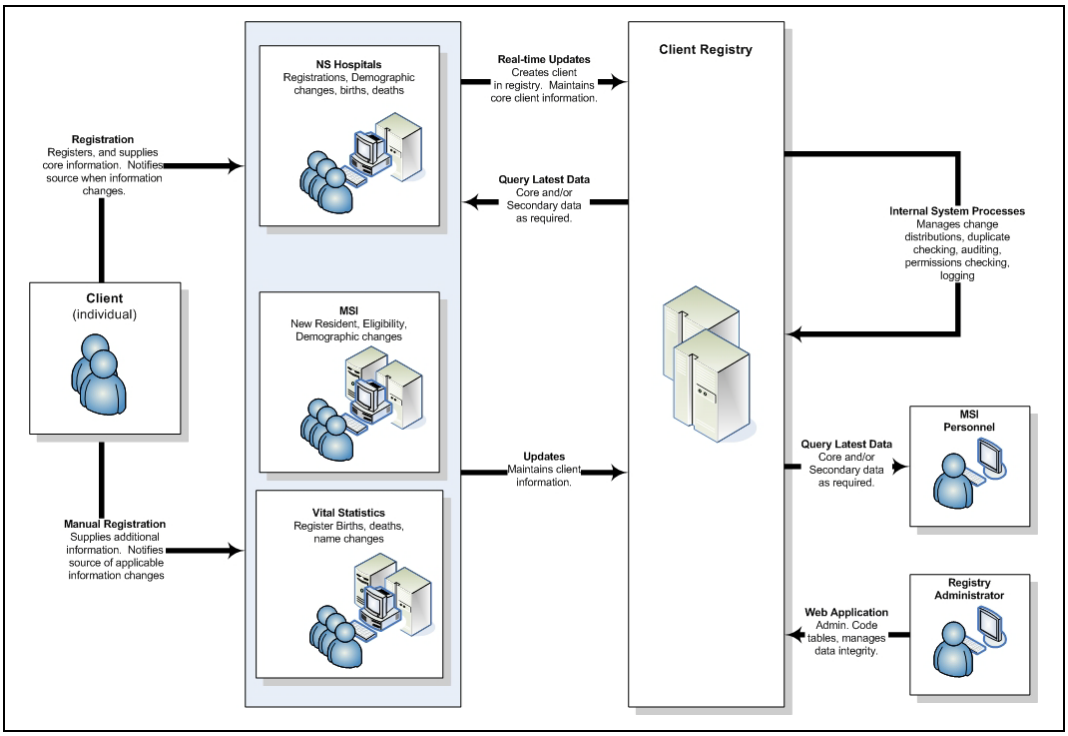


Figure 1 - Interaction model for the Client Registry in Nova Scotia [2].

The initial users of the Client Registry will come from four primary stakeholder groups, including acute care facilities throughout the province[1], Medavie (the province's publicly-funded health services insurer), the Nova Scotia Government's Department of Vital Statistics, and the NSDoH (who will be responsible for administration of the Client Registry). Each of the first three stakeholder groups currently relies upon its own local registry system (also called a Master Patient Index, or MPI) for identifying and managing patient demographic information.

The Client Registry (which is also referred to as an Enterprise Master Patient Index, or EMPI) will not replace any of the current stakeholder systems. Instead, interfaces will be developed linking it with these systems to facilitate the sharing of patient demographic information. Users of stakeholder systems will continue to use their local MPI, and may not even be aware that they are also accessing the Client Registry. A conceptual view of the architecture of the Client Registry is included in Figure 2.

[1] This includes all hospitals in District Health Authorities 1-8, the Isaak Walton Killam Health Centre (IWK), and the facilities comprising the Capital District Health Authority (CDHA).

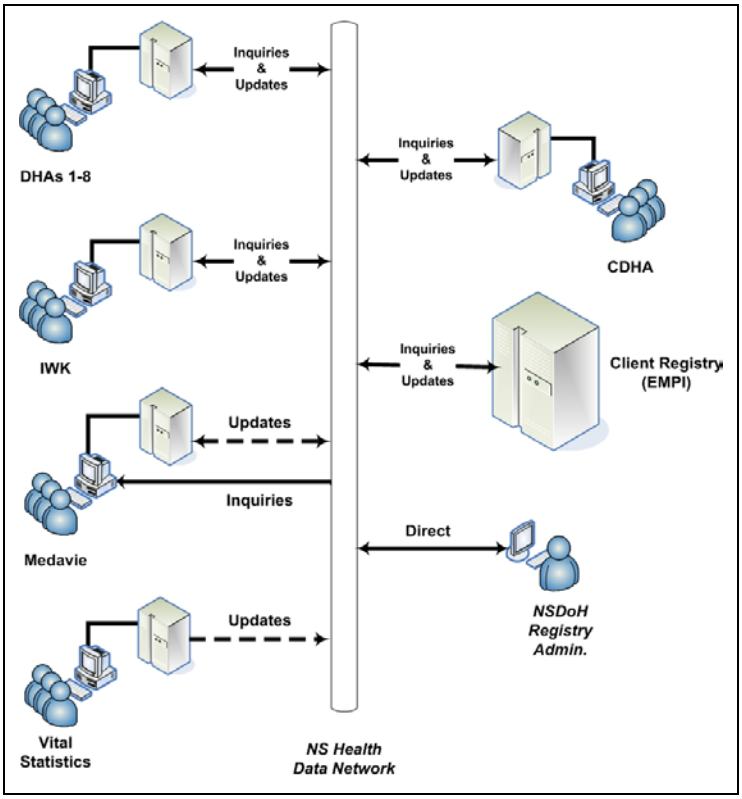


Figure 2 – Conceptual Technical Architecture for Nova Scotia Client Registry (adapted from [2]).

The primary difference between the existing and the "to-be" flow of information once the Client Registry has been implemented will be in the processes used by acute care centres to locate and update patient information. Consultations with vendors and other health jurisdictions (who have implemented a Client Registry) have helped to develop the process (Figure 3) which will be used to locate, add, and update patient demographic information once the Client Registry is implemented.

When locating a patient's demographic information, the local MPI system is always searched first. If the patient is found in the local MPI, then more detailed search information (e.g. Health Card Number or local Facility Unit Number) is used to acquire the patient's information from the Client Registry. If the patient is not located locally, then a more 'open ended' search (which could have multiple matches) is performed against the Client Registry.

This process is admittedly counter-intuitive, as one would expect to search the Client Registry before looking at the local registry. However it has been proven that in order to optimize the performance of the Client Registry application this process must be followed.

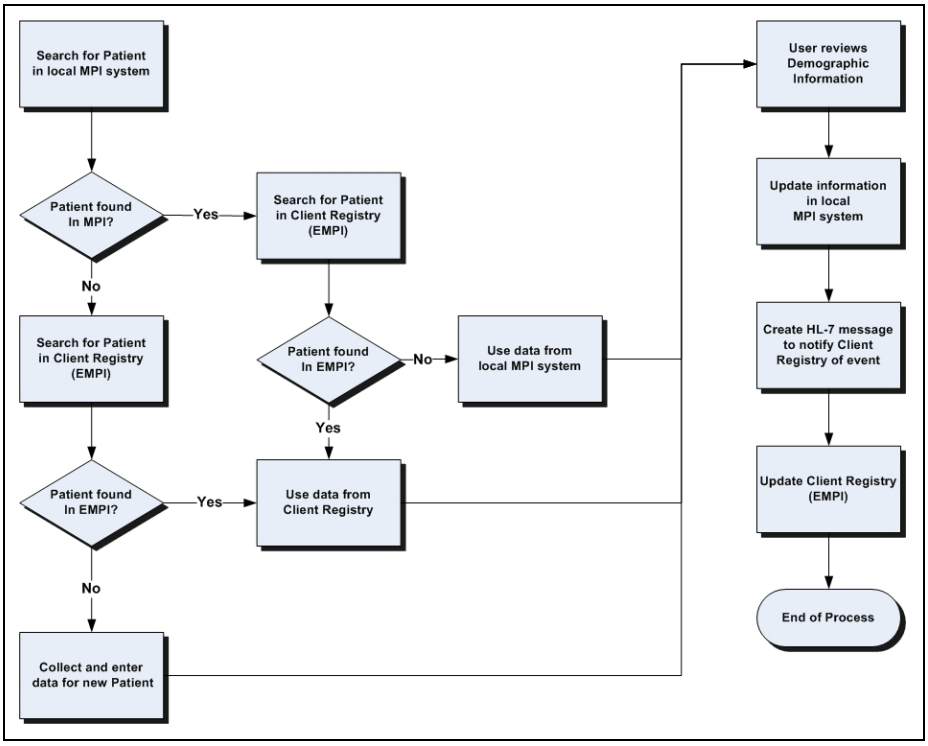


Figure 3 – Client Inquiry on a MEDITECH MPI system (adapted from [2]).

With numerous stakeholder systems acting as the sources for patient information, it is vital that the Client Registry include key fields from these systems to maintain linkages with them. A minimal set of fields (the Required Data Set) has been proposed containing only those data elements deemed essential for unambiguous patient identification and linkage with stakeholder systems. It should be noted that the Required Data Set will not include any clinical data. The design of the Required Data Set will allow stakeholder groups to maintain the same client identifiers they do today in their source systems, and to use this data to search the Client Registry.

4. Work Performed

The work completed by the author during this internship centered around three project areas:

- Assessment of privacy of information once the Client Registry is implemented;
- Data quality assessment and cleansing prior to loading data into the Client Registry; and,
- Business process analysis and Use Case creation.

4.1 Privacy Impact Assessment

Health care related projects deal with some of the most sensitive information currently being collected on individuals in Canada. Using the Privacy Impact Assessment (PIA) process, a health care organization can have assurances that privacy issues and impacts have been identified, and that compliance with privacy legislation and policies will be maintained before implementation begins. Canada Health Infoway now requires that a PIA be completed for all projects which it funds.

The process of completing this PIA was complicated by the fact that the Client Registry will be interfaced with five other pre-existing systems. The challenge was to determine how much scrutiny to pay to these stakeholder systems without conducting a PIA on each.

To complete the PIA for the Client Registry, the author:

- Met with the team Project Manager and the Manager of the Information Access and Privacy Unit to discuss requirements and receive guidance on completing the PIA, including discussion of the legislation pertaining to management of personal health information in the province;
- Conducted a self-study crash course on the Client Registry;
- Performed a gap analysis to determine what information existed, and what was outstanding;
- Developed a list of questions and scheduled interviews with stakeholders and team members to learn about privacy levels in stakeholder systems, and those proposed for the Client Registry;
- Studied all collected information, and generated a draft of the PIA Report;
- Reviewed PIA report and conclusions with the Manager of Information Access and Privacy Unit;
- Submitted draft PIA Report to key members of the project team for their input; and,

- Incorporated feedback from all reviewers and prepared final version of PIA Report for submission to CHI.

As a result of the PIA, two issues – one regarding the data contained in the Required Data Set, and one dealing with access to the Client Registry – were identified and will be dealt with before the next phase of the project begins. Had these issues been identified during implementation, they may have caused delays and potentially compromised the existing level of privacy of the data.

Unfortunately, due to the sensitive nature of the material which it contains, the entire contents of the PIA could not be included with this document. However, the Table of Contents from the document has been included as Appendix A.

4.2 Data Quality Assessment and Data Cleansing

Canada Health Infoway also required that the project provide a document detailing what data fields would be included in the Client Registry (the Required Data Set), how they would map to the data fields contained in each stakeholder system, the procedure used for the initial load of data into the Client Registry, and what standards would be used to impose order on this process. This document was known as the Data Mapping and Standards document.

The author was asked to submit an overview of Best Practices in Data Quality Assessment and Data Cleansing, and to provide an approach for the initial loading (seeding) of the Client Registry with data from each stakeholder system. Completing the Best Practices study involved reviewing both academic and business sources, including documents provided by CHI, and generating the overview for submission to the team Project Manager.

To complete the approach for seeding the Client Registry, the author:

- Performed a gap analysis to determine what information existed, and what was outstanding;
- Developed a list of questions and scheduled interviews with each stakeholder to determine the current level of data quality within their system, and the data quality practices used;

- Reviewed both academic and business sources, including documents provided by CHI and project documentation for information on pre-seeding, seeding, and ongoing approaches to data quality assessment and cleansing; and,
- Generated a document detailing the approach and submitted it to the team Project Manager.

Both the resulting overview (included as Appendix B in this document) and the approach to seeding were included in the Data Mapping and Standards document submitted to CHI. Unfortunately, due to the sensitive nature of the material which it contains, the approach to seeding could not be included in this document.

4.3 Business Process Design, and Use Case Analysis

One of the most important deliverables required by CHI was the Business Process Solution Design document. The goals of this document were to: [3]

- Describe the desired workflow for the Client Registry and stakeholder environments; and,
- Determine the impact on the affected stakeholder environments and existing business processes.

The author was asked to help capture and document workflow processes through the creation of Use Cases. This process included:

- Meeting with the Business Process Lead to review the Use Case analysis process;
- Locating and reviewing pertinent process documentation, including already constructed Use Cases;
- Attending and/or conducting interviews with stakeholders to capture business process flows;
- Providing both a textual description and a graphical depiction of each Use Case; and,
- Submitting Use Cases for inclusion in the final version of document.

All Use Cases constructed by the author (an example of which is included as Appendix C in this document) were included in the Business Process Solution Design document and submitted to CHI.

5. Internship Relevance to Health Informatics

> *"A cornerstone of the EHR will be the interactions and data required to register clients and maintain information that uniquely identifies them."* [4]

One of the overarching goals of Health Informatics is to support the development of a secure and effective Electronic Health Record (EHR). Canada Health Infoway, one of the primary sources of funding for Health Informatics projects in Canada has identified the Client Registry as a key component of the development of the EHR. Therefore work related to the implementation of the Client Registry in Nova Scotia has direct relevance to Health Informatics. All work performed by the author during this internship aligns with the goals of the Master's of Health Informatics program:

- Conducting a Privacy Impact Assessment on the Client Registry relates to the importance of privacy and security issues surrounding the management and use of electronic health information;
- The Best Practices study of Data Quality Assessment and Data Cleansing, and the development of an approach for Client Registry seeding relates to understanding the challenges of integrating existing databases of electronic health information, the importance of standards where multiple sources of information are being merged, and to the processes involved with implementing a new Health Informatics-based service; and,
- Working on Use Case analysis of Client Registry business processes relates to the human side of Health Informatics – the effects of new uses of electronic health information on patients and on the users of the systems which manage it.

6. Log File Auditing and Privacy Breach Detection

During the course of his internship, the author identified an opportunity for improvement in the handling of system log files. The following is a description of the problem and a framework for a solution.

6.1 Problem Definition

Most computer systems generate log files based upon user activity, and this is true of many of the systems in use in the healthcare industry today. These log files, containing a record for each access a user makes to a system or its information, are normally collected and reviewed by systems operations staff. The tools provided by the system for the management of these files (e.g. log file rotation, deletion, import of log data into a database, etc.) usually have limited filtering capabilities for locating certain important events (e.g. failed login attempt, denial of access to information, etc.). More rigorous methods, including those used to identify when breaches of privacy have been committed by users of a system, normally have to be developed by the system owner.

For the sake of this problem definition, the author is considering a breach of privacy to occur when a user of a system accesses a patient record for which they have no authorization. Whether for curiosity's sake (e.g. their neighbor is a patient in the facility), or perhaps for more devious reasons, any unauthorized access should be identified and investigated. Section 71 of the Nova Scotia Hospitals Act stipulates the confidentiality of, "…records and particulars … concerning a person or patient" [5], and guarantees patients access to information pertaining to them. This includes access to the information contained in these stakeholder system log files.

In order for medical records systems to be useful, they must provide the user with immediate access to any required information. Due to the nature of the job, it is often difficult to predict which patients' information a staff member will need to access over the course of a shift. To impose limits on access to this information, with the intent of making the system more secure, may hinder its usefulness and introduce delays in the provision of care. The best compromise in this situation may be to ensure that all user activities are logged, and then review the logs looking for "suspicious" accesses. But what constitutes a "normal" or "suspicious" access? The challenge with detecting this sort of breach by perusing log files is that it requires a more sophisticated filtering process.

6.2 Proposed Solution

Manual (human) review of system log files could be one approach to identifying this sort of breach, assuming the reviewer is familiar with the user population being studied. However with the magnitude of these system logs, the mundane nature of the task, and the continual competition for resources in the Canadian Healthcare System, this is not a practical solution to this problem. The solution is to automate the process of recognizing (classifying) these breach situations, with a high level of accuracy to help prevent false positive findings.

Initial investigations into a solution led the author to consider mechanisms used by credit card companies to approve or reject customer transactions. Renowned for their abilities to identify out of the norm cardholder behaviour (as an indication of fraud), this technology seemed applicable to this problem. Brause et al [6] describe just such a system, employing data mining and neural network techniques and offering highly sensitive detection with a low false positive rate. Chan et al [7] also use data mining, and even claim some good results when they apply their techniques in an Intrusion Detection System (IDS). Sadly, neither of these papers provides any details regarding the models or algorithms used for detecting abnormal (fraudulent) behaviour (for obvious reasons).

Lee et al [8] consider Intrusion Detection to be a data analysis problem, and proceed to apply data mining methods such as outlier detection to help classify ‘normal’ and ‘abnormal’ behaviours. Their motivation for using Association Rules to help build their system aligns well with the problem at hand:

- logs from stakeholder systems can easily be formatted for import into a database;
- the population of users on stakeholder systems is well known, and their behaviours will be quite consistent; and,
- aggregating rule sets may accommodate the complex nature of some healthcare jobs. For example, being able to correctly identify a user’s system accesses from a machine on a different network, at an unusual time of day, but with the regular volume of records processed as being a shift change as opposed to a breach.

The techniques used by Lee et al, including their statistical analysis methods, seem highly applicable to the problem at hand, but they utilize a data model which is not.

In his seminal work in the area of IDS, James Anderson identifies the various threats and attacks which can be committed against a system, how they are documented in audit data, and the classes of users who commit them. He refers to the author's concept of a breach of privacy as 'misfeasance', and defines it as an authorized user of a system, with authorized access to data, making deliberate unauthorized accesses to data [9]. Users committing misfeasance are defeating procedural controls over the data, not system controls, so detection is difficult. However users do, over time, establish patterns for their usage of a system, and this can be studied by monitoring behaviours of system usage. This early work in anomaly detection uses audit logs to help identify breaches through detecting changes in user behaviours. Anderson's solution (developed in 1980 for a customer, and including an estimated schedule and statement of work) describes a batch-mode system consisting of two sub-systems:

- a Security Surveillance Subsystem for identifying anomalous user behaviour; and,
- a Security Trace Subsystem which would provide a detailed activity log in support of an anomalous activity identified by the Surveillance Subsystem.

In another seminal work in the study of IDS, Denning [10] offers a system independent framework for real-time intrusion detection by identifying anomalous behaviour in a user population through audit log analysis. Her Intrusion Detection Expert System (IDES) framework was the basis for the development of SRI's Next-Generation Intrusion Detection Expert System (NIDES)[2] and the Event Monitoring Enabling Responses to Anomalous Live Disturbances (EMERALD)[3] system. It will also work well as the model for the author's proposed solution architecture.

Denning's data model consists of Subjects, Objects, Audit Records, Profiles for defining user behaviour, Anomaly Records, and Actions Taken [10]. Each of these relates well to the current problem:

Subjects

This includes the users of the stakeholder systems.

[2] For more information on NIDES see http://www.sdl.sri.com/projects/nides .
[3] For more information on EMERALD see http://www.sdl.sri.com/projects/emerald .

Objects

This would include personal health information records held in stakeholder systems, and the stakeholder systems themselves.

Audit Records

The NEMA/COCIR/JIRA Security and Privacy Committee provides some coverage of what events need to be captured in an audit log for a health-related system [11]. These events are broken down as interactions with personal health information (create, modify, view, and delete record) and other informative events (logins, location information, session auto-logouts, etc). It also advocates the inclusion of limited transaction data, for instance patient record ID, in the log file. This last item may not end up being considered depending upon what data it will hold (privacy advocates may not want too much patient-specific data being included in a system log). Audit Records for the proposed solution will consist of information culled from the various stakeholder system log files. An audit record will be a 7-tuple including: Subject ID, Action, Object, Location (e.g. IP address), Time-stamp, System ID, Transaction ID. The second last field, System ID, is included to allow for the provision of processing of log files from all interconnected systems (or systems which share a central data store) to enable detection of more sophisticated breaches. An example would be to detect simultaneous usage of the same logon ID on multiple systems.

Profiles

Profiles are created to define the normal behaviour of Subjects (with respect to how they interact with Objects). Normal subject activity can be learned through observation, or "experts" can create templates from known behaviours and then let them evolve[4]. Templates can be created for classes of subjects as well. An example of this would be to create a template for a staff position (e.g. Registration Clerk), and then use it to create a profile for auditing those staff. There is a

[4] Sundaram [13] cautions that while one of the strengths of statistical systems such as these is their ability to readily learn user behaviours, making them potentially more aware than human reviewers of log files, it is possible for intruders to introduce gradual changes in behaviour which would cause the system to ignore certain anomalous actions.

weakness with pre-created templates - they are biased by the level of knowledge or experience of the expert who creates them.

The sorts of behaviours to be profiled could include (but would not be limited to):

Category	Attribute(s)	Sample Behaviour Tracked
Temporal	Time-stamp	Accesses to records by hour of day (e.g. business hours) and day of week.
Locational	Location	Accesses from unusual location (IP address).
Quantity	Time-stamp	Number of records usually processed by user each day / week / month.
Quality	Action	Types of transactions being made by object, or types of transactions normally generated (e.g. the previous pattern of View only transactions suddenly become View and Modify).
Functional	Transaction ID, System ID	The population of patients accessed (e.g. discovery of Cardiac staff viewing orthopedic patient records.)
Security	Location, System ID	Multiple simultaneous logons / attempts.

Anomaly Records

When an anomaly is encountered, it will cause an Action to be Taken. An example of this would be to fire a procedure which writes that anomaly record (and supporting information) to a database table. Later this table can be used to generate reports, by System ID, to be sent to stakeholder security personnel.

Actions Taken

This can include specific actions to be fired depending upon the anomaly. An example would be that anomalies dealing with records belonging to patients who are considered to be "high profile" (e.g. celebrity) or that contain information which is sensitive in nature (e.g. HIV/AIDS positive) will cause the anomaly records and support information to be marked as PRIVATE within the reporting table.

It is beyond the scope of this paper to specify a detailed design of a system to be used to implement the above data model and study of user behaviours. However, a proposed high level solution architecture using Denning's [10] data model and based upon the work of Lee et al [8], would look something like the following:

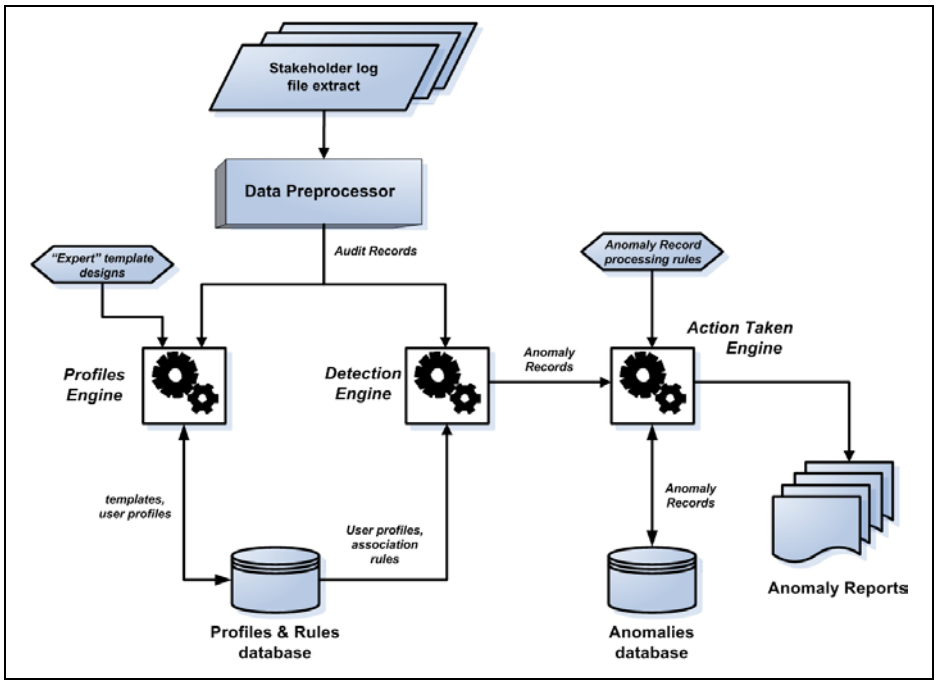


Figure 4 – Proposed solution architecture (adapted from [11]).

Log file extracts of predefined format and for a specified time period (e.g. a day, week, or month) would be provided by stakeholders. These files would be preprocessed to perform any required data cleaning and transformation of the records. The resulting audit records would be sent to both the Detection Engine, which would use Data Mining techniques to identify any breaches, and to a Profiles Engine responsible for using the audit records to maintain the database of user profiles and classifiers. The Profiles Engine would also accept template designs from "expert" users for creating new user profiles. As abnormal behaviours are identified, anomaly records are cut and stored in an Anomalies Database. An Action Taken engine would monitor the contents of the Anomalies Database, taking steps to process anomaly records according to specified rules. Anomaly Reports would be generated for each stakeholder, providing details for further investigation.

This high level architecture would accommodate log files from current and future stakeholder systems. It is conceivable that a system implemented using this architecture could also analyze log files created by intelligent biomedical devices. For example, if extracts of the log files from a digital radiography machine could be generated in the requisite format, they could be analyzed for breaches.

7. Conclusions

Canada Health Infoway considers the Client Registry to be an important foundational element of the Electronic Health Record [4], and for good reason. The benefits which could be realized by Nova Scotia's implementation of the Client Registry include:

- Increased efficiency in the system (reduction of manual processes and duplicate data entry);
- Increased accuracy of patient identification and information;
- Reduced duplicate / multiple identification information for the same patient; and,
- Evolution towards a provincial – and national – Electronic Health Record solution.

Any system implemented or modified to support this vision must not compromise the current level of data privacy guaranteed by current information systems. By requiring all projects which it funds to conduct a Privacy Impact Assessment, Canada Health Infoway is endeavouring to ensure that no patient loses the right to privacy of their data in the name of progress. The PIA for the Client Registry identified two issues which, if left undiscovered when implementation began, could have proved costly.

Implementing a new system which integrates data from multiple sources can be a complex undertaking. The current state of quality of each data source, the amount of data which needs to be cleaned, and the process for loading the data into the Client Registry both initially and then as updates on an ongoing basis are all important considerations.

When the implementation of any new system could affect the daily work processes of facilities throughout the province, it is important to understand which processes will change, which will remain unchanged, and what impact this will have on the personnel executing them.

There is a fine balance which must be struck between providing staff with the access they need to data on a system, and the rights of patients to have their information guarded from unauthorized accesses. Methods such as Data Mining can provide the sophisticated analysis needed to detect these breaches of privacy by analyzing system log files and identifying abnormal patterns in user behaviour with respect to system usage.

8. Recommendations

Pending NSDoH budget approval, it is hoped that the implementation of the Client Registry in Nova Scotia will start sometime in 2006. Once the technical design for the Client Registry has been finalized, and the required hardware and software products have been identified, Canada Health Infoway will require that a more detailed Privacy Impact Assessment be completed. This, however, should not be the final update made to the PIA for the Client Registry.

It is the intention of the PIA process that the PIA report be considered a living document; that it grows with the system which it describes. Therefore it is recommended that in the next phase of this project a resource be identified who will "own" the PIA report for the Client Registry. This resource, which will probably come from the NSDoH team responsible for maintaining and administering the Client Registry, will need to ensure that each modification or upgrade to the Client Registry system is duly documented in the PIA report and assessed for any possible privacy implications.

All data to be loaded into the Client Registry will be collected from each stakeholder and stored in a Client Registry Test Environment (a staging area). This staging area will be configured similarly to the Client Registry, and will allow personnel to test and prove the loading process. The following is the recommended approach (as included in [14]) for preparing and loading the Client Registry with stakeholder data in the next phase of the project:

1. Determine who at NSDoH is responsible for the data quality of the Client Registry.
2. Hold workshops with Data Integrity representatives from each stakeholder to:
    a. Present the data mapping and formatting standards as ratified by the provincial Admissions and Registration Policy and Standards Advisory group;
    b. Identify all types of data errors to be fixed before seeding the Client Registry;
    c. Determine when each data error type will be addressed (pre-seeding, during seeding, or just fix it on an ongoing basis);

    d. Discuss the design of the utility/environment which will be used to seed the Client Registry; and,
    e. Discuss strategies for making changes to existing business processes to prevent ongoing data corruption in stakeholder systems.
3. Support/aid stakeholders with their efforts to cleanse their data.
4. Collect all data in Client Registry Test Environment (staging environment) and run final data cleansing process.
5. Configure Integration Engine and/or Client Registry to format and cleanse data on import.
6. Perform a test seeding in the Client Registry Test Environment.
7. Address any problems, and seed Client Registry with cleansed data.

The log analysis architecture described in this document is a mere starting point for addressing the privacy breach detection problem facing Client Registry stakeholders. The next step in developing a solution would be to build a prototype system based upon this architecture, and to demonstrate it to interested NSDoH and stakeholder personnel. During the coming semester, the author intends to use the term project in HINF6210 (Data Mining for Health Informatics) to develop a proof of concept of the Data Mining component of a prototype system. It is hoped that sufficiently interested parties (either within the Province or perhaps in other jurisdictions) will make it viable for the author to explore the further development of this solution.

References

[1] http://www.gov.ns.ca/health/about.htm, About the Nova Scotia Department of Health (last accessed 21-SEP-2005).

[2] The Barrington Consulting Group. Client Registry Detailed Planning Phase 1 Project - Privacy Impact Assessment. Nova Scotia Department of Health, Halifax, Nova Scotia, Canada, Project Deliverable (2005).

[3] The Barrington Consulting Group. Client Registry Detailed Planning Phase 1 Project – Business Process Solution Design. Nova Scotia Department of Health, Halifax, Nova Scotia, Canada, Project Deliverable (2005).

[4] Canada Health Infoway. EHR : Canadian Client Registry Standards Project. http://www.infoway-inforoute.ca/ehr/stnd_prj_client.php?lang=en (last accessed 21-SEP-2005).

[5] http://www.gov.ns.ca/legi/legc/statutes/hosptls.htm, Hospitals Act Chapter 208 of the Revised Statutes, 1989 (last accessed 21-SEP-2005).

[6] Brause R, Langsdorf T, Hepp M. Neural Data Mining for Credit Card Fraud Detection. Proceedings 11th IEEE International Conference on Tools with Artificial Intelligence, 1999. pp. 103-106.

[7] Chan PK, Fan W, Prodromidis AL, Stolfo SJ. Distributed Data Mining In Credit Card Fraud Detection. IEEE Intelligent Systems, Vol. 14, Issue 6, pp. 67-74 (November-December 1999).

[8] Lee W, Stolfo S. Data Mining Approaches for Intrusion Detection. Proceedings 7th USENIX Security Symposium, 1998.

[9] Anderson JP, Computer Security Threat Monitoring and Surveillance, James P. Anderson Co., Fort Washington, PA (1980). http://csrc.nist.gov/publications/history/ande80.pdf (last accessed 21-SEP-2005).

[10] Denning DE. An Intrusion Detection Model. IEEE Trans. Softw. Eng., Vol. SE-13:2, pp. 222-232 (February 1987).

[11] NEMA/COCIR/JIRA Security and Privacy Committee. Security and Privacy Auditing in Health Care Information Technology. White Paper (November 2001). http://www.nema.org/prod/med/security/upload/Security_and_Privacy_Auditing_In_Health_Care_Information_Technology-November_2001.pdf (last accessed 21-SEP-2005).

[12] Vannan E. Quality Data - An Improbable Dream?. Educause Quarterly, No. 1, pp. 56-58 (2001).

[13] Sundaram A. An introduction to intrusion detection. Crossroads: The ACM student magazine, 2(4) (April 1996). http://www.acm.org/crossroads/xrds2-4/intrus.html (last accessed 21-SEP-2005).

[14] The Barrington Consulting Group. Client Registry Detailed Planning Phase 1 Project - Data Mapping and Standards. Nova Scotia Department of Health, Halifax, Nova Scotia, Canada, Project Deliverable (2005).

Appendices

APPENDIX A – Table of Contents from Client Registry PIA

The following is the Table of Contents from the *Privacy Impact Assessment for the Nova Scotia Client Registry Detailed Planning Phase 1 Project* [2].

Table of Contents

APPENDIX B – Best Practices in Data Quality Assessment & Data Cleansing

The following excerpt was written by the author and included in the deliverable *Client Registry Detailed Planning Phase 1 Project - Data Mapping and Standards* [14].

Data Quality Assessment and Cleansing

Goals and Objectives

The primary goal of the data quality assessment and data cleansing process is to ensure that only 'good' data makes it into the Client Registry. Data is considered to be 'good' if it meets the following criteria [12]:

- *Accurate* – the data does not contain any errors;
- *Complete* – all required fields in a record contain a value;
- *Consistent* – the data conform to a set of rules, and are managed in a consistent manner;
- *Timely* – the data is available when required and is not older than the defined history period; and,
- *Flexible* – the meaning of the data can be understood so that it can be analyzed.

By addressing the quality of the data coming from each data source (in this case, each stakeholder system) it is hoped that the following practices and procedures will produce a Client Registry which is loaded with – and continues to contain – only 'good' data.

Best Practices

Data quality issues are almost a given to any sufficiently large and complex database - especially when humans are entering and updating the data within. Problems usually compound when multiple systems are being used to feed data to a single system. In these cases, consensus must be achieved on the format of all common data elements for the new system. This will entail adopting a common standard for the formatting of all data fields managed by multiple sources. This standard will then be employed by the data cleansing process to ensure the consistency and integrity of all data.

The most common industry practice is to cleanse data in the source system. This will help to promote the modification of any faulty practices or processes which are introducing defects into the data. If data cannot be cleansed at the source, it should be assessed using the five criteria mentioned earlier to determine what level of cleansing is required before it is used. Cleansing can then take place to ensure that this data meets the same level of quality of other data being used.

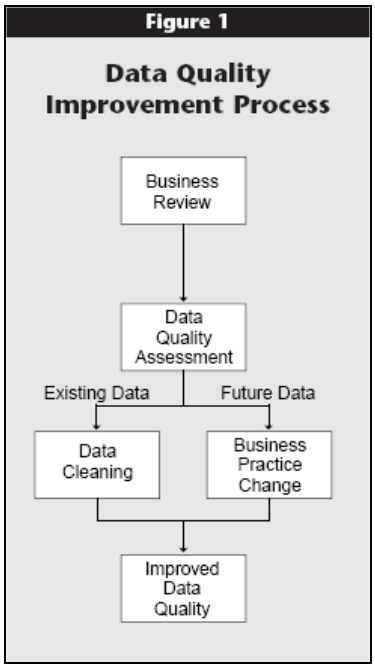
Regardless of where the data cleansing effort takes place, the first step in getting data ready for use is to identify the various error types to be rectified, and the volume of each error type present in the data. Next, all sources must agree on the errors which need to be rectified. Some errors may require much more time and resource to fix, and if their frequency (and severity) is sufficiently low it may be the case that they are left to be addressed later by a more advanced data cleansing effort.

When all error types have been identified, reports can be generated to show offending records and to support initial estimates of the magnitude of each error type. These reports will be re-run after the data cleansing effort to ensure that all errors have been fixed, and that no other data was harmed in the process.

There are three methodologies which can be used to cleanse the data, and the selection of one (or more) will be based upon the complexity and magnitude of the error type:

- **Automated cleansing** involves developing scripts or programs which are used to process the data. This methodology is most useful for high frequency errors or if resources are scant for the data cleansing effort.
- **Manual cleansing** is used for errors which require logical conclusions or more involved rule sets to fix. Users with knowledge of the data are used to manually process the data to fix the identified errors. This methodology can be resource intense if the magnitude of the errors being processed is great.
- It is quite common to use a **combination** of both automated and manual cleansing methods. In this scenario errors are categorized by magnitude and complexity to fix and are then cleansed using the most appropriate/efficient methodology.

Keeping data clean after a data cleansing effort involves a four step process, as follows [12]:



1. To improve data quality going forward, an institution must be prepared to thoroughly analyze existing business processes to identify all users of the data and to begin identifying problems with processes which may be causing data corruption.

2. The completed data cleansing process will have yielded valuable metrics on the current level of quality of the data. A data assessment must be repeated on a regular basis to track changes in data quality.

3. Once the source of an issue is determined, permanent changes must be made to business processes to maintain the quality of data going forward. To promote accountability, a "data custodian" can be assigned to maintain quality of portions of the information collected.

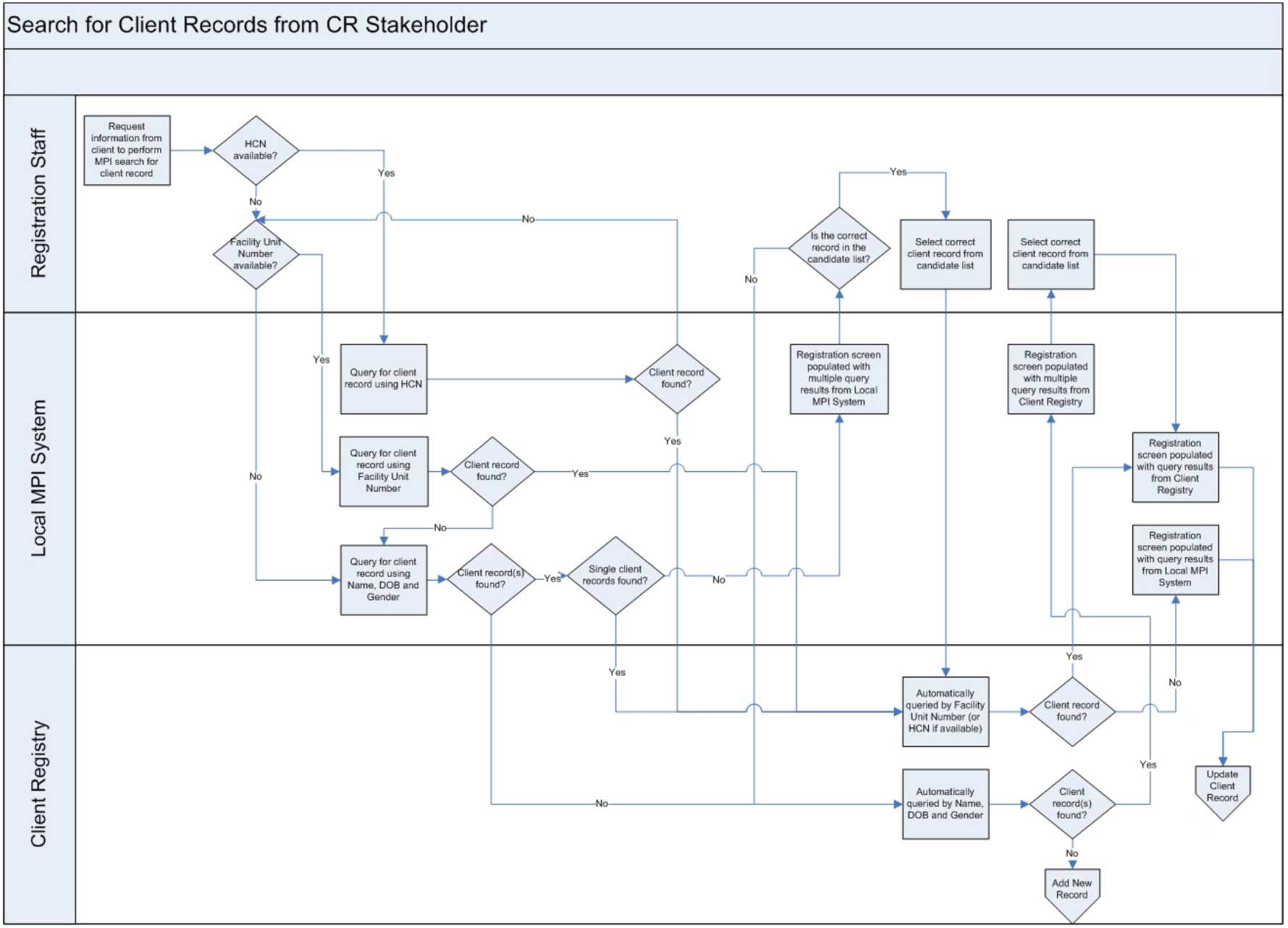
4. An institution must be dedicated to reviewing and improving data quality regularly. This can include appointing a resource responsible for monitoring overall data quality, creating policies and procedures for addressing data quality issues, and ensuring that all users are educated about the importance of maintaining the quality of data.

APPENDIX C – Sample Business Process and Use Case Diagrams

The following Use Case was constructed by the author and adapted from one included in the deliverable *Client Registry Detailed Planning Phase 1 Project – Business Process Solution Design* [3].

Use Case Name	**8.1.1.1 Search Client Records from CR Stakeholder**
Actors	Registration Staff
Summary	The goal of this activity is to search the Client Registry for a particular client through the local MPI system to support the process of client registration. Client searches conducted in support of Medical Records activities are described in section 5.4.4 Data Maintenance. The following use case describes the process for conducting a client search from the Inpatient Admission Routine, however the use case description is almost identical for the client searches conducted through the following routines and Menu Options: • Inpatient Admission; • Inpatient Pre-Admission Edit; • Clinical Registration (Long and Short Form); • Emergency Room Registration; • Referred Registration; • Recurring Registration; • Surgical Day Care Registration; • Observation Patient Admission; and, • Registration from Community Wide Scheduling. In each case, the routines are used to perform client searches using standard search criteria and will return the most current information for that client.
Pre-conditions	The client that is being searched must have had previous interactions with any of the Priority 1 stakeholder groups and is therefore present in the Client Registry.
Normal Flow of Events	**Search by HCN** 1. Local MPI system is searched by HCN (either by swiping HCN card or entering HCN into Patient field). 2. Client is found within local system (if client is not found, go to "Search by Facility Unit Number"). 3. Client Registry is automatically searched by HCN. 4. If the client is found within Client Registry, the local Inpatient Admission screen is populated with client demographic information from the Client Registry. 5. If the client is not found within Client Registry, the local Inpatient Admission screen is populated with client demographic information found in the local MPI system. **Search by Facility Unit Number** 1. Local MPI system is searched by Facility Unit Number (by entering it into Patient

	field). 2. Client is found within local system (if client is not found, go to "Search by Other Query Parameters"). 3. Client Registry is automatically searched by Facility Unit Number (and HCN if available in local client record). 4. If the client is found within the Client Registry, the local Inpatient Admission screen is populated with client demographic information from the Client Registry. 5. If the client is not found within Client Registry, the local Inpatient Admission screen is populated with client demographic information found in the local MPI system. **Search by Other Query Parameters** 1. Local MPI system is searched by Last Name, First Name, Date of Birth and Gender. 2. If a single record is found in the local system, the Client Registry is automatically searched by the client's Facility Unit Number. If multiple possible records are found, the list of possible candidate clients is presented to the Registration Clerk to select the correct record. When the correct, single client record is determined, the Client Registry is automatically searched by the client Facility Unit Number. 3. If the client is found within Client Registry, the local Inpatient Admission screen is populated with client demographic information from the Client Registry. 4. If the client is not found within Client Registry, the local Inpatient Admission screen is populated with client demographic information found in the local MPI system. 5. If no possible records are found with the local system, the Client Registry is automatically searched using the other query parameters. 6. If multiple possible records are found, the list of possible candidate clients is presented to the Registration Clerk to select the correct record. When the correct, single client record is determined within the Client Registry, the local Inpatient Admission screen is populated with client demographic information from the Client Registry. ***If no records are found in the local MPI system or the Client Registry, the "5.4.2 Add New Client Record" use case is employed for registering a new client.***


Search for Client Records from CR Stakeholder
Registration Staff
Local MPI System
Client Registry
Request information from client to perform MPI search for client record
HCN available?
Yes
No
Facility Unit Number available?
Yes
No
Query for client record using HCN
Client record found?
Yes
No
Query for client record using Facility Unit Number
Client record found?
Yes
No
Query for client record using Name, DOB and Gender
Client record(s) found?
Yes
Single client records found?
Yes
No
Is the correct record in the candidate list?
Yes
No
Select correct client record from candidate list
Select correct client record from candidate list
Registration screen populated with multiple query results from Local MPI System
Registration screen populated with multiple query results from Client Registry
Registration screen populated with query results from Client Registry
Registration screen populated with query results from Local MPI System
Update Client Record
Automatically queried by Facility Unit Number (or HCN if available)
Client record found?
Yes
No
Automatically queried by Name, DOB and Gender
Client record(s) found?
Yes
No
Add New Record