A Cloud-based Surveillance and Performance Management Architecture for Community Healthcare

Benjamin Eze

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements

For the Ph.D. degree in

Electronic Business

School of Electrical Engineering and Computer Science

Faculty of Engineering

University of Ottawa

Abstract

Governments and healthcare providers are under increasing pressure to streamline their processes to reduce operational costs while improving service delivery and quality of care. Systematic performance management of healthcare processes is important to ensure that quality of care goals are being met at all levels of the healthcare ecosystem. The challenge is that measuring these goals requires the aggregation and analysis of large amounts of data from various stakeholders in the healthcare industry. With the lack of interoperability between stakeholders in current healthcare compute and storage infrastructure, as well as the volume of data involved, our ability to measure quality of care across the healthcare system is limited.

Cloud computing is an emerging technology that can help provide the needed interoperability and management of large volumes of data across the entire healthcare system. Cloud computing could be leveraged to integrate heterogeneous healthcare data silos if a regional health authority provided data hosting with appropriate patient identity management and privacy compliance.

This thesis proposes a cloud-based architecture for surveillance and performance management of community healthcare. Our contributions address five critical roadblocks to interoperability in a cloud computing context: infrastructure for surveillance and performance management services, a common data model, a patient identity matching service, an anonymization service, and a privacy compliance model. Our results are validated through a pilot project, and two experimental case studies done in collaboration with a regional health authority for community care.

Acknowledgments

First, I would like to appreciate and thank my supervisor, Dr. Liam Peyton for his support, guidance, and encouragement throughout my research work. He encouraged me to enroll in the Electronic Business Technologies Ph.D. program. Without his continuous guidance, I would not have been able to complete this work. I am profoundly grateful.

I am also deeply grateful to Dr. Craig Kuziemsky for his breadth of knowledge and perspectives that continued to align this work in the right direction. His contributions are invaluable to the completion of this thesis. I also wish to express the sincerest gratitude to Dr. Bijan Raheemi for accepting to be a member of my Thesis Technical Advisory Committee and his invaluable support throughout this process.

Special thanks go to my family for their support and unconditional love. Thanks for believing in me and for your prayers. Those helped me pull through the difficult days. Dad, I know you are no longer with us, but this is for you!

I would also want to give special thanks to Jamie Stevens from the Local Health Integration Network (LHIN) for the opportunities he has given me and his openness to working with me. His constant desire to try new ideas and his trust in me contributed immensely to the success of the pilot project and the experiments described in this thesis.

I also want to say a special thank you to Paul Boissoneault for taking the great initiative to start the community support services project at the Champlain LHIN. Special appreciation goes to all the current and past members of the SSO project team at the LHIN - Munro Ross, Jim Brophy, Christian Gagnon, and Mana Azarm-Daigle for contributing great ideas to my research and facilitating a lot of the implementation work.

Ontario Graduate Scholarships (OGS), a University of Ottawa Excellence scholarship, and NSERC supported this research.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronym	Definition
ACS	Access Control Services
ADFS	Active Directory Federation Service
API	Application Programming Interface
AR	Action Research
BPMH	Best Possible Medication History
CCO	Community Care Organization
CDA	Clinical Document Architecture
CDM	Common Data Model
CPG	Clinical Practice Guidelines
CSV	Comma-Separated Values
CT	Clinical Terms
CQI	Continuous Quality Improvement
DCS	Data Collection Service
DSA	Data Sharing Agreement
DSR	Design Science Research
EHR	Electronic Health Records
EMR	Electronic Medical Records
ER	Emergency Room
ETL	Extract Transform Load

Acronym	Definition
GAV	Global As View
HIPAA	Health Insurance Portability and Accountability Act
IaaS	Infrastructure-as-a-Service
ICD	International Classification of Diseases
ISO	International Organisation for Standardisation
JSON	JavaScript Object Notation
KPI	Key Performance Indicators
LAV	Local As View
LDAP	Lightweight Directory Access Protocol
LHIN	Local Health Integration Network
LOD	Linked Open Data
LOINC	Logical Observations Identifiers, Names, Codes
PaaS	Platform-as-a-Service
PCDD	Privacy Compliance Definition Document
PHI	Protected Health Information
PHIPA	Personal Health Information Protection Act
PHR	Personal Health Record
PIPEDA	Personal Information Protection and Electronic Documents Act
QI	Quasi-identifier
QoS	Quality of Service
REST	Representational State Transfer

Acronym	Definition
RDF	Resource Description Framework
RHA	Regional Health Authority
RM	Reference Model
SaaS	Software-as-a-Service
SIN	Social Insurance Number
SLA	Service Level Agreement
SNOMED	Systematized Nomenclature of Medicine
SOA	Service Oriented Architecture
SOAP	Simple Object Access Protocol
SPARQL	Simple Protocol and RDF Query Language
SQL	Structured Query Language
SS	Shared Services
SSL	Secure Socket Layer
SSN	Social Security Number
SSO	Single Sign-On
TA	Triple Aim
TCO	Total Cost of Ownership
TLS	Transport Layer Security
URL	Universal Resource Locator
VPN	Virtual Private Network
VM	Virtual Machine

Acronym	Definition
XML	Extensible Markup Language
XPATH	XML Path Language
XSD	XML Schema Definition

Chapter 1. **Introduction**

1.1. Problem Statement

Governments and healthcare providers are under increasing pressure to streamline their processes to reduce operational costs while improving service delivery and quality of care. Systematic performance management of healthcare processes is important to ensure that quality of care goals are being met at all levels of the healthcare ecosystem. The challenge is that measuring these goals requires the aggregation and analysis of large amounts of data from various stakeholders in the healthcare industry. With the lack of interoperability between stakeholders in current health care compute and storage infrastructure, as well as the volume of data involved, our ability to measure quality of care across the health care system is limited.

The healthcare ecosystem is diverse with many stakeholders, including hospitals, physicians, long-term care homes, and small community care organizations. Regional health authorities charged with ensuring quality of care and good population health would like to measure, on a continuous basis, performance management across the entire healthcare ecosystem. Hospitals provide the initial source of surveillance data for community healthcare through discharge summaries, referrals, emergency room (ER) visits, and prescriptions. Continuity of care ensures that care processes extend beyond hospitalization to at-home care – the main focus of community healthcare (Roughead, Kalisch, Ramsay, Ryan, & Gilbert, 2011). This is usually coordinated through a Regional Health Authority (RHA). The overall goal is to have a healthcare system that provides a cost-effective and high-quality collaborative environment for efficient healthcare ser-

vice delivery (Sabooniha, Toohey, & Lee, 2012). Achieving this level of performance management of the healthcare system requires a continuous, systematic framework for surveillance and performance management of care processes.

1.2. Motivation

Achieving systematic performance management of care processes require an infrastructure that addresses interoperability and data standardization while fostering data governance and privacy compliance. However, heterogeneous healthcare data silos and inconsistent patient identity in the current health care system complicates matters. These result in the fragmentation of efforts and the inability of stakeholders to coordinate care delivery across the healthcare domain (Adler-Milstein & Jha, 2012). Sharing data electronically across healthcare organizations remains a substantial challenge (Adler-Milstein & Jha, 2012) with much of the data exchange still being done manually to feed static offline data analysis done on snapshots of summary data. As well, coordination and transformation of health care delivery often lead to unintended consequences (e.g., social, legal and workflow consequences) related to governance and behavioural issues that arise from technology-mediated connectivity (C E Kuziemsky, Randell, & Borycki, 2016).

Cloud computing is one potential infrastructure for achieving interoperable healthcare solutions (Andry, Ridolfo, & Huffman, 2015; Bhaskaran, Suryanarayana, Basu, & Joseph, 2013; Y. Li & Guo, 2015) and enabling performance management to validate improvements to the health care system. Many services today outside of the healthcare industry have moved to the cloud because of the significant benefits which include scalability, device and location independence, 24x7 support, lower total cost of ownership , reliability, scalability, agile deployment,

automation, lower capital expenditures and a single infrastructure to fulfill all computing and storage needs (Donnelly, Irving, & Roantree, 2014). Therefore, cloud computing is a potential infrastructure for addressing performance management challenges and supporting interoperable healthcare solutions (Andry et al., 2015; Bhaskaran et al., 2013; Y. Li & Guo, 2015), across multiple providers only if data model standardization and appropriate support for privacy compliance could be put in place to protect patient data.

1.3. Thesis Contributions

The main contributions of this thesis are the following:

1. A cloud-based surveillance and performance management architecture for community healthcare that provides:
    a. A cloud computing infrastructure that provides surveillance and performance management services using a multi-tenanted private cloud owned and operated by a regional health authority to host applications and operational databases (the entire server infrastructure of each organization) for community care healthcare stakeholders.
    b. A common data model (CDM) that provides a consistent view of information across multiple and disparate data sources with a collection of services, APIs, and tools for sharing (collecting, correlating, processing, analyzing, mining, and reporting) data.
    c. A patient identity matching service for correlating cloud-hosted data from multiple community care organizations into a common data model to support performance management of community healthcare.

d. A privacy compliance model based on declarative privacy compliance definition documents that capture Patient Consent forms and Organizational Data Sharing Agreements to configure the processes and services of the cloud-computing infrastructure, including anonymization to ensure legal compliance and a systematic approach to data governance.

2. A configurable patient identity matching algorithm that employs a weighted probabilistic matching system that can correlate data from a variety of data sources.
3. A configurable anonymization algorithm that uses the privacy compliance definitions to address the anonymization of Direct Identifiers (DI), Quasi-Identifiers (QI), and Sensitive Attributes (SA) in high-dimensional datasets.

The following publications have been published or submitted related to the thesis:

1. **Eze, Benjamin**; Kuziemsky, Craig; Peyton, Liam. "A Configurable Identity Matching Algorithm for Cloud-hosted Community Care". Journal of Ambient Intelligence and Humanized Computing (2019). pp. 1-14. DOI: 10.1007/s12652-019-01252-y.
2. **Eze, Benjamin**; Kuziemsky, Craig; Peyton, Liam. "Successful Deployment of Cloud-hosted Services and Performance Management for Community Care". 12th International Conference on Health Informatics (HEALTHINF), Prague, Czech Republic (Feb 2019).
3. **Eze, Benjamin**; Kuziemsky, Craig; Peyton, Liam. "Operationalizing Privacy Compliance for Cloud-hosted Sharing of Healthcare Data". Proceedings of the

2018 IEEE/ACM International Workshop on Software Engineering in Healthcare Systems (SEHS), Sweden (May 2018).

4. **Eze, Benjamin**; Kuziemsky, Craig; Peyton, Liam. "A Patient Identity Matching Service for Cloud-based Performance Management of Community Healthcare". Procedia Computer Science (2017) pp. 287-294. DOI: 10.1016/j.procs.2017.08.321. September 2017.
5. **Eze, Benjamin**; Kuziemsky, Craig; Peyton, Liam. "Cloud-based Performance Management of Community Care Services". Journal of Software: Evolution and Process, (2017). DOI: 10.1002/smr.1897
6. **Eze, Benjamin**; Kuziemsky, Craig; Lakhani, Rubina; Peyton, Liam. "Leveraging Cloud Computing for Systematic Performance Management of Quality of Care". Procedia Computer Science (2016) pp. 348-355. DOI:10.1016/j.procs.2015.08.353. September 19, 2016.
7. **Eze, Benjamin**; Peyton, Liam. "Systematic Literature Review on the Anonymization of High Dimensional Streaming Datasets for Health Data Sharing". Procedia Computer Science (2015) pp. 348-355. DOI:10.1016/j.procs.2015.08.353. September 15, 2015.
8. Stepien, Bernard; **Eze, Benjamin**; Peyton, Liam. "Testing Policy-based e-Health Monitoring Processes using TTCN-3". Springer International Publishing. Book Section – Lecture Notes in Business Information Processing. April 29, 2015.

1.4. Research Methodology

This work is an objective-centred solution that was carried out over a period of 4 years in collaboration with the Champlain Local Health Integration Network (LHIN), acting as the Regional Health Authority (RHA) and 54 community care organizations. Our approach is to take the RHA as a suitable context for observing the first iteration of our proposed solutions. Each solution is then further enhanced in the laboratory environment (with RHA data) for subsequent evaluation.

Our research methodology is based on design science research (Peffers et al., 2006; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008) as illustrated in Figure 1-1. Design Science Research (DSR) entails creating new knowledge through the design of novel or innovative artifacts. This type of research involves learning by designing and developing artifacts - prototypes, pilots, or experiments. The output of design research science are those intermediary steps or experiments as well as the analysis and evaluation of the artifacts' use and performance (Gregor & Hevner, 2017). This is then demonstrated, evaluated, and communicated through scholarly publications. DSR is relevant to this thesis at a high level. Using the RHA as context, artifacts are developed and operationalized using the RHA infrastructure, and subsequent experiments were done leveraging RHA data.

In addition, our research also involves some Action Research (AR) (Koshy, Valsa, & Waterman, 2010) since we collaborated with Champlain LHIN to tackle some of their community healthcare challenges.

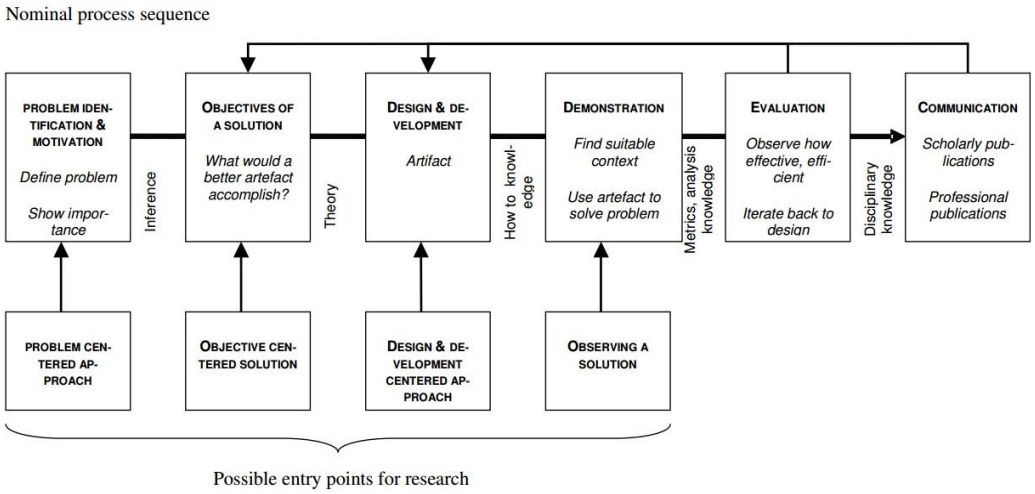


Figure 1-1 Design Science Research (Peffers, et al., 2006)

The Action Research component of our methodology is best illustrated using the action research spiral (Figure 1-2). The AR spiral is employed within each DSR iteration. At the onset of each DSR iteration, plans are developed and acted upon, results observed, reflected upon before proceeding with the DSR nominal process sequence. Since this work is done in collaboration with the LHIN to address their surveillance and performance management challenges with delivering community healthcare, AR is equally relevant to this thesis. Working with an RHA gave us the ideas for our solutions and allowed us to use these solutions to address problems identified within the RHA and across the partner community care organizations.

The following step by step descriptions show the specific iterations of our work.

1. Start with the AR steps - Observe, Reflect, and Plan using the current operations of the RHA and the community care organizations.
2. Use the outcome of step (1) to feed multiple DSR nominal sequence (Objectives, Design and develop, Demonstrate) iterations. The outcomes are usually new prototypes and artifacts or upgrades to existing solutions.
3. Deploy solutions – artifacts and prototypes at the RHA.

4. Evaluate solution and communicate results through peer-reviewed publications.
5. At this point, once again, AR is used to observe the enhanced solution, evaluation is done to identify and document gaps, plans are made on subsequent experiments to improve the existing outcomes using experimental data. We essentially use the AR loop to determine the gaps with the current iteration for the next iteration.
6. Repeat steps in (1) - (4) in the Lab using RHA data to develop enhanced prototypes, evaluate the efficiency of the lab experiment and finally update the existing deployed solutions at the RHA

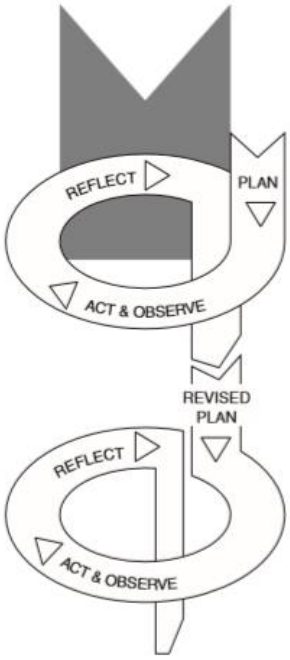


Figure 1-2 Action Research Spiral (Koshy et al., 2010)

As shown in Figure 1-3 , the AR loop – Act/Observe, Reflect, and Plan activities lead into each DSR iteration. The result of each iteration could either remain an artifact in the laboratory or would get adopted into the RHA performance management workflow.

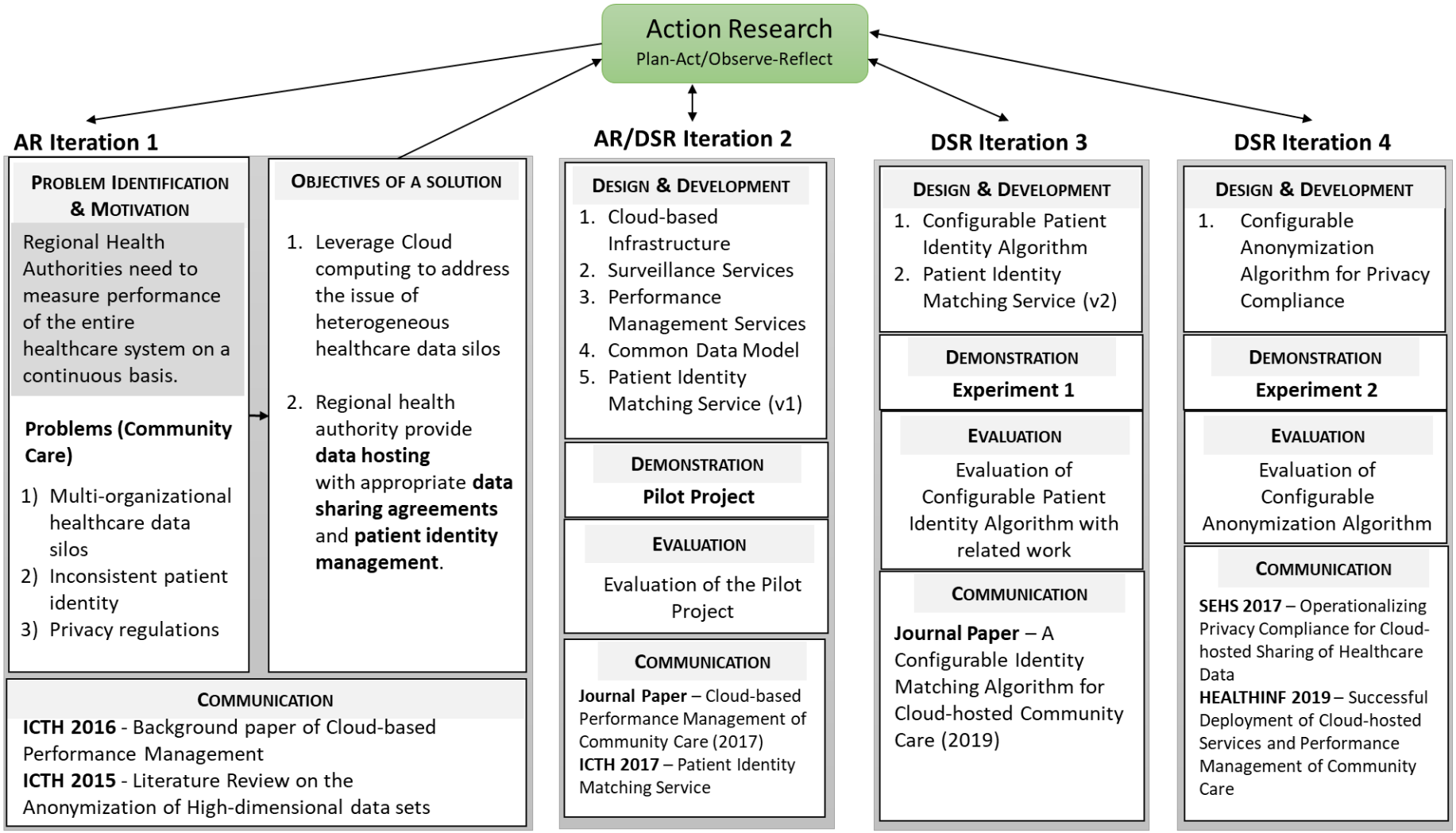


Figure 1-3 Thesis Research Methodology Mapped to DSR/AR Iterations

1.5. Thesis Organization

This thesis is organized as follows:

In Chapter 2, we provide an overview of healthcare performance management background materials such as Triple Aim objectives, healthcare interoperability, and a common data model. We also described various representations of health datasets, provided an overview of cloud computing and its various healthcare applications. We also provided some background on privacy and confidentiality as it relates to healthcare data. Finally, we described some related work to our approach.

In Chapter 3, we describe the problem of performance management through a community care example. We also described the result of our gap analysis on current practices and approaches used in the related work. The result of the gap analysis is used to define a set of criteria for evaluating our cloud computing performance management framework against other similar frameworks.

In Chapter 4, we present our cloud computing framework for addressing surveillance and performance management for a regional community healthcare authority. The main elements of our framework include the infrastructure for systematic data hosting and data collection, a common data model, a framework for legal compliance and governance, patient identity management, as well as performance management services.

In Chapter 5, we introduce a community care pilot project describing the application of our architecture to a regional health authority in Ontario, Canada. In this pilot project, we lever-

aged various components of our framework for systematic cloud hosting, systematic data collection, patient identity management, compliance and performance management services including data reporting, and subscription.

In Chapter 6, we introduce an additional experiment that focuses on operationalizing privacy compliance of cloud-hosted data using data sharing agreements in support of performance management of community healthcare. In particular, this experiment leverage anonymization methodologies to enable privacy-compliant data publishing data and performance management results for collaborating organizations in the context of data surveillance and performance management in a cloud computing environment.

In Chapter 7, we introduce another experiment on our configurable patient identity matching algorithm. This chapter leverages our probabilistic matching algorithm to correlate patient profiles across various community care organizations with varying levels of data quality.

In Chapter 8, we evaluated our framework against the other frameworks from related work. We also evaluated our framework against the pilot project and the two experiments, all leveraging the evaluation criteria defined in Chapter 3. We also discuss the limitations and assumptions made in the course of developing our framework.

Finally, in Chapter 9, we provide some conclusions and directions for future work in this domain.

Chapter 2. **Background**

In this chapter, we provide background on community healthcare surveillance and performance management, healthcare interoperability, and the nature of healthcare datasets for large-scale analytics and data exchange standards. We then introduce cloud computing, privacy compliance, and identity management and survey the relevant techniques and technologies. Finally, in the related work section, we identify other alternative approaches to cloud-based surveillance and performance management in the literature.

2.1. Community Healthcare Performance Management

Community-based healthcare aims at improving overall population health by managing chronic illness, providing rehabilitation support, nursing, physiotherapy, and end of life care to ageing patients (CIHR, 2017). Community healthcare is for people of all ages who need personal care or long-term healthcare assistance at home. The purpose of healthcare performance management in community care is to measure the extent quality of care goals are achieved with care processes through healthcare outcomes (Vanhaecht et al., 2007).

Hospitals provide the initial source of surveillance data for community healthcare through discharge summaries, referrals, emergency room (ER) visits, and prescriptions. Continuity of care ensures that care processes extend beyond hospitalization to at-home care – the main focus of community healthcare (Roughead et al., 2011). This is usually coordinated through a Regional Health Authority (RHA) but implemented through independent community care organizations. Community care organizations are usually small establishments, mostly non-profit, that provide various niche community care services to target populations within a community. These

community care services include but are not limited to personal support services, nursing services, occupational therapy, adult day programs, assisted living services, bereavement, crisis intervention, friendly visits, meals on wheels, mental health and addiction services, etc. (Boissonnealth & Lafreniere, 2014).

Patient needs are met through service-level planning and coordinated clinical care provided by the health care providers (primary care) and those in community settings such as public health units, community care organizations, workplace, etc. (CIHR, 2017).

Finally, continuous efforts are being made to provide good quality and cost-effective community-based care with support from government and non-profit organizations (Boissonnealth & Lafreniere, 2014). Most importantly, it is critical to building a strong collaboration platform among various aspects of community healthcare.

2.1.1 Complex Patients

In our ageing society, there is an ever-increasing complex patient population with multiple medical, social, and behavioural conditions, which is driving healthcare system transformation. Complex patients are often described as having comorbid health conditions that make the management of these patients very challenging (Grant et al., 2011). Providing care for complex patients is putting increased strain on healthcare budgets and service delivery performance goals (A. Sheikh, Sood, & Bates, 2015). As a consequence, there is a growing need to transform the health care system to more efficiently provide care for complex patients (Sabooniha et al., 2012).

Complex patient management is challenging and expensive as it requires care delivery and service provision from a variety of healthcare providers within a community (Mcgregor, Mercer, & Harris, 2016). Further, complex patients may be managed using multiple clinical

practice guidelines (CPGs), which may have conflicting recommendations about medications or treatments (Wilk et al., 2017). Thus, it is crucial that patient care is efficiently coordinated across all providers and care locations to prevent adverse interactions from conflicting medications and to ensure that all care providers are aligned with care delivery.

As part of healthcare transformation, government and healthcare organizations want better accountability for money spent on healthcare delivery (Bohmer, 2016), which requires performance management of care processes across all the stakeholders in the healthcare ecosystem (Berwick, Nolan, & Whittington, 2008). Achieving this accountability requires coordination and integration of data across disparate healthcare information systems (Sabooniha et al., 2012).

Systematic performance management is necessary to monitor progress in patient care delivery. However, heterogeneous healthcare data silos and inconsistent patient identity approaches, coupled with patient privacy regulations, limit our ability to correlate healthcare data for complex patients as part of performance management (B. Eze, Kuziemsky, & Peyton, 2017). These limitations result in fragmentation of efforts and the inability of stakeholders to coordinate care delivery across multiple healthcare domains (Adler-Milstein & Jha, 2012), with much of the data exchange being done manually. Attempts to address these factors individually often leads to unintended consequences (e.g., social, legal and workflow consequences) related to governance and behavioural issues that arise from technology-mediated connectivity (C E Kuziemsky et al., 2016).

2.1.2 Regional Health Authorities

Community care usually has a "regional" health authority (e.g., country, state, municipality, county). In Canada, healthcare is regionally managed by each province. In Ontario for example, the Regional Health Authority is the Local Health Integration Network (LHIN, 2018), with

the mandate to plan, integrate and fund local healthcare in the region. The LHIN operates through 14 local health authorities that target each sub-region within the province.

Norway employs a similar model with four main regional health authorities, each with many subsidiaries (Ringard, Sagan, Sperre Saunes, & Lindahl, 2013). For both models, the Regional Health Authority (RHA) is responsible for patient treatment, medical staff, healthcare planning, research, and development, as well as the support and training of patients and their caregivers. While RHAs are common with most developed countries, the USA uses a healthcare model managed using a combination of state and federal mandated authorities. RHAs and local health authorities are associated with improved healthcare outcomes, healthcare equality, increased life expectancy, improved coordination, and reduced cost to healthcare services (Vida, Lupse, & Stoicu-Tivadar, 2012).

The focus of this thesis is on performance management of all community healthcare services provided to patients in their home or a long-term care facility. Most of these patients have chronic and complex health conditions, and the RHA works with various community care organizations to provide needed services to this target population. This thesis may not apply if an RHA does not exist.

2.1.3 Triple Aim Objectives

Triple Aim for connected care is about improving patient experience, improving population health, and lowering the cost of care (Farmanova et al., 2016; Aziz Sheikh, Sood, & Bates, 2015). According to CDC (CDC, 2018), population health is the distribution of healthcare outcomes within a population, by considering a range of personal, social, economic, and environmental factors that influence the distribution of healthcare outcomes. According to this article, population health provides the opportunity for healthcare providers to work together to improve

the healthcare outcomes of the communities they serve. Factors that contribute to improved population health include increased life expectancy, reduced infant mortality, and improved general population health (Y. Hu & Bai, 2014).

Realizing the objectives of Triple Aim requires coordination of numerous stakeholders within and outside a patient circle of care with the overall goal of providing a cost-effective and high-quality integrated environment for efficient healthcare service delivery (Benjamin Eze, Kuziemsky, Lakhani, & Peyton, 2016; Sabooniha et al., 2012). Other important objectives include enabling safe care delivery, streamlining clinical and administrative tasks, and safeguarding patient data. One of the key challenges for Triple Aim is our ability to measure improvements at all levels of the healthcare ecosystem, especially as it relates to quality of care (Farmanova et al., 2016).

Consequently, the three objectives of Triple Aim – health care quality, cost, and population health status, are dependent on each other. According to Berwick et al. (Berwick et al., 2008), pursuing one goal without paying attention to the others could result in some unintended negative consequences. For example, cutting healthcare cost could result in limited care for the chronic and complex patients that need it the most. However, by paying attention to the quality and nature of care, cost savings can be achieved without compromising patient experience and overall population health (A. Sheikh et al., 2015).

Therefore, all the key objectives of Triple Aim need to be measurable in order to measure performance management across a healthcare system. Eliminating the opacity of performance provides the right environment to achieve Triple Aim objectives through proper decision-making processes, contracting and health care mandates, as well as legal and regulatory frameworks that

encourage cooperation among stakeholders. This role is played by an entity referred to as an integrator. An integrator is responsible for achieving the Triple Aim objectives for a target population (Berwick et al., 2008).

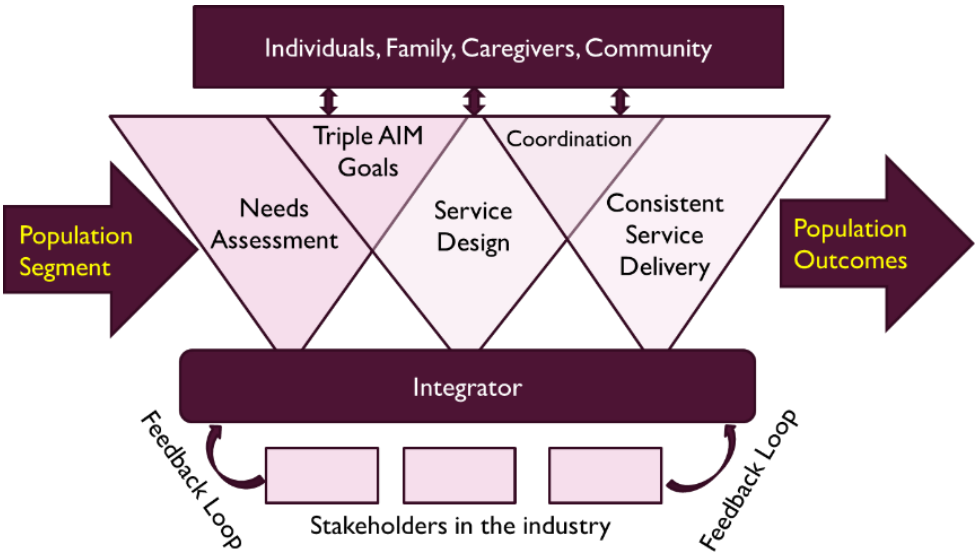


Figure 2-1 Triple Aim Integrator – Managing Services for a population

As demonstrated in Figure 2-1, it is the duty of the integrator to ensure that participating stakeholders coordinate the management of the sequence of steps: needs assessments, service design, consistency and scale of service delivery to improve population outcomes (IHI, 2016). In the context of this thesis, the RHA serves the role of a Triple Aim integrator.

2.1.4 Healthcare Interoperability

Collecting health data in a consistent, standardized, and timely manner is important in influencing healthcare decisions. Healthcare interoperability describes the difficulties in gathering data across organizations with disparate systems that otherwise will not communicate with each other. According to the Institute of Medicine (Miguel-Angel & Pablo, 2013), "Poor interopera-

bility leading to fragmentation of the health care system and poor data exchange and communication is a significant cause of medical errors." In healthcare, other than fax, there is no systematic approach to interoperability as with other industries. Rather, most operational, clinical information systems are still unable to exchange clinical data electronically and in a systematic manner (Chalasani, Jain, Dhumal, Moghimi, & Wickramasinghe, 2014; Gaynor, Yu, Andrus, Bradner, & Rawn, 2014; Guarrera et al., 2014).

Lack of interoperability in healthcare stems from a combination of factors such as heterogeneity in IT systems, data sources, data formats, non-compatible data ontologies, semantics issues, and general governance issues aimed at protecting patient privacy and confidentiality. Lack of interoperability has resulted in fragmentation of efforts and the inability of stakeholders to coordinate care delivery across the healthcare domain. The effects are service duplication, mistakes, and expensive administrative burdens and gaps that can be bridged easily with interoperability (B. Eze, Kuziemsky, Lakhani, & Peyton, 2016).

Healthcare interoperability can be grouped into these three broad categories - technical, semantic, and process (Benson, 2012; C. Kuziemsky, 2013).

Technical Interoperability relates to data exchange protocols between senders and receivers. It uses technologies to solve accessibility issues associated with integrating healthcare processes. The most common reason for this is because most health systems use non-compatible communication protocols, data ontology standards, as well as disparate underlying platform technologies (Dixon, Vreeman, & Grannis, 2014a; Gaynor et al., 2014). Though taken for granted, Technical Interoperability has resulted in fragmentation of efforts and the inability to coordinate care delivery across the healthcare domain (Benson, 2012).

Semantic interoperability challenge is one of the most investigated subjects in the healthcare interoperability domain today (Amato, Mazzeo, Moscato, & Picariello, 2013). It is about ensuring that senders and receivers understand the same data in the same way (Benson, 2012). Semantic interoperability guarantees that heterogeneous systems share, understand, interpret, and use data without ambiguity. There are many healthcare standards worldwide that address various aspects of semantic interoperability challenges in healthcare. In healthcare, semantic interoperability is addressed through ontologies (coding standards) and document exchange standards.

There are various coding standards that target various aspects of healthcare. SNOMED CT provides universal identifiers for organisms, substances, and diseases (Dixon, Vreeman, & Grannis, 2014b). It also provides a set of concepts and relationships as well as clinical reference terminologies. SNOMED concept table has about 344,000 entries organized in hierarchies with a description table of over 913,000 entries and up to 13 million semantic relationships (Chalasani et al., 2014). SNOMED CT is very comprehensive. But its complexity makes it very difficult to work with for realistic implementations (Tapuria, Kalra, & Kobayashi, 2013).

LOINC provides universal identifiers for laboratory tests and results, diagnostic study observations as well as other clinical observations (Chalasani et al., 2014; Dixon et al., 2014b; Gaynor et al., 2014). ICD-9 and ICD-10 are used primarily to perform diseases identification for billing purposes (Gaynor et al., 2014). DICOM is the standard for the transmission of medical imagery in radiology, cardiology, pathology, and dentistry (Gaynor et al., 2014). Being a worldwide standard, devices from various vendors can produce, transmit, and interpret DICOM data interchangeably (Gaynor et al., 2014; Noumeir, 2012). Despite these many available stand-

ards, most healthcare information systems are not based on these standards. The biggest challenge to adoption is encouraging clinicians to code the healthcare data they record to match the applicable standards (Gaynor et al., 2014). Unfortunately, most employ minimal coding that allows them to submit billing data to insurance companies, other stakeholders, and the government.

There are also the document exchange standards such as Health Level 7 (HL7) CDA and openEHR. HL7 is an XML-based messaging standard for health data exchange across providers (Chalasani et al., 2014; Gaynor et al., 2014). HL7 C-CDA("Consolidate CDA Overview," 2015) provides a common architecture, data coding, and markup language for creating electronic clinical documents and communicating clinical data (Chalasani et al., 2014). According to Gaynor et al. (2014), HL7 CDA is required for meaningful use of Electronic Health Records (EHR) in the US. openEHR ("openEHR Architecture," 2015) leverages domain-specific archetypes brought together through templates to describe health data.

The HL7 Fast Health Interoperability Resources (FHIR) is the most recent generation of HL7 standard ("Introducing HL7 FHIR," n.d.). FHIR includes the best features of the previous versions of HL7 – v2, v3, and CDA, while supporting the latest web standards to foster its implementability. Built on modular components or resources, FHIR reduces the complexity of applying HL7 to solving real-world clinical and administrative problems in a variety of contexts - mobile applications, RESTful architectures for cloud-based interoperability, EHR-based data sharing using XML, JSON, HTTP, OAuth, etc. It also provides many implementation libraries and examples to reduce the entry barrier to its integration with new and current healthcare applications (Berler & Apostolakis, 2014). While HL7 is used extensively in the US, openEHR is more popular in Europe and other countries around the world.

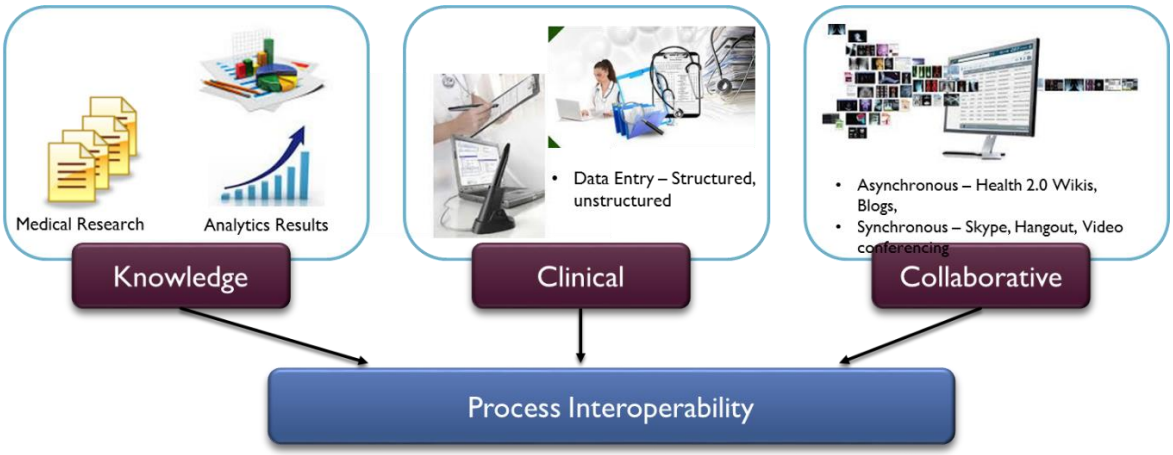


Figure 2-2 Process Interoperability Key Components

Finally, **process interoperability** is about ensuring that all actors (patients, clinicians, and decision makers and management) share a common understanding of the health system (Kuziemsky, 2013; Mouttham, Kuziemsky, Langayan, Peyton, & Pereira, 2011). Business and work processes need to be interoperable and coordinated across collaborating organizations for efficient and cost-effective service delivery.

While most interoperability discussions focus on the technical and semantic aspects of healthcare interoperability, process interoperability is the least discussed in the academic literature (Craig E. Kuziemsky & Peyton, 2016). Process interoperability looks at how technology, information, as well as guidelines and processes, affect the interactions amongst these actors as shown in Figure 2-2. It also measures how healthcare services, quality of care, patient hand-off are delivered and perceived - a major component of Triple Aim.

2.2. Surveillance and Performance Management

2.2.1 Continuous Quality Improvement in Health Care

In healthcare, Continuous quality improvement (CQI) is a structured organizational process for planning and executing the continuous flow of improvements to provide quality healthcare that meets and exceeds set expectations (McLaughlin & Kaluzny, 2004). CQI looks for ways to improve the output or the product of healthcare processes and workflows. It is the management philosophy that healthcare organizations use to reduce waste, increase efficiency, and increase patient satisfaction through continuous improvement to the quality of care. CQI is an on-going process that continuously evaluates how healthcare organizations work and how to improve their processes for better efficiency.

According to McLaughlin & Kaluzny (2004), three major performance improvement initiatives are usually associated with CQI. These are the include 1) localized improvement effects that investigate specific process problems and improvement opportunities, 2) organizational learning that comes from documenting these processes and being able to identify areas of improvement, and 3) process re-engineering that fosters performance improvement through investments in information systems using internal and external resources with the goal of improving organizational internal and external processes.

2.2.2 Healthcare Surveillance

Healthcare surveillance, according to World Health Organization (WHO) is the continuous, systematic collection, analysis, and interpretation of healthcare data in support of the planning, implementation, and evaluation of public health practices (WHO, 2018). The widespread

adoption of health information systems across healthcare organizations has created new care delivery models that are more patient-centred, allowing for care to be fine-tuned to match individual patient needs, therefore improving quality of care and patient satisfaction. However, without healthcare surveillance, measuring performance goals against broader public health initiates and mandates becomes very difficult and sometimes impossible (Adler-Milstein & Jha, 2012).

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is a multi-disease surveillance system based on primary care EMR data. CPCSSN data come from multiple EMR systems from physicians in 10 practice-based research networks across Canada. Data extraction from these sources is done quarterly and mapped to a common schema after standardization. Case detection algorithms are run against the dataset to identify patients with one or more of eight chronic conditions - diabetes, hypertension, osteoarthritis, depression, chronic obstructive lunch disease, dementia, Parkinson's disease, and epilepsy. The final datasets are then made available to researchers for further investigations and for performance management ("CPCSSN Data for Research," n.d.; Martin, 2018).

Many public healthcare surveillance initiatives target infectious disease outbreak by aggregating and mining hospital data with external data like social media feeds. However, In this thesis, community healthcare surveillance is seen as the process of continuous collecting and aggregation of patient data across various healthcare stakeholders solely to carry out continuous performance management of care processes (B. Eze et al., 2017).

2.2.3 Common Data Model

Every encounter of a patient with a caregiver or any diagnostic activity creates an Electronic Medical Record (EMR) that gets associated with the patient (Gaynor et al., 2014). EMRs

are created by physician medical practices, hospitals, diagnostic centers, community care organizations, and the regional health authority. Today, there is a plethora of EMRs from lab results and imaging results, to family doctors, to healthcare service and equipment providers, to therapists, etc. (Dixon et al., 2014a). Aggregating EMRs results in Electronic Health Record (EHR) (Hsieh & Chen, 2012). A patient EHR is the complete patient health information of care events not just from one healthcare provider but from all healthcare providers and institutions serving a patient (International Organization For Standardization, 2005).

A major challenge with integrating patient EMR data across healthcare organizations for large-scale surveillance is data quality. A Common Data Model (CDM) allows researchers and analysts to organize data in a standardized manner for a specific domain of information (Sabooniha et al., 2012; Sinaci & Laleci Erturkmen, 2013). With a CDM, analysis can be done on large scale data without constant data standardization and transformation each time, making data sets more reusable for continuous surveillance and analytics. According to IBM Knowledge Center (IBM, 2017), a CDM differs from a database schema because it includes both a logical model and the specification of how data maps to the physical model. Most importantly, CDM classifies and organizes commonly managed characteristics of data subjects, including patients, resources, services, information on processes and present them in a way that all applications can use.

In the healthcare industry, for example, different healthcare providers have different mandates, and that affects the type, level, and depth of data they can collect and hold on patients. Since each organization uses different data schemas, logical, and semantic relationships between data elements are usually inconsistent. A CDM for performance management is needed to ensure

a consistent view of information across all data sources (De la Rosa Algarín, Demurjian, Ziminski, Rivera Sánchez, & Kuykendall, 2014; Klann et al., 2014; Sabooniha et al., 2012).

Mapping schema from local data sources to the CDM schema follows two approaches - Global As View (GAV) or Local As View (LAV) (Katsis & Papakonstantinou, 2017). In GAV, the CDM schema, as the global schema, is expressed a function of the schema of the local databases. In LAV, the local schema is described as a function of the CDM schema. GAV-based systems do not facilitate adding a source to the system independently of other sources as every change could force changes to other mappings corresponding to the other sources. However, LAV mapping is declarative in nature as it describes the information within the global database that is contained within each local database. Also, LAV sources can be registered independently of each other (Katsis & Papakonstantinou, 2017).

2.2.4 Performance Management

Performance management is a systematic process for improving organizational effectiveness in achieving organizational goals and missions. Performance management involves planning, setting expectations, continuous monitoring of performance, developing the capacity to perform, and periodically rating and rewarding of performance (OPM.GOV, 2017).

Performance management provides a mechanism for translating strategic objectives and business goal to operational processes (Kemper, Rausch, & Baars, 2013) with a focus on identifying key performance indicators (KPIs) from these strategic objectives.

Therefore, the ability to keep performance management measures realistic comes from respecting the time sensitivity of KPIs through continuous monitoring. Continuous monitoring of KPIs is key to effective monitoring and management of strategic goals. However, each strategic

goal also needs to be linked to these KPIs to measure the extent the performance of the organization is far or near to its goals (C. Kuziemsky, Liu, & Peyton, 2010).

Systematic performance management is necessary to monitor progress in patient care delivery. However, heterogeneous healthcare data silos and inconsistent patient identity approaches, coupled with patient privacy regulations, limit our ability to correlate healthcare data for complex patients as part of performance management (Benjamin Eze et al., 2016). These limitations result in the fragmentation of efforts, the inability of stakeholders to coordinate care delivery across multiple healthcare domains (Adler-Milstein & Jha, 2012), with much of the data exchange being done manually. Attempts to address these factors individually often leads to unintended consequences (e.g., social, legal and workflow consequences) related to governance and behavioural issues that arise from technology-mediated connectivity (C E Kuziemsky et al., 2016).

2.3. Cloud Computing

Cloud computing is a style of computing in which virtualized resources and services are dynamically scalable and provided as a service over the Internet (Furht & Escalante, 2010). It can also be seen as a model that offers distributed, configurable, easily provisioned, on-demand, and elastic resources such as servers, storage, applications, and networks (Ma, Peng, & Chen, 2014). With the volume of data in most EHR systems today, many organizations are slowly hitting various thresholds to the amount of data they can handle within their IT infrastructure. Cloud computing can easily fill this gap since it offers “infinite” computing resources and capacity.

Cloud computing can also be described as a model for generating ubiquitous access to a pool of convenient, on-demand computing resources (compute, storage, platform, application and services) through a web interface with low administration overhead and the least intervention from a cloud service provider (Ochian, Suciu, Fratu, Voicu, & Suciu, 2014).

According to Li and Guo (Y. Li & Guo, 2015), implementing cloud computing technologies would aid healthcare providers in providing better and more effective quality of care. More importantly, it aids their ability to share information, improve collaboration, and reduce expenditures on infrastructure. If used as a data proxy, the cloud offers a consolidated view of patient-relevant data to healthcare providers (Bhaskaran et al., 2013).

The cloud paradigm provides the platform for regional, national and international data aggregation using a broad range of topologies that could integrate various devices, data sources and services very quickly in a scalable and cost-effective manner (Andry et al., 2015). A cloud environment offers a platform for developing complex applications capable of processing time-series data from different sources, providing scalability as requirements such as workloads continue to increase over time (Ochian et al., 2014). Bhaskara et al. (2013) corroborate that cloud infrastructure provides infinitely scalable storage for very data-intensive applications. Popular cloud providers such as Amazon, Google, and Microsoft offer Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) cloud packages to end users (Amato et al., 2013; Furht & Escalante, 2010).

Some of the significant benefits of cloud computing include device and location independence, 24x7 support, lower total cost of ownership (TCO), reliability, scalability and sustainability, easy and agile deployment, high level of automation, lower capital expenditure and a

single infrastructure to fulfill all computing, networking and storage needs for various applications (B. Eze et al., 2016; Furht & Escalante, 2010).

2.3.1 Types of Cloud Computing

There are four types of cloud computing – Private, Community, Public, and Hybrid Clouds. The differentiating factors for cloud types, according to Furht & Escalante (2010), are: 1) Where the infrastructure is located, 2) The user of the cloud infrastructure and 3) The entity that manages the infrastructure. A private cloud is operated by a single organization, which has full control over the infrastructure, data, security, and quality of service (QoS). The community cloud is similar to a private cloud but is shared by a group of organizations. The public cloud is operated by a 3rd party and can be used by any individual or organization with applications mixed on cloud servers, storage systems, and networks. Finally, a hybrid cloud is a mix of public and private/community clouds. In the hybrid cloud, data and applications are distributed across both public and private clouds using the appropriate secure data bridges.

A private cloud provides the owners with full control over everything – compute, storage, networking, as well as the quality of service. Having full control increases the complexity associated with the development and deployment of a cloud application and services. Compared to other cloud types, it offers the best security and confidentiality with user data. Unfortunately, compared to other types of cloud infrastructure, a private cloud is more expensive since the organization that owns the cloud infrastructure bears all the cost associated with setup and maintenance (Ma et al., 2014).

The public cloud provides the lowest Total Cost of Ownership (TCO) of the cloud types but offers the least control. Also, data security cannot be guaranteed since many organizations share cloud resources. Public clouds are also prone to resource contention issues, SLA breaches,

and service disruptions. For healthcare organizations with high volumes of highly sensitive data, this would not be acceptable since it violates data privacy laws in many countries (Furht & Escalante, 2010; Gazzarata, Gazzarata, & Giacomini, 2015).

For such scenarios, a hybrid cloud infrastructure would be preferable. Operational data is kept in the private cloud while pre-processed, anonymized data can be sent to a public cloud for analytical processes like patient profiling, clustering, association rules, and correlation mining, as well as descriptive and predictive analysis (B. Eze et al., 2016). Despite the benefits of a hybrid cloud, its complexity poses some technical, business, and management challenges.

2.3.2 Cloud Computing and Healthcare Performance Management

Utilizing cloud computing infrastructure in health care come with some challenges. Some of the characteristics of healthcare applications sometimes do not align well with cloud computing (Moumtzoglou, 2014). The healthcare industry is often characterized as high risk, highly regulated, process heavy with multiple stakeholders, slow in adopting new technologies, and tend to have long-term relationships with technology vendors (Moumtzoglou, 2014).

As a result, some of the shortcomings of cloud computing and the use of cloud services do not align with the structure and operations of the healthcare industry. For example, cloud computing requires data owners to relinquish control over compute and data resources, while healthcare regulations and policies require such controls and oversight. It is also difficult to ensure full accountability with cloud environments regarding patient privacy and confidentiality. Some of the other notable issues identified include functionality creep with cloud data management, monopoly and vendor lock-ins as well as privacy with cloud data since patient data could be stored in locations that are beyond the jurisdictional boundary of the care provider (B. Eze & Peyton, 2015).

Also, the high latency for accessing information is identified as one of the challenges with cloud computing in healthcare surveillance (Mendelson, Erickson, & Choy, 2014; Ochian et al., 2014). Likewise, for applications that require the storage and retrieval of medical imagery, the high bandwidth requirement of cloud services can be a challenge for health care providers (Mendelson et al., 2014). Also, the knowledge on the level of abstraction associated with cloud services, as well as the many tools required to develop new health applications, or integrate existing applications, can be quite daunting for many IT professionals; requiring special training (Bhaskaran et al., 2013) and specialization, which can be a steep learning curve for many developers and healthcare system vendors.

Despite these challenges, a "health cloud" (Moumtzoglou, 2014) would make it possible to:

1. Provide a unified patient medical record across all patient encounters, improve patient care, and is necessary for performance management for quality of care.
2. Create a collaborative economic environment because of the flexibility to pay for actual resource utilization.
3. Alleviate the scarcity of resources through dynamic resource allocation.

2.3.3 Cloud Computing and Healthcare Interoperability

Cloud computing is potentially becoming an interesting platform for national and regional healthcare systems interoperability. de la Torre-Díez et al. (2013), in their explorative work, investigated options for the Spanish Public Health National System and concluded that cloud infrastructure should be used for data sharing and service orchestration. One of the recommendations from this work is that cloud computing could be used for statewide health infrastructure and platform because of the many advantages - scalability, reach, extensibility, a low total

cost of ownership and availability (de la Torre-Díez et al., 2013). Other attractive features with the cloud include just-in-time scalable infrastructure, no upfront costs, and a usage model. Having such readily available scalable IT infrastructure would help reduce the deployment time for the cloud components of a solution (Bhaskaran et al., 2013)

Andry et al. (2015) review the layering of cloud infrastructure from IaaS through to SaaS through a prototype implementation for a home care delivery use case. In this work, custom PaaS containerization is used to abstract healthcare services. The bundled services include identity management, security (authentication, authorization, and single sign-on) as well as support for the management of cloud-based connected devices and clinical workflows. Patient EMR can be made available to patients while validation is done through traditional login and password credentials or online Single-Sign-On (SSO) identity providers, including Facebook, Google, LinkedIn, and Twitter.

Biswas et al. (2014) propose an "eHealth Cloud" platform for the government of Bangladesh. The proposed platform connects physicians, patients, hospitals, government departments, insurance companies, and pharmaceutical companies to the same platform that would give patients access to their EMR through federated credentials across all available healthcare services. Amazon cloud-based SimpleDB is used for data storage while all data exchanges are HL7 protected using appropriate data encryption technologies. Data mining techniques are used to measure various associations and correlations between care parameters - diseases, treatments, results, and procedures. Clustering is used for patient segmentation and investigating levels of severity of diseases from absence to severe disease conditions.

Person-Event Data Environment (PDE) (Vie, Griffith, Scheier, Lester, & Seligman, 2013) was implemented to unify disparate army and department of defense databases in a secure

cloud-based environment. The infrastructure allows researchers access to a repository of army data on corrections and legal issues, physical fitness tests, military service information, deployments, demographics, and a host of medical issues. The repository, totaling over six terabytes, is updated periodically. While there is no automated analysis of the data, researchers can log in to remote virtual machines (VMs) to use data management and analytics tools such as SAS, and SPSS to run statistical analysis on PDE data as needed.

2.3.4 Big Data Analytics

According to Gartner (2015), "Big data is high-volume, high-velocity, or high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, decision making, and process automation." For the three Vs that define big data (see Figure 2-3), Volume is influenced not only by the amount of data but also the dimensionality. Variety looks at data both in terms of heterogeneity of data formats - structured databases, unstructured and semi-structured data. Finally, velocity refers to static vs. dynamic for streaming and versioned data (Orit, 2013).

With the rapid explosion of health data in recent time, the need for the right infrastructure with the capacity to analyze such data in real-time is essential to providing needed operational and governance related statistics and key performance indicators for the healthcare industry (Sun & Reddy, 2013).

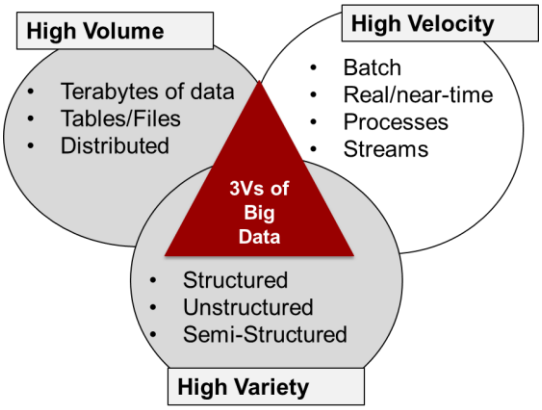


Figure 2-3 **Big Data 3Vs** (B. Eze et al., 2016)

2.3.5 Cloud Computing and Service Oriented Architecture

Service Oriented Architecture (SOA) and Web services are mature technologies that allow services to interact over the Internet. Cloud services at all levels – IaaS, PaaS, and SaaS, are typically designed as Web services accessible over the Internet (Furht & Escalante, 2010). While there are two types of Web services – SOAP and REST, most popular cloud services are built around REST Web services (Piyare, 2013). Cloud Web services are usually secured through Transport Layer Security encryption protocols like SSL, and Single Sign-On (SSO) authentication (Frontoni, Baldi, Zingaretti, Landro, & Misericordia, 2014).

2.3.6 Maintaining Privacy and Confidentiality with Cloud Computing

One of the common challenges with cloud computing for healthcare is privacy and confidentiality (Eze et al., 2016). While encryption helps secure data exchanges and storage, it does not protect the data from authenticated users that have access to the decrypted data. It also sacrifices the analytical utility of cloud data. Privacy measures, on the other hand, ensure confidentiality of patients and other members of the circle of care but still allow the data to be available for analytical processing (Mathew & Pillai, 2015; Spruijt-Metz et al., 2015).

The high sensitivity of healthcare data requires cloud implementation to address security and privacy-related issues (de la Torre-Díez et al., 2013). Ownership of cloud data and privacy law jurisdiction depend on the data custodian and privacy laws specific to the location of data storage and processing. These laws are discussed in section 2.4.1.

2.4. Privacy Compliance and Identity Management

Achieving systematic performance management of care processes requires an infrastructure that addresses interoperability and data standardization while supporting data governance and privacy compliance (Eze, Kuziemsky, & Peyton, 2017). Cloud computing is one potential infrastructure for addressing performance management challenges and supporting interoperable healthcare solutions (Andry et al., 2015; Bhaskaran et al., 2013; Y. Li & Guo, 2015), across multiple providers if data model standardization and appropriate support for privacy compliance are put in place to protect patient data (Fung, Wang, Chen, & Yu, 2010b; Perakis et al., 2013).

2.4.1 Privacy Laws and Regulations

Privacy concerns aimed at regulating the use, storage, and sharing of personal health records is one of the biggest impediments to healthcare interoperability. In Canada, there is an overarching Federal privacy act, the Personal Information Protection, and Electronic Documents Act (‘‘PIPEDA’’), which governs how private-sector organizations collect, use or disclose personal information, including healthcare data. PIPEDA is supplemented by provincial privacy legislation on healthcare data such as the *Personal Health Information Protection Act* (PHIPA) in Ontario (“Overview of privacy legislation in Canada,” 2014). Privacy legislation similar to HIPAA(“Summary of the HIPAA Privacy Rule,” 2015) in the US play a similar role and provide

the basis for protecting patient privacy while transmitting Personal Health Information (PHI). General Data Protection Regulation (GDPR) provides similar protection for all individuals in the European Union (The European Parliament & The European Council, 2016).

Using cryptographic techniques to achieve privacy techniques (Pang et al., 2013; Ribeiro, Viana-Ferreira, Oliveira, & Costa, 2014) may not always provide the necessary analytics utility that comes with data aggregation and consolidation. Chalasani et al. (2014) propose a framework that aggregates data but seeds an agent within each EHR system to anonymizes the data to the HIPAA compliant standard before transmitting it to an aggregation infrastructure. Unfortunately, HIPAA compliant datasets have very little analytical utility since all PHI in a dataset, including those with high analytical significance, would have to be removed before transmission. Nevertheless, HIPAA also allows for statistical validation of anonymized datasets that ensure the protection of patient privacy while preserving the analytical utility. It is essential that privacy protection is done alongside a robust compliance framework for implementing and enforcing data sharing agreements (Coats & Acharya, 2013).

Work exists that shows how the specification of data sharing agreements in a formal policy language can operationalize compliance in information systems (Swarup, Seligman, & Rosenthal, 2006). However, the conventional approach to privacy compliance in healthcare is an all-or-nothing strategy in which patient data that is excluded from access, is also excluded from aggregation. As a result, exclusion of patient data compromises the accuracy of performance management reports because the aggregate data is not precise (Benjamin Eze et al., 2016) even though there are techniques to allow aggregation while ensuring privacy (El Emam et al., 2009).

According to Weber et al. (Weber-Jahnke, Price, & Williams, 2013), healthcare systems should support data across the patient's life. Achieving that would require wide-scale surveillance and data sharing across various healthcare applications. Data surveillance in healthcare requires complex systems that can support multiple data models, ill-defined workflows, and information structures (Weber-Jahnke et al., 2013). However, what is prevalent today is that most healthcare systems either avoid wide-scale surveillance to ensure privacy protection and confidentiality or require extensive approval processes to share data. Even when such approvals exist, all patient identifiers may be stripped off, therefore limiting the analytics utility of such datasets (Cavoukian & Emam, 2011).

2.4.2 Data Sharing Agreements

One of the conventional methods of preserving privacy and confidentiality, especially with a 3rd party data custodian is through Data Sharing Agreements (DSAs). A DSA is a fundamental component of cloud-based solutions for supporting connected healthcare delivery (Ruiz et al., 2016). DSAs are very important for describing policies for maintaining privacy and confidentiality, especially with a third party data custodian (Matteucci, Petrocchi, Sbodio, & Wiegand, 2012; Navarro, 2008; Swarup et al., 2006). DSAs are an essential consideration in a circle of care (Mathew & Pillai, 2015; Spruijt-Metz et al., 2015) and a requirement for a cloud-based implementation to be HIPAA compliant (de la Torre-Díez et al., 2013; Mendelson et al., 2014).

A DSA is an agreement among collaborating data providers, that regulates the conditions for data sharing. It is also an agreed-upon mechanism for ensuring privacy with electronic data

exchanges (Matteucci, Petrocchi, & Sbodio, 2010). It can also be seen as a set of policies specifying what collaborating entities are allowed or denied access to data ownership and data use as covered by an agreement.

A DSA is a legal agreement among collaborating data providers, regulating the conditions for data sharing (Aziz, Arenas, & Wilson, 2011; Matteucci et al., 2012). A DSA can also be interpreted as a specification of the set of policies that determine what datasets collaborating organizations are allowed or denied access to data ownership and use (Aziz et al., 2011). A DSA specifies the purpose of use, participating organizations, prohibitions on secondary use, data elements to be extracted, formats, meta-data, data classification and organization, quality assurance, storage, security, data recipient responsibilities, intellectual property rights and legal requirements (Caimi, Gambardella, Manea, Petrocchi, & Stella, 2016; Waterloo, 2017). A DSA also provides an agreed-upon mechanism for ensuring privacy compliance with electronic data exchanges (Matteucci et al., 2010; Navarro, 2008). A key factor in DSA adoption is usability and applicability (Matteucci et al., 2012).

2.4.3 Patient Identity Management and Record Linkage

In community care, different organizations use a different identifier for their patients. Often, organizations that provide services not covered by government insurers use their own identity number. In some jurisdictions like Ontario, the use of a government health number for identification purposes is not allowed for many community care organizations, making it very difficult to create a consolidated view of patient services for performance management (B. Eze et al., 2017). Accurately identifying a patient is critical because identification affects clinical decision making when coordinating services. Improper patient identification negatively impacts man-

agement of patients through duplication of services, assessments and test results, and consequently increases the cost of care to both the healthcare system and the patient (Demster et al., 2011).

The most common attributes for matching or linking patient identities are names, date of birth, gender, address, phone numbers, and government-issued identifiers (Demster et al., 2011). Most interoperability solutions assume that each patient must have a government-issued id such as the Social Security Number (SSN) in the US or Social Insurance Number (SIN) in Canada. However, as it relates to community care services, many of the organizations involved may be prevented by law from requiring this identifier before providing services to those that need them. Also, without a mechanism for validating these identifiers, data attributes are prone to data entry errors. Patient identity matching, as identified by Mills (2006) is challenging because of the use of nicknames, hyphenated names, last name changes, unreliable personal identification, last name reversal, as well as frequent address and phone number changes. These inconsistencies in patient identification attributes across these organizations make consolidating health services data, privacy protection, and performance management very challenging, and sometimes impossible.

2.4.4 Identity Matching Algorithms

Record linking algorithms are usually categorized based on their complexity (Just, Fabian, Webb, & Hjort, 2009). The basic algorithms use deterministic approaches, while intermediate algorithms use probabilistic approaches that leverage fuzzy logic and weights. Advanced algorithms use automated weight allocation, as well as statistical, data mining, and machine learning approaches to record linkages.

An identity matching algorithm is deterministic when there is a unique identifier or key across all data sources. A deterministic identity matching algorithm is considered error-free since records are matched based on this identifier only. However, when error-prone identifiers are used for linkage or matches, decisions become probabilistic based on a level of confidence. Such algorithms use probabilistic record linkage techniques (Gu, Baxter, Vickers, & Rainsford, 2003). According to Gomatam et al. (2002), in a probabilistic model, record pairs, based on a set of attributes, can either be considered a match, possible match, or non-match, based on a weighted comparison.

Sachs et al. (2000) describe an algorithm that employs a variation of deterministic record linkage using a unique identifier and basic matching on other identifying attributes while employing phonetic roots of first names to reduce mismatches from data entry errors and misspellings. Sachs et al. (2000) may not perform well on datasets where patients do not share common identity attributes.

Sauleau et al. (2005) employ an approximate string-matching technique with clustering. This work also incorporates a data standardization phase for match attributes. It uses weighted blocking for matching by creating overlapping subsets called "canopies" based on records within a loose threshold distance from a cluster center computed from some fixed blocks, derived from substrings in the patient first name (FN) and date of birth (DOB). While this algorithm is relatively complex, it supports only FN and DOB values in building its match clusters. This dependency on FN and DOB attributes makes the algorithm susceptible to collisions since FNs are pretty common in any population.

Three anonymous, population-based Dutch perinatal registries were linked using a combination of deterministic and probabilistic record linkage techniques (Méray, Reitsma, Ravelli, &

Bonsel, 2007). It also employed the Expectation Maximization algorithm (Dempster, Laird, & Rubin, 1977) in calculating the weights of the match blocks. One of the conclusions from this work is that deterministic approaches produced considerably worse results compared to probabilistic approaches on error-prone data. This difference is corroborated by Zhu et al. (2015) with the additional fact that the differences between these two approaches diminish if the datasets have few missing values and errors (validated data). If that is not the case, then probabilistic record linkage is preferable, more efficient, and produces better results.

Record linkage systems require all stakeholders involved to share identifying data on very sensitive patient attributes such as names, date of birth, gender, address, etc. Therefore, there are concerns about patient privacy and confidentiality (Gu et al., 2003; Mills, 2006). Some record linkage approaches perform matching on encrypted attribute values – allowing the linking of databases between entities that otherwise would not share such data. These approaches are usually referred to as "Privacy-preserving record linkage (PPRL)" (Vatsalan, Christen, & Verykios, 2013). The biggest shortcoming with these approaches is scalability since cryptographic matching is computationally expensive, and the associated algorithms can be very complex to implement.

2.4.5 Attribute categorization for anonymization

Protected Health Information (PHI) sometimes referred to as attributes, are categorized under one of four categories:

1) **Direct Identifiers** – attributes that singularly identify an individual in the dataset. These attributes include names and government-issued identifiers such as social insurance numbers, social security numbers, and driver license identifiers (Samarati & Sweeney, 1998).

2) **Quasi Identifiers** –attributes that on their own cannot identify an individual (Samarati & Sweeney, 1998). However, when quasi-identifiers (QI) are combined, they behave like direct identifiers. Most re-identification efforts link QI values to publicly available data repositories to re-identify individuals in an anonymized dataset. A popular study (L Sweeney, 2000) carried out by Samarati showed that 87% of the US population could be identified uniquely by their gender, date of birth, and their 5-digit ZIP code. The more QI values an adversary knows, the higher the level of re-identification that can be carried out on a target data set.

3) **Sensitive Attributes** – attributes that are usually public data on their own but sensitive if associated with an individual. In the healthcare domain, sensitive attributes represent health information associated with a patient such as procedure and drug codes, as well as disease conditions (Fung, Wang, Chen, & Yu, 2010a). They are not considered quasi-identifiers since they are not necessarily the type of data that is out there and known to a de-anonymization adversary. However, knowledge of QI values could reveal these sensitive attributes (Fung et al., 2010b; Ninghui, Tiancheng, & Venkatasubramanian, 2007), therefore helping an adversary gain more background knowledge on their victims.

4) **Insensitive Attributes** – attributes that are not sensitive and would not reveal any sensitive information about individuals in the dataset.

Figure 2-4 shows sample records with various attribute categories. The target of anonymization is to ensure that both identifiers and sensitive attributes are protected from a de-anonymization adversary.

	Direct Identifiers		Quasi-identifers			Sensitive Attr.
s/n	Name	Phone No	Zip Code	Age	Nationality	Condition
1	Vladmir Arnonova	6131569222	13053	28	Russian	Heart Disease
2	Tom Green	6482652145	13068	29	American	Heart Disease
3	Chu Lee	5052692312	13068	21	Japanese	Viral Infection
4	Johanna Marer	2022614623	13053	23	American	Viral Infection
5	Maria Durhame	4562356021	14853	50	Indian	Cancer
6	Masha Apostolova	2580264566	14853	55	Russian	Heart Disease
7	Helen Tulid	3625698233	14850	47	American	Viral Infection
8	Tim Cook	2364586523	14850	49	American	Viral Infection
9	Jason Borner	2589335422	13053	31	American	Cancer
10	Kristin Mushaf	2653154523	13053	37	Indian	Cancer
11	Mi Wong	1545210112	13068	36	Japanese	Cancer
12	Jim Washington	2410251236	13068	35	American	Cancer

Figure 2-4 Attribute Classification illustrated

There are three types of disclosures: identity, attribute, and membership disclosures. Identity disclosure occurs when the record of an individual or an entity in the dataset is re-identifiable. It is also referred to as record linkage since a record in a dataset can be linked to the actual record belonging to the individual (Fung et al., 2010a). Attribute disclosure occurs when new information can be gained on the sensitive attributes by an attacker or adversary. It is essential to understand that attribute disclosure is often a consequence of identity disclosure while membership disclosure is a probabilistic measure of the presence or absence of an individual in a dataset (Eze & Peyton, 2015). This knowledge changes the behaviour of an adversary towards de-anonymization. Both identity and attribute disclosures are important in our approach to protecting patient privacy in this thesis.

2.4.6 Adversary and Privacy Models for Risk Determination

Privacy models are the core tenet of anonymization. Various research efforts are geared toward applying various privacy models to ensure proper anonymization of their target datasets. *k*-Anonymity is the most popular measure of anonymity. Introduced by Samarati and Sweeney (Samarati & Sweeney, 1998; Sweeney, 2002b) in 1998, this algorithm ensures that each record in a dataset has at least *k*-1 indistinguishable records. We simply identify each group of records

with the same values in all their quasi-identifiers to be equivalent or in the same equivalence class. If all the equivalence classes in a dataset satisfy *k*-Anonymity, the target dataset is *k*-Anonymous. *k*-Anonymity protects against identity disclosure, but it cannot guarantee protection against attribute disclosure of the sensitive attributes (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006; Ninghui et al., 2007).

A re-identification adversary can discover the values of sensitive attributes when the diversity is low after anonymization. Say there are at least *k* records in the same equivalence class and we assume that our adversary knows only the QI attributes but not the sensitive attributes, identifying the equivalence class for an individual would reveal the rest of those sensitive attributes not known originally if they happen to be the same values across the equivalence class. From Figure 2-5, we can easily figure out that any male between 30 and 40 that lives in the area with Zip Code that starting with "130" has cancer. Of course, we have to be sure that our victims are in the anonymized dataset.

Basically, even when an equivalence class satisfies *k*-anonymity, it may not satisfy *l*-diversity for the sensitive attributes if they are not diverse enough (Machanavajjhala et al., 2006; Ninghui et al., 2007). *l*-diversity protects against sensitive attribute disclosure by requiring every equivalence class to have at least *l* well-represented values for each sensitive attribute (Ninghui et al., 2007). *t*-Closeness (Kohlmayer, Prasser, Eckert, & Kuhn, 2013; Ninghui et al., 2007) takes this a little further by requiring that the distribution of these sensitive attribute in each equivalence class be close to the distribution of the attribute in the entire data set.

	Quasi-identifers			Sensitive Attr.
s/n	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Quasi-identifers			Sensitive Attr.
s/n	Zip Code	Age	Nationality	Condition
1	130*	20-29	****	Heart Disease
2	130*	20-29	****	Heart Disease
3	130*	20-29	****	Viral Infection
4	130*	20-29	****	Viral Infection
5	1485*	>39	****	Cancer
6	1485*	>39	****	Heart Disease
7	1485*	>39	****	Viral Infection
8	1485*	>39	****	Viral Infection
9	130*	30-39	****	Cancer
10	130*	30-39	****	Cancer
11	130*	30-39	****	Cancer
12	130*	30-39	****	Cancer

Figure 2-5 Attribute disclosure and inference attack illustrated

Table linkage shows those scenarios where the adversary can infer the presence or absence of an individual's record in a data set or table. *d*-Presence is the privacy model that protects against such adversarial knowledge. According to Nergiz et al. (2007), *d*-Presence indirectly protects against identity and attribute disclosures because if the adversary has a *d*% percent confidence that an individual record is in a data set, then the probability that the adversary can identify or link the individual record is at least *d*%.

e-Differential privacy is the privacy model that focuses on how an adversary would change the probabilistic belief on the sensitive information of a victim after accessing the published anonymized data. According to the author of e-Differential privacy (Dwork, 2006), the risk of an individual's privacy should not substantially increase when the database is queried or used for statistical analysis. A dataset meets differential privacy concerns if this risk remains relatively unchanged by the presence or absence of an individual record in the dataset (Dwork, 2006).

2.4.7 Anonymization Techniques for Healthcare datasets

These four basic models form the basis for other privacy models as well as most anonymization algorithms and frameworks. Generalization is the most popular anonymization technique (Ninghui et al., 2007; Sweeny, 2002a; Sweeney, 2002) for satisfying *k*-Anonymity. It uses forms of aggregation and clustering to group individuals into equivalence classes with the same QI values. However, real datasets tend to produce unique groups, putting most of the equivalence classes or quasi-identifier groups at risk of re-identification. The generalization of an attribute is a transformation of its data to yield larger clusters with the same data values. For example, a Postal Code value "*K2E 4E6*" can be transformed into "*K2E**" to increase its cluster size. Anonymization processors usually require a generalization tree or generalization hierarchies (Sweeny, 2002a) for each quasi-identifier. These hierarchies are then combined to form a generalization lattice. It is usually necessary to search this lattice for optimal generalization set for all quasi-identifiers that satisfy *k*-Anonymity and sensitive attributes satisfying *l*-Diversity while providing the least information loss and the greatest data utility. This process of finding this optimal solution from a generalization lattice is NP-Hard.

There are many algorithms for effectively searching the lattice for an optimal solution, notably among them are Samarati's k-minimal generalization algorithm (Samarati & Sweeney, 1998), Sweeney's Datafly (Sweeny, 2002b), Kirsten's Incognito (LeFevre, DeWitt, & Ramakrishnan, 2005), El Emam's Optimal Lattice Anonymization (OLA) (El Emam et al., 2009), Flash from Kohlmayer et al. (2012) and many others.

Most generalization algorithms use global recoding of attribute values. That means the same transformation is applied to each quasi-identifier value. However, there are efforts to develop algorithms that perform this action locally for each equivalence class (J. Li, Wong, Fu, &

Pei, 2008; Xu et al., 2006). Local recoding increases data utility but also drastically increases the complexity of the generalization process. This study considers any research that looks into local recoding on high dimensional datasets as very relevant. The anonymized dataset employs local recoding.

Generalization is a great anonymization technique for achieving *k*-Anonymity and sometimes *l*-diversity. However, it is not always enough because it degrades the utility of a dataset. Generalization decreases the number of equivalence classes in a dataset but increases the size of each equivalence class. Suppression, on the other hand, strikes a balance between utility and availability. Equivalence classes whose sizes are less than *k* need to be suppressed to keep the entire dataset k-Anonymous. Algorithms such as OLA (El Emam et al., 2009) and Flash (Kohlmayer et al., 2012) try to determine the optimal level of generalization that would result in the minimal level of suppression. Suppression can be done at a record or cell level. Record level suppression simply deletes those equivalence classes that are not *k*-Anonymous. Unfortunately, suppressing those records could be overly excessive. Cell level suppression, on the other hand, determines those quasi-identifier values that make a tuple identifiable and removes them. When used appropriately, it provides the least information loss for anonymization. There are algorithms that use perturbation to supplement generalization instead of suppressing identifiable attributes. Usually applied to sensitive attributes, perturbation techniques either shuffle at-risk quasi-identifiers and sensitive attributes or merely replace them with fake masks of the original.

2.5. Related Work

This section surveys related work in cloud computing or similar technologies that were used for surveillance and performance management of healthcare. These frameworks can be divided into the following four categories:

1) Cloud-based Software-as-a-Service (SaaS)
2) Cloud-based Peer-to-Peer
3) Cloud-based Containerization
4) Semantic Web and RDL Type Frameworks

2.5.1 Cloud-based SaaS Frameworks

Cloud environment offers a platform for developing complex applications capable of processing time-series data from different sources, providing scalability as requirements such as workloads continue to increase over time (Ochian et al., 2014). One such system, Cloud Health Information Systems Technology Architecture (CHISTAR) uses asynchronous communication through loosely coupled components to achieve semantic interoperability, data integration, and security (Bahga & Madisetti, 2013). CHISTAR was deployed to Amazon EC2 in the following phases: tier-1 - web servers, load balancers, tier-2 - cloud-based distributed batch processing infrastructure like Hadoop. The target objective of CHISTAR Framework is to create an aggregate patient EHR in the cloud.

CHISTAR uses Hadoop/HBase Cloud storage for data management and relies on web servers and apps for service orchestration. Services are presented to clients as REST API web service interfaces. CHISTAR's archetype and template model make it easy to add or make

changes to data structures used in message exchanges between applications. They can be deployed at runtime and facilitate data validation during data capture, import, and querying. Data from different EHR systems gets converted to flat files that get stored in the Hadoop File System (HDFS) distributed storage. It then uses the MapReduce-based bulk loader to load processed data into HBase. Hive, the data warehouse system for Hadoop is used for analysis. Authentication is done through SAML Single Sign-On (SSO) with Federated Identity Management. Data at rest is encrypted using AES-256 and data exchange is done using SSL over HTTP. Connectors are built and configured against data sources. For example, there is an HL7 connector for reading/writing HL7 files.

CHISTAR uses a metadata repository for looking up data from various sources to an intermediate XML file with all data elements from the source data but eliminating the need for understanding the source syntax for various analysis. Semantic matching and relations are determined, ensuring that the data from the different sources are semantically identical. This process provides a mapping XML file used within the import process.

Though CHISTAR has a patient portal for viewing aggregate data, it has no support for dynamic analytics result push to the participating organizations and external stakeholders. In this thesis, performance management frameworks similar to CHISTAR will be referred to simply as **Software-as-a-Service**.

2.5.2 Cloud-based Peer-to-Peer Frameworks

Many government legislations do not yet permit the storage of health data on the cloud mostly because of jurisdictional restrictions. One approach is to store data in the cloud but maintain all identifying data locally. The PACE healthcare architecture (Donnelly et al., 2014) uses a combination of cloud and peer-to-peer technologies to model healthcare units or clinics where

Personal Health Information (PHI) is stored in off-cloud data storage while non-identifying data kept in the cloud. According to the authors, each PACE user acts as a peer, which is connected to other peers over a P2P connection for exchanging patient data over a secure connection.

This was demonstrated through collaboration with a number of dementia researchers in Ireland. This framework uses a hybrid application that anonymizes local patient identifiers while using anonymous identifiers on the cloud. PACE provides a useful framework for protecting patient privacy, but this can have a considerable impact on performance since each query must transverse all the peers.

In this thesis, this type of performance management framework will be referred to simply as **Peer-to-Peer**.

2.5.3 Cloud-based Containerization Frameworks

These are health information systems that use cloud-based containers or dockers for each participating organization. The authors of this paper (Andry et al., 2015) review the layering of cloud infrastructure from IaaS through to SaaS through a prototype implementation for a Home Care Delivery use case. This work identifies the need to build a cloud architecture not from the bottom up or top down but from inside out by making Platform as a Service (PaaS) the central critical layer for an elastic and extensible framework for a Home Care Delivery use case.

In this work, custom PaaS containerization is used to abstract healthcare services. The authors propose creating custom containers for healthcare services. It creates a generic open PaaS infrastructure that includes identity management, security (authentication, authorization, and single sign-on) and support for the management of cloud-based connected devices and clinical workflows.

Patient EMR is made available to patients with profile authentication done through tradition login and password credentials or through online identity providers such as Facebook, Google, LinkedIn, and Twitter.

In this thesis, this type of performance management framework will be referred to simply as **Containerization**.

2.5.4 Semantic Web and RDF Type Frameworks

Resource Description Framework (RDF) is a powerful tool for federated querying and heterogeneous data integration. The SPARQL query language for RDF is used to query RDF based documents for patient, clinical, and diagnostic data. This World Wide Consortium (W3C), Linked Open Data (LOD) community project aims at publishing various open datasets as Resource Description Framework (RDF) on the Web (Pathak, Kiefer, & Chute, 2012). The LOD project publishes various "fake" open datasets as RDF and extends those with links to actual data items from different data sources containing data on genes, proteins, care pathways, diseases, and drugs.

This related work (Amato et al., 2013) describes a prototype framework that supports document exchanges between heterogeneous data sources over the cloud using Semantic Web Resource Description Framework (RDF) and Web Ontology Language. Agents stationed close to each data source encode outgoing data attributes to semantic web XML/RDF documents and the agent at the destination decodes the RDF documents back to the local format. Supported data extraction techniques include XPATH (for XML documents), SPARQL (based on RDF) and SQL (for legacy relational databases).

This work suggests that semantic web techniques and data format can be used for encoding data for heterogeneous applications using cloud infrastructure for data transmission between

various on-premise applications. It leverages the cloud as a medium for exchanging RDF documents using a common semantically annotated model, requiring data owners to incorporate these agents within their data domain. Unfortunately, it fails to provide interoperability between existing health applications.

A framework (Poulymenopoulou, Papakonstantinou, Malamateniou, & Vassilacopoulos, 2015) for obesity surveillance, tackled the semantic interoperability challenge using HL7-CDA data from heterogeneous sources using a semantic Extract Transform Load (ETL) service. Extracted data is in the form of RDF documents stored in NoSQL databases such as MongoDB and HBase. Similarly, this work (Minutolo, Esposito, Ciampi, Esposito, & Cassetti, 2014) describe a mechanism for deriving OWL ontologies from XML schemas to make XML documents available for semantic queries. By providing mapping for each XML type to RDF ontology, the translation could be done on the fly.

Pathak et al. (2012) share early experiences in applying Linked Data principles towards representing patient data from EHRs at Mayo Clinic in RDF. Most clinical data with the clinic EHR systems are available as unstructured data in narratives from transcribed physician dictations. Unfortunately, unstructured data does not owe itself well to searches, summarization, decision support, and statistical analysis. Natural Language Processing (NLP) techniques are used to extract structured data from unstructured and semi-structured clinical narratives. Data collected from various sources containing demographics, diagnoses (ICD and SNOMED CT), procedures (Current Procedural Terminology - CPT codes), lab results, and various reports for cardiology, microbiology, and pathology get converted to RDF for cloud storage.

A case study for the cloud-enabled search of disparate healthcare data was carried out by Bhaskaran et al. (2013) using Microsoft Cloud – Azure. Crawlers were set up at each collaborating organization to push data to the Azure platform from each local Health Information System (HIS). A free-text based search solution indexes patient data from different internal IT systems to provide a consolidated view of the patient within each HIS. Data from various sources are then accumulated in the cloud and subsequently indexed for quick information retrieval and reporting.

Large-scale interoperability involves many players and different data sources. Sinaci and Erturkmen (2013) propose a federated semantic metadata registry (MDR) framework to address semantic interoperability for data created through the many semantic standards in healthcare. This central metadata registry /repository would maintain a set of common data elements in the domain. This way, data attributes that are semantically similar are not ambiguous between the data sources and requestors. Supported data extraction techniques include XPATH (for XML documents), SPARQL (based on RDF) and SQL (for legacy relational databases).

In this thesis, these types of surveillance and performance management frameworks will be referred to simply as **Semantic Web**.

2.6. Chapter Summary

This chapter provided background content on the various concepts that are fundamental to this thesis. We provided some background on healthcare performance management, introducing concepts like Triple Aim, healthcare interoperability standards, and the Common Data Model. We then described various forms of healthcare data sets. We introduced cloud computing, relating it to the healthcare industry. We also discussed some of the concepts related to big data analytics before introducing privacy legislation, compliance techniques, patient identity

management, adversarial models, and anonymization techniques for high-dimensional data datasets.

Finally, we introduced related surveillance and performance management frameworks for healthcare, providing the necessary background on how these frameworks work and the technologies that were employed.

The next chapter will describe the problems and challenges associated with leveraging cloud computing for wide-scale surveillance and performance management of healthcare. We provided a set of evaluation criteria that can be used to evaluate any framework or approach intended to provide systematic surveillance and performance management for healthcare.

Chapter 3. **Problem Definition**

In section 3.1, we define the problem of surveillance and performance management in healthcare using examples from current practice. In section 3.2, we identify gaps in current approaches and related architectures or frameworks as identified from the relevant literature, domain experts, and approaches used in many healthcare organizations. We then leverage a gap analysis to understand where and how existing approaches are problematic or deficient. Finally, in section 3.3, we discuss a set of evaluation criteria as identified in the academic literature, domain experts, and some of our research outcomes.

3.1. Current State of Community Healthcare Performance Management

The current state of performance management in community healthcare is depicted in Figure 3-1. Typically, a regional health authority (RHA), will contract community care services to smaller community care organizations that target specific populations with niche community care services. Usually, because these smaller community care organizations have a very limited budget, enterprise IT systems are not deployed. As a result, there is minimal interoperability and minimal ability to do performance management. Most data collection is in the form of ad-hoc organization-specific invoices (typically in Excel format) for services rendered by each Community care organization that was submitted to RHA financial officers for payment. Norway, for example, does not contract out community care services but track community care budget through local health subsidiaries (Ringard et al., 2013).

In this current state, Business intelligence officers leverage this data to create data reports and do have a limited ability to analyze performance and respond to ad-hoc requests.

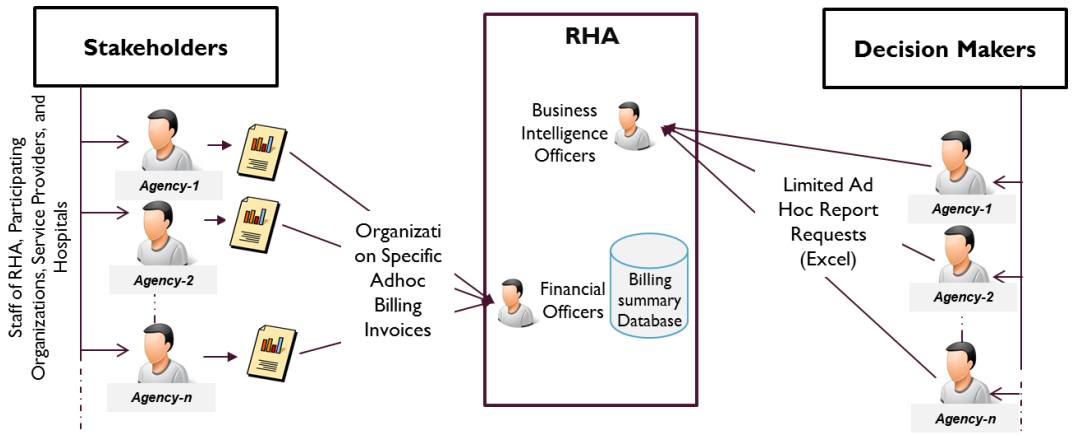


Figure 3-1 Current State of Performance Management with RHAs Depicted (B. Eze et al., 2017)

Most reports are manually created, often after intensive manual data collection specific to a data request. In the current state, the RHA relies almost entirely on ad-hoc reports, mostly prepared for funding and budgeting needs.

Aggregating data for performance management based on the current practice is:

1) Very time-consuming and can only be done a few times in a calendar year;

2) Highly error-prone and cannot provide answers quickly to population health question;

3) Susceptible to duplications since activities from healthcare service providers are usually not well coordinated;

4) Too high-level to provide consolidated care picture on individual patients;

4) Cannot be used dynamically, since it could take months or years to answer basic questions on patient care or population health.

There are three main challenges or gaps in attempting to address this situation: the lack of a common model, patient identity management, and regulatory compliance.

3.1.1 Lack of a Common Data Model

Performance management across independent community care organizations can be painful as data exchanges are ad-hoc – mostly through faxes, with minimal useful analytics applications. Healthcare applications use different data representations and storage schemas. Even within the same healthcare organization, it is sometimes impossible to reconcile data representation strategies as a different vendor develops each application. In community care, most community health providers are smaller organizations with a limited budget, and with the least automation, they can get away with.

Data quality is equally a challenge. Since the individual mandates for these organizations differ widely, the level of data they can collect and hold on a patient also varies greatly. Also, there are considerable differences in ontology used in various healthcare applications, as well as data representation, and interpretation. This lack of data standardization significantly limits the level and frequency of performance management that an RHA can undertake because 1) Local expertise is required for each local database translation, making surveillance and analytics error-prone; 2) Identifying events associated with an individual across all data sources becomes very difficult or impossible; 3) Compliance rules and Privacy protection through anonymization is very difficult to realize because of standardization and profiling barriers.

3.1.2 Patient Identity Management

It is easy to assume that a patient is identified similarly across the healthcare system, but it has been shown that it is hardly the case in practice (see section 2.4.1, 2.4.3). Each healthcare organization tends to identify a patient differently. The use of a government-issued healthcare

identification number is mostly inconsistent and mostly plagued with data quality issues if not validated electronically.

In Canada for example, community health care providers don't necessarily identify the patient by the government issued health card numbers since some patients do not have those because of homelessness, excessive drug user, or simply to protect their privacy like mental health patients and those living with diseases such as HIV/AIDS. Instead, each organization uses some internal identifiers within its healthcare infrastructure to identify patients. Health card numbers only come into play when the provincial government is being billed for services provided to the insured patients. In some cases, patients receive treatment for services without any form of formal identification. What we see is that many community care providers do not consistently capture important patient identification information such as patient Health Card Number, demographics and contact details, making it impossible to create an aggregate view of patient electronic health record across a health region.

For wide-scale surveillance, a patient profile must be identified across all the community care organizations in the health region. Accurately identifying a patient is critical because identification affects clinical decision making when coordinating service delivery. Improper patient identification negatively impacts the management of patient care. These are seen through the duplication of services, health assessments and test results, and increases in the cost of care delivery to both the healthcare system and the patient. This challenge poses a significant bottleneck to wide-scale surveillance and performance management.

3.1.3 Regulatory Compliance

Compliance with privacy laws can be challenging to implement operationally. Usually, data sharing agreements are drawn up and signed on paper, making them very difficult to operationalize. However, if compliance is designed into a surveillance and performance management infrastructure, it would make it possible to have solid privacy protection done alongside a robust compliance framework for implementing and enforcing data sharing agreements across participating organizations.

Privacy laws like the US HIPAA and Canada PIPEDA aim at regulating the use, transmission, and storage of personal health records. While these policies protect patient privacy and confidentiality, the implementation has become one of the biggest impediments to wide-scale surveillance and performance management in healthcare.

What is prevalent today is that most healthcare systems either avoid wide-scale surveillance to protect privacy protection and confidentiality or require extensive approval processes to share data. Even when such approvals exit, all patient identifiers may be stripped off, therefore limiting the analytics utility of such datasets. Some of the critical compliance concerns include:

1. Ensuring that only the caregivers that require access to a patient have access to their health record and only for the time when such access is required.
2. Ensuring that both patient and organization consents are applied during analytics. For example, a patient that fails to consent to data sharing must be excluded from all patient-level analytics. However, this doesn't preclude appropriate privacy protection for patients that consent to data sharing as is required by privacy laws with the jurisdiction. The sensitivity of healthcare data requires that such protection is given to all patients if analytics data is to be made public.

3.2. Gap Analysis

In this section, we analyze how the related work discussed in section 2.4 addresses the gaps identified in section 3.1.

3.2.1 Cloud-based SaaS Frameworks

The CHISTAR system (Bahga & Madisetti, 2013), a type of SaaS framework achieves cloud-based data integration and semantic interoperability using archetype models. Services are presented to clients as Web service interfaces. Data from different EHR systems gets converted into flat files to be stored in the Hadoop File System (HDFS) distributed storage. It uses the MapReduce-based bulk loader to load the data into HBase. Hive, the data warehouse system for Hadoop is used for analysis.

The strength of this type of framework is its ability to create an aggregate patient PHR in the cloud. There are two major gaps associated with SaaS frameworks. First, it impedes on organizational autonomy since all participating partners must use the same patient management system and database. It essentially forces all partners to use the same model. Secondly, there is a lack of support for dynamically pushing analytics results to the participating organizations and external stakeholders that are not part of the cloud-based SaaS application. The assumption is that stakeholders need to go through the portal to access reports and analytics results.

3.2.2 Cloud-based Peer-to-Peer Frameworks

The primary objective of peer-to-peer data sharing frameworks like the PACE healthcare architecture (Donnelly et al., 2014) is to use a combination of cloud and peer-to-peer technologies to model healthcare units or clinics where PHI is stored in off-cloud data storage while non-

identifying data is kept in the cloud. This has been demonstrated through collaboration with a number of dementia researchers in Ireland (Donnelly et al., 2014). This type of framework requires a hybrid application that uses actual patient local identifiers within each participating organization infrastructure while using anonymous identifiers on the cloud. The strength of peer-to-peer frameworks is its strong privacy protection since each organization maintains controls over PHI associated with its patient.

The significant gap with this type of framework is that privacy protection through anonymization can be quite complex to implement for data sharing. Also, simply replacing identifiers is never enough for protecting patient privacy, especially in the healthcare domain with high-dimensional events (B. Eze & Peyton, 2015). Also, analytics require proper identification of the patient across all participating organizations. Another major gap with these frameworks is that they assume that all participating organizations provide data of the same quality and that all patients can be accurately identified across all participating organizations. In community healthcare, this is usually not the case.

3.2.3 Cloud-based Containerization Frameworks

These frameworks use cloud containers or virtual machines to create separations for each tenant. For these frameworks, Platform as a Service (PaaS) is the central critical layer for an elastic and extensible framework for hosting and delivering healthcare services on the cloud (Andry et al., 2015).

PaaS containerization frameworks simply target technical interoperability. Unlike SaaS models, PaaS maintain organizational independence and autonomy while providing a platform that

facilities service sharing and collaboration. Organizations may use the same or similar data schemas or models. However, the major gap is that it doesn't address data or semantic interoperability. Data quality issues could limit the analytical utility of aggregate data. Therefore, their effectiveness for continuous performance management is highly questionable.

3.2.4 Semantic Web and RDL Type Frameworks

These frameworks use the cloud as a medium for exchanging RDF documents using a common semantically annotated model and requires data owners to incorporate these agents within their data domain. Semantic web RDF and SPARQL query language are mature frameworks. The strength of these frameworks is that RDF is a powerful tool for federated querying and heterogeneous data integration. Efforts similar to the Linked Open Data (LOD) community project that aims at publishing various open data sets as Resource Description Framework (RDF) on the Web (Pathak et al., 2012) showcase the potentials of RDF. A central metadata registry/repository would maintain a set of common data elements in the domain. This way, data attributes that are semantically similar are not ambiguous between the data sources and requestors (Sinaci & Laleci Erturkmen, 2013).

The major gap with this type of information exchange scenario is that they require specialized customizations for local EMR or CCIS applications within each organizational domain. In addition, the ability to query remote RDF sources as remote data sources means that designing a Performance Management Infrastructure around such a model can be very complex since data is not cached or preprocessed for analytics.

3.3. Evaluation Criteria

This thesis proposes a systematic surveillance and performance management architecture for achieving interoperable healthcare solutions and for validating progress on Triple Aim objectives. To evaluate this architecture, it is important to have evaluation criteria for determining how well the infrastructure supports Triple Aim objectives (Verma & Bhatia, 2016).

The set of evaluation criteria identified in this section came from:

1. A careful review of various related work where relevant criteria were identified. We cite the relevant literature for these criteria.
2. Feedback and discussions from our interactions with domain experts in the healthcare industry (listed below).
3. Our own experiences when evaluating our candidate solutions. Occasionally, there was a criterion not mentioned in the literature or articulated by the domain experts that was crucial for communicating why our solution was or was not addressing the problem effectively.

Our domain experts are:

1. **Jamie Stevens** is the Director of Business Intelligence and Performance at the Champlain Local Health Integration Network, Ottawa, Canada. Jamie has over 15 years experience managing BI and Performance Management teams in community healthcare. He knows what it means to operationalize performance management data to describe an RHA performance goals. Jamie's expertise is on community care data collection and aggregation, and business intelligence.

2. **Paul Boissonneault** is the Director of Information Systems and CIO, Performance, and Strategy at the Champlain Local Health Integration Network. Paul has led the IT initiatives at the LHIN for 25 years. Paul has been championing interoperability among the community care organizations in the Champlain region. He leads a highly technical team that has championed many integration efforts at the Champlain Local Health Integration Network, the community care organizations, and hospitals in the region. Paul's expertise is in cloud infrastructure – systematic data hosting, cloud infrastructure security, privacy, patient identity management, and legal compliance.

Sections below describe each of the evaluation criteria.

3.3.1 Triple Aim Objectives

This evaluation criterion uses three important metrics: 1) Improvements to patient experience to care delivery, 2) Improvements in overall population health, and 3) Cost savings to the healthcare system.

3.3.1.1 Improvements to Patient Experience

Patient experience improvements are difficult to measure because it builds on a combination of factors. Stakeholders in a patient circle of care are responsible for creating care plans using specific guidelines. Paul Boissonneault identifies that patient experience can also be measured through a patient questionnaire or simply by measuring the incidence of complaints. If those decrease over time, then we assume that patient experience has improved. Jamie believes that measuring metrics such as time to first service after a patient seeks community care services is a good measure of the overall patient experience to care because it reduces wait times.

3.3.1.2 Population Health Improvement

Jamie Stevens identified a few measures for population health. A reduction in the emergency room (ER) admissions over time is an indication that community healthcare intervention is reducing the need for patients to go to the ER. Other measures include better awareness of inefficiencies in community care so providers can focus on providing a better care landscape for patients. Professor Peyton highlighted the fact that continuous performance management is the only way the health care system can track and maintain good population health. In this thesis, the focus is on providing an architecture for continuous data integration and performance management. This way, the RHA can measure healthcare outcomes. However, because of the complex mix of factors influencing this criterion, we will not be able to provide a formal evaluation of population health improvements in this thesis.

3.3.1.3 Reduction in Health Care Cost

With very little outcome data, it becomes difficult to measure most Triple Aim objectives for the region. We have through the course of this research identified this metric as an important measure of cost savings to the health system. It also allows us to measure how well our architecture supports better coordination and monitoring of care delivery at both the individual patient and population levels. Jamie Stevens sees this as an important measure since the reduction in healthcare cost means more care is being provided to more patients at the same budget.

3.3.2 Surveillance Services Interoperability

This evaluation criterion targets all levels of interoperability (section 2.1.4) and measures how much impact our architecture has at addressing interoperability challenges with data surveillance services. Interoperability will be evaluated in the following three sub-categories - Technical, Semantic, and Process.

1. **Ease of Interoperability** - This component of technical interoperability measures the level of complexity with integrating each participating organization PHR data into a common infrastructure. The more complex this process is, the more likely some providers would not participate because of technology and skillset gaps. Paul Boissonneault identified this as a major criterion in healthcare since smaller healthcare providers such as physician practices and community care organizations usually don't have the expertise or funding to undertake major technical integration projects.
2. **Efficiency of Data Exchange –** This semantic interoperability metric measures the ease of communicating data across providers. If communication is plagued by service disruptions and outages, then this would rank very low on the scale (Hincapie & Warholak, 2011).
3. **Efficiency of Data Encoding** – Clinicians, community healthcare administrators, and care coordinators, usually, do not have the time to encode clinical notes in a standard format except for billing purposes. Depending on the organization, patient demographic details may have data entry errors if not validated. Data coding delays, improper encoding, and sometimes lack of encoding could affect the interpretation and usefulness of sections of a patient PHR. Jamie Stevens identifies data encoding as a major impediment to operationalizing data for an RHA.

4. **Efficiency with data translation** - Translation errors can occur as data is translated from a health information system internal format to those of general standard and vice versa. According to Benson (2012), the choice of interchange language alone is not sufficient. Each message exchange needs to be translated correctly without errors, so data exchanges are consistent, coherent, and computer readable.
5. **Collaborative communication –** This metric is identified from academic literature in support of Process Interoperability (C. Kuziemsky, 2013). All participants in a patient "Circle of Care" need to be able to use the cloud platform to communicate effectively. Therefore, data or instructions from a physician or a community care coordinator need to be received in the right mode by other participants. Reporting it alone is not enough. For example, a patient that visited the ER should trigger notifications to all participants in the circle of care.
6. **The efficiency of collaborative decision-making** – This Process Interoperability metric, identified in C. Kuziemsky (2013) measures how effective collaborative decisions are being made on behalf of the patient, and the correlation between the effectiveness in the decision-making process to patient quality of care in terms of timeliness of interventions?

3.3.3 Performance Management Services

This evaluation criterion measures the ease of setup of Performance Management Services across the domain, including data feed management, patient profiling, data clustering, as well as reporting and data subscription. It is important that this can be done on a large scale as required by an RHA.

The need for an analytics Infrastructure that supports performance management services was identified by our domain expert, Jamie Stevens. This included the need for dynamic data and report subscriptions.

1. **Support for an analytics infrastructure** – This metric determines the nature and type of analytics infrastructure supported for performance management. This is identified in the literature as an important requirement for performance management workflow automation (Mouttham, Peyton, Eze, & El Saddik, 2009).
2. **Dynamic Analytics Report Generation** – This metric measures how dynamic analytics reports can be generated or produced from aggregate data by healthcare providers. Jamie Stevens sees this metric as very important in a cloud-based infrastructure that supports process interoperability as the framework needs to have to capacity to trigger, create, and stream reports as required by users and external processes.
3. **Data/Report Subscription** – This metric measures the architecture's ability to provide support for subscriptions to data or analytics reports. Jamie Stevens identified this as an important instrument for increasing internal efficiencies with health care providers. If staff can receive and review the data they need at the time they need it through data and report subscriptions, they can do their work more efficiently at less administrative cost.

3.3.4 Common Data Model

A common data model is the end product of data integration across disparate data sources. A common data model for performance management is needed to ensure a consistent view of information across all data sources (De la Rosa Algarín et al., 2014; Klann et al., 2014; Sabooniha et al., 2012).

1. **Data Structure Definition** - This measures the ease of defining a common data model for the healthcare domain. In the course of this research and corroborated with concrete examples from the LHIN, we identify that healthcare data come from various sources and in various formats. Jamie Stevens and Paul Boissoneault believe that being able to have surveillance done on structured, semi-structured, and unstructured datasets is important for a regional performance management infrastructure.
2. **Support for batch and Streaming Datasets** - This measures the support for data of various velocities – static, batched, and streaming data. This evaluation metric was identified from our literature survey (de la Torre-Díez et al., 2013; Ochian et al., 2014).
3. **Ease of change to model** – In healthcare, one thing that is common is the constant change in data collection and types of analysis being requested by caregivers. This metric measures the rigidity of a common model where it exists. Can the model adapt easily to changes in requirements, new data elements, etc.? Jamie Stevens pointed out that it is a major challenge with the Business Intelligence team at the LHIN. Changes to data elements can cause many reports to error out, and in many cases, the team is unable to find and fix such issues until they are reported by users.
4. **Scalability** – While having a common data model is great, one important evaluation metric for this model is scalability. Can the data collection process scale to support many parallel ETL processes at the same time? This is the main reason why our research primarily targets the use of Cloud infrastructure for this type of Performance management since it offers infinitely scalable storage for very data-intensive applications (Bhaskaran et al., 2013).

3.3.5 Patient Identity Management

This evaluation criterion ensures that patient data across all data sources are correctly identified and associated with the right patient in creating the common data model. This criterion was identified in the course of soliciting requirements for integrating various community care organizations in the Champlain region. According to Jamie Stevens, it became obvious early in the process that we could not rely on the community care organizations to identify a patient EHR using the government-issued health care number. Many of the patient records from these organizations do not have this important identifier, and for those that have it, the data is not validated, so are prone to data entry errors.

1. **Type of Matching** – This metric measures if the matching system is deterministic or probabilistic. Probabilistic matching is preferable when data quality issues exist (Gu et al., 2003).
2. **Support for Phonetic Roots** - Phonetic roots links first and middle names from the various forms to a phonetic root based on their pronunciation. This is much more efficient than simple misspelling correction (Sachs et al., 2000).
3. **Support for Declarative Match Definitions** - Measures how adaptable the matching algorithm is with external declarative definitions for tweaking its precision and general performance. Paul's team sees this as a critical requirement for ensuring consistency accuracy of the matching system and the ability to respond quickly to changes to data structure changes or matching requirements.

3.3.6 Privacy Compliance Model

This criterion evaluates the solution's ability to sustain data surveillance while supporting both organizational data sharing agreements and patient consent. This is necessary because binding agreements that protect patient privacy and confidentiality need to be signed between these organizations to allow their data to be part of the systematic data collection and analytics processes described in the thesis.

This evaluation criteria is supported by academic literature as an important consideration in a circle of care (Mathew & Pillai, 2015; Spruijt-Metz et al., 2015) and identified as a necessary component for a cloud implementation (de la Torre-Díez et al., 2013) to be HIPAA compliant (Mendelson et al., 2014).

1. **Nature of Data Sharing Agreement** – This metric measures the nature of data sharing agreement signed by data custodians and how it can be operationalized. Data sharing agreements can be in three forms: paper-based, electronic but not integrated, and fully integrated DSAs. Professor Peyton identified this as an important metric since the more automated it is, the easier it is to operationalize in a cloud computing environment. It is important that privacy protection be done alongside a solid compliance framework for implementing and enforcing data sharing agreements (Coats & Acharya, 2013)
2. **Patient Consent** - This measures the efficiency of incorporating patient content in a performance management framework. Is it integrated and how granular is the definition? We have identified in the course of our research and from the literature (J. Hu, 2011), that patient consent is not always an on or off switch. It depends on the context of use, the level of details being exposed and the nature of the data recipient. For the LHIN in its role as the data custodian for the region, this is an important requirement for any performance management system.

3. **On-demand Anonymization** – This metric measures if the architecture supports some form of on-demand anonymization to preserve data utility in conjunction with DSAs and Patient Consent (Cao, Carminati, Ferrari, & Tan, 2011). Some of the other important features for an anonymization tool required for addressing privacy compliance in a cloud-based surveillance and performance management infrastructure identified in the course of our research are described below.
    a. **Declarative attribute definition –** For anonymization to be done in a distributed manner, datasets need to be described and annotated declaratively.
    b. **Direct Identifier, Quasi-Identifiers, and Sensitive Attributes –** Declarative attribute definitions must support the classification of attributes based on the type of sensitive data they hold.
    c. **Resource-based libraries** – These are libraries for supporting anonymization processes like generalization hierarchies for various types of quasi-identifiers and lookup gazetteers for direct identifier masking.
    d. **Risk Assessment and Measurement** – Anonymization algorithm needs to be able to assess and measure re-identification risks, including prosecutor, and journalist risks on both cross-sectional and longitudinal, high-dimensional data sets.
    e. **Anonymization Approaches** – These include generalization with or without local recoding, tuple suppression, and date shifting. For high-dimensional datasets, traditional anonymization tends to fail, so a tool needs to support data model complexity reduction techniques.
    f. **Selective Anonymization** - Finally, for mixed data sets with varying sources and target recipients with varying risk profiles, tools must be able to carry out both selective and

none selective anonymization as required to keep the target data set anonymous to an adversary while maintaining high analytical utility.

3.4. Chapter Summary

In this chapter, we identified the key challenges with surveillance and performance management in healthcare, especially as it relates to community care. We then analyzed the current approaches to performance management for regional health authorities and their limitations. We also analyzed the current approaches used in related work and their limitations. Finally, we developed a set of evaluation criteria drawn from the following sources – review of academic literature, analysis of related work, gap analysis, discussion and feedback from domain experts, and finally from experiences drawn from our research.

The next chapter introduces our proposed surveillance and performance management architecture for regional health authorities. This architecture is designed to address the key gaps identified and discussed in this chapter.

Chapter 4. Surveillance and Performance Management Architecture

In this chapter, we present our cloud-based surveillance and performance management architecture for community healthcare. These are the four critical components of our architecture:

1. A Cloud Computing Infrastructure for surveillance and performance management that enables a multi-tenanted private cloud provided by a regional health authority to host applications and operational databases (the entire server infrastructure of each organization) for community care healthcare stakeholders. This infrastructure integrates the following services to enable a regional health authority to effectively monitor and manage community healthcare:
    a. Surveillance services for hosting community care data and collecting it into a single comprehensive common data model.
    b. Performance Management Services for continuous on-demand and privacy compliant analytics, reporting, and subscription in support of process interoperability.
2. A Common Data Model for addressing semantic interoperability and for creating a consolidated and consistent view of a patient profile for performance management.
3. A Patient Identity Matching Service for identifying patient profiles across multiple community care organizations with various internal identities for a patient.
4. A Privacy Compliance Model based on the Privacy Compliance Definition Document for operationalizing Patient Consent and organizational data sharing agreements.

4.1. Architecture Overview

Figure 4-1 provides an overview of our architecture. A data custodian, typically a Regional Health Authority (RHA), provides a multi-tenanted private cloud infrastructure for surveillance and performance. The role of the RHA can be played by a national health authority if all the regions in a country are hosted in a single cloud owned and operated by this organization. The infrastructure could also be owned and managed by a regulated third-party commercial entity (e.g., in Canada, a Crown Corporation such as Canada Post or Air Canada). In the context of this thesis, we assume that an RHA plays the role of the data custodian for all participating community care organizations.

The first component of our architecture is the cloud computing infrastructure for providing surveillance and performance management services for small community care organizations. Our architecture supports two Surveillance Services - a Systematic Data Hosting Service and a Systematic Data Collection Service. The Systematic Data Hosting Service is a cloud IaaS service for hosting data from various community care organizations in the cloud. This service moves the burden and responsibility of financing and managing health system applications from the community care organizations to the RHA. It also offers additional benefits of providing a common operating platform for these independent community care organizations while removing the limitations associated with moving large volumes between these organizations and the cloud infrastructure. Please see section 4.2.1 for more details on this service. As the data custodian, the RHA must respect and enforce data sharing agreements (DSA) signed by each collaborating Community care organization.

The Systematic Data Collection Service supports heterogeneous data sources through data aggregation PaaS containers. It supports the collection of data into a Staging database where it can be processed into a Common Data Model to support community-wide performance management. The behaviour of this service is controlled by the status of each Community care organization's DSA in terms of Organization Consent for data sharing. Please see section 4.2.1 for more details on this service.

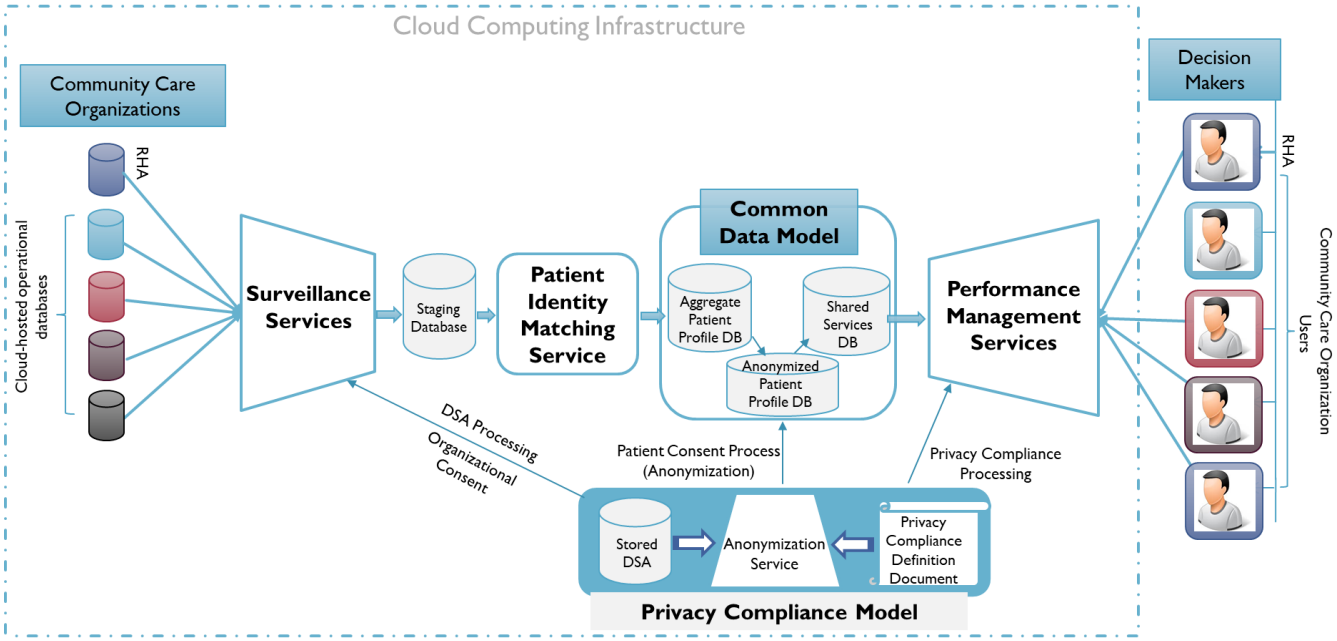


Figure 4-1 Cloud Computing Architecture for Surveillance and Performance Management

The Performance Management Services supported by the cloud computing infrastructure includes – analytics, reporting, and subscription services. Analytics services manage data mining processes set up to profile patients through clustering, classifications, and association rules. Reporting services use report templates for generating various reports for the collaborating community care organizations that contribute data to the cloud infrastructure. Subscription Services allow users to subscribe to various reports and processed data feeds that get generated automatically and

pushed to each subscribing organization and their users. Please see section 4.2.2 (c) for more details on this service.

The second component of our proposed architecture is the Common Data Model (CDM). It represents the minimum view of data across all the attributes of the aggregate data set required for performance management while ensuring a consistent view of information across all data sources. To migrate data from the Staging Database to a Common Data Model, the transformation component of the Systematic Data Collection Service is used to addresses semantic interoperability issues with staged data. It also eliminates ambiguities in data elements for each data source. Semantic matching of data attributes must be done to translate data elements from various data sources to fit the CDM. Transformed data become semantically equivalent, with each data source updating only the applicable sections of the CDM. Please refer to section 4.3 for more details on the CDM. In this architecture, the CDM is also the central database that feeds the downstream performance management services.

The third component of our proposed architecture is the Patient Identity Matching Service. It performs the important role of ensuring that patient data across all data sources are correctly identified and associated with one unique aggregate patient profile in the CDM. Patient profiles in the Aggregate Patient Profile Database of the CDM represent the full Patient EHR across all participating community care organizations.

The fourth component is the Privacy Compliance Model. This component models the privacy compliance framework for operationalizing both organizational Data Sharing Agreements (DSA) and Patient Consent definitions through a Privacy Compliance Definition Document (PCDD). The organizational DSAs and explicit Patient Consent forms must be operationalized to ensure the compliance of the entire infrastructure to the laws governing the disclosure and use of

personal information. The Anonymization service uses the cloud infrastructure PCDD to enforce organizational DSAs and Patient Consent definitions on the CDM.

4.2. Cloud Computing Infrastructure

This section describes the architecture of the cloud computing infrastructure required to support surveillance and performance management. A related work (Andry et al., 2015), identifies the need to build a cloud architecture not from the bottom up or top down but from inside out by making Platform as a Service (PaaS) the central critical layer for an elastic and extensible framework. By leveraging PaaS containerization, we are able to build various IaaS models and offerings in very generic ways. The multi-tenanted private cloud uses custom PaaS containerization to create abstractions of the various services in our architecture.

The infrastructure is based on a cloud containerization abstraction that supports:

1. A systematic data hosting service for hosting and centrally managing the community care patients for each participating community care organization.
2. A systematic data collection service for supporting a variety of data source ETL drivers. In this thesis, this service supports drivers to the Community care organization cloud-hosted data sources, RHA internal databases, external RESTful Web services, and SharePoint data. It can be also be extended to support external sources that belong to stakeholders that have on-premises IT infrastructure. This service also supports most necessary transformations for incoming data streams to a common data model.

3. Performance Management PaaS containers that run micro-services for the performance management services – anonymization, analytics, reporting, and subscription.

Our approach is to host a resource library with multiple container images representing one or more surveillance and performance management functions. PaaS container images for these functions are developed by experts in the component area and hosted in the Resource Library. The PaaS container abstraction allows cloud developers to build cloud applications for performing roles like data collection, transformation, analytics, reporting, anonymization, and subscription - all leveraging a common set of tool and functions. This also offers a generic mechanism for scalability and high availability.

The Resource Library is central to all cloud infrastructure functions. In addition to PaaS container images, it also hosts meta-data for data migration, data transformation, patient identity matching definitions, privacy compliance definition documents, anonymization meta-data, and report templates and data subscriptions.

The basic container depicted in Figure 4-2 shows the basic functionalities of each service container. A Provisioner is a service that triggers the instantiation of a PaaS container image as a running node - virtual machine or a docker. After the initial bootup, the Provisioner configures and furnishes the node with various resources for its initialization. It also monitors the health of all running nodes. For each cloud function type, there could be many nodes running and providing various data surveillance and performance management services, all at the same time. For consistency, subsequent sections will refer to running nodes as Virtual machines.

Each container also hosts a Web service that allows it to interact with other containers and external applications. PaaS containers host a configuration cache for keeping internal configurations required for running specific tasks or jobs. Finally, each virtual machine incorporates an internal database for caching intermediate results while processing various tasks. Most containers are configured to connect to the Staging database and the Common Data Model to pull/push data and for in-place data transformations.

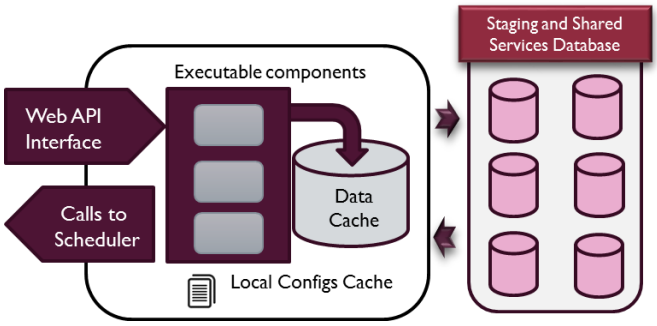


Figure 4-2 Basic PaaS Cloud Container Components

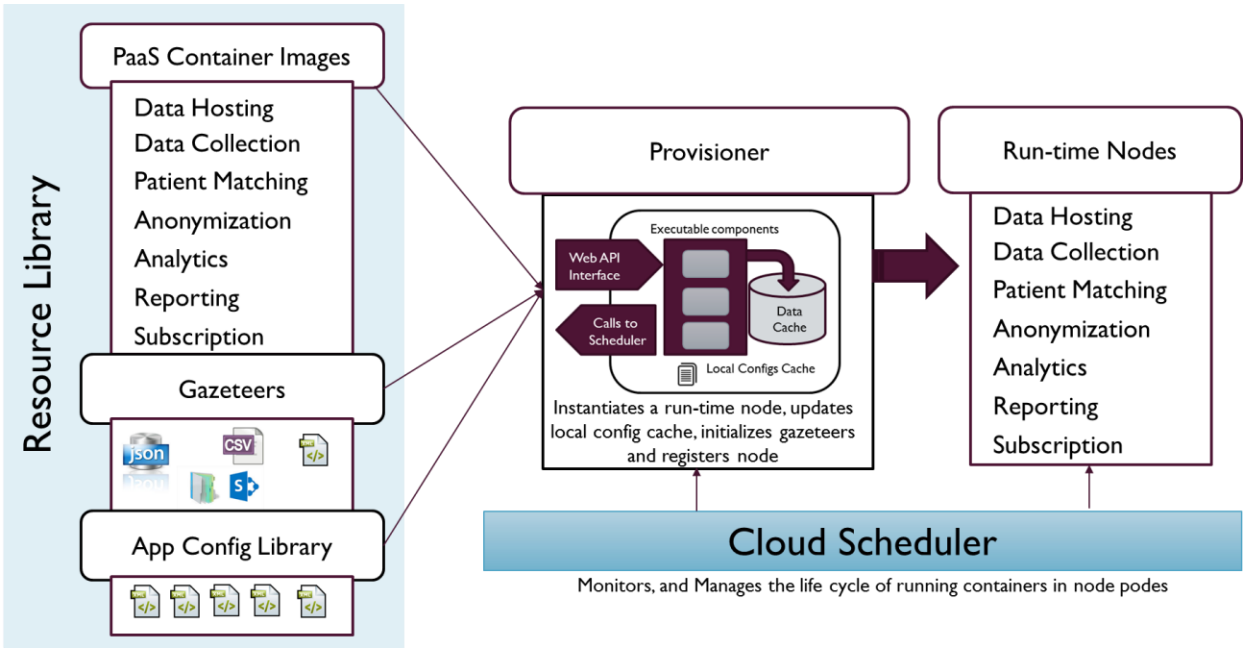


Figure 4-3 Role of the Infrastructure Provisioner depicted

Developers leverage this model in developing new container models for various functionalities in the performance management pipeline. Applications are developed to leverage this infrastructure in support of specific data integration, transformation, anonymization, subscription, and reporting requirements.

Within each PaaS container, high availability of resources such as infrastructure data caches, and the clustering of the staging databases allow nodes to scale to incoming tasks and restart tasks or jobs if failures are detected.

The containers can also receive explicit calls from the Cloud Scheduler to perform specific tasks based on specific configurations and instructions (Figure 4-3). Finally, the base container model incorporates an automatic update feature that checks and updates internal executable applications, so administrators don't have to perform this task manually for bug fixes and when installing new features.

4.2.1 Surveillance Services

The two Surveillance services supported by our architecture are Systematic Data Hosting and Systematic Data Collection services.

a) Systematic Data Hosting Service

The main purpose of this service is to host applications and operational databases (the entire server infrastructure of each participating Community care organization if required). With a multi-tenanted private cloud infrastructure, the RHA is able to host the Patient Information Management System and operational databases for each of the collaborating community care organizations on a common infrastructure while maintaining their autonomy. This essentially removes the burden of the management of this infrastructure from the smaller organizations that may not

have the resources – financial and technical expertise required to integrate with a cloud-based performance management infrastructure.

In this setup, IaaS VM images prebuilt with the Patient Management Information System required to set up each Community care organization are archived as ISO images within the resource library. For each new organization that needs to be cloud-hosted, the Systematic Data Hosting Service uses the Provisioner to apply the configuration set up by administrators. Depending on the size of the organization, one or more nodes are spun up using the appropriate ISO image. After provisioning, the running VM is linked to an organization-specific database instance from a clustered database server – serving as the operational database for the new Community care organization. This setup allows the Provisioner to set up and configure the applications and databases required by each participating Community care organization in an automated manner.

Administrators then create domain accounts for the users using LDAP or Windows Active Directory and assign them rights to their cloud applications and VM instances. Each of these Virtual Machines is monitored for overall system health through the Cloud Scheduler. Those that have higher loads would trigger the load balancer to spin up more VMs for the same function, therefore allowing users and applications to enjoy a seamless experience. The staff at each organization access the infrastructure through a secure Virtual Private Network (VPN) tunnel over the Internet.

The data tier is managed using a clustered database sitting on cloud-based storage for all the participating organization's operational databases. Therefore, database size is never an issue as the storage is simply elastic. By creating a custom database instance for each participating

Community Care Organization, organizational autonomy, privacy, and confidentiality are maintained within each operational database. When a new organization joins the network, data analysts use custom ETL processes to load/transform and import existing data into the new database instance from their legacy or in-house operational database. This way, they can access to all historical data in the new platform. To maintain organizational autonomy and security, data-at-rest is encrypted and will not accessible to both the cloud infrastructure administrators and the data custodian (RHA).

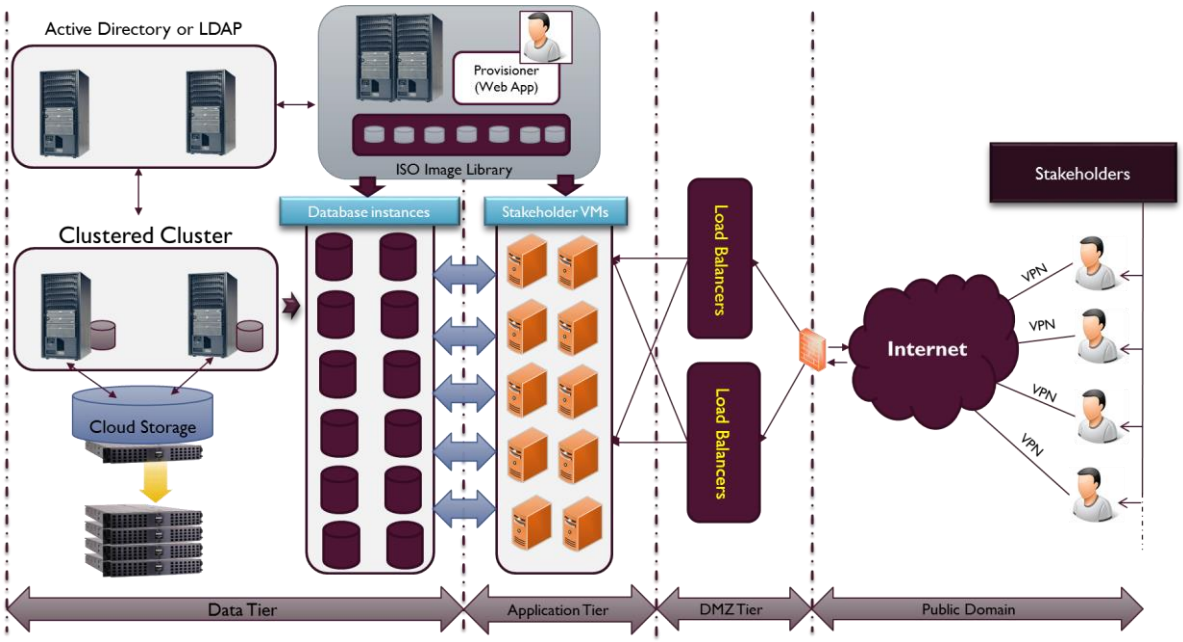


Figure 4-4 Private Cloud Architecture for Systematic Hosting Service

Leveraging cloud computing provides infrastructural scalability and elasticity at all the layers. At the data tier, high availability clustered database ensures that the database engine never fails. Cloud-based storage ensures that the databases for each organization can be expanded as needed without causing the Patient Management Information System to fail. The application tier VMs are also sitting on a cloud infrastructure that ensures high availability. The architecture of

these VMs makes them interchangeable, and with a load balancer, users are spread across multiple VMs to ensure the best performance and experience for all users.

It must be noted that while this setup primarily solves most technical interoperability challenges through the multi-tenanted private cloud infrastructure from the RHA, it does not address interoperability issues as each Community care organization operational database is still a silo. Also, while this architecture encourages a uniform platform for community care organizations, it does not mandate these organizations to use the same platforms. Nevertheless, if they use a different platform, a generic PaaS container must be created for the specific Community care organization. Further, patient identifiers within each database are different, and there is no consistency in the data elements that gets collected on patient and services, hence the requirement for data staging.

b) Systematic Data Collection Service

In our architecture, data collection is done through the data aggregation PaaS containers. Since each Community care organization data is different, custom data drivers are needed for various data sources. These include drivers for structured sources such as SQL Server/Oracle/MySQL database sources, MS Excel and CSV dumps; semi-structured sources like XML and JSON for specific hospital applications; healthcare exchange standards like HL7 and OpenEHR, and drivers to pull XML data from SharePoint document libraries. Our approach is to use custom data drivers for each type of data source. Each systematic data collection container type is then customized with a specific data source driver. For example, if there are five types of data sources across a hundred community care organizations, then only five drivers need to be supported. More drivers can be developed to support new data sources as needed.

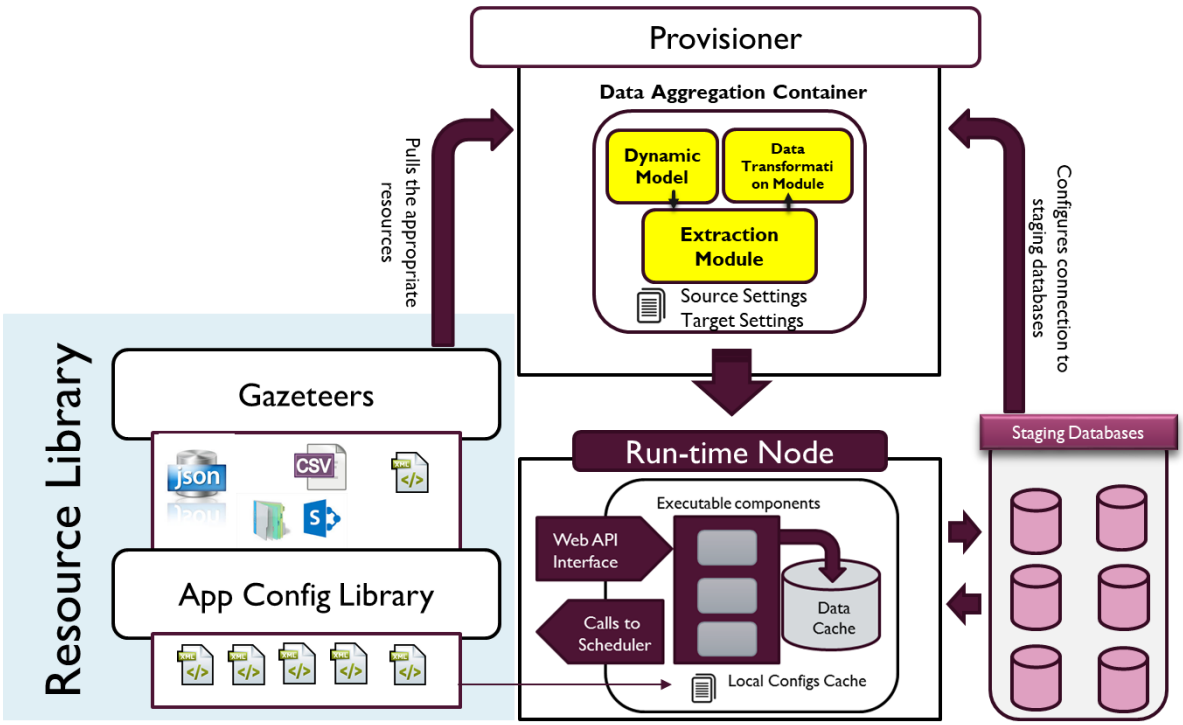


Figure 4-5 Data aggregation container provisioning process depicted

Each data source requires specific configuration settings; for example, connection settings for data sources - SQL Server, SharePoint library Web Services URL, and document library names, etc. For XML and JSON data, the required settings include the link to the external Web services and their authentication details. Aside from the data connectors, other data aggregation configurations that are supported include the list of data entities (tables), fields to exclude from, and target databases/tables within the cloud infrastructure. Data surveillance adapts to changes in the source scheme by reading and importing all fields in the source schema. The declarative configuration only needs to specify the field exceptions – for example, those fields that should be dropped from their source entities.

Like other container types, the data aggregation PaaS containers are archived as iso images in the resource library but are managed within the Provisioner. For each new Community care organization that become part of the infrastructure, one or more container images could be

developed and deployed to match specific data aggregation requirements for the healthcare application and staging database.

To connect to a data stream, the Provisioner spins up the appropriate container image and configures the source and target database connection settings. Data aggregation containers support two interfaces. The westbound interface is very specific to each application or data source while the eastbound interface points to the staging database. The infrastructure uses various tables in the staging database to cache data from the data collection services for data transformation to a common data model. Each container is configured to push data to specific database tables within the staging database. Staged data is destroyed after transformation and migration to the CDM.

Data Standardization

Data aggregation is a continuous process of creating and updating the staging database. Data in the staging database, being in a more structured form makes it more amenable to applying transformation rules. However, the data attributes at this point are still very much in their original forms from the source stream. The data standardization component of the data collection service uses context-based rule engines for various transformations to source data for the CDM database. These transformations include data conditioning, semantic matching, and other necessary transformations required to ensure the consistency of data in the CDM. This process ensures that each data element, irrespective of the source is semantically equivalent.

To support basic attribute transformations, data analysts that are subject matter experts research the data from each Community care organization to develop the transformation rules that map staged data to a version that is consistent with the CDM. Mapping rules specific to each

incoming data stream are then created and uploaded to the resource library with each entry mapping a data attribute to the generalized model and where necessary provide a set of transformation rules for the attribute. Transformations rules can be explicit or based on external gazetteer in the resource library. The following data standardization processes are supported:

1. **Replace Transforms**: Find specific patterns in one or more data fields and replace those with the replacement text. These are used to remove database defaults, replace non-standard texts to a more standard text. For example, "Male", "Female" to "M" and "F" respectively.
2. **Truncate Transforms**: Allows texts to be shortened through truncation. This allows retrieving a shorter version of names or the first 3 characters of a postal code or a shorter version of a street address.
3. **Split Transforms**: It is not uncommon to have fields that contain catenated texts. For example, a patient name field that includes either first name, middle name or last name separated by some separator like "John Mitch" or "John (Jack)". These patterns reduce the accuracy of the matching process. The Split transformation defines the rule for splitting such texts and how to treat its various components.
4. **In-place Transforms**: The matching process allows users to define any in-place SQL compatible data transformation functions. These functions can be used to catenate texts (CONCAT), convert dates to a common format (CONVERT or CAST), extract year of birth (YEAR, MONTH), get substrings (SUBSTRING), etc.

4.2.2 Performance Management Services

There are three main services for performance management – Analytics, Reporting, and Subscription. These services leverage the CDM to provide various data, analytics, and reporting

needs for all participating community care organizations and are controlled by the privacy compliance model. In this architecture, performance management services must not have access to the Staging Database.

a) Analytics Service

The analytics service provides continuous performance management of care processes by analyzing and discovering various patterns and relationships from the CDM. In this architecture, data analysts can define elements of the data feed that are interesting for analysis using a set of big data analytics templates. For example, the data stream on heart attacks to hospital emergency rooms (ER), analyzing the correlation between community care discharges and visits to the ER, or to determine if wait time to admissions has any correlation with patient comorbidity over time. Some of the models that can be applied to the data include:

- Classification models
- Time Series Analysis
- Association Rules and Correlation
- Sequence Discovery
- Clustering and Patient Profiling

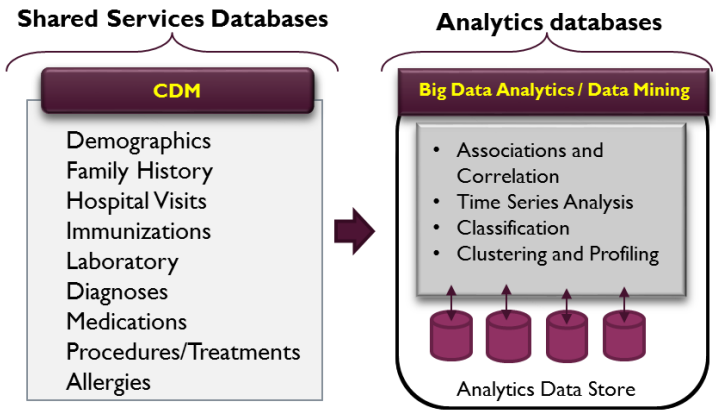


Figure 4-6 Analytics Services interaction with the Common Data Model

Each analytics model is packaged so it can be triggered declaratively to run on the CDM at set intervals. When deployed, each analytical model creates tables within the analytics database for the resulting data set. This database is refreshed daily, hourly, or as needed to for various analytical needs. The cached results are saved and monitored for interesting patterns (Figure 4-6).

These patterns can be monitored by data analytics and based on outcomes, triggers for "interesting" patterns can be set. When such patterns are detected, the analytics module will notify the relevant resource through either an email notification or a report or data subscription. For example, the correlation between medication history and certain diagnosis could trigger a subscription when it reaches a particular threshold. This subscription then pulls a full report on patients that have such diagnosis and immunization in the last 6 months for review by experts.

Finally, it is important that data analytics reports and indicators are fed back to the healthcare stakeholder to influence decision making and healthcare governance. These alerts can be in the forms of emails, analytics reports, data push to remote services, and updates to dashboards for various health providers. Operational systems can subscribe to advisory alerts that help guide physicians, nurses, community care coordinators, and other care providers in their day-to-day decisions on patient care, therefore improving the quality of care.

b) Reporting Service

The reporting service provides the infrastructure for publishing performance management reports that would be viewed by the Community care organization decision makers. It requires a distributed, and Web services enabled set of reporting servers that mirror each other. Each analytics report created by analysts is published to all reporting servers on the cloud computing infrastructure. Personnel from participating Community care organization use the reporting portal

to access the standard reports, dashboards, and other data visualization objects published on the reporting portal.

The subscription services use the Web APIs support available within the reporting service to dynamically create and stream data reports, a very important feature for cloud-based integration. Unlike the traditional report generation services that work with static data sets, and parameters, the reporting services are triggered with references to a data set, a report viewer, and a list of parameter values for filtering the report. This is designed to allow subscription or analytics requests to trigger the anonymization service first to create the needed anonymized data sets which are then sent to the reporting service as an input to the report creation process.

c) Subscription Services

The Subscription service is the component of our architecture that closes the loop regarding pushing information from the CDM to the stakeholders and decision makers at all levels across all collaborating organizations and their partners. It is also the component of the performance management framework that supports process interoperability described in (Benson, 2012; C. Kuziemsky, 2013). The subscription service leverages data from the CDM, other Performance Management services, LDAP, and Active Directory to provide knowledge, collaborative, and operational data needed for performance management of community care processes. The requested data sets are anonymized to meet their risk profile of the target recipient.

The key features of the subscription service that makes it adaptable to a cloud computing environment include:

1. Support for dynamic data-driven subscriptions with declarative and SQL executable definitions for subscribers and report parameters. All settings for subscriptions and reports are in XML.

2. Ability to package and deliver multiple reports for users in these different formats – HTML, MS Word, MS Excel, and Acrobat PDF.
3. Support for multiple delivery modes – email, file system, and calendar appointments.
4. Dynamic scheduling – daily, day periods, weekdays, weekly, monthly, and quarterly, specific days of the week, month, quarter.
5. Support for rich failure notifications for administrators.

For each report developed and published, a custom report definition file that allows the subscription service to stream the report in many formats is also published to the resource library. Figure 4-7 shows the main sections of a Subscription definition file. This is an XML document that allows users to specify reports and other data elements they require from the CDM.

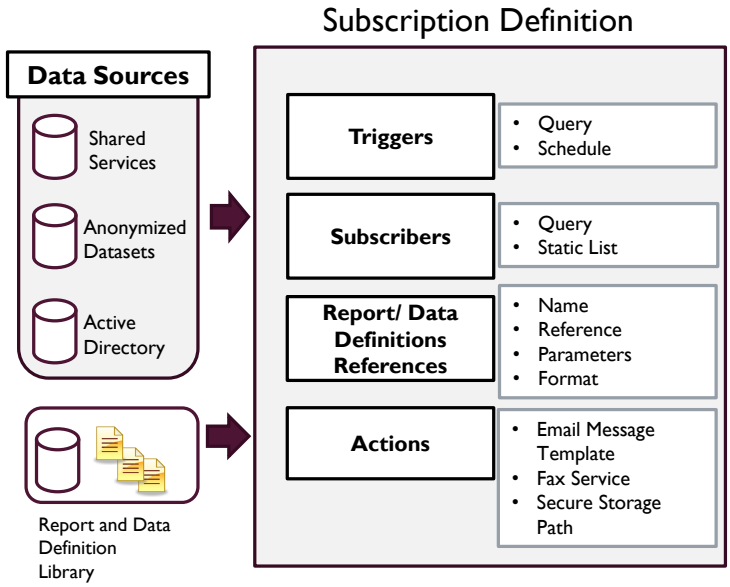


Figure 4-7 **Subscription Definition** (B. Eze et al., 2017)

A subscription definition is comprised of 5 different sections that work together with varying levels of interdependencies. The first component links each subscription to one or more data sources. The second component defines the triggers that must be met for a subscription to run.

The trigger condition could be a set time schedule, or a query result meeting a baseline condition. The third component defines a static or query generated list of subscribers. Subscriber definition can be static or dynamic, but for the best flexibility, dynamic query-defined subscriber list is preferable since it could be used to bring in a rich set of information attributes on each subscriber that can be used as source parameters for report definitions and action components of a subscription. Dynamic subscription query must be processed first by the subscription service to retrieve attributes of the subscriber that would be used as parameters for the reports and subsequent actions.

At run-time, the subscription service packages all the report for each subscriber in the specified formats based according to the subscription definition and then delivers them based on the action definition. For the current iteration, the actions supported include:

- **Email**: The reports are emailed using a template email message in the subscription definition.
- **Fax**: The reports are faxed to a fax number associated with the subscriber organization

Secure Storage: The reports are pushed to a cloud-based share that is only accessible to a defined set of stakeholders.

Subscriptions are created and managed by Business Intelligence officers with in-depth knowledge of the CDM data. The key utility of the subscription service is its ability to dynamically package and deliver multiple reports in various formats (HTML, MS Word and EXCEL, PDF, XML, and JSON) through multiple delivery mechanisms to data recipients.

4.3. Common Data Model

A Common Data Model for surveillance and performance management is needed to ensure a consistent view of information across all data sources (De la Rosa Algarín et al., 2014; Klann et

al., 2014; Sabooniha et al., 2012). We evaluated the following three options in addressing the CDM problem include:

1) Force all data streams from participating organizations to be transformed into the target database schema for the regional health authority or the data custodian;

2) Force all data streams to conform to a generalized standard such as openEHR ("openEHR Architecture," 2015), or HL7 CDA ("Introduction to HL7 Standards," 2016));

3) Convert all data streams to a logically and semantically equivalent common model for each health domain.

The first option limits the generality and the applicability of our approach. The second option would have been a more acceptable approach, but these standards relate more to clinical EMR systems with little support for non-clinical services (such as meals on wheels, social support, etc.) important for community care and are often coordinated by the RHA. The last option was preferred as it allowed us to define a common data model that embodies the minimum representative set of attributes for systematic surveillance and performance management.

4.3.1 Important features for a Common Data Model

The following considerations are recommended in designing a Common Data Model:

1. The design requires input from domain experts that know what data elements and attributes that are considered important for performance management of the target health region as regards to community care.
2. Data from each organization needs to be mapped into this model where applicable.
3. In achieving item (2), data elements must be organized into multiple hierarchical levels.

a. The 1st level is patient-centric data corresponding to patient identification and demographic data, contact details – phone numbers and addresses, personal and emergency contacts, caregivers, family physicians. Patient Identity Matching (section 4.4) is carried out on all patient-level records across all the incoming data streams based on published matching rules. At the end of the process, a global identifier is issued to each cluster of patient profiles belonging to a single patient across the collaborating organizations. This identifier is then used to map the rest of patient data (at the episodic and longitudinal levels) to the model.

b. The 2nd or episodic level captures data on episodes of care, such as service referrals, care plans, and medication history. Other important episodic data include Service Authorization (when and who admitted and authorized care), Care Pathways (Plan for care and/or treatment of each patient, as well as the frequency and limits), and Long-term Care Placements (Seniors admitted to long-term care placements). An important level 2 data is the patient population. In community care, patients are categorized into specific populations based on the identified care needs. Complex patients tend to belong to multiple populations groups and are identified differently by each participating Community care organization. Some common population classifications examples are 1) population by age include Seniors, Infants, Teenagers, Middle-aged; 2) population by the length of stay and needs severity for Long-term-Chronic, Short Stay, Short Stay-Rehabilitation, Palliative, Acute Palliative, and At-Risk Seniors.

c. The 3rd and lowest level captures longitudinal events and maps to data on the episodes defined at the 2nd level. These include event data such as patient visits, consultations, treatments, service visits, and billing details.

4. Support for data primitives, and user-defined types in the forms of privacy tags like firstname, lastname, city, country, postalcode, etc.. These privacy tags allow the model to perform automatic validation of all incoming data streams. Table 4-1 contains the full list of supported privacy tags.
5. The CDM must allow for updates to the model without disrupting existing services that are dependent on the model. Therefore, support for data versioning and the enforcement of versioning rules is highly desirable.
6. It must be both human and machine readable. The representation can be in XML/XSD, or any of the common data modeling notations with support for attribute validation, type checking, support for reusable types and groupings.

Table 4-1 Common Privacy Tags for Data Primitives

Privacy Tags	Meta	Description
Lastname		Patient Last names
Firstname		Patient First names
Othername		Other names
SIN		Social Insurance Number
SSN		Social Security Number
Identifier		Medical Record Number, Account Number, Driver License
Address		Full Address
StreetAddress	City	Street address. Requires an optional city parameter
City	State	City. Requires an optional state/province parameter
State/Province	Country	State or Province. Requires an optional country parameter.

PostalCode	City, State	Postal Code. Requires an optional city and/or state parameter.
ZipCode	State	US Zip Code. Requires an optional state parameter.
ZipCodeExtended	State	US Extended Zip Code. Requires an optional state parameter.
Country		
Hospital	Country	Name of country
WebUrl		Website or web resource URL
IpAddress		IP address
DateOfBirth		Date of birth of a patient
DateOfDeath		Date of death of a patient
EventDate		Admission date, discharge date, Assessment date
Age		Age of a patient
Height	Standard, Metric	Height defined in metric scale or static scale.
Weight	Standard, Metric	
Gender		Assumed gender of the patient.
Sex		Sex at birth
CPT		Current Procedural Terminology (CPT) Code
ICD	9,10	International Statistical Classification of Diseases version 9 or 10.
SNOMED		SNOMED CT Code
NBC		NBC Code
Unstructured		Unstructured data

4.3.2 Populating the Common Data Model

In a cloud-based Performance Management infrastructure, the CDM can be implemented as a structured or semi-structured database. If structured, this database should be hosted as a relational database such as with MS SQL Server, Oracle, MySQL. If semi-structured, then document or column-based databases like MongoDB or Cassandra should be used. What is important is that it must maintain a "virtual" tree-like representation that allows various sections of the data tree to

be updated from various data sources coming from the systematic data collection service (Figure 4-8).

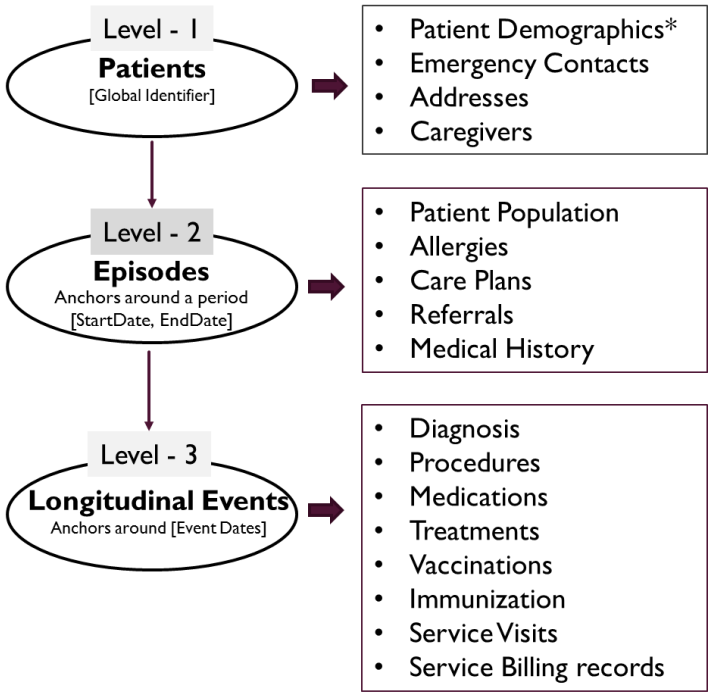


Figure 4-8 Common Data Model Hierarchical Levels Illustrated

Another important feature of the CDM is its ability to merge data from various sources without creating erroneous duplicate entries. For example, the data collection service from a Community care organization could push a patient profile home address but without the city or postal code attribute values. If another service provides these missing attribute values for the same address, the CDM should update the particular address in the tree with all the details without creating duplicate address entries. This feature is possible because each entry in the CDM must have a key field or set of fields whose values uniquely identify the record. A unique identifier can be used if such an attribute value exists in the incoming stream. When new entries match existing key records, they are automatically merged with the existing record.

4.4. Patient Identity Matching Service

Patient identity management is important with cloud-hosted data because, in some jurisdictions like Ontario, patient identifiers like the HCN are restricted, and each Community care organization is required to use an internal identifier, they control for patient identification. This presents some challenges to patient identity matching for cloud-hosted data, making it very difficult to create a consolidated view of patient service for performance management. In jurisdictions where a common patient identifier is enforced, the role of the patient identity matching is simplified or unnecessary. In addition, this service plays an important role in data collection by ensuring that patient data across all data sources are correctly identified and associated with the single patient profile record in the CDM. This is important in populating the CDM as it represents the full patient EHR across the community care organizations. It is also important in enforcing privacy compliance on CDM dependent performance management service.

For patient identity management, we assume that:

1) Each Community care organization has an internal identifier specific to their database for each patient,
2) Community care organization data could have data entry errors and may not consistently validate identifiers such as government-issued social insurance numbers even when they are consistently captured.
3) Identity matching should leverage the obvious patient identifiers and broader attributes of the patient, including current and historical addresses, phone numbers, and details of the patient personal contacts to fine-tune and verify matches.

4.4.1 Matching Algorithm

A generic record linkage system has the following components as identified by Gu et al. (2003):

1. **Data Gathering Component** – pulls and consolidates data from all sources of patient data.
2. **Match Attribute Identification** – determining those attributes of a patient that should be part of matching.
3. **Standardization Component** – formats data attributes across data sources to fix data quality issues.
4. **Blocking Strategy** – matching strategy for reducing the number of comparisons of record pairs.
5. **Decision Component** – determines how to conclude if a matching pair is a full match, non-match, or a possible match.

The choice of a matching algorithm determines the flexibility and complexity of components 2 to 4. The following options apply:

- For ***deterministic*** matching, there are a few uniquely identifying attributes that can is used for matching. Data standardization is minimal, and blocking strategy is simply a direct match of those attributes.
- For ***probabilistic*** matching, there can be a number of matching rules or blocks using one or more attributes of the patient. Match determination is a probabilistic combination of these rules with a fairly complex decision component.
- Finally, probabilistic matching can be done using ***machine learning techniques***. This may involve a complex set up for the matching processes for each dataset.

In the context of this thesis, the choice of deterministic vs. probabilistic identity matching is very dependent on the RHA jurisdictional authority to community care and those of all the collaborating community care organizations. These choices are dependent on the following conditions:

1. If the community care organizations share a common and well-validated identifier such as a government-issued identifier for all patients, then a special matching algorithm is not required as patient profiles can be consolidated deterministically based on this identifier only.
2. If the community care organizations do not share a common identifier but capture rich, valid patient profile data including first name, last name, date of birth, gender and addresses, then identity management can be done deterministically using these patient attributes.

Finally, if conditions (1) and (2) are partially fulfilled, with community care organizations contributing data with varying richness, and quality, then a probabilistic matching algorithm or more advanced machine learning techniques must be employed for patient identity matching.

Probabilistic Matching Algorithm

In this thesis, our proposed patient identity matching algorithm derives from existing works in the probabilistic record linkage domains such as the expectation-maximization (EM) algorithm (Dempster et al., 1977), as well as the theories of record linkage (Fellegi & Sunter, 1969). Ideally, record linkage requires a large number of record comparisons, a very expensive and inefficient process. For n patient profiles, m attributes for each patient, the complexity is of

the order $O(m * n^2)$. We chose a probabilistic matching algorithm because it is the best approach for the type and nature of datasets associated with community health care – disparate, sparsity, non-homogenous and very error-prone.

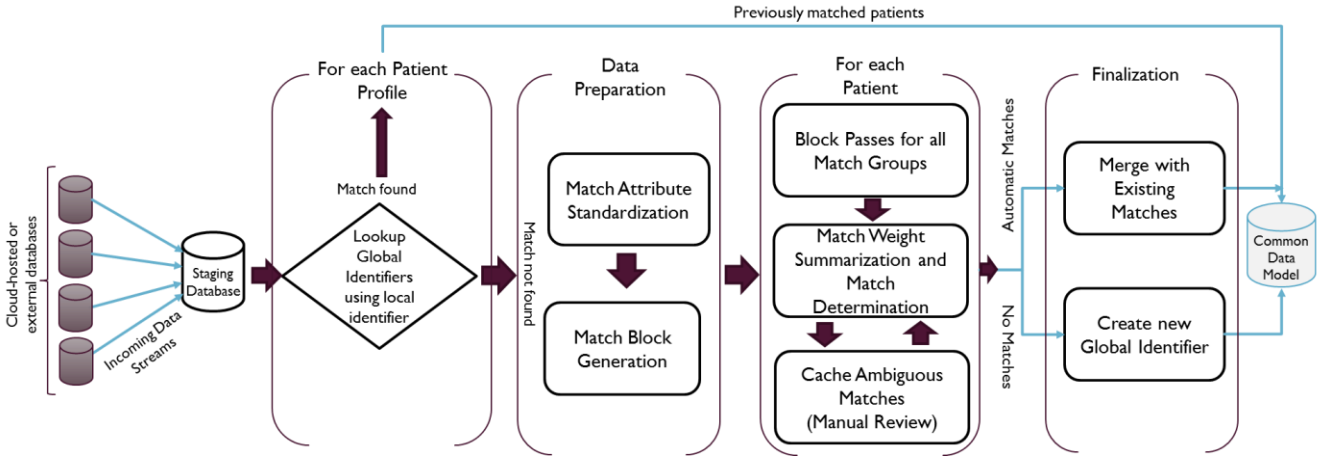


Figure 4-9 **Probabilistic Matching Algorithm** (Benjamin Eze et al., 2017)

The probabilistic patient identity matching algorithm is illustrated in Figure 4-9. Each patient profile from an incoming data stream goes through a lookup process using its local identifier. If it was previously matched, the patient profile would skip the matching process and proceeds to get processed into the common data model. If not, then it is a good candidate for the probabilistic matching process.

The precursor to matching is data preparation or standardization. Each patient profile goes through a data standardization process for the match block attributes. After this data standardization, match blocks are generated from these attributes based on the match configuration. In generating the match blocks, the matching algorithm addresses data sparsity and errors in these attributes by dividing incoming data into blocks to minimize record comparisons to only records within the same block (Fellegi & Sunter, 1969).

A block is a combination of one or more patient identifying attributes with an associated weight or probability. The weights used need to be heuristic and "tuned" specifically to reflect

the properties of the data set that requires matching. These weights reflect the level of confidence of the business analysts for matches on the block with the data used for matching. The higher the weight, the higher the level of confidence. For well-validated, error-free attribute values like government-issued identifier, a maximum weight of 1.0 can be assigned. Using multiple patient personal attributes with various weights provides more context for richer probabilistic matches.

4.4.2 Matching Process

The algorithm requires two match thresholds, one for full matches and the other for possible or ambiguous matches. These thresholds can be chosen arbitrarily or algorithmically. Matches with weights over the match threshold are automatically considered full matches while those within the possible matches range are saved for manual review. The rest of the matches is rejected.

Figure 4-10 illustrates this process. The algorithm performs and accumulates match results from multiple block passes. Matches from each match block pass are assigned the weight of the block in the match definition. At the end of all block passes, matched patient weights are combined across all matching blocks to compute their final composite weights. This process produces a composite weight accumulated from all block passes for each matched profile. This composite weight is compared to the threshold weights to determine if it is a match, non-match, or a possible match.

Each matched set of profiles is assigned a global identifier for the aggregate patient profile. To determine the global identifier for the matched set, a scan is done across all matches to determine if any of the candidate patient profiles are already matched to an existing match group. The highest weighted match group from this scan is then selected. At this point, all patients within this match set are assigned the global identifier from the highest weighted match group,

thereby increasing the size of this existing aggregate patient profile. However, if none of the candidate patient profiles is previously matched, then a new global identifier is generated and assigned to the new match set, representing a new aggregate patient profile.

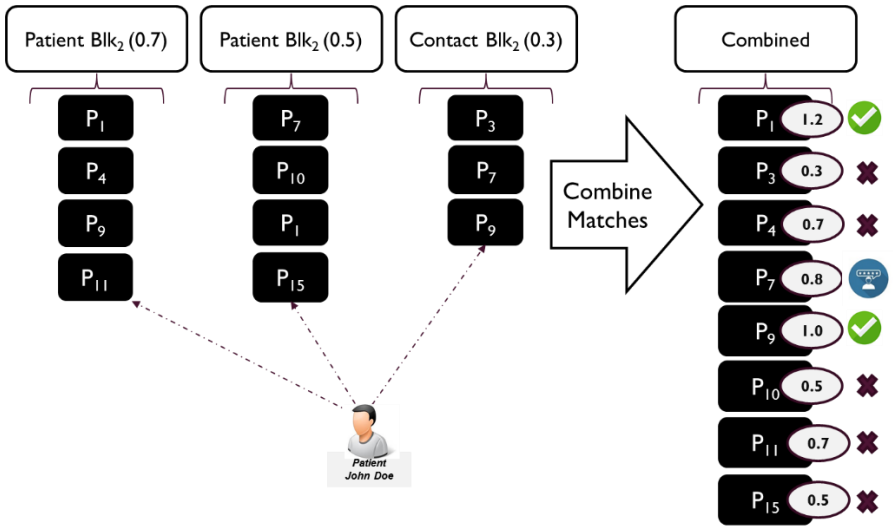


Figure 4-10 **Combining Match Blocks Illustrated** (Benjamin Eze et al., 2017)

The output of the matching process updates the Patient Match database while those matches within the lower threshold go to the possible or ambiguous matches list. Those can be reviewed by a subject matter expert through a web application and can subsequently be accepted or rejected as matches. Most ambiguous matches require data entry fixes. After these fixes are done, subsequent matches of these profiles would push these entries to the appropriate aggregate patient profile identified using the appropriate global identifiers.

4.5. Privacy Compliance Model

Privacy of patients is an important consideration for large-scale data surveillance and performance management, especially in the healthcare industry. In a cloud computing infrastructure, concerns over the nature and pattern of data sharing and confidentiality of patient sensitive health

data must be addressed (El Emam et al., 2009; Ma et al., 2014). Large-scale systematic data processing to support surveillance and performance management requires a compliance framework that operationalizes both organizational Data Sharing Agreements (DSA) and Patient Consent definitions. This ensures the compliance of the entire infrastructure to the laws governing the disclosure and use of personal information. Consequently, binding agreements that protect patient privacy and confidentiality should be signed by each participating Community care organization to allow their data to be part of the surveillance and performance management services.

4.5.1 Data Sharing Agreements

The two major sections of a DSA for each participating Community care organization are the Patient and Organization Consent forms. In our architecture, DSA applies to 1) Incoming data streams, 2) Data belonging to each organization within the CDM, and 3) Analytics and Reporting services outputs.

Organization Consent represents each participating organization's consent and permissions for accessing and using data within its custody. Organization Consent allows each participating organization to describe what data elements are allowed or not allowed expressively in conformance with the law and their level of comfort regarding utility vs. privacy and confidentiality. Organization Consent can be in the form of an explicit opt-in, opt-out or anonymize settings for each Community care organization that apply to 1) an entire data feed, 2) specific entities in the data feed, 3) specific attributes in the data feed, and 4) specific community programs (cuts across multiple entities). Therefore, none of the performance management results would be made available to the public without thorough privacy considerations.

As shown in Figure 4-11, the Organization Consent section allows each organization to include/exclude specific fields and attributes from the data that go into the CDM like those with

sensitive financial records and audit data. Each organization is also associated with a defined risk profile for the data they receive from the cloud infrastructure.

While Organization Consent has local significance for each incoming data stream, Patient Consent can have both global (across all data streams), local (specific to each organization), or partial (apply to select data entities and attributes) significance. Enforcing these consents could result in complete removal of patient data from the common data model. In some cases, there could be full or partial anonymization (data masking, generalization, suppression) of patient data to meet set risk thresholds for the infrastructure.

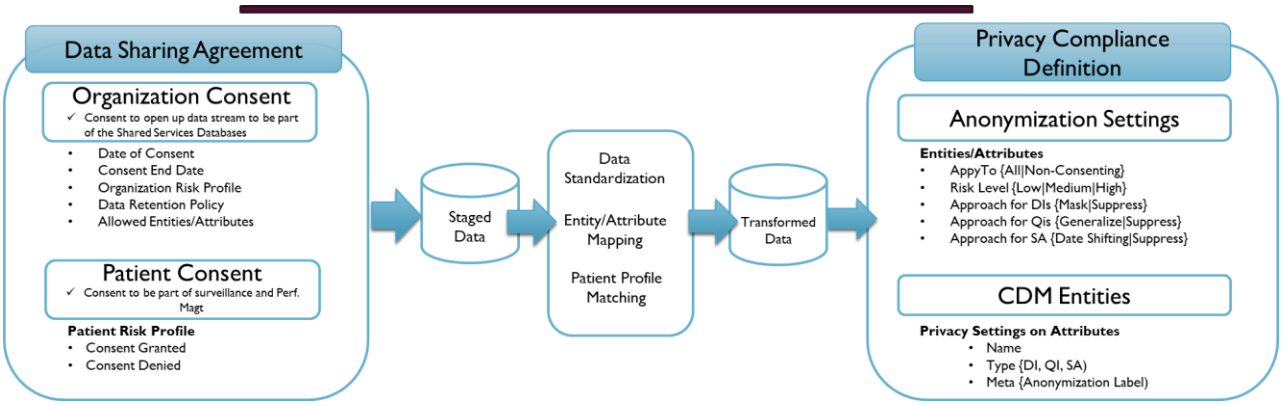


Figure 4-11 Privacy Compliance Data Flow

Patient Consent can be explicitly listed or pulled from a participating Community care organization database instance. This simply indicates whether consent is granted or denied for surveillance and performance management. The context of a Patient Consent is fully described within the Privacy Compliance Definition Document (Figure 4-12). Patient consent determines if a patient data is excluded completely or anonymized for downstream performance management services. It could result in the patient identifying records being stripped from incoming data streams or trigger anonymization of the patient complete or partial profile data.

If Patient Consent definition is not found for a patient, the default consent for patients in the Privacy Compliance Definition Document is applied. This could either be a "***Consent***

Granted", "***Consent Denied"*** or "***Anonymized"***. Patient consent applied at this point has a local significance to each Community care organization data. This is because some patients may choose to deny access to performance management services on some of the community care services they receive while granting consent to others. For example, deny consent to sharing their mental health data but allowing their nursing and other therapy data to be used for performance management. The framework can be configured to exclude non-consenting patients at this point or allow their data through but have them anonymized before it hits the CDM.

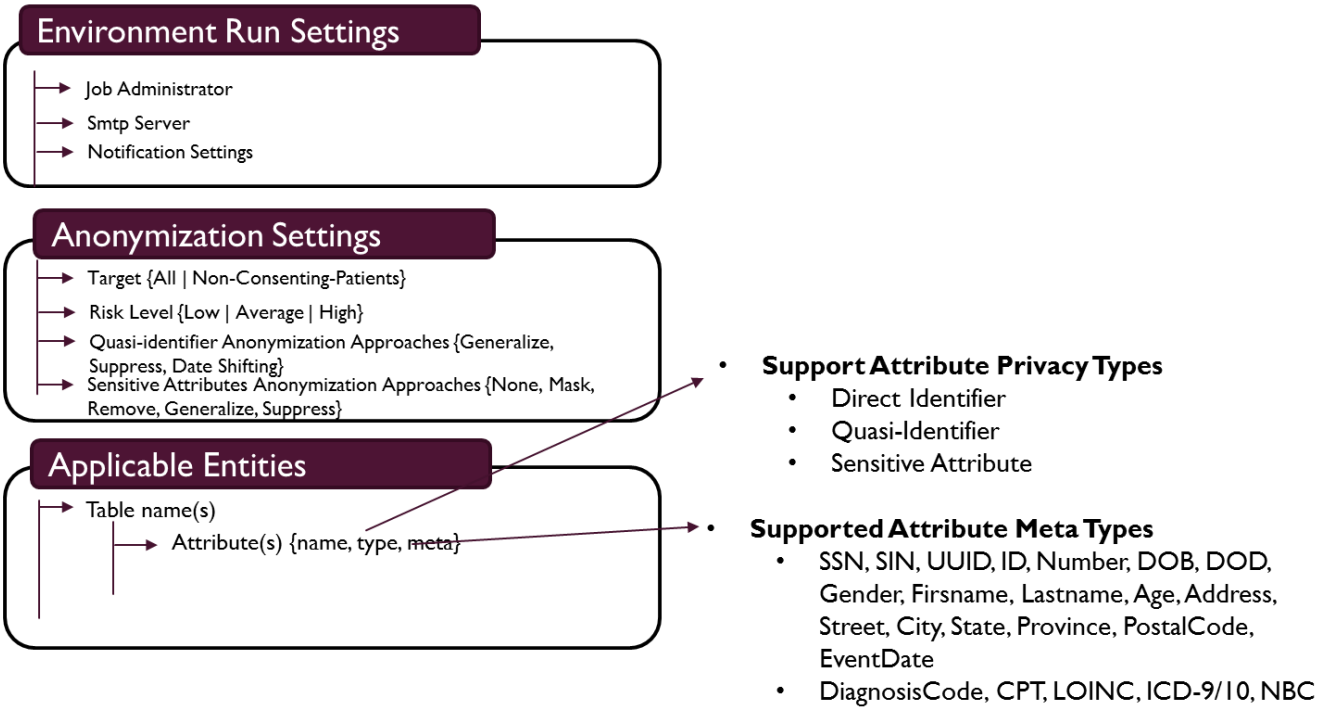


Figure 4-12 Privacy Compliance Definition Components Depicted

The Privacy Compliance Definition Document applies to transformed data. It has two major sections – Anonymization settings and CDM Entities definitions (Figure 4-11). The anonymization settings provide details on patient profiles to target, risk levels, approaches to anonymizing the direct identifiers, quasi-identifiers, and sensitive attributes. The CDM Entities definitions classify each attribute that requires anonymization based on their privacy type and

other details such as the levels of generalization for quasi-identifiers. Please refer to section 6.2.4 for more details on the process.

In line with this model, our architecture ensures that various levels of anonymization can be applied to data as needed to protect patient privacy and confidentiality as well as to ensure that only users within the patient circle of care or with the proper authorization have access to data reports built from the common data model.

Anonymization setting is set up for the entire infrastructure. By centralizing the anonymization settings, we ensure that patient consent is applied consistently in the context of their profiles across all services that they participate in across the health region, therefore ensuring the global protection of patient data across the entire infrastructure.

4.5.2 Privacy Compliance Definition Document

The Privacy Compliance Definition Document is the anonymization configuration for the CDM. Figure 4-12 shows the structure of this document. While it is XML-based, it does not follow a formal language and does not have a formal semantics. The ***Environment Run Settings*** are used to capture privacy compliance job management settings. It also captures the job admin email, the SMTP server settings for relaying anonymization result summary, as well as notification settings. These ***Anonymization Settings*** controls the behaviour of the anonymization service within the cloud infrastructure.

The ***Applicable Entities*** definitions are used to determine the tables/attributes defined in the CDM that should be considered for various anonymization processes.

Anonymization processing depends on a complex set of definitions that must be based on the expected risk associated with the data recipient. In the proposed infrastructure, there is no

public release of data. Nevertheless, anonymization is necessary to adhere to patient consent while maintaining high analytical utility for the analytics generated from the CDM.

The anonymization settings are described in Table 4-2 below. The default behaviour is to apply anonymization to all incoming data from patients that refused to consent to data sharing. However, it could also be applied to the entire data set when needed, say to release data to an external partner that is not one of the participating community care organizations. One important contribution of this thesis to privacy compliance is that anonymization is not always applied to the entire data set but to select patient profiles as stipulated by the Patient Consent forms. But the risk measurement that determines the level of anonymization to apply to these select patient profiles takes into account the data distribution of all patient profiles in the CDM.

Table 4-2 Anonymization Setting

Setting	Operation	Details
apply_to		
	All	Applies to the entire dataset in the CDM
	Non-consenting-patients	Applies to only the patients that refused to consent to data sharing
Risk level		
	High	Assume a high risk of re-identification (k=10)
	Average	Assume an average risk of re-identification (k=5)
	Low	Assume a low risk of re-identification (k=2)
Approaches		
	Mask	Applies to DIs only. Masking is done based on the identifier type meta attached to the attribute
	Generalize	Applies to QIs and SAs only. Generalization is based on the type of attribute.
	Suppress	Applies to QIs and SAs only. Suppression is applied to Qis and SAs after generalization where applicable.
	Date Shifting	Applies to QI event dates. This process ensures that date QI values associated are shifted to a period that ensures their anonymity.

Irrespective of the risk level, direct identifiers are always masked on anonymization of a patient profile. Quasi-identifier risk mitigation is carried out using a combination of generalization/suppression for a k-anonymity value determined by the risk level setting. If the risk level associated with the infrastructure is considered average, a k-anonymity value of 5 is applied. If it is high, then a k-anonymity value of 10 is applied. These thresholds must be set as configurable options for the anonymization service. Quasi-identifiers generalization options are set based on the attribute type as identified by the meta attribute definition. Where applicable, the anonymization engine could apply algorithms such as OLA (El Emam et al., 2009) to choose the right generalization for each quasi-identifier. Date Shifting is used to anonymize patient event dates like nursing visits, activity visits, assessment dates while preserving the inter-event time intervals (Liu et al., 2009).

4.5.3 Anonymization Service

The anonymization service addresses the most critical privacy compliance requirements by ensuring that PHI and sensitive data never makes it to unintended data recipients. The analytics and reporting services usually provide an aggregate view of data, but if the granularity is low on quasi-identifiers and sensitive attributes, then anonymization must be carried out on the results. The anonymization service is also triggered to run on out-going datasets for the reporting and subscription services.

The anonymization service addresses four types of anonymization – data masking for direct identifiers, generalization with local recoding for quasi-identifiers, data suppression for quasi-identifiers, and data shifting for patient profiles.

1. **Data Masking** – This simply replaces each direct identifier with a fake but realistic looking identifier from the applicable resource library gazetteer. This resource library

includes gazetteers for first names, last names, addresses, and postal codes, cities, countries, diagnosis. Additional gazetteers can be added to the resource library as needed.

2. **Generalization** – This process uses defined generalization hierarchies for quasi-identifier including postal-code, date of birth or age, height, weight, and blood pressure to ensure that equivalence classes from these attributes in an outbound dataset provide anonymity for all patients. Local recoding ensures that for each equivalence class, various levels of generalization is applied, making large equivalence classes less generalized than smaller ones within the same data set. If needed for consistent reporting, generalized attributes could be returned in the original forms by selecting a random value within the range of the generalization. For example, a random age within an age bracket, a random date of birth within the year of birth, or a random postal code within an area.
3. **Suppression** – This applies to both quasi-identifiers and sensitive attributes. Suppression ensures that attributes of a patient that are too identifying, such as mental health status, HIV diagnosis, addiction profiles, etc. are automatically deleted from an anonymized dataset. Suppression can be explicit or determined after generalization. When it is explicit, identifying attributes are hidden from other community care organizations except for the data owners and the data custodian in all circumstances. When done as part of anonymization, then it is used to manage the level of generalization in making a dataset anonymous. In this second scenario, if exposing an attribute value would result in data for an equivalence class getting over-generalized, then

it would be suppressed to preserve the integrity and analytics utility of the affected equivalence classes.

4. **Data Shifting** – This ensures that longitudinal event dates are not correlated by an adversary to re-identity patients. For those within the patient circle of care, this last level of anonymization is never required since they need to correlate patient services by the dates they occurred. However, for the rest of the data recipients, date shifting is used to move all patient events or profile dates to a meaningful, realistic, but a different time period to throw off an adversary. Usually, data shifting is never used on its own. Rather, it should be used in conjunction with generalization and suppression.

The anonymization service is configured using the data recipient risk profiles associated with each Community care organization as defined in the PCDD. Since the community care organizations all use cloud-hosted applications and databases, the risk profiles for data releases is considered low. For external stakeholders, risk profiles are dependent on the result of a security assessment of the organization IT infrastructure security as well as their data management practices. Those that score low in the security assessment will have a high-risk profile. The higher the risk profile, the higher the level of anonymization and vice versa. Risk profile is zero for the data owner or the data custodian.

This is illustrated in Figure 4-13. If the data recipient is also a data owner of the target dataset, then anonymization is not required. On the other hand, if the data recipient is part of a patient circle of care, then only service-level data and direct identifiers associated with the Community care organization should be exposed. For external stakeholders, all identifiers should be anonymized for patient-level data releases. Therefore, only aggregate summary data should be released to the stakeholders outside a patient circle of care.

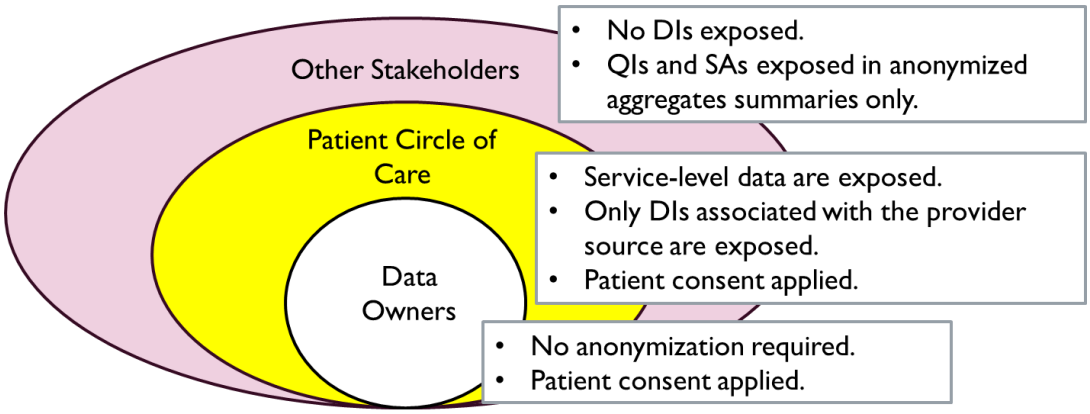


Figure 4-13 Anonymization vs. Data Recipient Category

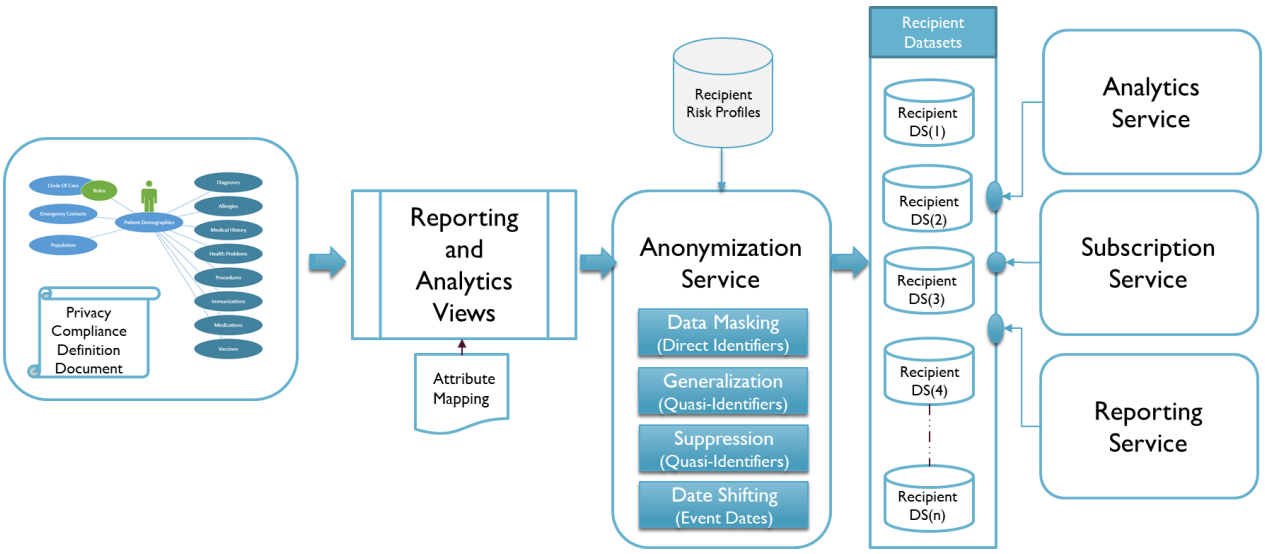


Figure 4-14 Privacy Compliance Workflow for Performance Management Services

The role of the anonymization services within the performance management services infrastructure is depicted in Figure 4-14. For each report and analytics request, a data view from the CDM is created. This is then associated with an attribute mapping that links the output of this report to the attributes of the CDM. If a computed field is added to the view, it must be annotated with its anonymization type and other needed details.

When a data request is made to the CDM for any of the published views, the anonymization service looks up the risk profiles for each target recipient. Following the workflow in Figure 4-15, it then triggers anonymization for each unique recipient profile. The target anonymized data sets are then published but also maintained by the anonymization service every time the CDM is refreshed.

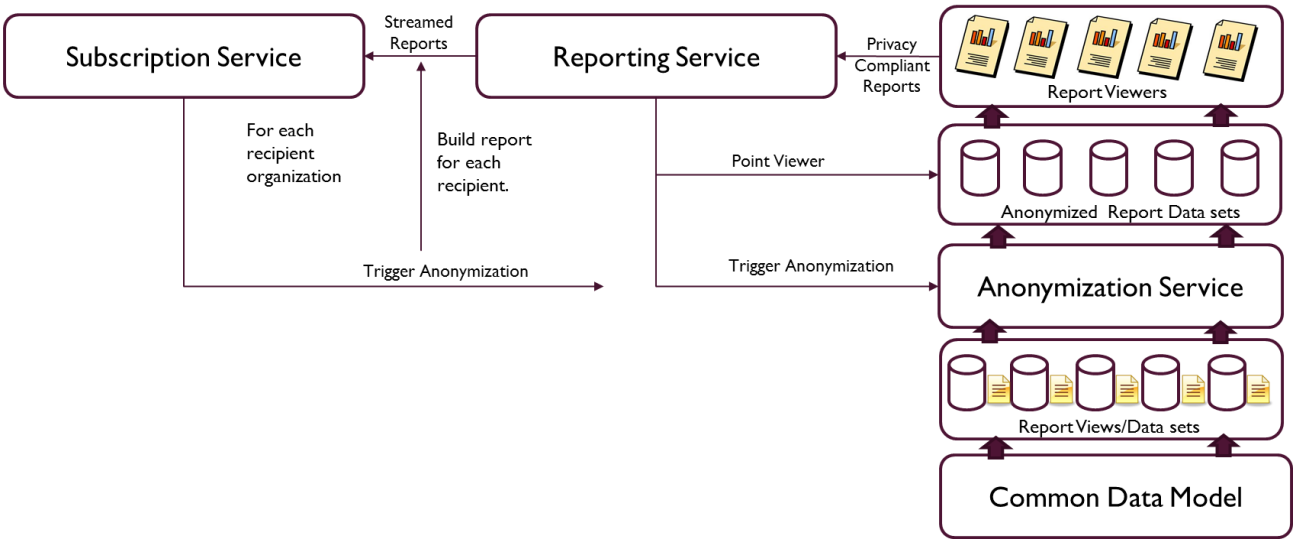


Figure 4-15 Subscription Service Interaction with Anonymization and Reporting Services

As shown in Figure 4-7, when a subscription is triggered, the Subscription Service first looks up data recipient organizations, pull the report view associated with each subscription report and calls the anonymization service to anonymize the data sets. After this process completes, the target anonymized data sets are then used to trigger the reporting service to create the reports for each data recipient. These reports are then streamed back to the subscription service for delivery to the target recipients.

4.6. Chapter Summary

In this chapter, we proposed a surveillance and performance management architecture that addresses the gaps identified in Chapter 3. The regional health authority provides a multi-

tenanted cloud infrastructure for systematically hosting collaborating Community care organizations' Patient Management Information Systems, and operational databases. Leveraging cloud PaaS containerization through the Systematic Data Hosting Service enables this architecture to adapt to the dynamic nature of healthcare data and scale to support various data formats and volumes. This tremendously reduces the cost of funding the implementation and maintenance of on-premises server infrastructure to host these applications for each participating organization.

Further, the Systematic Data Collection and Patient Identity Matching services aggregate and populate patient profile and services data across all the community care organizations to a CDM which is subsequently used to feed the performance management services. The Patient Compliance Model uses the Anonymization Service to address Organization and Patient Consents, thereby ensuring that data for Performance Management Services meet the required risk profile for each data recipient.

In the next three chapters, we will present a pilot project using our architecture, an experiment on addressing Privacy Compliance through anonymization, and an experiment on a Configurable Patient Identity Matching Service.

Chapter 5. Pilot Project – Surveillance and Performance Management of Community Healthcare

In this Chapter, we present a pilot project that was used to validate the initial version of our Surveillance and Performance Management architecture. The feedback from the project helped us develop the final version described in chapter 4. The pilot project was completed in collaboration with a Regional Health Authority (RHA) that provides and coordinates home care amongst 54 community care organizations in the region surrounding the greater metropolitan area of Ottawa, Canada. The RHA is known as the Champlain Local Health Integration Network (Champlain LHIN). This pilot project is a combination of Design Science Research (DSR) and Action Research (AR) and corresponds to AR/DSR Iteration 2, as shown in Figure 1-3 of section 1.4 Research Methodology. We participated in the pilot project to implement a cloud computing infrastructure to better support performance management across these Community care organizations (CCOs). These CCOs provide community support services to the Champlain region. This chapter describes this pilot project in terms of our architecture defined in Chapter 4.

5.1. Champlain Local Health Integration Network

The RHA in our cases study is the Champlain Local Health Integration Network. The mandate of the RHA is to help people in the region to 1) live independently at home; 2) apply and receive day programs for supportive housing or assisted living; 3) apply and receive long-term care; and 4) to provide palliative care to patients nearing the end of their lives. On average, the RHA has about 60,000 active patients annually that receive over two dozen community care

services from the RHA and the 54 community care organizations that it funds to provide healthcare services in the Champlain region.

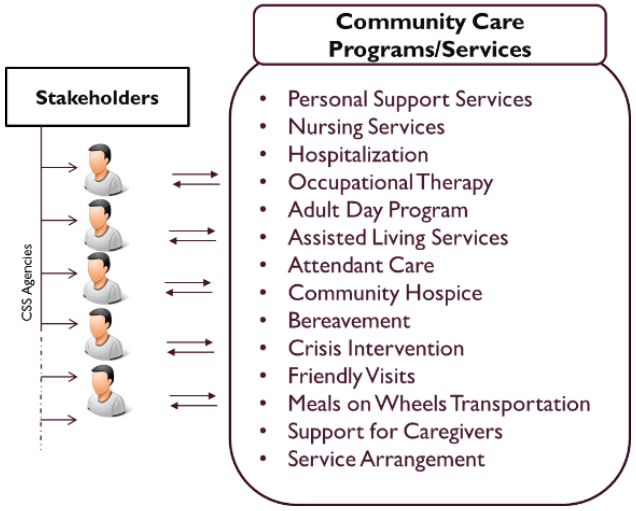


Figure 5-1 Spectrum of Community Care Services

Figure 5-1 shows a sample of the care services provided by these community care organizations. As depicted in Figure 5-2, before the pilot project, there was minimal interoperability and limited ability to do performance management. Data collection was typically in the form of ad-hoc organization-specific invoices (typically in Excel format) for services rendered that were submitted by email from financial officers of the community care organizations. RHA financial officers manually processed these invoices and maintained a database of summary information extracted from the invoices.

RHA business intelligence officers had a limited ability to analyze performance and respond to ad-hoc requests. All reports were manually created, often after intensive request specific data collection. Each Community care organization and RHA had their own data silo resulting in service duplication and very little coordination of care delivery efforts and consistency in the quality of care provided.

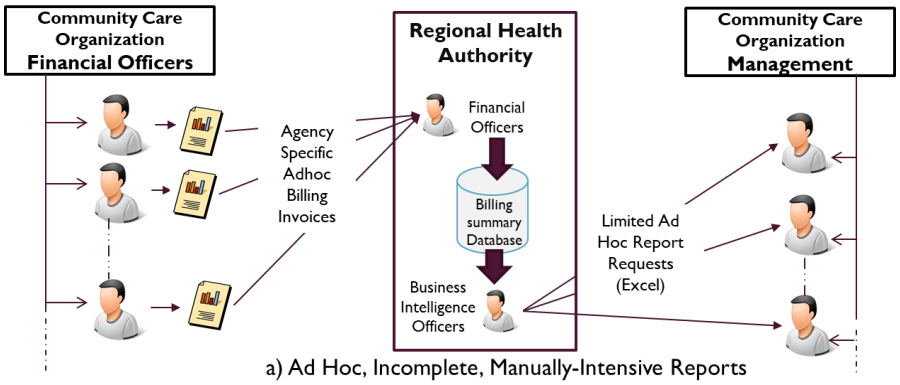


Figure 5-2 Ad Hoc Performance Management

Also, government-issued patient identifiers were not being used consistently for patient identification across these community care organizations. Except for the RHA, where patient health card numbers are validated using an external government Web service, the health card number entries from the 54 community care organizations are prone to data entry errors. The inability to have consistent patient identifiers across these community care organizations and the RHA limited the ability to perform population-level analysis for the health region.

5.2. Cloud Computing Infrastructure

Figure 5-3 shows the multi-tenanted private cloud infrastructure for supporting the surveillance services (data hosting, data collection, and a common data model) and performance management services (reporting and subscription services) that were implemented in the pilot project to support performance management.

The RHA funded a private cloud which provides both compute and storage capabilities for the community care organizations. The multi-tenanted private cloud infrastructure hosts each Community care organization internal patient management application, the operational database, as well as data integration and performance management infrastructure.

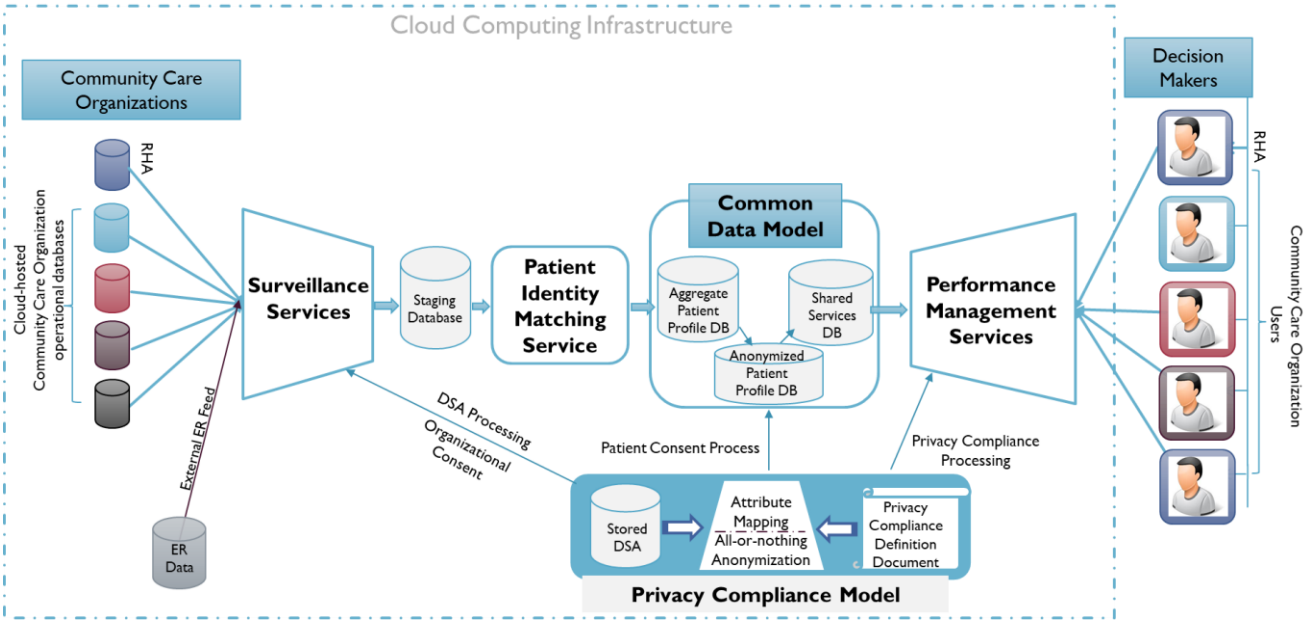


Figure 5-3 Cloud computing infrastructure for the RHA/Community care organization Pilot Project

The cloud computing infrastructure is an implementation of the general architecture presented in section 4.1. However, it allows external data streams from the RHA and the local ER in the Champlain region. Also, the Privacy Compliance Model is based on an anonymization service that uses the all-or-nothing approach to privacy compliance.

5.2.1 Surveillance Services

a) Systematic Data Hosting Service

Community care organization legacy platforms are migrated to the RHA funded cloud. Each organization data hosting is carried out following the workflow described in section 4.2.1a. Additionally, Administrators create migration scripts to move Community care organization data from the previous platform to a new cloud-hosted database instance. Afterward, user acceptance testing is done before the Community care organization goes into production.

For this pilot project, cloud Provisioner described in Chapter 4 is not used to manage the private cloud infrastructure. Rather, infrastructure administrators created VMs as needed for each role as required for surveillance and performance management services setup and deployment. Each organization's patient management application is hosted in a cloud-hosted Virtual Machine while using a cloud-hosted Microsoft SQL Server database cluster to manage the instances of each community care organization local database as they migrate to the cloud infrastructure. Users from each organization are created on the cloud domain Active Directory (AD) and subsequently given access to their VMs. These users access their respective application instances through a (Virtual Private Network) VPN Client.

b) Systematic Data Collection Service

In this pilot project, the Systematic Data Collection Service pulls data from two categories of data sources (see Table 5-1 and Figure 5-4). The RHA and Community care organization databases are structured data sources. The pilot project also included Emergency Room (ER) data from local hospitals as semi-structured data sets pulled from a remote Web service. The Systematic Data Collection Service supports heterogeneous data sources through data integration PaaS containers customized for these different data sources.

Table 5-1: Systematic Data Collection with disparate data sources

Data Source	Type	Hosting
Community Care Organization Data	Structured data (SQL Server)	Systematically hosted within the private cloud infrastructure.
RHA Data	Structured data (SQL Server)	Hosted outside the Community care organization private cloud infrastructure. Special network configurations were set up to allow the systematic data collection service to access this database.
ER Data	XML	ER Coordinated Care Documents (CCPs) are systematically collected nightly and staged in the Staging Databases.

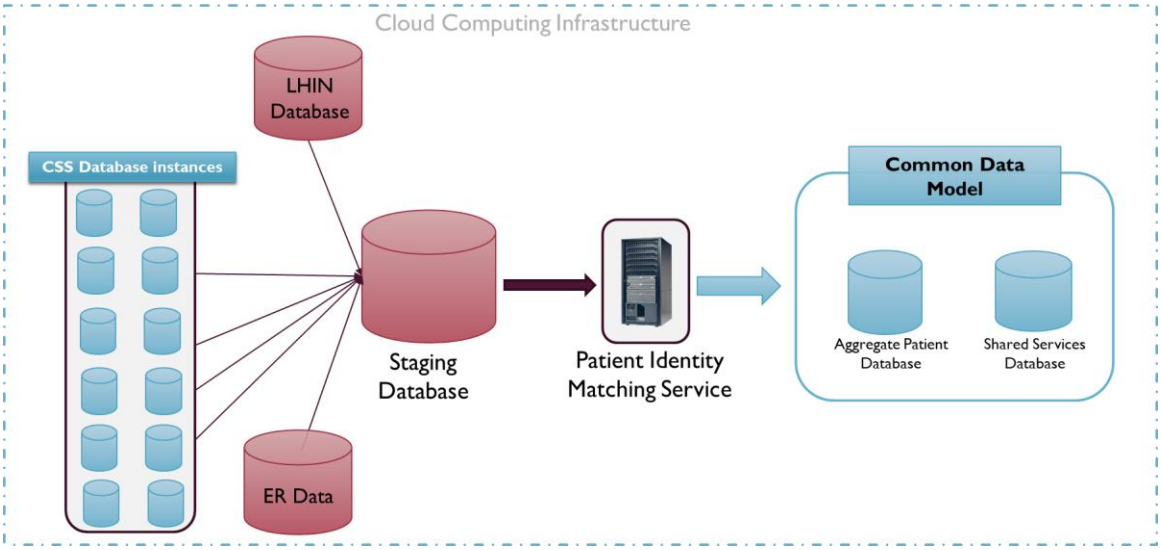


Figure 5-4 Systematic Data Collection Service for Pilot Project Depicted

Data Collection

Data collection is set up for all sources. It is controlled by the status of each Community care organization' DSA. Data from those community care organizations that opted out are automatically eliminated from the Staging Database and any downstream data integration for the performance management services.

Systematic data collection for the community care organizations is provided through the data collection service that connects to the MS SQL Server cluster and systematically collects and streams data from each database instance belonging to a Community care organization that has signed the DSA to the Staging Database.

Data Collection Service leverages sections of the PCDD (See Appendix A). For this Pilot project, the PCDD contains the "*RunSettings*" section that controls the run-time behavior of the data collection nodes. The "*OrganizationConsent*" section contains the list of the cloud-hosted database instances belonging each of the CCOs and the table definitions that would be imported

to the CDM. The "*PatientConsent*" section contains the reference definitions for determining if patient consent is "granted" or "denied" and the consent date.

The "*Databases*" section of the organization consent definition (see Figure 5-5) contains the databases definitions for each Community care organization. Each entry has "active" flag that indicates if the DSA is signed. The "Table" definition provides the details for the database objects that would be migrated to the Shared Services database. These settings include flags for ignoring entities or fields from data migration and details for fields that need to be anonymized before being allowed into the performance management infrastructure. In this pilot project, all operational databases share the same table definitions file for surveillance services.

```
<?xml version="1.0" encoding="utf-8"?>
<Databases>
  <Database server="agency_db_cluster" target_database="AGENCY_A_DB_PRODUCTION" agency_code="10001" active="true" />
  <Database server="agency_db_cluster" target_database="AGENCY_B_DB_PRODUCTION" agency_code="1002" active="false" />
  <Database server="agency_db_cluster" target_database="AGENCY_C_DB_PRODUCTION" agency_code="1003" active="true" />
  <Database server="agency_db_cluster" target_database="AGENCY_D_DB_PRODUCTION" agency_code="1004" active="false" />
</Databases>
```

Figure 5-5 Data Collection Service Definition

After all Community care organization data is streamed to the Staging Database, Patient Identity Matching Service is then used to populate the Aggregate Patient Database. Each data source must also go through the Patient Identity Matching Service to ensure the records for the same patient across all data sources are pulled together into a single aggregate patient profile. Subsequently, after the identity matching, the Systematic Data Collection Service kicks in one more time to use this data to populate the Share Services Database with aggregate service-level data from the staged datasets. All data migration apply transformation rules associated with each organization table definition. While all community care organizations' applications share the

same schema, data quality across these operational databases vary widely. So each staged database must go through various forms of standardization to ensure that migrated data elements are semantically the same as the target CDM Shared Services Database.

RHA Data Migration

Transformation functions are used for basic transformation definitions to be carried out by the ETL processor within the data collection service. Sometimes, certain fields need to be seeded into the imported data. For example, a source field that acts as a record identifier for each new record in a staging table. Those are supported by the "Seed" definition. The "Set" function maps value to a CDM field from the source. It can either be assigned an explicit value or gets mapped from a source field.

```
<?xml version="1.0" encoding="utf-8"?>
<Mapping>
  <Map source="stg_Adhoc_Clients" target="Patient">
    <Set field="Source" value="LHIN" />
    <Set field="Patient Identifier" from="Client_Number" />
    <Set field="Surname" from="Surname" />
    <Set field="Firstname" from="Firstname" />
    <Set field="Othernames" from="Preferred_Name" />
    <Set field="Birth Data" from="DOB" />
    <Set field="Gender" from="" "Gender">
      <Replace search="Male"  replace="M" />
      <Replace search="Female" replace="F" />
      <Replace search="Transexual" replace="TS" />
    </Set>
    <Set fields="Health Card Number" from="HCN">
      <Replace search="00000*|11111*|99999*,''" replace="" />
    </Set>
    <Set field="Citizenship" from="Nationality" />
    <Set field="Service Language" from="Service_Language" />
    <Set field="First Language" from="First Language" />
  </Map>
</Mapping>
```

Figure 5-6 Data Translation Mapping for RHA Data Feed for Patient Demographics

Replace function allows for basic search/replace transformation and supports SQL supported wildcards such as '*' (substitutes for zero or more characters) and "?" (substitutes for a single character). It also allows for a "|" separated list of search words for a single replacement.

External ER Data Migration

ER Data requires a systematic approach that declaratively maps this data to the CDM. The data collection service incorporates an XML data mapping and translation utility that uses declarative definitions to enables the data migration service to create or map data attributes to structured table fields in a relational database. This XML data is then applied to the mapping definitions and is processed and fed into a staging relational database.

The mapping process starts with a definition that links attribute within the XML data with attributes of various entities of the CDM. For each applicable table in the CDM, the definition provides a "Table" definition that links to a node or node list in the ER data. Data for each field is populated using a match field definition or attribute from the model.

```
<DocumentMap>
  <VersionControl schema="s1" />
  <Section>
    <Table name="Patient" path="patient/personalDetails">
      <Field name="Source" value="s1" />
      <Field name="Patient Identifier" path="Surname" isKey="true" />
      <Field name="Firstname" path="First" />
      <Field name="Birth Date" path="DOB" />
      <Field name="Gender" path="Gender" />
      <Field name="Ethnicity" path="Race"  />
      <Field name="Birth Place" path="Place of Birth" />
      <Field name="Nationality" path="Nationality" />
      <Field name="Consent" path="Patient_Consent" />
      <!-- Mapping for patient contact fields -->
      <Table name="Patient Contact Details" path="patient/personalDetails">
        <Field name="Home Phone Identifier" path="HomePhone" />
        <Field name="Mobile Phone" path="Mobile" />
      </Table>
      <!-- Mapping for patient addresses -->
      <Table name="Patient Addresses" path="patient/Addresses">
        <Field name="AddressId" value="repeat-index" />
        <Field name="Address" path="Street" />
        <Field name="City" path="City" />
        <Field name="Postal Code" path="PostalCode" />
        <Field name="Province" path="Province" />
      </Table>
    </Table>
  </Section>
</DocumentMap>
```

Figure 5-7 Simplified Data Translation Mapping for ER Data Feed Patient Demographics

This process that executes the data translation mapping generates a relational database schema that automatically persists the data from semi-structured data sources to structured tables. The full mapping definitions and mapping rules are described in detail in Appendix B.

5.2.2 Performance Management Services

For this pilot project, two performance management services were leveraged – reporting and subscription services. All performance management reports are made available to only those community care organizations that have signed the DSA. There was no anonymization service as an all-or-nothing approach was used in which patients either consented to the use of their data or they did not. If they did not, the data was not included.

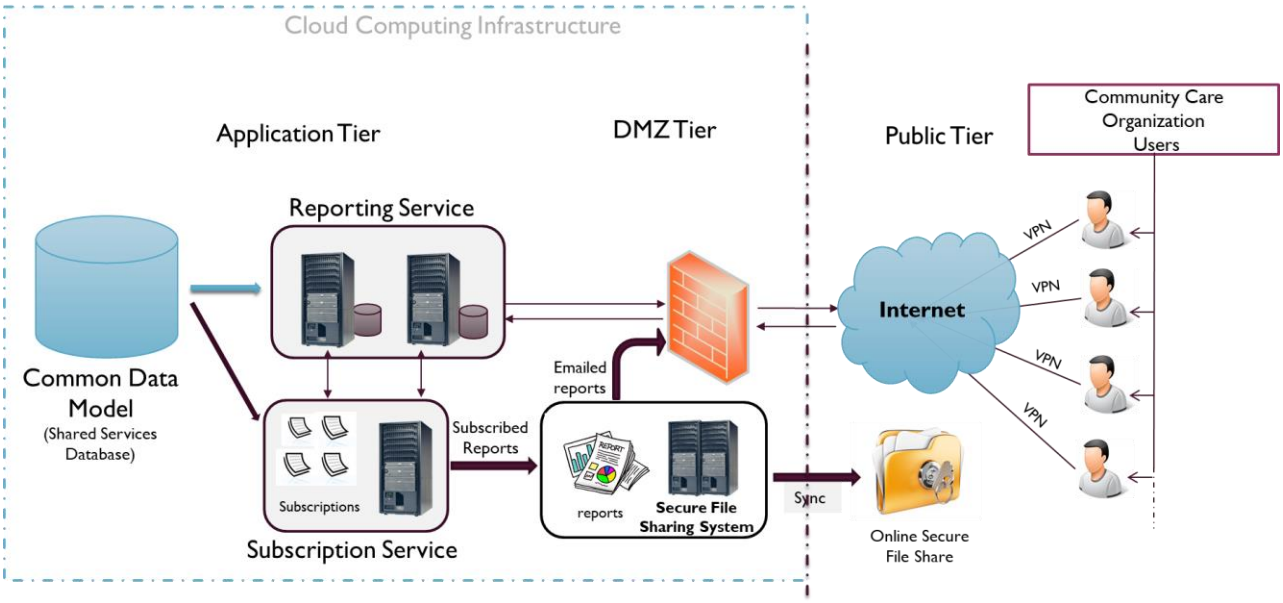


Figure 5-8 Pilot Project Performance Management Services

Figure 5-8 depicts the interaction of the performance management services with the CDM Shared Services Database. The reporting services use two MS SQL Server Reporting Services

(SSRS) that mirror each other. Each analytics report created is published on both servers. Personnel from the RHA and the community care organizations use the reporting portal to access the standard reports made available to them.

For this pilot project, the Subscription Service interacts with the report services APIs to process subscriptions and dynamically build reports in various formats (refer to 4.2.2 (b) for more details). For each report developed and published, a custom report definition file that allows the subscription service to stream the report in many formats is also published on the subscription server.

```
<?xml version="1.0" encoding="utf-8"?>
<Subscription
  name="BPMH Reminders"
  enabled="true"
  owner = "Benjamin.eze@sample.email.com"
>
  <Schedule repeat="WeekDays" period="morning" />
  <Subscribers type="query">
        SELECT Distinct [Team], [AssessedBy]
        FROM sso.BPMH_Reminders
        ORDER BY Team, [AssessedBy] DESC
  </Subscribers>
  <Reports>
    <Report reference="BPMH_Reminder_Client_List_All.xml" format="html">
      <Parameters>
        <Parameter name="Team" type="list">
          @Team
        </Parameter>
        <Parameter name="AssessedBy" type="list">
          @AssessedBy
        </Parameter>
      </Parameters>
    </Report>
  </Reports>
  <Actions>
    <Action>
      <Email subscriber_email="E-mail">
        <Schedule repeat="WeekDays" period="morning" start="2015-09-06" />
        <Subject>BPMH Reminders for @report_date</Subject>
        <Message>
          <span style='font-size:10.0pt;font-family:"Verdana","sans-serif";'>
            Hi,
            <p>
              Please find attached report showing clients that have overdue BPMH after recent assessments.
            </p>
            @signature
          </span>
        </Message>
      </Email>
    </Action>
  </Actions>
</Subscription>
```

Figure 5-9 Sample Subscription definition

Figure 5-9 shows a sample subscription used to send Best Possible Medication History (BPMH) reminder to RHA Care Coordinators. This subscription generates the list of subscribers

dynamically using a SQL query to the Common Model. Resulting record set is subsequently used as parameters for streaming the BPMH report in HTML format for each subscriber entry.

A subscription can either be delivered to the subscribers directly by email or through a secure file sharing service. For email delivery to subscribers, custom reports for each subscriber are packaged as attachments using the email template in the subscription definition. For the secure file share type delivery, the report is generated locally within the subscription service. Based on the target recipient, it is then transferred to the target local share for the target recipient. This local share is set up to automatically sync with an online remote file share that each Community care organization can access using a unique security key. Therefore, while the reports in this share are automatically refreshed and updated based on the subscription, the file synchronization system, updates the remote share with the latest files. Community care organization users can download the file-sharing software and map the remote share to a local folder or simply go online to a management portal similar to Dropbox and OneDrive to access and download the latest versions of the reports.

5.3. Common Data Model

The common data model is implemented as a database. Data from each organization needs to be mapped into this database where applicable. Data collected is structured into 3 hierarchical levels as outlined in section 4.3.

5.4. Patient Identity Matching Service

In this pilot project, the following assumptions were made in designing this service:

1. All patients within the RHA database have valid health card numbers (HCNs) in the profiles. Since the RHA validates HCNs through a validation Web Service, they can be trusted as accurate. Therefore, a match by the HCN attribute is considered a deterministic match.
2. Community care organization databases have unique local patient identifiers for each patient. In matching patient profiles, duplicate entries within the same database, where they exist, also need to be identified.
3. Match results are kept across data collections runs in the Shared Services database. Fresh entries are added as needed.
4. The matching algorithm needs to incorporate a feedback mechanism for addressing and resolving ambiguities.

5.4.1 Matching Component

Matching was implemented using static rules derived from studying the characteristics of the data sets. Matching rules are defined in the following two hierarchies:

A) Patient Matching Rules

- **Level 1**: Matches based on HCN and Last name. (10 pts)
- **Level 2**: Matches based on the Last name, First name[1], DOB, and Gender[2] (10pts)
- **Level 3**: Matches based on the Last name, 1st two characters of the First name, DOB, and Gender (5pts)
- **Level 4**: Where Gender field does not exist, match patient based on Last name, 1st two characters of the First name, and Year of Birth (YOB) (2pts)
- **Level 5**: Where ambiguities exist – Use the patient contact matches to resolve those.

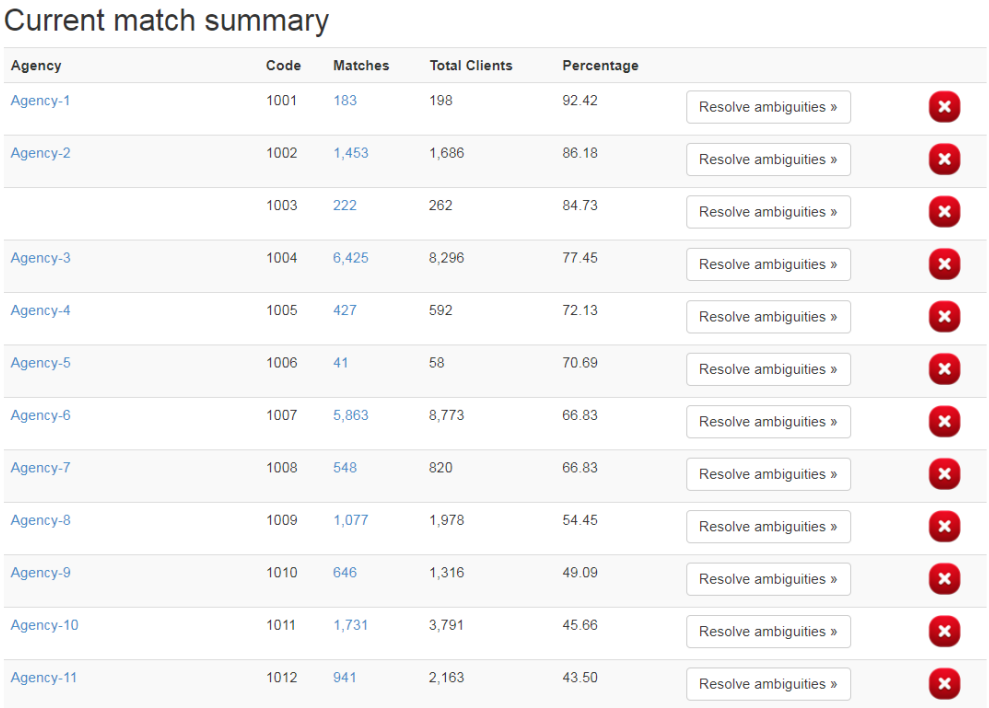
[1] The accuracy for First names is very low since variants of the names could be recorded if a government issued ID is not used for the original data capture.
[2] Gender data is not always clean since they are not selected from a dropdown but typed. Transformation is necessary.

B) Patient Contact Matching Rules

- **Level 1**: Matches based on Last name, First name, DOB, and Gender (10pts)
- **Level 2**: Matches based on the Last name, 1st two characters of the First name, YOB, and Gender (5pts)
- **Level 3**: Matches based on Last name, Postal Code and Contact Number (5pts)

A match is valid if the cumulative weight across all levels is at least 10pts. As such, a match on either of levels 1 and 2 is enough to match the two different patient profiles to the same person. However, levels 3 and 4 cumulatively will not provide enough weight and would require patient contact weights to confirm the matches. Where the patient contacts do not exist, the matches would be left in an ambiguous status. For example, three patients share the same last names, but there are typos in the first name, and the DOB of one of them is also entered incorrectly. In this scenario, they will end up with a cumulative match weight of 2.

The match results are monitored through a web application using a summary page similar to the one below in Figure 5-10.



Current match summary

Agency	Code	Matches	Total Clients	Percentage		
Agency-1	1001	183	198	92.42	Resolve ambiguities »	
Agency-2	1002	1,453	1,686	86.18	Resolve ambiguities »	
	1003	222	262	84.73	Resolve ambiguities »	
Agency-3	1004	6,425	8,296	77.45	Resolve ambiguities »	
Agency-4	1005	427	592	72.13	Resolve ambiguities »	
Agency-5	1006	41	58	70.69	Resolve ambiguities »	
Agency-6	1007	5,863	8,773	66.83	Resolve ambiguities »	
Agency-7	1008	548	820	66.83	Resolve ambiguities »	
Agency-8	1009	1,077	1,978	54.45	Resolve ambiguities »	
Agency-9	1010	646	1,316	49.09	Resolve ambiguities »	
Agency-10	1011	1,731	3,791	45.66	Resolve ambiguities »	
Agency-11	1012	941	2,163	43.50	Resolve ambiguities »	

Figure 5-10 Match Summary Page

Historical match results can be reviewed using a daily trend report (Figure 5-11).

Match History

Match Date	Matches	Total Clients	Percentage
4/15/2018	21,606	248,632	8.69
4/14/2018	21,606	248,632	8.69
4/13/2018	21,606	248,632	8.69
4/12/2018	21,593	248,632	8.68
4/11/2018	21,580	248,632	8.68
4/10/2018	21,565	248,632	8.67
4/9/2018	21,543	248,632	8.66
4/8/2018	21,532	248,632	8.66
4/7/2018	21,532	248,632	8.66
4/6/2018	21,532	248,632	8.66
4/5/2018	21,524	248,632	8.66
4/4/2018	21,501	248,632	8.65
4/3/2018	20,810	248,632	8.37

Figure 5-11 Continuous Match Summary

It must be noted however, that this summary is seen from the context of the RHA comparing the matches from the community care organizations that signed DSA against their entire patient profile – current and historical. Therefore, the 8.60% average match percentage for the RHA is an average of 65% matches for the participating community care organizations.

5.4.2 Dealing with Ambiguous Matches

Ambiguities are identified and cumulated after each batch processing. Ambiguities are determined by comparing matched profiles from various community care organizations and those of the RHA. Since the RHA profile is validated, it is taken as the source of truth.

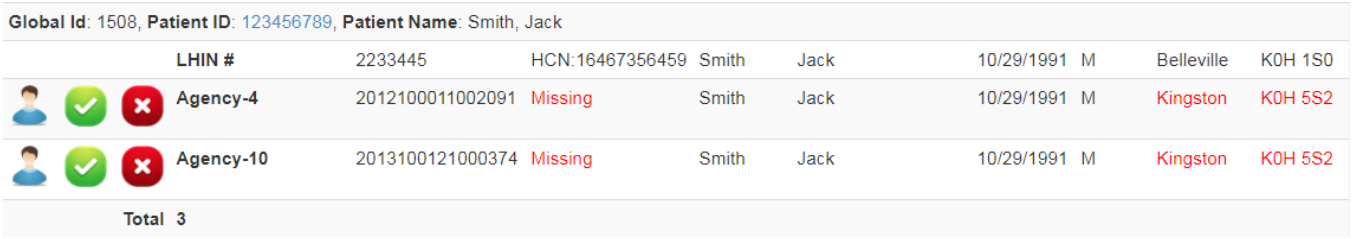


Figure 5-12 Sample Ambiguous Profile

Each patient profile attribute is compared with the RHA profile - with missing and wrongly entered values highlighted in red, so it is easy for a reviewer to accept or reject each match. If the matches are rejected, the rejection is captured, so the ambiguity is not repeated with subsequent matches. If a match is approved, the profile is then associated with the global id for the match group. Subsequent matches only handle new or unmatched patient profiles. This ensures that global identifiers for matched profiles remain unchanged across match runs.

5.5. Privacy Compliance Model

For this pilot project, an all-or-nothing type of privacy compliance was implemented. It means that data is either imported or dropped. For example, if an Organization Consent specifies that certain programs are sensitive to share with other community care organizations, all data elements associated with this program across their database is dropped. The same goes with a Patient that failed to consent to data sharing. Also, their entire profile data is dropped across all entities in the incoming data stream.

The data collection services leverage the Privacy Compliance Definition Document associated with each organization (See Appendix A). Organization Consent provides the details of the consent definition for each CCO. Each data entity is automatically imported with all its

fields. Exceptions can be made to drop specific fields such as login details, passwords, etc. Explicit anonymization can be applied at the source to drop sensitive fields that these organizations wouldn't want to share, such as administrative and internal accounting data sets. Changes to this document automatically affect the behaviour of the data collection, patient matching, reporting, and analytics services.

The Privacy Compliance Definition Document can provide a type reference on data elements and fields collected from each data source. Each organization can specify fields that need to be ***nulled***, ***removed***, ***masked, or anonymized,*** so their data never makes it to the common data model. An important consequence of this level of consent is that it is reciprocal. A participating Community care organization is unable to access data from other participating community care organizations that it is uncomfortable sharing within its own data stream. For example, failure to share patient demographic details means that reports and analytics or data subscriptions that return data on patient demographic information from other participating community care organizations are explicitly denied from the requesting organization.

All stakeholders that signed the DSA, including the RHA, have access to all aggregate reports. For patient-level reports, each Community care organization can see PHI on their patients as well as data on services received from all other organizations within the patient circle of care. The RHA is also able to see patient-level reports across all organizations except for those patients that opted out of data sharing.

While the all-or-nothing approach provides good privacy protection, it is not optimal because it does not incorporate anonymization processes. The consequence is huge data losses to

the CDM after addressing Organization and Patient Consents. In Chapter 6, we use an experiment to discuss how the anonymization service can be leveraged to address privacy compliance and reduce data losses in addressing Privacy Compliance.

5.6. Results

- The pilot project successfully demonstrated our architecture using a private cloud infrastructure for data collection, hosting, aggregation, and support performance management of community care organizations in a health region. The estimated implemented cost is about $32k per Community care organization and a projected annual operating cost of $6.5k per Community care organization.
- Cloud infrastructure set up at the RHA took about 6 months. The setup and deployment of the Patient identity matching service and some of the initial performance management reports took about 4 months.
- The Common Data Model has about 120 relational tables, 250k patients with over 95k patients from the community care organizations, and relational data on over two dozen services.
- 48 of the 54 community care organizations with over 135,000 patients migrated to the cloud-hosted platform. 17 community care organizations with about 28,000 patients signed the DSA and participated in the performance management infrastructure.
- There are up to 8 active report subscriptions set up for the RHA and community care organization contacts that publish and emails various reports. These reports provide ER admission notifications, Patient aggregate profile changes, eReferral data, and

Match Profile reports for showing patient aggregate services across all community care organizations.

- Identity matching results show that at least 19,500 patients have matches from another Community care organization or the RHA. 1,948 of these patients (4%) are ambiguous, while only 43(0.1%) did not consent to data sharing. Only 10 matches have so far been identified as wrong.
- The pilot project implementation uses an all-or-nothing approach to Privacy Compliance, resulting in the removal of about 18,500 patient records, a significant data loss to Performance management services. To put this in perspective, these patient records were excluded because of denied patient consents and withdrawn organization consent on sensitive services that resulted in the exclusion of the associated patients. In Chapter 6, we will show how anonymization allows us to anonymize these records, so they are available for performance management.

Chapter 6. Experiment on Configurable Anonymization Algorithm for Privacy Compliance

In this chapter, we present an experiment that focuses on operationalizing privacy compliance of cloud-hosted data using DSAs in support of performance management of community healthcare. In Chapter 5, privacy compliance was implemented using an "all-or-nothing" approach to privacy that eliminates data from non-consenting patients from downstream performance management services. This experiment is a new Design Science Research (DSR) iteration and is designed to show that a Privacy Compliance Definition Document combined with DSAs can be very instrumental in operationalizing privacy compliance for collaborating community care organizations. It corresponds to DSR Iteration 3, as shown in Figure 1-3 of section 1.4 Research Methodology.

This experiment provides various processing workflows that leverage anonymization methodologies to enable privacy-compliant data publishing data and performance management results for collaborating organizations in the context of data surveillance and performance management in a cloud computing environment.

6.1. Architecture used in Experiment

The configurable anonymization experiment leverages the cloud-based architecture described in section 4.2 and enhances the cloud-based infrastructure for the pilot project (see section 5.2). Figure 6-1 depicts the cloud infrastructure and the various components of the architecture leveraged in this experiment. The focus of this experiment is to extend the overall architecture to support anonymization in addressing Organization and Patient Contents.

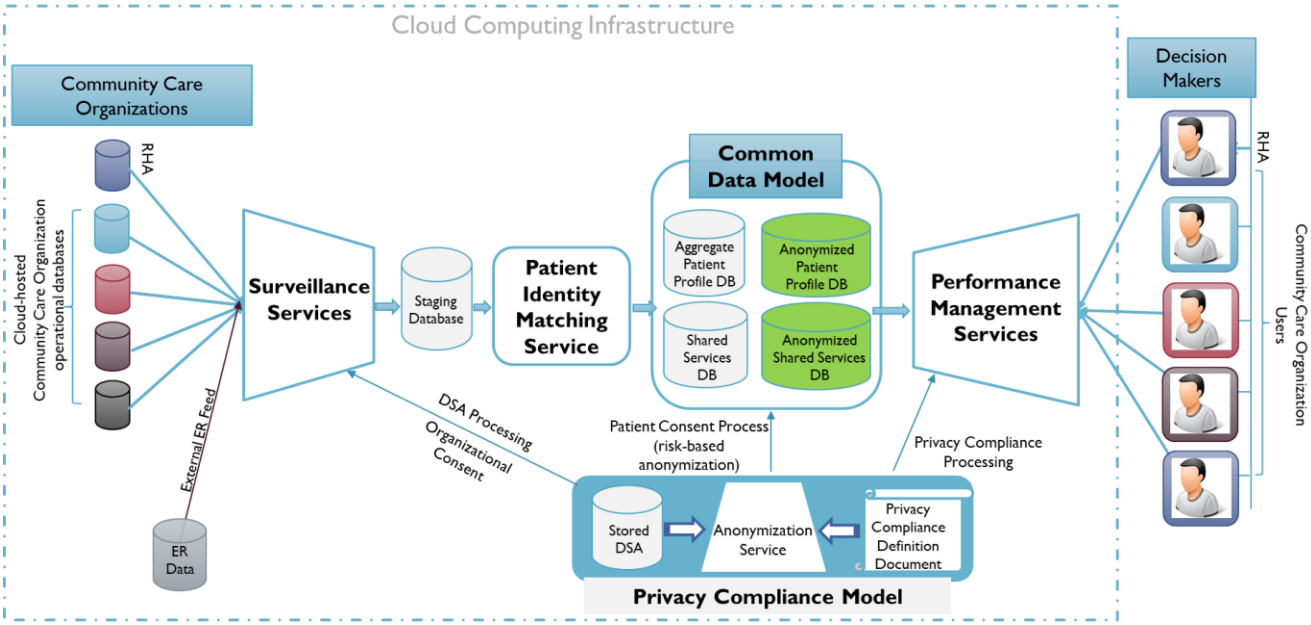


Figure 6-1 Cloud-based Infrastructure for operationalizing privacy compliance using DSAs.

6.2. Configurable Anonymization Algorithm for Privacy Compliance

This experiment is designed to help us depict the use of DSAs and Privacy Compliance Definition documents for populating a CDM. The objective is to show how DSAs can be used to operationalize privacy compliance for a cloud-hosted surveillance and performance management infrastructure by leveraging selective anonymization based on both Organization and Patient Consents. We also show how one can achieve anonymization on longitudinal and high-dimensional data sets that are common in the healthcare domain. These approaches enable a cloud-computing infrastructure to configure processes and services, including anonymization, to ensure privacy compliance and a systematic approach to data governance. The algorithm presented in this section is developed specifically to address anonymization of high dimensional datasets for performance management reporting, analytics, and subscriptions. It assumes that data recipients from performance management services could have different risk profiles and therefore, would receive anonymized data that fits their specific risk profile.

The Common Data Model is described graphically in Figure 6-2. Patient demographic details are at the core of this definition.

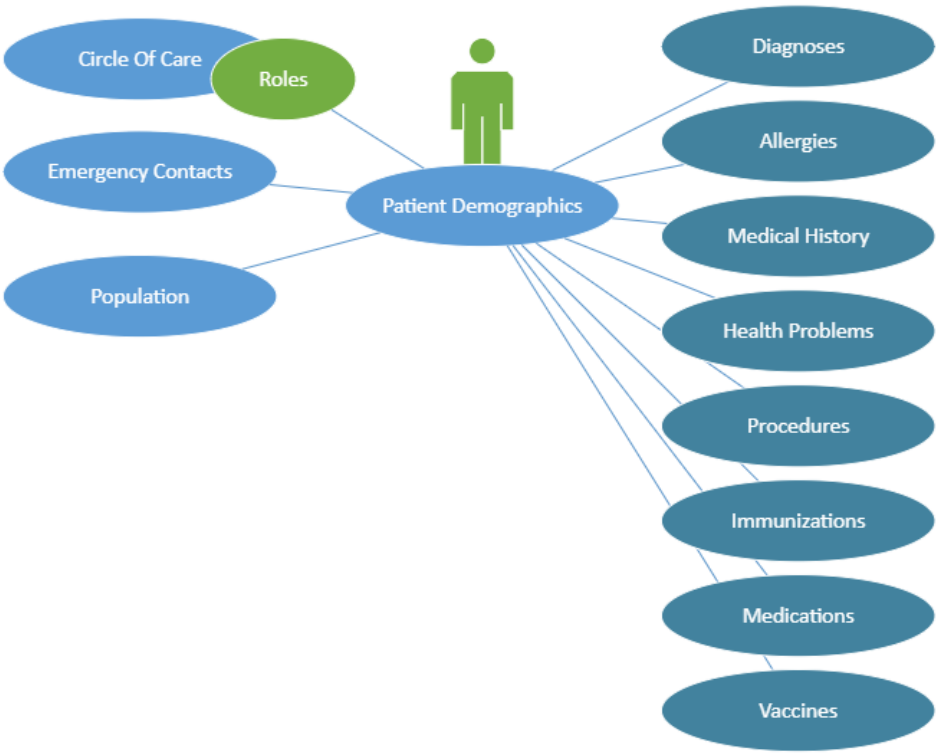


Figure 6-2 Common Data Model for an aggregate patient PHR database

On the left are entities that contain attributes that categorize patients in the health system. These include patient population, the patient circle of care definitions and roles for caregivers, and patient emergency contacts. On the right are the entities that describe patient health problems, treatment history, interventions, and other treatments. Table 6-1 summaries the target entities of the CDM from data streams each data source.

In this experiment, all CDM data entities are longitudinal, including patient demographic details. Each patient profile from the data sources is a new record in the appropriate CDM entity.

Each record is identified by either the global patient identifier, the local identifier, or the data source. They also have additional meta-data such as the date the record is created and last modified.

Table 6-1 Target Entities Data source update entities in the Common Data Model.

Data Source	Type	Target Entities in the CDM
RHA	Structured data (SQL Server)	Patient Demographics, Circle of care, Emergency Contacts, Patient Population, Services, Diagnoses, Allergies, Health Problems, Allergies.
ER Feed	XML	Patient Demographics, Circle of care, Diagnoses, Allergies, Procedures, Immunizations, Medications, Vaccinations.
Community care organizations	Structured data (SQL Server)	Patient Demographics, Circle of care, Emergency Contacts, Diagnoses, Allergies, Health Problems, Allergies.

Data from the CDM is highly confidential. It is not to be exposed to downstream performance management services without anonymization.

6.2.1 Privacy Compliance Model

The Privacy Compliance Definition Document is the anonymization configuration for the common data model. This definition applies to the data in the aggregate patient profile database, the outcome of the patient identity matching service, as well as the Shared Services database. Figure 6-3 depicts a simplified sample of this document – showing only the patient and patient address entities. The ***<RunSettings>…</RunSettings>*** show the job management settings. It sets the job admin email, the SMTP server, and notification settings. The ***<AnonymizationSettings>…</AnonymizationSettings>*** definition controls the behaviour of the anonymization service within the cloud infrastructure.

Anonymization approaches depend on the ***<Entities>…</Entities>*** definitions to determine the tables/attributes defined in the CDM that should be considered for various anonymization processes. Anonymization processing depends on a complex set of definitions that must be based on the expected risk associated with the data recipient. In this experiment, there is no public release of data, so the release of the entire CDM is very unlikely. Nevertheless, anonymization is needed to enforce patient consent while maintaining a high analytical utility of the target data set.

```
<?xml version="1.0" encoding="UTF-8"?>
<PrivacyComplianceDefinition>
   <RunSettings
    job_admin="etljobadmins@rha.ca"
    smtp_server="198.22.0.3"
    send_notification="onerror"
    set_defaults="false" />
   <PatientConsent
    reference="consentTable.Status" consent_values="Consent Received"
    denied_consent_values="[NULL]"  consent_date="consentTable.Consent_Data" />
   <OrganizationConsent>
      <Databases>
         <Database server="source_cluster" target_database="db1" org_code="1111" active="true" />
         <Database server="source_cluster" target_database="db2" org_code="1112" active="true" />
         <Database server="source_cluster" target_database="db3" org_code="1113" active="false" />
      </Databases>
      <Tables />
   </OrganizationConsent>
   <AnonymizationSettings apply_to="non-consenting-patients">
      <Risk level="average" />
      <DirectIdentifiers approach="Mask" />
      <QuasiIdentifiers approach="Generalize,Suppress" />
      <SensitiveAttributes approach="none" />
      <Entities>
         <Entity name="Patients">
            <PrivacySettings>
               <Attribute name="Health Card Number" type="direct-identifier" meta="identifier" />
               <Attribute name="Patient Identifier" type="direct-identifier" meta="identifier" />
               <Attribute name="Surname" type="direct-identifier" meta="lastname" />
               <Attribute name="Firstname" type="direct-identifier" meta="firstname" />
               <Attribute name="Birth Date" type="quasi-identifier" meta="dateofbirth" />
               <Attribute name="Gender" type="quasi-identifier" meta="gender" />
               <Attribute name="Death Date" type="quasi-identifier" meta="dateofdeath" />
               <Attribute name="Ethnicity" type="quasi-identifier" meta="race" />
               <Attribute name="Occupation" type="quasi-identifier" meta="occupation" />
               <Attribute name="Nationality" type="quasi-identifier" meta="" />
            </PrivacySettings>
         </Entity>
         <Entity name="PatientAddress">
            <PrivacySettings>
               <Attribute name="Address" type="direct-identifier" meta="AddressCanada" />
               <Attribute name="Postal_Code" type="quasi-identifier" meta="PostalCodeCanada" />
               <Attribute name="City" type="quasi-identifier" meta="cityCanada" />
               <Attribute name="Country" type="quasi-identifier" meta="country" />
               <Attribute name="Start_Date" type="quasi-identifier" meta="eventdate" />
            </PrivacySettings>
         </Entity>
      </Entities>
   </AnonymizationSettings>
</PrivacyComplianceDefinition>
```

Figure 6-3 Section of a Privacy Compliance Definition Document for a Common Data Model

Attributes that make up each entity are decorated with privacy type definitions and a "meta" flag that describes their privacy tags. While privacy type definitions identify each attribute as either a direct-identifiers, quasi-identifiers, sensitive attributes or non-identifier, privacy tags associate these attributes with the type of anonymization that should be applied to them on exposure to an untrusted party. Based on the privacy tag definition, the anonymization service will automatically apply the type of masking, pseudonymization, or generalization applicable to the attribute.

Most anonymization approaches apply the same type of anonymization to shared data. Our approach does not apply any anonymization to the CDM. Rather, the anonymization service creates recipient targeted data views from the CDM based on the intended data recipient while adhering strictly to the rules described in Figure 4-13. Anonymization is then applied to the data views, and the result moved to an anonymized table that is exposed to downstream performance management services – reporting, or subscription.

For this experiment, as described in Figure 4-13, a data recipient is categorized as 1) a data owner (if a target dataset has patients from the organization associated with the data recipient), 2) patient circle of care (if the recipient organization is providing services to the patient) or 3) other stakeholders (if the recipient organization is not part of the patient circle of care and is not one of the data providers). Aggregate data needs to be anonymized as well. However, since most aggregate data show summaries across the health region for a cohort of the population, anonymization is needed to protect the privacy of the patients within each of these cohorts.

6.2.2 Anonymization Consideration for Direct-Identifiers

Direct identifiers are masked based on their privacy tags. The anonymization service includes a rich library of gazetteers for data masking. Gazetteers are a list of names with grouping

details as needed for data masking. With these gazetteers, masked values though fake, look very similar to the originals in structure and format. As shown in Table 6-2, each gazetteer is either based on built-in types or resource hosted CSV file that is associated with a privacy tag.

Table 6-2 Gazetteers definitions within the resource library

Gazetteers	Privacy Tags	Type	Values
Gender	Gender	Built-in	Male, Female, Transsexual, Others
Blood-Groups	Blood-group	Built-in	A, B, AB, 0
Marital-Status	Marital-Status	Built-in	Single, Married, Divorced, Separated, Widowed
Countries	Country	Resource	Countries.csv
Languages	Language	Resource	Languages.csv
Firstnames	Firstname	Resource	Firstnames.csv
Surnames	Lastname	Resource	Surnames.csv
Cities	City	Resource	Cities.csv
Allergies	Allergy	Resource	Allergies.csv
Ethnicities	Ethnicity	Resource	Ethnicities.csv
Religions	Religion	Resource	Religions.csv
Vaccines	Vaccine	Resource	Vaccines.csv

Identifiers are masked using a format preserving pseudonymization process. Basically, the identifier is replaced with a similarly looking pseudonym of the original data.

6.2.3 Anonymization Consideration for Quasi-Identifier and Sensitive Attributes

In anonymizing quasi-identifiers, they are matched to a level-based profile categorization that associates them either directly with the patient, a care episode, or a care event. Since the CDM data is inherently longitudinal, traditional anonymization approaches that work for cross-sectional datasets fails in this scenario. Therefore, anonymization is done using a hierarchical anchor structure, as described in Figure 6-4.

Entities at level 1 anchor solely on the patient. Those in level 2 anchor around an episodic period identified by a start and end date. Those in level 3 are longitudinal events to episodic periods but anchor around an event date. This is better illustrated by a narrative.

Bob Richie is a 65 years old construction worker that lives in Napean, Ottawa. Until the last decade of his life, he was healthy and had regular annual visits with his family doctor. At 55, John started suffering from chronic back pain that required him to seek help from the RHA. He was admitted into the community care program and was subsequently offered and received some physiotherapy sessions for a 3 months period.

Eventually, he recovered fully and went back to work. After a few months, his pain returned, so he visited the ER for treatment. He was admitted and eventually had back surgery. To facilitate his recovery, the ER once again referred John to the RHA for more physiotherapy treatments. The RHA also contacted a couple of community care organizations to support John with some Occupational Therapy sessions. In addition, John also received some Assisted Living Services to help him with some modifications to his home to help with his mobility issues.

Because Bob lived alone, he also received some friendly visits from volunteers in his area, coordinated through a Community care organization in the local area. He continues to receive personal support services and a few therapy sessions from the RHA as he recuperates.

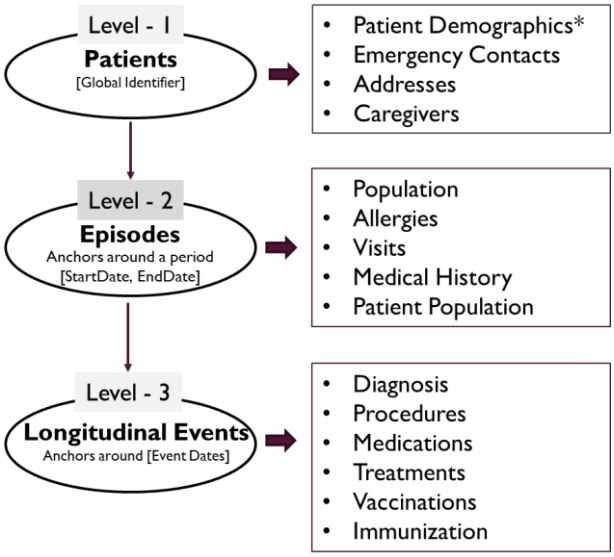


Figure 6-4 Categorization of the Common Data Model Entities to facilitate anonymization

In the scenario described in the narrative, Bob will have a least 5 profiles in the CDM. One or more from the RHA, 1 ER record, and 3 Community care organization records. Clearly, there are 3 service episodes. The first resulted in him receiving some physiotherapy sessions. The

second episode resulted in him visiting the ER again, and received more physiotherapy treatments, occupation therapy, and assisted living services. He also had back surgery.

Bob's case, as illustrated in Figure 6-5 is typical of many patients in the health system. Care episodes can last a single day, a week, months, and even years. The frequency of care events is very dependent on the age and health status of the patient.

Grouping quasi-identifier dates around profiles, episodes, and event dates (see Figure 6-5), ensure that anonymization produces results that respect these natural groups and preserve important analytical measures and event sequence associated with the dataset.

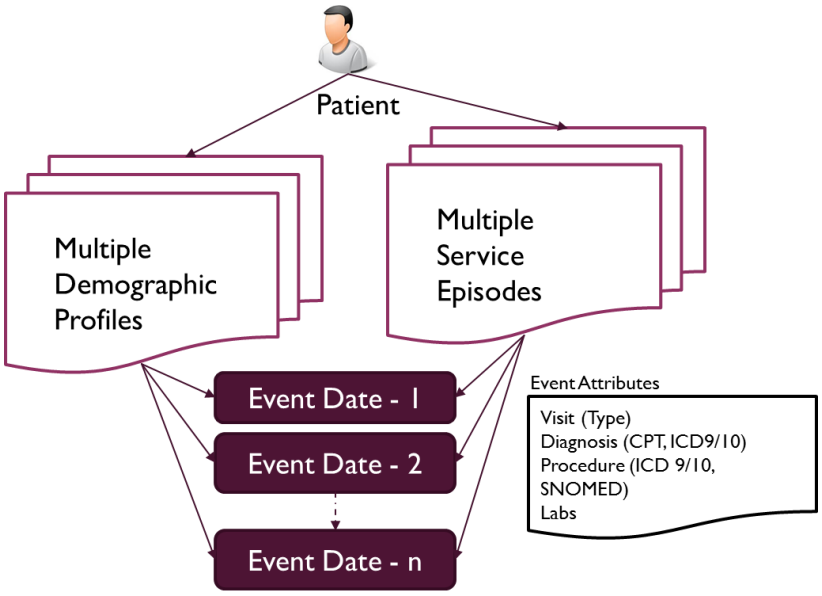


Figure 6-5 Grouping Quasi Identifiers around profiles, episodes and event dates

6.2.4 Anonymization Processing Workflow

The anonymization workflow typically starts with a data request from either the reporting or subscription service to the CDM. The first step is to identify the dataset view for the target recipients for anonymization. To anonymize this data set, the ETL processor pulls together the

CDM attributes from the view attribute mapping. As depicted in Figure 6-6, this first process categorizes the attributes by their anonymization types. Direct identifiers are pulled together into a separate subset of the target recipient dataset for masking. Quasi-identifiers and Sensitive Attributes are separated into a separate subset for risk-based anonymization.

The dataset recipient risk profiles are then looked up. For each unique recipient profile, the QI and SA in the data set are then anonymized to meet the risk profile for that category of intended data recipient, if anonymization is required. Afterward, the masked DI data set is then stitched together with each data recipient anonymized data set to produce the final fully anonymized data set.

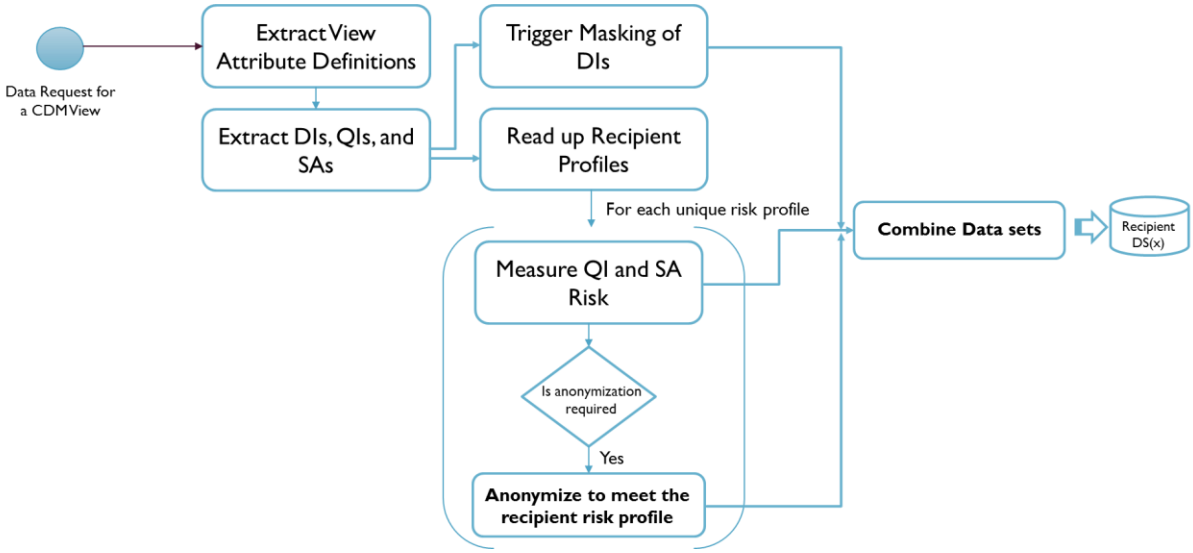


Figure 6-6 Anonymization workflow

This workflow is better illustrated with a real-world example:

community care organizations need to know when their patients have been admitted to the ER. There are many expensive administrative hurdles that must be followed if caregivers should show up at a patient's residence and not find their patient at home for booked appointments. Since they couldn't just go away, they would need to call the police to determine what happened to the patient and it would usually take hours and significant resources to determine the status of the client - a very expensive process.

As a result, the community care organizations have asked that a daily report to be sent to them on their patients that were admitted to the ER. This report should have the following fields:

- Date report is generated.
- Patient Local Identifier
- Surname
- Firstname
- DOB
- Hospital Name
- ER Visit Date
- ER Event Description
- Complaint Presented
- Any addition information

An evaluation of the data request shows that each Community care organization is allowed to receive Circle of Care data from the ER feed on their patients only. Our approach allows for the sharing of service details to data recipients in the patient's circle of care. The attribute mapping for the daily ER notification report is illustrated in Figure 6-7. The report pulls data from level – 1 patient demographics entities, addresses, and level 2 visit details. Level-1 data contain the following Direct Identifiers (Patient Number, Surname, Firstname, Gender, and Address). The level-1 anchor is the patient number (global identifier), and the level-2 episode is the

"ER Visit Date", a period of one day. Quasi-identifiers are "DOB", Gender, and "ER Visit Date". Other visit attributes including "Hospital Name", "ER Event Description", "Complaint Presented", and "Any Additional Information" are considered Sensitive Attributes.

If we assume 14 unique recipient profiles, one for each participating Community care organization. Since report recipients are Community care organization contacts, and the report view is filtered to show only data from the recipient Community care organization. Our CDM reporting view is then mapped to each recipient profile to create 14 recipient data sets. Each record within each recipient data set is then evaluated for risk against the target recipient profile. Therefore, since each recipient risk profile matches their target data set in the "Owner" category, the target data is not at risk and will not be anonymized for the Community care organization data recipients at the demographic levels. Since ER data comes from the Visit tables, they are considered service-level details, and since the recipients are in the circle of care of the patient, anonymization is also not required if the patient consented to data sharing.

For this experiment, re-identification risk is dependent on the data recipient. A recipient that is associated with any of the collaborating service organization is considered part of a patient circle of care if the patient is actively receiving community care services from this service organization. For these patients, service-level data from other data feeds can be shared with the exception of cases where the patient explicitly denies consent to data sharing through the Patient Content definition. For example, say a patient went to one of the community care organizations for a sensitive service such as addiction counseling, domestic abuse, or treatment for a mental health condition. If the patient also explicitly denies consent to data sharing for such services, then the

service level data for this patient is marked at risk for any caregiver outside of the custodian service organization. Therefore, this data must either be excluded from such reports or must be fully anonymized if patient-level data needs to be included.

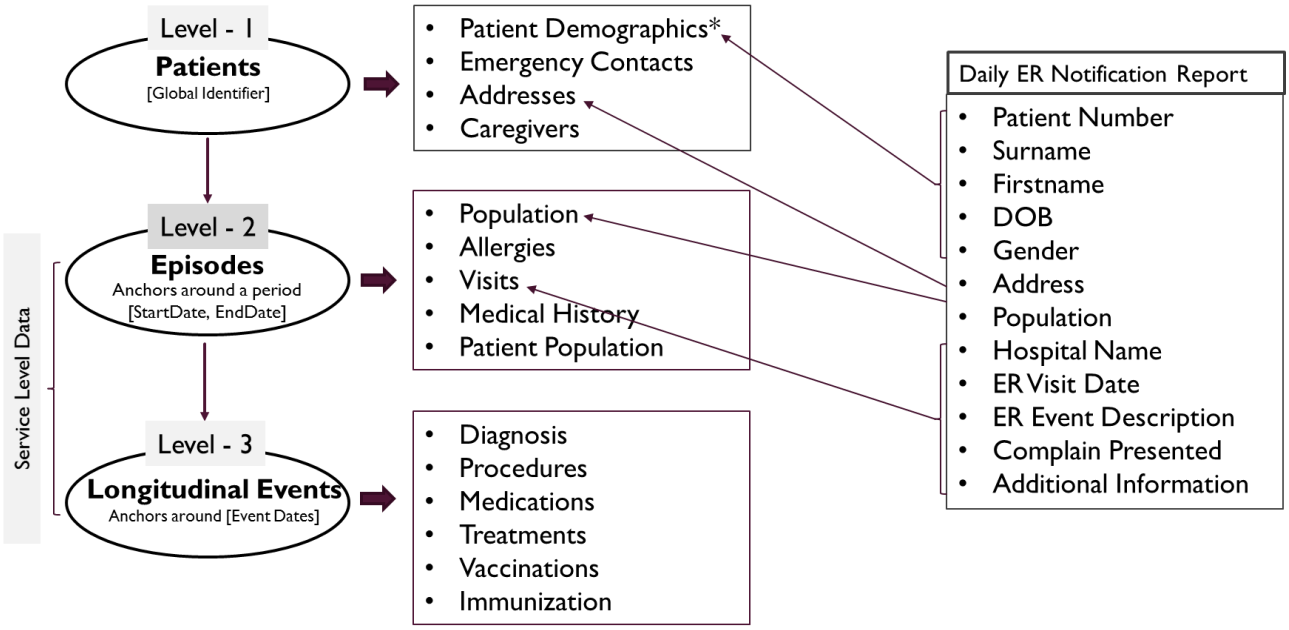


Figure 6-7 Attribute mapping illustration for the daily ER notification report

If only aggregate data is to be provided and the recipient is outside the collaborating community care organizations, the aggregate data must be checked for at-risk cohorts, and those cohorts belonging to at-risk patients must be anonymized before the target dataset can be released to the target recipient.

6.2.5 Re-identification Risk Measurement

In this experiment, re-identification risk is dependent on the data recipient. Table 6-3 depicts this process clearly. The anonymization service creates multiple datasets from the reporting view for each data recipient category. If these datasets have patient-level data, then the patient

consent flag needs to be included in the data set. If the data is simply an aggregate data set, patient consent flag will only apply to at-risk patient cohorts.

These recipient data sets need to be evaluated for re-identification risk while respecting patient consent to the data. In this experiment, Prosecutor risk is addressed on target targets. Other types of risks, such as journalist risk are not applicable since it is assumed that the data recipient knows if a patient is part of the resulting dataset or not.

Table 6-3 Applicable risks for each recipient type category

Type of data set	Risk Type	Recipient Type	Risk Threshold
Patient/Episode/Event level data sets.	Prosecutor	• Owner • Circle of Care • Other Stakeholders	• 1 (No anonymization applied) • 3 (Only applies to patients outside the CoC) • 5 (All patients)
Aggregate data	Prosecutor	• Owner • Circle of Care • Other Stakeholders	• 1 (No anonymization) • 1 (No anonymization) • 5 (All patients)

Risk measurement processing steps include:

1. **Patient-level Risk Measurement**: This is done on the patient-level quasi-identifiers (QI) values, including DOB, Gender, and Postal Code. Patient-level data is considered longitudinal as data for the patient comes from multiple data sources. Patient-level quasi-identifier aggregation is done across all available profiles. If re-identification risk at this level is LOW for the target recipient, then the data set proceeds to the episodic and eventually the event level risk measurements. If the patient-level risk measurement result is HIGH, then the episodic and event-level data are automatically considered at risk.

2. **Episodic/Event-level Risk Measurement**: Episode and event-level data are usually at risk since it is almost impossible to have two patients with a similar pattern of behaviour. Therefore, episodic data needs to be anonymized for high-risk recipients.
3. **Aggregate Reporting Data**: Aggregate target data for report generation tend to hide patients within cohorts. However, the aggregate weights can still reveal at-risk patients if the data set is not anonymized. For example, for a population breakdown of patients by region, groups with at-risk patients with a support that is less 5 are at-risk of re-identification and must be excluded from these reports if not anonymized.

6.3. Discussion and Other Considerations

6.3.1 Risk-based Anonymization Workflows

Risk-based anonymization is carried out on quasi-identifiers only. Direct identifiers are always masked for target data recipients outside a patient CoC. For risk-based anonymization, quasi-identifiers in the recipient data sets need to be identified and categorized according to its associated levels (levels 1 -3). Those that are directly associated with the patient are classified as level-1. Those at the episodic level are level-2 while those that represent activities or events associated with these episodes are at the level-3.

In the basic form, anonymization uses the generalization and sometimes data suppression to address re-identification risk associated with the target recipient. Generalization is carried out using both built-in and resource library archived hierarchies (Table 6-4). In this experiment, generalization uses lattice-based solution determination using OLA (El Emam et al., 2009). When a

solution is determined, the anonymization service walks in reverse and does a random replacement of the original value with another value within the generalization range. For example, if a patient age is 34, but generalized to [35-40], the patient age in the anonymized data set could end up being 39.

Table 6-4 Resource library gazetteers for quasi-identifier.

Quasi-identifier	Type	Hierarchies
DOB	Built-in	Male, Female, Transsexual, Others
DOD	Built-in	A, B, AB, 0
Gender	Built-in	Single, Married, Divorced, Separated, Widowed
Ethnicities	Resource	Countries.csv
Languages	Resource	Languages.csv
Event Date	Resource	Firstnames.csv
Episode	Resource	Surnames.csv
CPT	Resource	Cities.csv
ICD9	Resource	Allergies.csv
Ethnicities	Resource	Ethnicities.csv
Religions	Resource	Religions.csv
Vaccines	Resource	Vaccines.csv

Suppression is done to balance generalization results to keep the analytical utility of the resulting data set. We generalize to a level that minimally impacts the analytical utility of the data set while suppression is then used to clean up attribute values that remain at risk after generalization.

6.3.2 Patient-level Anonymization Approach

Data from the various data sources usually provide some patient-level personal and demographic details data. As illustrated in Figure 6-8, patient-level data masking is done longitudinally with the patient global id (derived after identity matching) serving as the patient profile anchor. If

the target recipient is the data owner, patient-level data associated with the recipient is returned unchanged. If the target recipient requires data from other sources, patient-level data is replaced with a mask of the original. However, masking is done such that all the resultant profiles of a patient have the same mask values for the same attribute value. This is illustrated in Figure 6-8. For recipient 1, we can see that the patient with Global Id "Id001" in the source dataset from the CDM has multiple profiles for the same patient, but the target has the same masked details for those profiles that belong to this patient. For the other recipients, anonymization is not required since they will receive only the patient-level data from their organization internal database for their patients in this example.

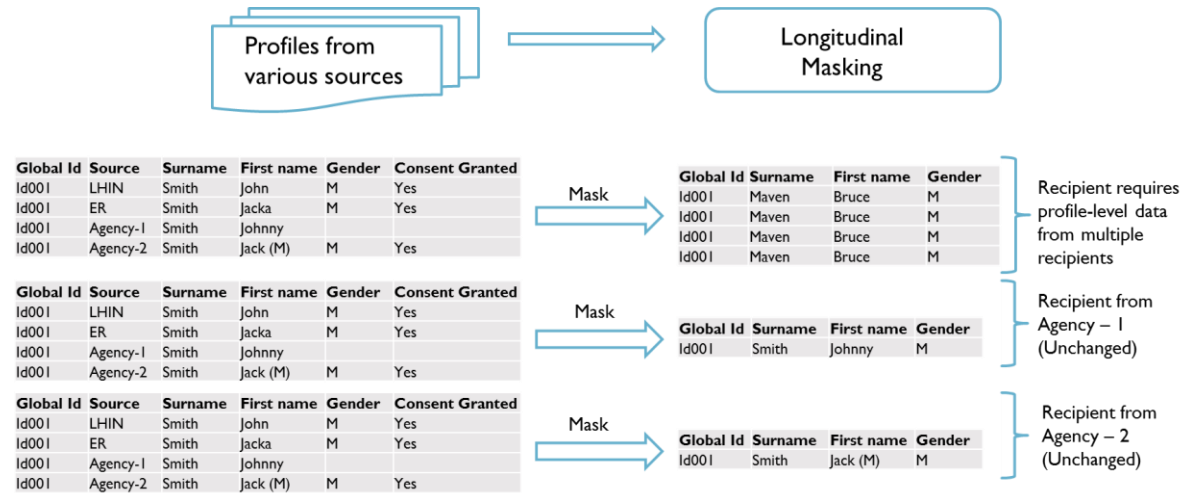


Figure 6-8 Patient-level Data Masking illustrated

Patient-level quasi-identifiers are considered differently for anonymization for a target recipient. If the recipient profile requires anonymization, then the profile needs to be merged for each quasi-identifier value. If two different values are detected, then the first non-null or empty value is picked for the attribute. At the end of the process, there will be a single cross-sectional record for each set of patient profiles that share the same global identifier. Patient-level QI values are the level 1 supergroups for episodic level-2 and event-based level-3 data.

6.3.3 Care Episode/Event Anonymization Approach

Care events are usually captured in episodes representing admissions and discharges to community care services. However, for patients with comorbid and complex conditions, care episodes can span months or years. Nevertheless, typically, what we usually see is that a patient gets admitted, go through some rounds of visitation and treatments and then eventually gets discharged. For our sample patient Bob, within one care episode, we have care events for a consultation, assessment, therapy session, visitation, treatment, surgery, laboratory work, etc. These events are identified by a sequence of dates, all within the care episode duration. It is essential that the anonymized dataset episodes and events follow a similar sequence as the original data set

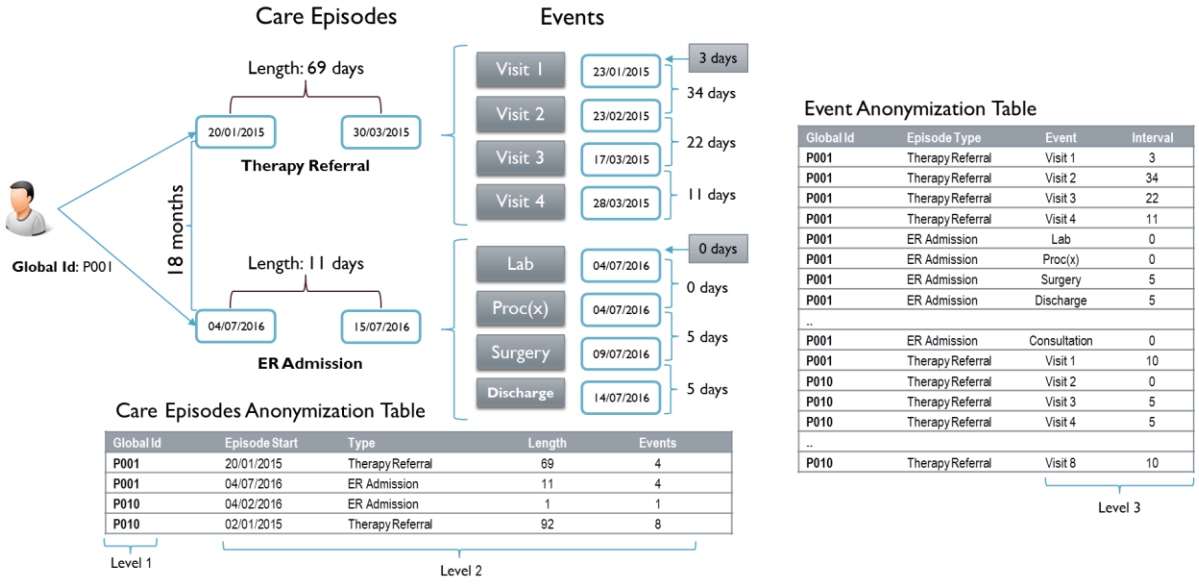


Figure 6-9 Patient Care Episodes Pre-processing Steps for Anonymization

Anonymizing this type of data is complex. Also, most common anonymization algorithms work well with cross-sectional and simple longitudinal data sets. Our approach to solving

this problem of reducing the complexity of episodic and event data is to transform the data into a new data set that can be anonymized using available anonymization tools. In Figure 6-9, we show the two care episodes that patient "P001" received within an 18 months period. First, there was a therapy referral that started off from the RHA in January 2015. Within this period, the patient had 4 therapy visits. The next year, the patient was in the ER, ended up getting admitted and staying in the hospital for surgery but was discharged on the 11th day.

Our approach is to use the start date of each episode as an anchor date. We then compute the length of the episode and the number of events for each episode type. These are then compiled into a staging table for the anonymization tool. A second table contains the event data and the inter-event intervals. Most importantly, the events are sequenced, from the first to the last for each patient. The inter-event intervals are the days from the last event to the next one within the episode. The patient global id is the level 1 identifier, the episode start date, type of episode, episodic length, and number of events are at level 2. Finally, the event details and inter-event intervals are at level 3.

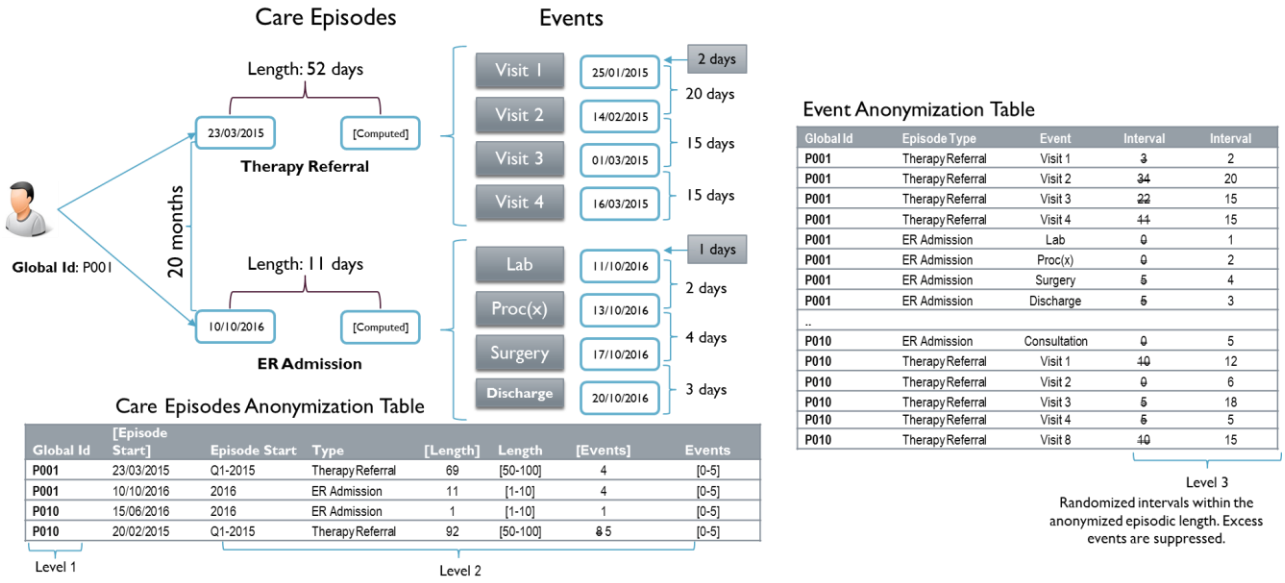


Figure 6-10 Patient Care Episode Post-anonymization processing steps

After anonymization, the resulting dataset contains both the generalization and the random replacement for each quasi-identifier. In this example (Figure 6-10), the episode start date for Therapy referrals has a quarter-year generalization. Within this period, a random start date of 23/03/2015, which is in Q1-2015 is set for this patient. The episodic length is set to be between [50-100] days. The event size is generalized to [0-5], meaning that the original 4 events are within the allowed limit for this patient. However, for patient "P010", the therapy session events need to be limited to a maximum of 5. We also see that the event table has new inter-event intervals. Based on these anonymized data values, the anonymized care episode and events for the patient is then reconstructed and returned as the anonymized care episode

6.3.4 Impact of All-or-Nothing Approach to address Patient Consent

To illustrate the privacy compliance model processes for addressing patient consent through anonymization, we will use a small data set of seven patients with three of these patients not-granting consent to data sharing (Figure 6-11). Figure 6-11a shows the records for those patients with the Consent Granted column indicating if they granted or denied consent to data sharing. In Figure 6-11b, an all-or-nothing approach to processing patient consent is applied. In this scenario, all the 3 records belonging to those non-consenting patients are removed. In Figure 6-11c, the records belonging to the non-consenting patients are anonymized instead of being removed. For clarity and space constraints, these examples included direct and quasi-identifiers from a single table, with no sensitive attributes.

Anonymization applied to these records includes data masking of the direct identifiers such as Patient Surname and First name; generalization for the quasi-identifiers such as birthdate and postal code attributes. Because of the size of the sample data set, we set the *k-anonymity* threshold value to 2. Based on this risk setting, the birthdate data is generalized to YOB, and the

Postal codes are generalized further as FSA with the last 3 digits of each postal codes suppressed. Suppression is then applied to the equivalence classes of the quasi-identifiers at a *k-anonymity* value of 2. The anonymization component uses the entire data set to build equivalence classes but applies anonymization only to the non-consenting patient records.

a) Original Dataset

Surname	First name	Gender	Birth Date	Postal Code	Consent Granted
Smith	John	Male	1965-08-25	K2A 5N6	Yes
Blake	Trevor	Male	1980-10-25	K1B 6N4	Yes
McGregor	Hilary	Female	1965-10-08	K2A 4B6	No
Elliot	James	Male	1980-10-01	K1B 5N4	No
Jones	Fatima	Female	1984-03-10	K2E 2P6	Yes
Wright	Martha	Female	1945-08-23	K1N 5N6	Yes
McCarthy	Terry	Male	1945-06-20	K2A 4N5	No

b) All-or-nothing approach: Non-consenting patients removed.

Surname	First name	Gender	BirthDate	Postal Code	Consent Granted
Smith	John	Male	1965-08-25	K2A 5N6	Yes
Blake	Trevor	Male	1980-10-25	K1B 6N4	Yes
Jones	Fatima	Female	1984-03-10	K2E 2P6	Yes
Wright	Martha	Female	1945-08-23	K1N 5N6	Yes

c) Selective anonymization: Non-consenting patient profiles anonymized

Surname	First name	Gender	BirthDate	Postal Code	Consent Granted
Smith	John	Male	1965-08-25	K2A 5N6	Yes
Blake	Trevor	Male	1980-10-25	K1B 6N4	Yes
Doe+	Jane+	Female	1965*	K2A*	No
St-Pierre+	Peter+	Male	1980*	K1B*	No
Jones	Fatima	Female	1984-10-03	K2E 2P6	Yes
Wright	Martha	Female	1945-08-23	K1N 5N6	Yes
Pit+	Patrick+	***	1945*	***	No

Figure 6-11 All-or-nothing vs. selective anonymization approaches to addressing patient consent

We see that the record belonging to *Terry McCarthy* ended up with a new name of *Patrick Pit*. The equivalence class formed using the Gender, YOB, and FSA in the database was too unique for a *k-Anonymity* value of 2. So, the Gender and Postal code values were both suppressed. For *James Elliot* and *Fatima Jones*, the quasi-identifiers were generalized but kept in-place because the equivalence classes have adequate support in the data set.

6.4. Results

In this experiment, we introduced an anonymization algorithm that addresses privacy compliance for both surveillance and performance management services in a cloud-computing infrastructure.

- Our anonymization algorithm showed that by categorizing and classifying privacy attributes at the patient, episodic, and event levels, we are able to address complex, longitudinal health records. This experiment took 2 months to set up using a dataset of select 240k patients with millions of records of services level data. To achieve continuous anonymization of data streams, the implementation should be significantly longer.
- In this experiment, by applying anonymization to the records of the non-consenting 18,500 patients, overall data loss in the common data model was reduced by up to 75%.
- To demonstrate the impact of anonymization to data missingness vs. all-or-nothing, anonymization is then applied to the entire patient demographic data set of 240k patients while simulating the increasing impact of no consents to attribute value missingness. The result is summarized in Figure 6-12. Our findings show that if the number of non-consenting patients is low, both approaches yield about the same level of attribute value missingness. However, if one assumes a maximum 5% suppression, we see that anonymization consistently reduced the overall missingness of the data set even in scenarios with over 80% of non-consenting patients.

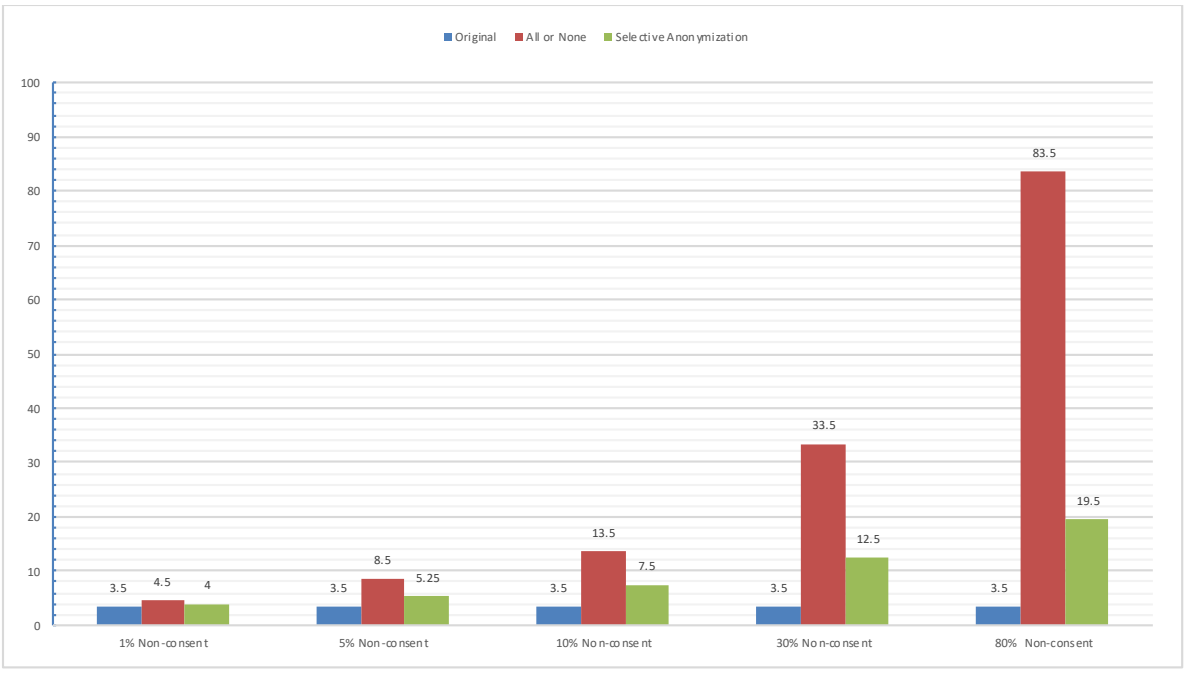


Figure 6-12 Data Missingness and the percentage of patients consenting to data sharing

- In the final analysis, our conclusions are that healthcare data should not be released publicly except for certain privacy-secure statistics or if the entire data set is anonymized appropriately. Therefore, irrespective of the patient consent, it is still the data custodian's responsibility to anonymize all publicly released data sets to safeguard patient privacy and confidentiality and to conform to existing privacy compliance regulations.

Chapter 7. Experiment on a Configurable Patient Identity Matching Algorithm

In Chapter 5, we discussed a pilot project for implementing a cloud-based infrastructure for supporting surveillance and performance management between community care organizations, and Champlain Local Health Integration Network (Champlain LHIN), acting as the Regional Health Authority (RHA). Aggregating data from various providers into the common data model requires a consistent and dependable mechanism for identifying a patient across all the data sources. However, this is not always the case with various community care scenarios. We saw from this pilot project that each Community care organization identifies the patient differently and may not capture the validated identity details of the patient. Since these community care organizations are not mandated to validate each patient profile in their database, direct deterministic matching of patient profile based using a government identifier, or their personal details is not always possible.

This experiment is a new Design Science Research (DSR) iteration and is designed to address the identity management limitations of the pilot project through an experiment that leverages a configurable identity matching service on the same data set used in the pilot project. It corresponds to DSR Iteration 4, as shown in Figure 1-3 of section 1.4 Research Methodology. We reiterate that a configurable identity matching service is required for community care identity management where 1) There is no universally accepted and properly validated identifier for each patient across the data providers, 2) Patient personal and demographic details are not captured consistently and are prone to data entry errors.

7.1. Architecture used in Experiment

In this experiment, Figure 7-1 shows the cloud-based infrastructure which hosts all the community care organizations local databases. It also correlates the data from these databases into a CDM for performance management. The Systematic Data Collection Service pulls data from the community care organizations cloud-hosted database instances and the RHA internal database into the Staging Database.

The focus of this experiment is on the configurable Patient Identity Matching Service. The Patient Identity Matching Service correlates patient profiles across these databases using declarative XML Match Definitions, to populate the Aggregate Patient Profile database. Each set of patient profiles across all community care organizations that belong to a single individual is assigned a globally unique identifier. The global identifier ensures that all the services provided to a patient by each of the stakeholders are grouped under this common patient identifier.

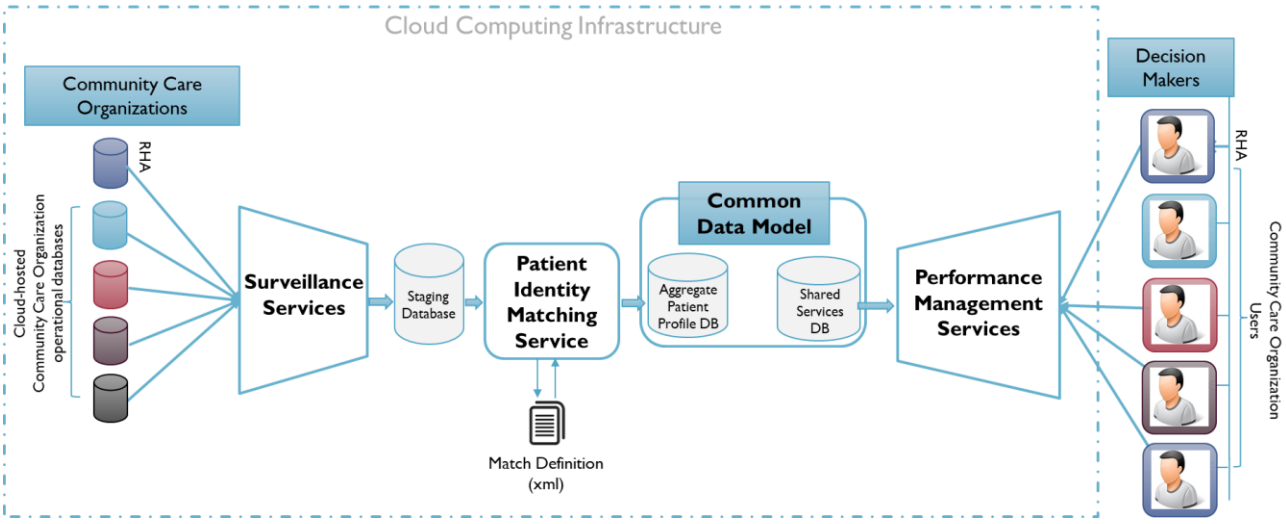


Figure 7-1 Cloud-based Infrastructure for Patient Identity Management.

7.2. Configurable Patient Identity Matching Algorithm

The configurable patient identity matching algorithm correlates cloud-hosted data from multiple stakeholders into a common data model in support of performance management of community healthcare. Our choice of a probabilistic matching algorithm is because it is the best approach for the type and nature of patient profiles in the datasets used for identity matching in our experiment. This is explained in detail in section 4.4.

7.2.1 Design Considerations

The primary challenge with patient identity matching for patient profiles across the community care organizations internal databases is that these organizations do not share a common patient identifier. As well, other attributes of the patient including first and last names, date of birth, gender, phone numbers, address, and personal contact details are not collected consistently and could have data entry errors. Also, government-issued Health Card Numbers (HCN) entries in these databases usually contain data entry errors since validation is usually not done or enforced at the time of data entry. As a result, it is a challenge to identify a patient profile across these data sources consistently.

In this experiment, the identity matching algorithm is extended from the version in the pilot project to:

- Support the ability to perform identity matching without a reference superset database of validated patient profile records. In the pilot project, the target was to match Community care organization patient profiles against a reference master database from the RHA. Consequently, some active Community care organization patients that were never admitted to the RHA could not be matched across these

organizations. In this experiment, matches are done against community care organizations internal databases without consideration for any master database from the RHA.

- Support for declarative and configurable definitions for attribute selection, transformation, match block group definitions, and the management of the match processes. Configurable definitions are an important requirement for our cloud-based surveillance and performance management architecture.
- Support for optimized match block processing with full and incremental matching, parallel asynchronous match processing, and improved overall performance.
- Matches are to be carried out on all patient-level records across all the incoming data streams. Since match datasets are continually changing, matching needs to be done on entire datasets or incrementally on new patients.
- Finally, in matching patient profiles, duplicate entries within the same Community care organization, where they exist, also need to be identified.

As described in section 4.4, the matching algorithm is broken into the following five components: 1) attribute identification and match block generation, 2) data standardization, 3) blocking strategy and match weight summarization, 4) decision, and finally 5) generation of a global identifier for each group of matches and transfer to the Shared Services Database.

The patient identity matching algorithm is described in section 4.4.1 and employs a weighted probabilistic matching system that derives from existing works in the probabilistic record linkage domains.

7.2.2 Attribute Identification and match block generation

Figure 7-3 shows the match blocks and weights used in this experiment. The weights were heuristic and "tuned" specifically to reflect the level of confidence of the business analysts for matches on the block with the selected data set. Match blocks build from direct and indirect patient identity attributes. Match blocks are grouped into match groups that exist within each applicable table in the Staging Database. Match groups are defined in a hierarchical pattern starting from a top-level group defined with direct patient attributes. The top-level group defines match blocks based on the direct attributes of the patient only, while subsequent groups define secondary identifying attributes like addresses and patient personal contacts. A match group definition includes a name, reference to the source table and the unique local identifier or key associated with the patient (if the group is at the patient level) or a join expression to the patient table if it is dependent on the patient table. The "JOIN" definitions help associate each patient record from this table to the primary patient table. Additionally, each block includes an identifier or index, a field list, and a user-assigned weight.

Each block definition can also include references to standardization and transformation functions such as computing the year of birth from the patient date of birth or extracting only the first character in each gender attribute. Defining match blocks is an iterative process that requires a thorough review of the attribute values, levels of missingness, data consistency, and data entry errors. Blocks with attributes that are very common, like first names, tend to have many matches and therefore need to be assigned low weights. As a result, the more unique the combination of attributes that define a match block, the higher the confidence weight. For this experiment, weights were not computed algorithmically. Rather, they were determined by iteratively weighing the impact of each matching block based on the attributes in its definitions. Those with a

higher likelihood of generating ambiguous matches have lower weights while those with lower likelihood were assigned higher weights.

7.2.3 Data Standardization

Data standardization is required for attribute-level data reconciliation, without which true matches could be wrongly designated as non-matches simply because identifying attributes do not have sufficient similarity (Gu et al., 2003). Data standardization definitions can be applied to an attribute across all match blocks or specifically to select match block. Each block definition can include standardization definitions for search/replace, data splitting, and data truncation functions. These transformations help data analysts fix data semantic issues like dealing with hyphenated names declaratively, on-the-fly data conversions (for example Gender values "M" or "Male" to semantically mean the same), removing Nicknames from names (for example, retrieving the first name from name values in the form of "John (Jack)"), removing database constants and placeholders, and processing matches with transformed dates such as year of birth. The supported standardization functions are supported through the Systematic Data Collection Service, as listed in Table 7-1.

Table 7-1 Supported data standardization definitions

Standardization Functions	Description	Examples
Search and replace (<Replace>)	Use to Search and replace text in an attribute. This function is used to standardize misspellings, remove attribute default values, and harmonize data formatting and semantics.	<Replace table="PATIENTS" fields="HEALTH_NUMBER" search="00000*|11111*|99999*" replace="" />
SQL Date functions	Dates can be in various formats. Transformations using SQL functions can generate computed values, including age, year of birth, month/year, and quarter/year of birth.	<Seed table="PATIENTS" field="YOB" function="YEAR([DATE_OF_BIRTH])" CONVERT(varchar(7), [DATE_OF_BIRTH],'mm-yyyy') As MOB
Data Splitting (<Split>)	Use to split text, so a part of an attribute text is used with CDM, while the rest is discarded. For examples,	<Split table="PATIENTS" fields="PATIENT_GIVEN" split_by=" " merge_with="" />

	to extract only the first name within a patient given name even if multiples names are listed.	
General SQL functions	In-place SQL transformation functions such as CONCAT can be used within block definitions.	<Seed table=”PATIENTS” field=”PhoneNo” function=”CONCAT([PHONE_AREA_CODE],[PHONE_EXCHANGE],[PHONE_NUMBER])” />
Truncate (<Truncate>)	Use to retrieve a substring of a text attribute for matching. For example, the first three characters of a patient Postal Code, or the first or last four digits of a phone number.	<Truncate group=" PATIENTCONTACTS " blocks="1|3" fields="POSTAL_CODE" param="1|3" />

Figure 7-2 and Figure 7-3 show some example of explicit standardization and in-block standardization functions used in the declarative XML Match Definition file for this experiment. Standardization is applied to remove fake Health card numbers sometimes used as placeholders during data entry, fix hyphenated names, remove constant values used to define unknown gender, default dates, split patient surname in some match block, truncate postal codes, compute year of birth and to concatenate data field attributes.

The set of transformations depicted in Figure 7-2 are designed to standardize and sometimes generalize data from the Community care organization datasets. The same set of rules can be applied to other databases with specific transformation definitions that suit the dataset used for identity matching. Determining the data standardization functions to use is an iterative process that requires a thorough review of the attribute values, levels of missingness, data consistency, and data entry errors within each dataset used for identity matching.

```
<Transformations>
        <!--Patient table transformations -->
        <Replace group=" PatientGroup" fields="HCN " search="00000*|11111*|99999*" replace="" />
        <Replace group=" PatientGroup" fields="Lastname |Firstname " search="'|-| " replace="" invalue="true" />
        <Replace group=" PatientGroup" fields="Sex" search="MK|NULL" replace="" />
        <Replace group=" PatientGroup" fields="DOB " search="1753-01-01" replace="NULL" />
        <Split group="PatientGroup" fields="Othernames " split_by=" " merge_with="" />
        <Truncate group=" PatientGroup" levels="2" fields="Lastname" param="1|5" />
</Transformations>
```

Figure 7-2 Sample Standardization Transformation Definitions

Explicit transformations shown in Figure 7-3 can apply to one or more fields within a match group, as well as specific blocks containing such fields – providing more flexibility with the definitions. If a block attribute definition is added, then the transformation would only apply to fields within the specified block index.

```
<!--Patient-level match blocks-->
<Match_Group name=" PatientGroup" table="Patients " matching_key="PatientId ">
    <Block index="1" weight="1.0" fields="HCN| Lastname | DOB| Sex"/>
    <Block index="2" weight="0.7" fields="Lastname | Firstname | DOB | Sex"/>
    <Block index="3" weight="0.3" fields="Lastname | YEAR([DOB]) AS YOB| Sex | PostalCode" />
    <Block index="6" weight="0.2" fields="Lastname | Firstname | Sex" parent="2|5"/>
    <Block index="7" weight="0.2" fields="Lastname | Firstname | Sex| PostalCode " />
    <Block index="8" weight="0.2" fields="Lastname | Firstname " parent="2|5|6" />
</Match_Group>
<Match_Group name="PatientRelations" table="Relations " join="PatientGroup.PatientId = PatientRelations.PatientId">
    <Block index="1" weight="0.2" fields="Lastname| YEAR([DOB]) AS YOB| Sex | PostalCode"/>
    <Block index="2" weight="0.2" fields="Lastname| Sex| MobileNo"/>
    <Block index="3" weight="0.1" fields="Lastname| sex | PostalCode"/>
</Match_Group>
```

Figure 7-3 Sample In-block standardization functions

As shown in Figure 7-3, our definition also supports transformations applied in place within a block definition. A block can contain a single field, but most contain one or more field attributes separated by a pipe ("|"). Each block must have a probability weight that would be associated with their matches. The *parent* attribute in each match block defines block dependencies to another with more attributes. This provides a mechanism for avoiding weight padding to matches based on blocks with a subset of attributes from other blocks. For example, the field attributes for patient-level block 6 is a subset of blocks 2 and 5. This way, a match on the parent block (2 or 5) will force the matching algorithm to ignore matches on dependent blocks like block 6.

7.2.4 Blocking Strategy and match weight summarization

String comparison across attribute original or transformed values that make up each block is an expensive process and usually do not to scale for large datasets. As a result, in processing the blocks, the algorithm generates hash values for each block based on patient attributes within its definition. These hash values can be indexed and sorted for quick searches and filtering. For each new patient profile and for each match block definition, a hash is generated and placed in the appropriate hash table for the applicable block. This process essentially flattens the set of attributes that define a match block into a hash value that describe unique combinations of values for the block within each dataset. Hash values are only re-generated when data attributes of a patient change in an incoming data stream.

Figure 7-4 describes the block processing strategy. Block processing is done once for each patient. For each incoming data stream, the algorithm first checks if the patient profile already has a global identifier. If the identifier is found and the patient match is a full match, then matching is ignored for the patient. If the global identifier is found, but the match is either a possible or non-match, then matching is initiated for the new patient profile.

The algorithm does multiple blocks passes for each new patient profile. For each block pass, the match weight associated with the block is assigned to each matched patient record for the block pass. To improve matching speed and accuracy, block records with missing values in their block attributes are excluded. For example, a record with these attributes - LASTNAME: Smith, FIRSTNAME: John, DOB: NULL will be excluded when processing a block that requires LASTNAME, FIRSTNAME, and DOB. The more support a block has, the higher the impact on the match outcomes.

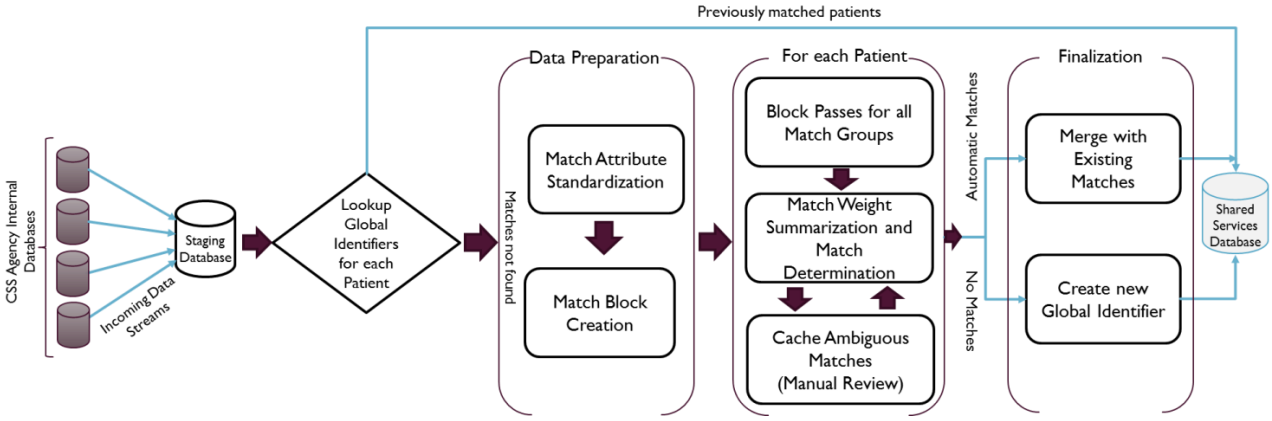


Figure 7-4 Patient Profile Matching from Staging to the Shared Services Database Illustrated.

7.2.5 Patient Identification Decision

The process for determined matches is described in 4.4.2. For this experiment, the composite weight of each matched patient (across all match blocks) is a minimum of 0.8. Those with weights from 0.6 – 0.7999 are seen as possible matches (ambiguous). Matches below 0.6 are rejected.

7.3. Discussion and Other Considerations

The Identity Matching Service used in our infrastructure helps consolidate complex patient records across multiple data sources to facilitate interoperability in a common model for performance management of patient services. The matching algorithm used by the Identity Matching Service is in the intermediate category since it uses a probabilistic approach that leverages fuzzy logic and weights. It also incorporates a user management component for reviewing full and possible matches.

7.3.1 Configurable Match Definition

Most matching algorithms are based on fixed definitions regarding attributes, rules, and weights. One of the major contributions of our work is that it is driven entirely by the Match Definition file. The declarative Match Definition file defines the components of the patient matching algorithm - attribute identification, data standardization, block definition, and the behaviour of the decision components. This XML document also includes definitions for the source and target databases, success and error notification email settings, match groups and block definitions, a wide range of standardization functions that apply to attributes in general or specifically to attributes within match groups and match blocks. It also allows one to define match blocks using multiple attributes across multiple tables. Also supported are decision settings for full matches, possible matches, and non-matches. This level of configurability is required because, in the context of cloud-hosted data, configurable definitions make it less tedious to change match behaviour or to tweak weights to align with data changes. This definition also determines how matching is carried out for incoming data streams– full or incremental.

The match definition document (see Figure 7-5) was employed for this experiment. This document consists of:

1. **Notification Section** – Defines the administrators that would receive an email notification if the matching process fails.
2. **Match Groups/Blocks** – One of more match blocks that represent matching probabilistic matching rules. Being declarative, they can be tweaked as needed to achieve the best results. Blocks are grouped into match groups that exist within each table in the Staging Database. A match group definition includes a name, reference to the source table and the unique local

identifier or key associated with the patient (if the group is at the patient level) or a join expression to the patient table if it depends on the patient table.

3. **Standardization rules** - These are rules for ensuring that data elements are semantically equivalent. The example in Figure 7-5 uses some replace, split, and truncation rules.

```
<?xml version="1.0" encoding="utf-8"?>
<Patient_Matching match_weight="0.8"
    possible_match_weight="0.6"
    clean_before_match="true"
    update_field="LAST_EDIT_DATE">
    <Data_Source server="Staging_Cluster" database="Agency_Staging" />
    <Match_Database server="Shared_Svc_Cluster" database="Agency_Matching" />

    <!--Pipe separated list of recipients. Use mode="onerror" to send log if client matching fails -->
    <Notification smtp_server="192.168.58.54" recipients="admin@regional-health-canada.ca" mode="onerror" />

    <!--Patient-level match blocks-->
  <Match_Group name="PATIENTS" table="dbo.C3CLINT" matching_key="PATIENT_NUMBER">
    <Block index="1" weight="1.0" fields="HEALTH_NUMBER| PATIENT_SURNAME| DATE_OF_BIRTH| GENDER" />
    <Block index="2" weight="0.7" fields="PATIENT_SURNAME| PATIENT_GIVEN| DATE_OF_BIRTH| GENDER" />
    <Block index="3" weight="0.3" fields="PATIENT_SURNAME| YEAR([DATE_OF_BIRTH]) AS YOB| GENDER| POSTAL_CODE" />
    <Block index="4" weight="0.3"
        fields="PATIENT_SURNAME| CONCAT([PHONE_AREA_CODE],[PHONE_EXCHANGE],[PHONE_NUMBER]) As PhoneNo|POSTAL_CODE| GENDER" />
    <Block index="5" weight="0.3"
        fields="PATIENT_SURNAME| PATIENT_GIVEN| GENDER| CONCAT([PHONE_AREA_CODE],[PHONE_EXCHANGE],[PHONE_NUMBER]) As PhoneNo" />
    <Block index="6" weight="0.2" fields="PATIENT_SURNAME| PATIENT_GIVEN| GENDER" parent="2|5" />
    <Block index="7" weight="0.2" fields="PATIENT_SURNAME| PATIENT_GIVEN| GENDER| POSTAL_CODE" />
    <Block index="8" weight="0.2" fields="PATIENT_SURNAME| PATIENT_GIVEN" parent="2|5|6" />
  </Match_Group>
  <Match_Group name="PATIENTCONTACTS" table="dbo.C3CLCONT" join="PATIENTS.PATIENT_NUMBER = PATIENTCONTACTS.PATIENT_NUMBER">
    <Block index="1" weight="0.2" fields="SURNAME| YEAR([DATE_OF_BIRTH]) AS YOB| GENDER| POSTAL_CODE" />
    <Block index="2" weight="0.2" fields="SURNAME| GENDER| CONCAT([HP_AREA_CODE],[HP_EXCHANGE],[HP_NUMBER]) As PhoneNo" />
    <Block index="3" weight="0.1" fields="SURNAME| GENDER| POSTAL_CODE" />
    <Block index="4" weight="0.1" fields="SURNAME| GIVEN_NAME" />
    <Block index="5" weight="0.2" fields="SURNAME| GENDER| CONCAT([CP_AREA_CODE],[CP_EXCHANGE],[CP_NUMBER]) As PhoneNo" />
  </Match_Group>
    <Transformations>
        <!--Patient table transformations -->
        <Replace group="PATIENTS" fields="HEALTH_NUMBER" search="00000*|11111*|99999*" replace="" />
        <Replace group="PATIENTS" fields="PATIENT_SURNAME|PATIENT_GIVEN" search="'|-| " replace="" invalue="true" />
        <Replace group="PATIENTS" fields="GENDER" search="UK|NULL" replace="" />
        <Replace group="PATIENTS" fields="GENDER" search="UD" replace="" />
        <Replace group="PATIENTS" fields="DATE_OF_BIRTH" search="1753-01-01" replace="NULL" />
        <Split group="PATIENTS" fields="PATIENT_GIVEN" split_by=" " merge_with="" />
        <Truncate group="PATIENTS" levels="2" fields="PATIENT_SURNAME" param="1|5" />
        <Truncate group="PATIENTS" levels="3" fields="POSTAL_CODE" param="1|3" />
        <Truncate group="PATIENTS" levels="8" fields="PATIENT_GIVEN" param="1|3" />
        <Replace group="PATIENTS" fields="YOB" search="1753" replace="NULL" />
        <!--Patient Contact table transformations -->
        <Replace group="PATIENTCONTACTS" fields="YOB" search="1753" replace="NULL" />
        <Truncate group=" PATIENTCONTACTS " blocks="1|3" fields="POSTAL_CODE" param="1|3" />
        <Truncate group=" PATIENTCONTACTS " blocks="4" fields="GIVEN_NAME" param="1|3" />
    </Transformations>
</Patient_Matching>
```

Figure 7-5 Match definition file for the Experiment.

7.3.2 Managing Organization and Patient Consents

The matching algorithm also integrates with Privacy Compliance Filtering component of the cloud infrastructure to ensure that community care organizations that have not signed a data

sharing agreement are excluded automatically from data migration to the Shared Services Database as well as subsequent analytics in the framework. Similarly, patients that refused consent to data sharing are eliminated from the Shared Services Database. Attribute disclosure is one challenge with this level of data aggregation. This "all-or-nothing" approach to privacy compliance is a limitation but not the focus of this experiment. This is addressed in Chapter 6 by incorporating anonymization with privacy compliance.

The reporting and subscription services refer to the local patient profile and identifier associated with each Community care organization when returning identifying details on a patient for organization-specific reports, ensuring privacy and confidentiality protection in the process. All patient-level reports and analysis include the global identifier which uniquely identifies a patient across all participating organizations.

7.3.3 Impact of Match Block Distribution

For this experiment, Patient Identity Matching was carried out on all the cloud-hosted databases. These have 135k patient records of which 25k patient profiles have confirmed matches with almost 3k possible matches. About 20% of the matches came from the first two patient level match blocks because of their high confidence weights. Table 7-2 shows the support and match percentage breakdown for each of the match blocks. We can see that while 99.7% of patients have both first and last names (Patient level block 8), while only 31% has their HCN on record. As many as 12.92% of the available HCN have matches. The match percentage for each Community care organization ranges from a high of 72.3% to as little as 0.9%, with an average of 15.6% across the community care organizations. Our observation is that the organizations that

offer more services tend to have higher match rates. Also, the location of a Community care organization affects the likelihood of matches. The larger organizations with multiple locations tend to have more matches than those that exist in a single location.

We can also see the impact of using secondary attributes of the patient, such as the personal contacts on the matches. Most patient contact entries have few missing values for names but show a high-level of attribute value missingness with address details, date of birth, and gender attributes hence the low support for the blocks with those attributes. However, patient contacts match blocks provided significant support and validation for many of the low weight patient-level blocks.

Also, we reviewed the matches based on their match contexts. A match context is a combination of match blocks rules used to arrive at matches based on the composite weights of the associated match blocks. We found that the top 10 of 129 match contexts represents over 67% of the matches. What is more interesting is that the top 2 represent about 25% of the matches. As shown in Table 7-3, these top 2 popular contexts built up weights from all patient-level blocks. These represent very richly captured data with few missing values with the all patient identifiers used for matching.

Table 7-2 Match block summary for the community care organizations

Level	Matching Blocks	Weights	# of Profiles	# of Patients with matches	Block Match %
Patient Level Blocks					
1	Health Card Number, Last name, DOB, Gender	1.0	41, 582	4,383	10.54%
2	1st five characters of Last name, First name, DOB, Gender	0.7	97,322	7,889	8.11%
3	Last name, Year of Birth, Gender, 1st three characters of Postal Code	0.3	89,856	8,662	9.64%
4	Last name, Phone Number, Postal Code, Gender	0.3	93,451	7,737	8.28%
5	Last name, First name, Gender, Phone Number	0.3	101,573	7,780	7.66%

6	Last name, First name, Gender	0.2	108,784	13,092	12.03%
7	Last name, First name, Gender, Postal Code	0.2	97,858	7,056	7.21%
8	Last name, 1st three characters of First name	0.2	137,455	19,991	14.54%
Patient Contact Level Blocks					
1	Contact Last name, Year of Birth, Gender, 1st three characters of Postal code	0.2	1,001	30	3.0%
2	Contact Last name, Gender, Home Phone Number	0.2	16,047	968	6.03%
3	Contact Last name, Gender, 1st three characters of Postal Code	0.1	4,361	286	6.56%
4	Contact Last name, 1st three characters of First name	0.1	105,869	16,851	15.92%
5	Contact Last name, Cell Phone Number	0.2	5,783	366	6.32%

We could also see the importance of patient contacts in contexts 3 and 6. Patient blocks 3, 5, and 7 with patient postal codes and contact number fields also contributed in determining many of the matches. In addition, at least 25% of matches include the first patient block. It shows the importance of deterministic attributes to the matching process. However, we would have lost 75% of the matches with a deterministic approach using only those patient block 1 attributes.

Table 7-3 Top 10 Match Contexts

Index	Match Contexts	% of Matches
1	Patients:2[0.7],Patients:3[0.3],Patients:4[0.3],Patients:5[0.3],Patients:7[0.2]	12.61%
2	Patients:1[1],Patients:2[0.7],Patients:3[0.3],Patients:4[0.3],Patients:5[0.3],Patients:7[0.2]	12.40%
3	Patients:1[1],Patients:2[0.7],Patients:3[0.3],Patients:4[0.3],Patients:5[0.3],Patients:7[0.2],PatientContacts:4[0.1]	9.61%
4	Patients:4[0.3],Patients:5[0.3],Patients:7[0.2]	7.00%
5	Patients:2[0.7]	6.31%
6	Patients:2[0.7],Patients:3[0.3],Patients:4[0.3],Patients:5[0.3],Patients:7[0.2],PatientContacts:4[0.1]	5.72%
7	Patients:3[0.3],Patients:4[0.3],Patients:5[0.3],Patients:7[0.2]	3.46%
8	Patients:1[1],Patients:2[0.7],Patients:3[0.3],Patients:5[0.3]	3.32%
9	Patients:2[0.7],Patients:3[0.3],Patients:7[0.2]	2.70%
10	Patients:2[0.7],Patients:5[0.3]	2.42%
		67.90%

The results presented in this thesis came from many iterations of tweaking attribute definitions and the weights of the blocks. In the first iteration, one assumption that resulted in some erroneous matches came from the perfect confidence given to a patient block 1 with HCN and Surname only. However, we found a few matches on some family members that received services under the same HCN. Technically, the matches were correct, but it prompted us to add additional attributes such as Gender and Date of Birth to this block to address such errors. It is also important to point out that most blocks included Surname and Gender attributes. These attributes help reduce false positives in the matches.

In the final analysis, the distribution of the matches shows that 82% have a support of 2 (2 patient profiles matches). 15% has 3 matches, 2.5% has more than 4 matches. Only 56 patients have 5-7 matches – that means they received services from more than 5 community care organizations.

7.3.4 Dealing with Ambiguous Matches

Ambiguous or possible matches occur when a patient record has matches only on the weaker match blocks (those with low probability weights), ending up with a composite match weight that is less than the required threshold. For example, a patient profile with the patient blocks 3 and 4 matches in our experiment would have a composite weight of 0.6, 0.2 points below the threshold of 0.8. These matches would require additional personal contact block weights to become confirmed matches. If the patient record has no personal contacts, these matches would be left in an ambiguous state. Such ambiguous entries are pushed to a temporary staging table for a human to review and approve or disapprove the matches (Figure 7-6). If approved, a

global identifier is issued, and those matches are pushed to the Shared Services Database. If un-approved, the matches get excluded from the ambiguity database and subsequent ambiguity identification.

Status	Agency	Patient Id	HCN	Surname	Firstname	Other names	Date of Birth	Gender	City	Postal Code
Global Id: 5679, **Regioinal Authority Patient Id**: 1111111, **Client Name**: Doe, Jane										
	RHA Id	111111	222222	Doe	Jane		10/31/1974	F	Orleans	K0A2X0,K1H8M2,K1N8B7 K2L4H8
	Agency - 1	20131234000231	Missing	Doe	Jane		10/30/1974	F	Ottawa	K0A2X0
	Total 2									

Figure 7-6 Possible Matches with Review and Resolution

Ambiguity data helps the RHA create a report on patients with ambiguities for Community care personnel to investigate and resolve. For example, data entry errors with health care numbers, names, or patient contact details. Fixes are made in the source databases so that they get matched with subsequent data streams.

The standard quadrant accuracy report (Figure 7-7), shows the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FP) percentages in this experiment. All the FP and FN numbers fall under the possible or ambiguous matches. These numbers are very dependent on the thresholds for full matches and possible matches. When thresholds are lowered, more ambiguous entries are moved to TP, reducing the TN numbers and increasing the FP numbers. When the thresholds are very high, the FN numbers go up since many true matches are missed while FP becomes very small. Determining the right threshold requires a thorough analysis of the match contexts and the impact of the weights assigned to the match blocks.

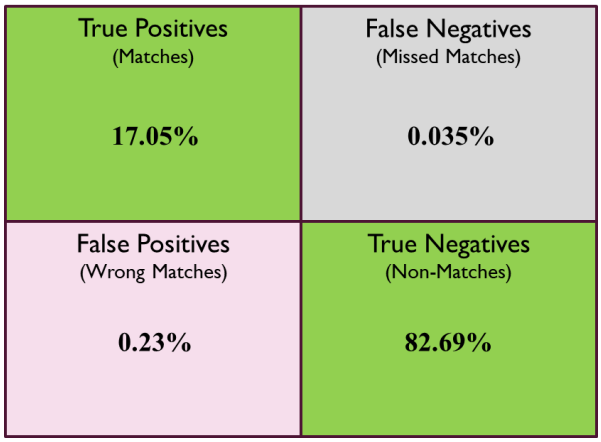


Figure 7-7 Quadrant Accuracy Report

7.3.5 Performance

Traditionally, deterministic record linkage on a single identity attribute on n patient profiles has the complexity $O(n^2)$. In this experiment, the performance of the identity matching algorithm is dependent on the number of available patient profiles and the number of match blocks m. The maximum match block passes $T(M) = n * m$. Since $m \ll n$, block passes do reduce the match complexity significantly. To speed up the block passes, the attributes that make up each match block are encoded into an indexed hash table containing the entries for all patients with non-null values across all block attributes. Therefore, the support for all blocks, Support(B) = $\sum_{i=1}^{m}(|B(i)|) < n * m$. Since the support for each block determines how quickly matching within the table completes, this optimization results in a significant reduction in the number of passes. For this experiment, $n = 137{,}743$. The total number of blocks $m = 13$. Therefore the maximum block passes T(M) = 1.8 million instead of 246 billion. Another optimization we added to the matching process is to run up to 100 patient profile matches in parallel through asynchronous threads. To eliminate concurrency issues with asynchronous threads, when the matches for a thread are being committed to the database, it automatically blocks other threads.

This way, matches do not get overwritten by results from other threads. Running concurrent patient matches help maximize compute resource utilization.

The matching service was hosted on a 4 CPU Virtual Machine (VM) with 4 GB RAM running MS SQL Server 2014. A full profile matching run of the total Community care organization patients of about 135k takes about 35 minutes process. However, since matching is done incrementally for only new or updated patient profiles, the process completes in less than 5 minutes for each daily ETL run.

7.4. Results

In this chapter, we used an experiment from a healthcare region in Canada to develop and evaluate a probabilistic patient identity matching algorithm for correlating patient records across multiple organizations in support of performance management for community care services.

- While match statistics show that about 25k patient of Community care organization patients profiles belong to patients that receive community care services from multiple organizations, the comparison of data within each aggregate patient profile has become a useful mechanism for addressing data quality issues with these organizations. The community care organizations use the daily ambiguities report to identify data quality issues in their patient that require some investigation.
- The Patient Identity Matching algorithm and monitoring infrastructure implementation took about 4 months for a team of one 1 developer, two Business Intelligence officers, and a Project manager. The testing and tweaking of match blocks took about 4-6 weeks.

- Patient identity matching service runs unattended and requires little or no maintenance after its initial configuration.
- Data on possible or ambiguous matches are currently being sent to the community care organizations to help them fix erroneous patient data in their operational database so subsequently matches come up as full matches. The distribution of error sources for ambiguous aggregate profiles are shown in Table 7-4. Most sources of errors come from data entry issues with health card numbers and address details. The least errors are from the surname or last names.

Table 7-4 Error distribution for ambiguous matches

Attributes	
Health Card Number (HCN)	97%
Postal Code	84%
City	39%
First name	38%
DOB	28%
Surname	1%

Chapter 8. **Thesis Evaluation**

In this chapter, we evaluate the evolution of our surveillance and performance management architecture in section 8.1 based on the core components of the architecture described in chapter 4. We then evaluate our architecture in section 8.2 in terms of the success of the pilot project described in chapter 5. Section 8.3 evaluates our configurable anonymization algorithm for privacy compliance based on the experiment described in Chapter 6. Subsequently, section 8.4 evaluates the configurable patient identity matching algorithm based on the experiment described in Chapter 7. It also compares our algorithm with similar algorithms and systems from the literature. Section 8.5 evaluates our architecture against related work based on the evaluation criteria defined in section 3.3. Finally, section 8.6 discusses the assumptions, limitations, and threats to the validity of our thesis research.

8.1. Evaluation of Surveillance and Performance Management Architecture

Table 8-1 summarizes the evaluation of evolving iterations of our work, from current practice to our pilot project to our final architecture defined in chapter 4.

Table 8-1 Framework evaluation

Components	Current Practice (RHA before Pilot Project)	Initial Architecture (Pilot Project)	Complete Architecture (after experiments)
Cloud Computing Infrastructure for Surveillance and Performance Management	• Surveillance infrastructure non-existent • Performance Management was manual and ad-hoc, mostly to reconcile billing data. • Shared excel files for funding and budgetary needs.	• Private Cloud Infrastructure • Systematic Data Hosting • Systematic Data Collection • Supports for external data sources – Structured and Semi-Structured (XML and JSON data) • 8 Reports delivered through the email subscriptions • Dynamic Subscription Service for emailing reports. • Supports secure cloud-based file share report delivery.	• Private Cloud Infrastructure • Systematic Data Hosting • Systematic Data Collection • 8+ Data reports through a reporting portal and subscriptions. • Dynamic Subscription Service for emailing reports. • Support for secure data file transfer through configurable subscriptions. • Supports secure cloud-based file share for report delivery. • Support for addressing privacy compliance through anonymization. • Support for recipient-specific anonymized reports.
Common Data Model	• None existed	• Implements a Shared Services Database CDM Data aggregation depends on the same schema across data providers.	• Implements a Shared Services Database CDM • Support disparate schemas from structured and semi-structured data sources.
Patient Identity Matching Service	• Deterministic • Based on Health Card Numbers. • No matching of patient profiles across community care organizations.	• Probabilistic Matching using 6 demographic attributes. • Weights are not heuristically tuned. • The matching algorithm is not configurable • Depends on a reference set of RHA patient records for matching Community care organization profiles. • No support for inter Community care organization patient matching	• Probabilistic matching • Dynamic match definition using over 12 patient attributes • Weights are heuristically tuned. • Matches independently of RHA patient records. • Supports inter Community care organization patient matching

Privacy Compliance Model	• No data sharing	• All-or-nothing approach to compliance	• Supports full or selective anonymization on all identifiers • Supports recipient report views for recipient risk-driven anonymization

8.2. Evaluation of Pilot Project

We evaluate the success of this pilot project through quantitative statistical indicators collected and feedback from both operational and management staff of the RHA and the Community care organizations (CCOs).

8.2.1 Evidence of the success of the pilot project

Participation in the cloud-hosted environment was quickly adopted by many of the community care organizations. There are currently 54 community care organizations participating in this project. At the time of this writing, 90% of these organizations in the Champlain region have migrated to the cloud-hosted platform. Participation in performance management requires each community care organization to sign a DSA. Performance management services include identity matching, reporting, and subscription services.

Table 8-2 Pilot Project success metrics

Metrics	Statistics		Details
	Total	Percentage	
Community care organizations Participating in the Project	54	100%	All community care organizations that participated in this pilot are from the Champlain Region.
Community care organizations that migrated to the Cloud-hosted Platform	48	89%	These are the organizations that have cloud-hosted their Client/Patient Management Information System.
Community care organizations that have signed the DSA and	17	35%	These are the organizations that currently have their data going

are currently participating in Performance Management Services			through the entire workflow described in the pilot project.
Total patients across all community care organizations	151,144	100%	This is the aggregate count of patient profiles across the CCO local databases.
Total Patients currently active in the RHA Database	52,000	35%	The RHA has about 52k active patients. Not all of them receive services from community care organizations in the Champlain region.
Total patient from the community care organizations currently included in Performance Management Reporting.	28,013	18.5%	These are the patients from the 17 community care organizations that signed the DSA. They represent a little over 50% of active RHA clients.
Total Unique Patients across the community care organizations	135,706	90%	This represents a unique patient profile across community care organizations. While there are 135k unique patients, almost 25k patients are getting services from 2 or more community care organizations.
Total Patients with matches	24,988		
Total Patients with Ambiguous Matches	2,928	10% of matches	This represents the count and percentage of patient profiles that seem to be a match but with a low enough probability that a manual verification is required

The pilot project is considered a success because the following are now supported (it was not possible before the pilot project):

- There are nightly data collection and aggregation across the Community care organization operational databases to the Common Data Model.
- There are nightly patient identity matching and progressive clustering of patient profiles across the community care organizations.
- There are active report subscriptions set up for the Community care organization contacts that publish and emails various reports. Some of these reports are described in Table 8-3.

- The subscription service has been adopted by the RHA and is used for packaging and sending via email and file transfer, hundreds of packaged reports to RHA employees and external partners multiple times a day.
- Identity matching results are being used to provide data quality feedback to community care organizations on various data quality issues with patient profiles that need addressing.

Table 8-3 Popular Performance Management Reports

CCO ED Notification Client Report • This report provides the community care organizations with daily Emergency Room (ER) admission feeds from the local hospitals in the Champlain area. It leverages the results of the patient identity matching service to link ER records with Community care organization local profiles. The report is filtered by each Community care organization and provides a view with the local patient identifier, the name of hospital ER, the visit date, type of visit, complaints presented and any additional info. • For the community care organizations, the report provides important daily planning data on their clients or patients. By knowing that a patient has been admitted to or visited the ER, they are able to plan appointments to clients' homes or meetings with clients better. Courtesy calls can be made to client home resulting in the more efficient use of the employee time and increased the overall satisfaction of the patient.
Match Comparison Report • The identity matching service compares patient profile attributes across Champlain Community care organizations to find matching profiles of each patient across other community care organizations. Our approach to identity management came out of the necessity to match patient profiles across these organizations and because they do not always share a common identifier for patients in the health region. Since most community care organizations do not validate patient demographic data, many patent attributes may be incorrect. This report compares each matched profile to the validated patient record within the RHA (where it exists), and other matched profiles to highlights data entry errors in each organization local database instance. • It also has a view that provides the list of ambiguous records with sparse data so those can be adjusted or corrected. • The data provided in the report include the Agency Name, Patient Id, Patient HCN (only provided to those CCOs allowed to access patient HCNs), Surname, First name, other names, Date of Birth, Gender, Address details like City, and Postal Code. For privacy reasons, the report only highlights issues in attribute values. It doesn't provide suggestions. CCOs are responsible for finding and updating the attribute with the correct values.
Agency Match Profile Report • One of the important utilities of the patient match result is the consolidation of care services. With patient profiles matched, service data can be aggregated to show a complete picture of the services being received by a patient. This report provides those aggregate services for each Community care organization on their clients or patients. The report has two views: The first view shows the

aggregate services of each active patient for the Community care organization while the second view provides the aggregate services for all clients within the organization's local database – active and inactive.

- The service level data provided include – Patient local identifier, the global identifier, Community care organization providing the service, provider or caregiver organization, the type of referral, type of service provided, the method of service delivery, the current state of the referral (active, on hold, completed) and the date of admission and first service.
- This report is also part of the subscribed reports accessible to Community care organization users. Each organization is provided with a cloud-based secure share to access these reports. The report is updated nightly.

8.2.2 Feedback Collected

To evaluate the success of the pilot project, we interviewed the 3 RHA project employees (Project Manager, Business Intelligence Partner, and Program Coordinator) that interact directly with the community care organizations on a day to day basis. We also spoke with Jamie Stevens, the Director of Performance Management at the RHA. According to them, quality and consistency of care is the number one priority. It ensures that each patient in the region will receive the same frequency and level of care for the same care needs. These are some of the indicators used to measure the impact of this project:

1) **Ease of the Admission Process**: Ensuring that every admission is followed up by an assessment which is used to classify patients into the appropriate population groups that best suits their care needs. It is also important that these assessments don't become a bottleneck since that requires caregivers to give the patient a call before traveling to their homes for a booked appointment. In addition, it is important that Community care organization staff can lookup and retrieve similar assessments from other organizations in the area if that was done in the recent past. Before the pilot project, this information was not readily available.
2) **Ease of Referral by the Community care organizations**: The referral process for community care should be easy, with support for proper electronic trails. These organizations should be able to send electronic referrals to each other on behalf of the patient.

3) **Wait Times**: Long wait times mean patients wait too long for the care they need. This is measured from the time of admission into a service to the initial service appointment or visit for the patient. Wait times need to be kept as short as possible if the funding and care resources required are readily available.

4) **Missed Visits**: Missed visits means either the healthcare service provider missed an appointment with the patient or when a patient is unable to make or be available for this appointment with the healthcare service provider. This is usually a result of factors such as urgent hospitalization of the patient, administrative errors such as communicating patient locations and logistical constraints like bad weather conditions.

5) **Hospitalizations**: The biggest measure of the success of community care, according to Jamie is to reduce hospitalization. In Ontario, it costs over $3,000 CAD every day a patient is admitted to the hospital. This is a huge expense to the healthcare system. While hospitalization may not always be avoided, the target is to reduce the incidence of those by ensuring the proper care of the chronic and complex patients from the comfort of their homes.

6) **Rich Service Level Data**: Before the pilot project, the community care organizations worked in silos. There is no visibility into the services their patients receive from other organizations. However, reports such as the "Agency match profile" report currently provide a rich context of service level data across the community care organizations for each patient. Healthcare service providers are better equipped to ask the right questions and provide better care to their patient or clients.

The table below summarizes the feedback, in terms of how well the evaluation criteria were addressed, from those who participated in the pilot project.

Table 8-4 Home and Community Care Evaluation Criteria

Evaluation Criteria	Measure Impact	Description
Ease of the Admission Process	High	• Having the community care organizations use the same Patient Information System means that the admission processes – questionnaire, forms are very similar. Also, community care organizations are now able to share completed assessments with ease since they share a common application and data formats. • These impact the patient positively as it reduces duplication of efforts during admissions. CCO staff can see the last assessment across all the CCOs and can easily seek the transfer of these forms where needed. • The overall patient experience is more positive because the length of the admission process is drastically reduced as a result.
Ease of Referral by the community care organizations	High	• By using a common platform, referrals are sent electronically through the Patient Management System instead of faxes. The old method requires Community care organization staff to fax a referral, which then prompts fresh data entries and assessments for the patient. With electronic referrals, many of these manual steps are eliminated.
Wait Times	Fair	• Many factors affect wait times. However, the impact of the pilot project on wait times could not be measured easily. However, we could infer that by streamlining and shortening the admission process, the administrative component of wait times is greatly reduced.
Missed Visits	Very High	• Missed visits usually occur because either a patient isn't available because of an urgent hospitalization or they are unable to make an appointment. By incorporating Emergency Room data feed from the RHA and matching those to CCO clients, the CCOs receive daily status reports that tell them the patients that are unavailable for the day. Today, Community care organization staff uses data provided to follow-up with recently hospitalized patients to determine if their appointment should be kept or be canceled.
Reduced Hospitalizations	Not Sure	• Hospitalizations occur for various reasons. While it is hard to determine the direct impact of the pilot project on hospitalizations, we believe that it has greatly improved the care processes for the community care organizations which indirectly impacts quality and nature of patient experience to care. If a patient outlook is improving, those might have some impact on the frequency of hospitalization and early discharge of patients.
Rich Service-Level Data	High	• CCOs employees have access to performance management data for each of their patients, across the region. For those 17 CCOs

		that have signed the DSA, the impact of the rich service-level data is high. Many of these CCOs have come to depend on this data for the management of their patients.

8.3. Evaluation of Configurable Anonymization for Privacy Compliance Experiment

The configurable anonymization for privacy compliance experiment described anonymization methods and workflow for supporting automated, configuration-driven anonymization for cloud-hosted data in support for performance management for data recipients with varying risk profiles.

Our evaluation uses this experiment to compare the approaches used in our Pilot project and with ARX (Prasser, Kohlmayer, Lautenschläger, & Kuhn, 2014), the most popular open source anonymization utility available for biomedical healthcare data. ARX is a powerful open source anonymization tool with rich privacy compliance and anonymization features. It is fast, supports fairly large datasets (a few million records), rich GUI and very flexible API support in Java. ARX also supports automated data import/export from CSV and relational data sources. It also supports generalization and suppression for high-dimensional data with up to 30 quasi-identifiers (Prasser et al., 2016).

Table 8-5 below evaluates the features and requirements of our pilot project and anonymization experiment with ARX based on the evaluation criteria identified in section 3.3.6.

Table 8-5 Evaluation of Privacy Compliance approaches and Tools

Evaluation Criteria	Pilot Project	Anonymization Experiment	ARX Tool
Electronic DSA Support (Organization Consent)	• Yes – Applied at the time of data import. Non-consented data	• Yes – This is fully supported at the time of data import.	No – DSA is implied. Data is processed outside the tool, and only those that

	sources are not imported or processed.	Non-consented data sources are not imported	require anonymization are imported.
Patient Consent	• Yes – Non-consenting patient records are dropped.	• Yes – Non-consenting patient data/records are anonymized	No – Patient Consent is implied. All records are anonymization.
Declarative attribute mapping	• Yes – Configuration resource files	• Yes – Configuration resource files	Yes – through the maintenance API.
Supported Identifiers – DI, QI, and SAs	• No (all-or-nothing)	• Direct Identifiers, Quasi-identifiers, Sensitive Attributes	Quasi-identifiers, and Sensitive Attributes
Support for resource-based libraries	• No (all-or-nothing)	• Yes – XML Resource files for QI hierarchies, and DI gazetteers	Yes – External CSV files
Support for DI masking/Pseudonymization	• No (all-or-nothing)	• Yes – Supports masking, pseudonymization, and format preserving encryption	None
Privacy Criteria	• None implemented	• k-anonymity, l-diversity	k-anonymity, l-diversity, and t-closeness.
Support for High-dimensional datasets	• No anonymization	• Yes – Supports both cross-sectional data sets with many QIs as well as high-dimensional longitudinal data sets	No – Only cross-sectional data sets with many QIs (up to a maximum of 15 attributes)
Support for risk assessment and measurement	• No	• Prosecutor Risk • Journalist Risk	• Prosecutor Risk • Journalist Risk
Supported Anonymization Approaches	• All-or-nothing	• Generalization • Generalization with an automatic optimal solution • No local recoding (Desirable if supported by the anonymization tool) • Tuple Suppression • Date Shifting	• Generalization • Generalization with an automatic optimal solution (Lightening, Flash) • Support for local recoding • Support for tuple suppression
Support for selective anonymization	• No	• Yes	No
Support for data model complexity reduction	• No	• Yes	Limited

In summary, ARX has flexible data import/export features with support for some RDMS data sources. But it works best with CSV data. Our anonymization algorithm works with similar data sources as well as semi-structured data imports and transformations. In the pilot project, anonymization was not supported. Non-consenting patient data is simply dropped. ARX has very strong API support for most of its features. Therefore, one can programmatically initialize the tool with the right import source/destination and attribute definitions.

In addition, we have observed is that most anonymization tools tend to focus solely on either direct identifier masking or quasi-identifier anonymization. ARX follows this trend and supports the anonymization of quasi-identifiers and sensitive attributes. Our anonymization algorithm supports all identifiers and anonymization processing for those identifiers.

The ability to support externally loaded hierarchies is very important for automated cloud-based anonymization. ARX allows configuration using a set of maintenance APIs. Using the API interfaces to ARX, these resources can be loaded declaratively or dynamically. Both our algorithm and ARX have full support for the major privacy criteria for triggering the anonymization modules. ARX also supports datasets with many generalization hierarchies.

Finally, our algorithm ensures that data sets are anonymized not only by its content but also takes into account the intended recipient as defined in the Privacy Compliance Definition Document. This is not a standard anonymization behaviour since most anonymization tools work on entire datasets, not cohorts, and individual patient profiles. Basically, it means the target data set can include records from some patients from the data recipient organization, and others from other data streams, from other community care organizations. ARX is a tool for anonymizing whole data sets with the assumption that the recipient is an external stakeholder. ARX supports most of the

required anonymization functions required for our architecture. However, it has limited support for complex data models such as those required for defining care episodes and events. It also has no support for selective anonymization.

8.4. Evaluation of Configurable Patient Identity Matching Experiment

This section evaluates the Patient Matching Service experiment described in Chapter 7 with similar systems based on their matching algorithms. Record linking algorithms are categorized based on their complexity as Basic (uses deterministic approach only), Intermediate (uses fuzzy logic and weights), and Advanced (uses automated weight allocation, statistical, data mining, and machine learning approaches) (Just et al., 2009). While the matching algorithm for our identity matching service is in the intermediate category, this study corroborates our work by pointing out that record linking algorithms must also incorporate mechanisms for using manual matching to handle ambiguities and reduce false positives.

A summary comparison of algorithms in each class is presented in Table 8-6. While various algorithms have various strengths regarding accuracy and precision, given the appropriate conditions, our evaluation compares the algorithms on their flexibility for adaptation to environments similar to those required for a cloud-hosted service as described in our experiment.

Table 8-6 Patient Identity Service comparison with related algorithms.

Evaluation Criteria	Our Algorithm	Sachs et al. (2000)	Sauleau et al. (2005)	Méray et al. (2007)
Algorithm Complexity	Intermediate (Probabilistic)	Basic (Deterministic)	Advanced (Clustering)	Intermediate (Deterministic & Probabilistic)
Support for Identity Attribute Standardization	Yes – Definition Driven	Yes – Phonetic Roots	Yes	Very minimal and offline (Requires error-free data)
Block definitions	Yes – Definition Driven	No match blocks	Static	Static

	Supports multiple attributes		Single attributes per block	Single attribute per block
Weight determination	Definition-driven. Weights determined through offline analysis	N/A	Static	Algorithmic
Support for Multi-Criteria Matching or Match Blocks	Yes – Definition Driven Supports multiple match blocks	No – Uses a unique identifier and one criterion for other attributes	No – Only one criterion based on FN and DOB	Yes – Static Blocks No support multiple match blocks
Use of Phonetic Roots	Not supported	Yes	No evidence	No evidence
Management of ambiguities	Yes – Web-based online management interfaces	No evidence	No evidence	Yes – Offline review
Distributed Match Processing	Limited – Cloud Containerization	No evidence	No evidence	No evidence
Organization and Patient Compliance	Data Sharing Agreements (DSA) – Organization and Patient Consent	No evidence	No evidence	No evidence

Sachs et al. (2000) describe an algorithm that employs a variation of deterministic record linkage using a unique identifier and basic matching on other identifying attributes while employing phonetic roots of first names to reduce mismatches from data entry errors and misspellings (Sachs et al., 2000). The use of phonetic roots for first names is one improvement that could also be made to our data standardization process. Sauleu et al. (2005) employ an approximate string matching technique with clustering. Like our identity matching algorithm, it employs a data standardization phase and uses a weighted blocking technique for matching. However, this blocking technique creates overlapping subsets called "canopies" based on records within a loose threshold distance from a cluster center computed from some fixed blocks, derived from substrings in the patient first name (FN) and date of birth (DOB). While this algorithm is relatively complex, it is not production ready because it is specifically built to use only FN, and DOB values for building its match clusters.

If we compared these three systems, Sachs et al. (2000) would not perform well since it is deterministic, and patients across our data streams do not share a common identity attribute. Sauleu et al. (2005) use an advanced record linking algorithm but depends on FN, and DOB attributes only. With only two identity attributes, it would also perform poorly for large datasets and for healthcare environments with very little tolerance for errors. Our algorithm uses a large list of patient personal attributes with various weights, providing more context for richer probabilistic matches. Nevertheless, the matching system described by Méray et al. (2007) is the closest to our approach. It employs match blocks but with only one attribute per block with weights computed algorithmically. Our algorithm needs to be extended to use algorithmic techniques to determine the appropriate weights for each match block using data characteristics such as attribute missing value rate, error rate, value distribution, and uniqueness. Using phonetic roots with first names is also one area of improvement that would help increase the match rates for blocks with first name attribute values.

8.5. Evaluation of Related Work

This section describes four alternative solutions to the problem addressed by the Cloud-based Surveillance and Performance Management Architecture for Community Healthcare. Like our architecture, these are other approaches to cloud-based connected healthcare as identified from related work (section 2.5). They are also representative of the popular approaches that are being adopted in academic literature to address interoperability and performance management in a cloud computing environment. These approaches, introduced in section 2.5, are summarized below:

- **Software-as-a-Service** - The main objective of the Software-as-a-Service approach (described in section 2.5.1) is to create an aggregate patient EHR on the cloud, requiring providers to adapt their datasets to match the interfaces of the SaaS infrastructure. As described in Bahga & Madisetti (2013), this approach provides little support for dynamic analytics result push to the participating organizations and external stakeholders. Users are always required to access the portal to find important details on their patients, a practice that hasn't work too well with most caregivers.
- **Peer-to-Peer** – This approach described in detail in section 2.5.2 provides a good framework for protecting patient privacy since patient data is kept at the source until it is queried. The work by Donnelly et al. (2014) is the use case for evaluating this approach. Replacing identifiers is never enough for protecting patient privacy, especially in the healthcare domain with high-dimensional events. On the other hand, our architecture considers organization and patient compliance in the implementation of privacy protection for data elements. Essentially, full anonymization is applied to data returned to external partners that are not data custodians. For those that are, partial anonymization and sometimes only patient profiles within their organization with local identifying details are released alongside the shared services details from other participating organizations.
- **PaaS Containerization** – PaaS containerization approach (described in section 2.5.3) is a very scalable and configurable approach to data collection using cloud infrastructure. The work by Andry et al. (2015) illustrates the benefits of this approach. Our framework takes this idea a little further by leveraging PaaS containerization to build various IaaS models in very generic ways for systematic data collection, reporting, and subscription services.

- **Semantic Web** - Semantic web technologies use the cloud as a medium for exchanging RDF documents through common semantically annotated models that requires data owners to incorporate data transfer agents within their data domain that respond to data requests from other agents. The work by Sinaci & Laleci Erturkmen (2013) describes this approach in more details (see section 2.5.4). This is a scenario we avoided in our architecture because it requires 1) specialized customizations for local EMR applications within each organizational domain and 2) use the cloud as a proxy for data exchange. Instead of using the cloud as a proxy for interoperability, our approach leverages the cloud as the platform for data surveillance and performance management using a CDM. This way, data attributes that are semantically similar are not ambiguous between the data producers and requestors.

Sections below show the general evaluation matrix for comparing the approach taken in this thesis and those of the four related frameworks. Our evaluation compares these frameworks on the basis of their approaches to meeting triple aim objectives, supporting performance management services, achieving interoperability, implementing a common data model, supporting patient identity management, and achieving privacy compliance.

8.5.1 Triple Aim Objectives

Our architecture support for measurable performance management goals for a health region, directly and indirectly, provides the means for operationalizing the Triple Aim objectives. We evaluate the four identified approaches against our approach on measurable performance management goals for quality of care processes in support of Triple Aim.

Table 8-7 Related Work Comparison based on Triple Aim Objectives

Evaluation Criteria	Our Experiment	Software-as-a-Service (Bahga & Madisetti, 2013)	Peer-to-Peer (Donnelly et al., 2014)	Containerization (Andry et al., 2015)	Semantic Web (Sinaci & Laleci Erturkmen, 2013)
Improve Patient Experience/Care Delivery	Very likely – More streamlined health processes, ease of collaboration and coordination of caregivers	Possibly – Patient portal contain aggregate data for patients and caregivers.	No evidence	Likely based on the architecture but no evidence in the literature.	No evidence
Improvements to overall population health	Possibly – It is difficult to determine the impact of this experiment on population health.	Possibly – Patients and caregivers have access to data, and that must improve their overall awareness and knowledge	No evidence	No evidence	No evidence
Cost savings reduction to the healthcare system	Yes - reduced service duplication and improved system level coordination.	Yes – Eliminates data silos.	Likely – No evidence of implementation.	Likely – No evidence of implementation.	Likely – Based on the architecture.

Table 8-7 shows the evaluation of these approaches to meeting Triple Aim objectives. Meeting Triple Aim objectives is a significant component of our pilot project (Chapter 5). We recognize that it may not always be feasible to measure objectives directly like patient experience. However, by reducing or eliminating the factors that trigger negative patient experiences, for example, the length of time a patient is waitlisted for community services or notifying caregivers about changes in patient care status, so they can respond faster to patient needs, we are indirectly impacting patient experience positively. Also, providing the RHA and the community care organizations with the capacity to consolidate duplicate services and stream their processes through efficient collaboration, do result in huge cost savings to the health care system,

therefore freeing up funds needed to provide more services to those patients with more complex and chronic health conditions.

8.5.2 Surveillance Services Interoperability

Software-as-as-Service efforts such as CHISTAR (Bahga & Madisetti, 2013) achieve cloud-based data integration and semantic interoperability using archetype models. Services are presented to clients as Web service interfaces. Data from different EHR systems gets converted into flat files that get stored in the Hadoop File System (HDFS) distributed storage. It uses the MapReduce-based bulk loader to load data into HBase. Hive, the data warehouse system for Hadoop is used for analysis.

Peer-to-peer frameworks such as the PACE healthcare architecture (Donnelly et al., 2014) uses a combination of cloud and peer-to-peer technologies to model healthcare units or clinics where Personal Health Information (PHI) is stored in off-cloud data storage while non-identifying data kept in the cloud. This framework uses a hybrid application that anonymizes local patient identifiers while using anonymous identifiers on the cloud.

Table 8-8 Related Work Comparison based on Interoperability Benefits

Evaluation Criteria	Our Experiment	Software-as-a-Service (Bahga & Madisetti, 2013)	Peer-to-Peer (Donnelly et al., 2014)	Containerization (Andry et al., 2015)	Semantic Web (Sinaci & Laleci Erturkmen, 2013)
Ease of Data Exchange (Technical)	Yes - PaaS Containers for applications, Support for Web Service for all components.	Yes – Web Service Interfaces	No - data storage using P2P technologies.	Yes - PaaS containers for data aggregation.	Yes- Semantic Web RDF.
Efficiency of data encoding and	Yes (Limited) - does not use healthcare data	Yes - openEHR("openEHR Architecture," 2015) and	No	No	Yes (Limited) - Federated Metadata

Evaluation Criteria	Our Experiment	Software-as-a-Service (Bahga & Madisetti, 2013)	Peer-to-Peer (Donnelly et al., 2014)	Containerization (Andry et al., 2015)	Semantic Web (Sinaci & Laleci Erturkmen, 2013)
data translation (Semantic)	interoperability standards.	HL7("Introduction to HL7 Standards," 2016)			
Efficiency of collaborative decision making (Process)	Yes - Reporting Portal and Subscriptions.	Somewhat - Patient Portal	No	No	Somewhat - XPATH, SPARQL, SQL

Containerization framework (Andry et al., 2015), make Platform as a Service (PaaS) the central critical layer for an elastic and extensible framework to abstract healthcare services. Semantic Web frameworks (Amato et al., 2013) support document exchanges between heterogeneous data sources over the cloud using semantic web Resource Description Framework (RDF) and Web Ontology Language. Supported data extraction techniques include XPATH (for XML documents), SPARQL (based on RDF) and SQL (for legacy relational databases).

As described in section 2.1, process interoperability is a needed area of research as it is under-researched compared to technical and semantic interoperability. Our framework enables technical and semantic interoperability but more importantly provides support for specific processes (i.e., reporting and subscription) as part of operationalizing performance management.

8.5.1 Performance Management Services

Performance management services include support for analytics infrastructure, dynamic report generation and delivery, and data/report subscriptions.

Table **8-9** summarizes the evaluation of the related work approaches with our approach.

Table 8-9 Related Work Comparison based on Performance Management Services

Evaluation Criteria	Our Experiment	Software-as-a-Service (Bahga & Madisetti, 2013)	Peer-to-Peer (Donnelly et al., 2014)	Containerization (Andry et al., 2015)	Semantic Web (Sinaci & Laleci Erturkmen, 2013)
Support for an Analytics Infrastructure	Yes – BI reports and general analytics. Supports data anonymization for privacy compliance.	Yes – Reports on Patient Portal	No evidence	No evidence	No evidence
Dynamic Analytics Report Generation	Yes	No evidence	No evidence	No evidence	No evidence
Data/Report Subscription	Yes	No evidence	No evidence	Somewhat - Event notification	No evidence

Our evaluation shows that related work lack support for performance management exception for the SaaS use case. However, patient reporting is hosted and derived through the patient portal. We have no evidence in the literature to confirm that the Peer-to-Peer, PaaS Containerization, and Semantic Web approaches support these services since they are designed for data querying across peers with built-in support for data transformation.

8.5.2 Common Data Model

In this section, we evaluate these approaches on the features of the CDM and approaches to populating one. Our approach, SaaS, and containerization employ a database-driven common data model. Peer-to-Peer and Semantic Web assume a distributed model where data is spread out across collaborating stakeholders and put together in response to queries.

Therefore, some of the evaluation criteria evaluation shown in Table 8-10 may not apply to these two approaches. However, for the approaches that support a CDM, using a loose architecture helps in incorporating new data streams into the model.

While our architecture uses a clustered SQL Server database (RDMS), it isn't necessarily the best if the data sources are based on unstructured or semi-structured datasets. Rather, for the scenario used for our pilot projects, an RDMS provides the best utility for the performance management services.

Table 8-10 Related Work Comparison based on the use of a Common Data Model

Evaluation Criteria	Our Experiment	Software-as-a-Service (Bahga & Madisetti, 2013)	Peer-to-Peer (Donnelly et al., 2014)	Containerization (Andry et al., 2015)	Semantic Web (Sinaci & Laleci Erturkmen, 2013)
Data Collection.	Conversion to a community health common data model (Structured RDMS)	Conversion from other standards to openEHR.	N/A – Data is distributed and resides within each PACE peer.	No evidence	Conversion to RDF/XML
Data Structure Definition	Yes – Declarative XML mapping definitions	Yes - Supports multiple drivers. These definitions are embedded in archetype models	Limited – Depends on each peer.	None – Data is stored in blobs.	None
Support for batch, and streaming Datasets	Batch Processing Limited Stream processing – Pilot projects updated the CDM nightly. However, this can be done incrementally to provide support for streaming data.	Batch Processing. No evidence of stream processing	No Batch or Stream processing. Real-time response to queries from peers is supported.	Limited data aggregation but supports real-time analytics of aggregate log events.	No – Focus is on aggregating semantic metadata registries.

	Ease of change of model	High - Average – CDM is model-driven. Surveillance processes are declarative.	Average – CDM is fixed but based on a NoSQL structure that can be changed as needed to support downstream processes.	Low – Model changes require changes to the peers and their associated drivers. That can be complex and expensive.	Low – Logic is built into each container. Changing that can be complex.	Average – CDM Definition-driven.
	Scalability of CDM	Average to High – Uses a clustered relational database	High – Hadoop data warehouse system.	High – there is no limitation to the number of peers	High – Uses a blob store deployed in a Cloud Foundry (Open PaaS)	Average to High – But dependent on the environment that is used to host the Master Data Records (MDR)

In our experiment, changes to the schema of the data sources to the Systematic Data Collection Service do not affect the CDM since it employs the LAV approach (Katsis & Papakonstantinou, 2017) with declarative mappings. However, changes to the CDM would require modifications to LAV mappings for all data sources.

8.5.3 Patient Identity Management

Patient identity management is necessary when a patient cannot be identified by a generally recognized identifier. In a hospital setting, most patients are identified by a government identifier, but this is not always the case in community care. As shown in Table 8-11, the use of probabilistic matching is not very common. The SaaS uses cryptographic hashes as checksums to patient records but assumes a uniform identity across all data sources.

For Peer-to-Peer approaches, a globally recognized identity is required to identify each patient data across collaborating peers. The approach completely fails if patient identity is ambiguous, as is typically seen in community health care.

Table 8-11 Related Work Comparison based on Patient Identity Management approaches

Evaluation Criteria		Our Experiment	Software-as-a-Service (Bahga & Madisetti, 2013)	Peer-to-Peer (Donnelly et al., 2014)	Containerization (Andry et al., 2015)	Semantic Web (Sinaci & Laleci Erturkmen, 2013)
	Type of Matching	Probabilistic matching of Patient profiles	Cryptographic Hashes	No evidence	Yes - Assigned Patient Identifier	None
	Support for Phonetic Roots	No – Planned for the future	No – Assumes a government issued identifier for each patient	No evidence	No evidence	No evidence
	Support for Declarative Match Definition	Yes – Identity Matching, Transformations are declarative	No evidence	No evidence	No evidence	Yes – SPARQL queries are declarative.

8.5.4 Privacy Compliance Model

A Privacy Compliance Model is an important component of any cloud-based surveillance and performance management architecture. We evaluate related work approaches to privacy compliance by reviewing their support for data sharing agreements, patient consent, and anonymization of patient records during data sharing.

Evidently, our approach incorporates privacy compliance in all the stages of the process and performance management workflows. Other approaches such as SaaS and containerization have security measures for data exchanges and data at rest but lack privacy compliance measures to protect patient identity with data sharing.

Table 8-12 Related Work Comparison based on Privacy Compliance approaches

Evaluation Criteria		Our Experiment	Software-as-a-Service (Bahga & Madisetti, 2013)	Peer-to-Peer (Donnelly et al., 2014)	Containerization (Andry et al., 2015)	Semantic Web (Sinaci & Laleci Erturkmen, 2013)
	Nature of Data Sharing Agreements	Supports explicit electronic DSAs	Implied	Implied	Implied	Implied
	Patient Consent	Yes – Supported in patient records	No evidence	Implied	Yes	Implied
	On-demand Anonymization	Yes - Done for external data recipients	No - attribute encryption for data exchanges	Limited – Identifier Masking	No	No

8.6. Assumptions, Limitations, and Threats to Validity

This thesis proposes a cloud computing architecture for surveillance and performance management for community healthcare. Sections 8.6.1 and 8.6.2 describes some of the assumptions and limitations of our research, while section 8.6.3 discusses potential threats to the validity of our work.

8.6.1 Assumptions

Our approach to surveillance and performance management works under the following assumptions:

1. **Regional Health Authority** - There is a central body responsible for healthcare in the region that either coordinate care or regulates care or both. It has the authority to host and integrate data for performance management (as defined by appropriate data sharing agreements). If such a body does not exist there may be legal, privacy or trust issues that make it impractical for the RHA to provide a private cloud infrastructure in which data from community care organizations can be hosted and shared.

2. **Anonymization Algorithm** – Assumption that anonymization is sufficient to address privacy and re-identification concerns. Risk-based anonymization does not result in zero risks but reduces re-identification risk to an allowable threshold. The configurability of our approach ensures that the RHA can control this threshold. In some jurisdictions where zero risk is required, the all-or-nothing approach must be used to address privacy compliance.
3. **Identity Matching Algorithm** - The algorithm presented as part of our architecture is only required where patients cannot be identified uniquely by a singular identifier across all participating community care organizations in the health region. If each patient can be identified uniquely, then identity matching service is reduced to a simple deterministic matching service. However, it is still useful to verify that each Community care organization has correct patient information (age, date of birth, contacts, etc.) and flag them for correction if they do not.
4. **Adoption of Architecture** – The architecture described in this thesis addresses data integration for performance management of community healthcare and configuring a declaratively configured pipeline that handles identity matching and privacy compliance in that context. We do not claim that we have the best algorithms for identity matching or for anonymization, but our architecture and declarative configuration gives sufficient flexibility for RHAs to optimize for their particular context. Other contexts, settings, and domains outside community healthcare may find one or more components of this architecture interesting for their adoption.

8.6.2 Limitations

The following limitations are identified:

1. **Cloud Computing and the Nature of Healthcare data** – Healthcare data is very sensitive. In the pilot project and all our experiments, we worked with data in a private cloud funded and hosted by the RHA for the community care organizations. While hybrid and public cloud infrastructures are common outside healthcare, they could not be used for our work because of the sensitivity of healthcare data and privacy protection laws and regulations.
2. **Community Care Model**– Our work is based on performance management of shared healthcare services - in the community healthcare as is typical in Ontario, Canada. There may be other approaches to organizing and managing community healthcare that is different for which this approach does not apply. Components of this work also apply to other similar surveillance and performance management needs. However, Patient Identity Management may not apply to surveillance infrastructure within the same organization with multiple branch office that shares a common patient database or in countries where every patient is issued a common and easily validated identifier.
3. **The maturity of Privacy Compliance Tools** – While our work presents an approach to managing privacy compliance using anonymization, we have not recommended specific anonymization tools for achieving privacy compliance. Achieving our anonymization model requires a tool that supports various anonymization functions such as risk measurement, generalization with random replacement, suppression, Pseudonymization, various masking functions, and date shifting on cross-sectional, longitudinal, and high dimensional data sets. In addition, the tool must have simple to use API support to its features so it can be customized to work with a cloud computing infra-

structure and workflow. Finally, such tools must be able to run in a distributed fashion for the most part. There is currently no anonymization tool in the market today that meets all these requirements. The closest is ARX (Prasser et al., 2016, 2014); an open source project can be extended to support the features introduced in this thesis.

4. **Incorporating an accountability framework to Data Sharing Agreements** – The DSAs used in this thesis are operationally tuned for the pilot projects and the experiments. It doesn't use proper accountability frameworks such as the eXtensible Access Control Markup Language (XACML) that allows one to provide broader privacy definitions in terms of data usage including the purpose of use, obligations associated with the use and retention conditions.

8.6.3 Threats to Validity

The following threats to the validity of our work are identified:

1. **Reproducing results** – Most important threat to validity is that it was done in one health region and only by our team at the LHIN. It is not obvious that anyone else could do the same. More case studies are needed to see if others can repeat what was done.
2. **Type and Nature of data** - There are a few assumptions regarding the type and nature of data used and the nature of the community care organizations involved. For example, if some of the community care organizations decided to host their operational database locally, that would increase the complexity of the surveillance infrastructure.

3. **Privacy Laws and Legislations** – Privacy legislation in most countries dictate the methods for handling and sharing healthcare data. These laws and approaches to compliance, especially with respect to anonymization are not the same everywhere (variation of the one region limitation), but most importantly, they could change over time.

Chapter 9. Conclusions and Future Work

This thesis was motivated by and benefited from years of work on a regional initiative of the Champlain Local Health Integration Network to bring together the community care organizations in the Champlain region to a common platform while solving the problems associated with interoperability, data transformations, patient identity management, and performance management. The Cloud-based Surveillance and Performance Management Architecture for Community Healthcare presented in thesis provides the path to addressing these challenges.

The work done in setting up both the cloud and data infrastructure provided us with important insights into the unique interoperability challenges with community healthcare. In Chapter 3, we reviewed the state-of-the-art in community healthcare, identified some of the gaps between existing frameworks from related work, and used those to identify the evaluation criteria for our contributions. Chapter 4 introduced our Surveillance and Performance Management architecture. The pilot project described in Chapter 5 demonstrates a successful implementation of this architecture. We then leverage the experiments in Chapter 6 and Chapter 7 to drive our gap analysis on some of the unique challenges of patient identity and privacy compliance with community healthcare.

9.1. Recap of Thesis Contributions

In summary, we wish to recap the major contributions of our thesis from Chapter 1 while highlighting the significance of each contribution to surveillance and performance management of community health care.

- **A cloud computing infrastructure that provides surveillance and performance management services using a multi-tenanted private cloud owned and operated by a regional health authority that can host applications and operational databases (the entire server infrastructure of each organization) for community care healthcare stakeholders.**

  We are able to show that a multi-tenanted cloud infrastructure provides the environment for supporting systematic hosting (of small organization IT infrastructure and applications), systematic data collection, and systematic performance management, as long as the infrastructure is owned by the Regional Health Authority or a trusted 3rd party to all participating community care healthcare stakeholders. Leveraging a private cloud infrastructure is ideal since it is scalable, easy to manage, and addresses issues with data security and privacy concerns usually associated with public cloud infrastructure.

  One of the key advantages of such a regional performance management infrastructure is that it is attractive for small Community care healthcare organizations and can easily be adopted at a little cost, allowing them to operationalize insights to regional performance management data.

- **A Common Data Model (CDM) that provides a consistent view of information across multiple and disparate data sources with a collection of services, APIs, and tools for sharing (collecting, correlating, processing, analyzing, mining, and reporting) data.**

  A common data model fills the gap for leveraging surveillance data. It provides a common, minimal representative dataset that is required for performance management. The

common data model ensures that 1) Surveillance processes transform incoming data streams to a common format and structure, 2) Downstream performance management services are designed with a common model in mind, thereby increasing sustainability and re-usability of the data.

The performance management services - analytics, reporting, subscription, and anonymization, all leverage a CDM instead of having very complex logic for dealing with each data source. By decoupling data collection, transformation, and some privacy compliance requirements from these services, we are able to use generic designs for these performance management services.

- **A Patient Identity Matching Service for correlating cloud-hosted data from multiple stakeholders into a common data model to support the performance management of community healthcare.**

  Identity Management is an important component of our architecture. Without identifying patient profiles across all stakeholders, it is impossible for privacy compliance, reporting, and subscription components to operate as expected. The Patient Identity Matching Service fills this gap, ensuring that patient profiles across community care organizations are correlated correctly, even with various data entry errors in the input data.

- **A declarative privacy compliance definition document linked with Patient Consent forms and Organizational Data Sharing Agreements that configure the processes and services of the cloud-computing infrastructure, including anonymization to ensure legal compliance and a systematic approach to data governance.**

Privacy Compliance is a challenge with many data surveillance and performance management infrastructures. This challenge is magnified by the use of cloud services. We provided an anonymization framework for implementing organization and Patient Consents. This architecture embeds privacy compliance in all its components.

- **A Patient Identity Matching algorithm that employs a weighted probabilistic matching system that can correlate data from a variety of data sources.**

  The patient identity matching algorithm presented in our Chapter 7 experiment applies a probabilistic matching algorithm with automatic data transformation to correlated patient profile and demographic data across multiple input streams using heuristically tuned weights.

- **A configurable anonymization algorithm that uses the privacy compliance definitions to handle the anonymization of Direct Identifiers (DI), Quasi-Identifiers (QI), and Sensitive Attributes (SA) in high-dimensional datasets**

  The configurable anonymization algorithm presented in Chapter 6 handles anonymization for high-dimensional data sets with mixed patient contents, instead of the traditional all-or-nothing or full anonymization of entire data sets that are commonly seen with other approaches. Also, this algorithm handles the anonymization of all identifiers – DIs, QIs, and SAs based on the infrastructure Privacy Compliance Definition Document.

9.2. Future Work

The following areas identified in the course of writing this thesis could be addressed in future work:

9.2.1 Distributed Data Custodian Model

In this thesis, community care organizations have a central funding body that provided the cloud infrastructure for application hosting. If that didn't exist, or if some of the community care organizations are profit-making private vendors, collaboration may be difficult or impossible. Future research could look at a distributed data custodian model where various data custodians can push data to a centralized cloud infrastructure hosting the CDM and other downstream performance management services while allowing each organization to manage the hosting and transformation of their data to match the CDM.

9.2.2 High-dimensional data Anonymization tools

In this thesis, achieving the anonymization model described in Chapter 6, requires a tool that supports various anonymization functions for high-dimensional data sets including re-identification risk modeling, risk measurement, data attribute generalization with support for local recoding and random replacement, attribute suppression, Pseudonymization, various masking functions, and date shifting on regular and high dimensional data sets with episodic and event data. Future work can address this by integrating various anonymization algorithms and techniques into a single anonymization tool that supports these functions in a distributed fashion as required to support a cloud-based performance management infrastructure and workflows.

9.2.3 Support for Unstructured clinical notes

The architecture described in this thesis has no support for unstructured clinical notes in the data extraction processes. Future work can look into incorporating machine learning

techniques for extracting structured data from clinical notes and providing support for that in the CDM and all downstream performance management services.

9.2.4 Analytics Services

A major component of the performance management architecture described in this thesis is the analytics services. This service should leverage the CDM and the identity management service to support patient profiling, clustering techniques, association rules, attribute correlation, and other data mining descriptive and predictive analysis techniques to support healthcare providers and improve health care decisions and outcomes. Developing various components of this service should be addressed in future work.

9.2.5 Algorithmic Approaches for Calculating Match Block Weights

The Configurable Identity Matching Service uses rules or match blocks to configure the behaviour of the probabilistic matching algorithm. However, the weights associated with these match blocks are determined heuristically. Future work can investigate the use of algorithmic approaches to assigning weights to match blocks using known characteristics of match block attributes such as data missingness, uniqueness, distribution, and error rate.

9.2.6 Incorporating Accountability Framework to Data Sharing Agreements

The DSAs used in this thesis are operationally tuned for the pilot projects and the experiments. This is a limitation that should be addressed in future work. The accountability framework should provide a framework for supporting broader privacy definitions in

terms of data usage, including the purpose of use, obligations associated with the use and retention conditions.

Bibliography

Adler-Milstein, J., & Jha, A. K. (2012). Sharing clinical data electronically: A critical challenge for fixing the health care system. *JAMA - Journal of the American Medical Association*, *307*(16). https://doi.org/10.1001/jama.2012.525

Amato, F., Mazzeo, A., Moscato, V., & Picariello, A. (2013). A Framework for Semantic Interoperability over the Cloud. *2013 27th International Conference on Advanced Information Networking and Applications Workshops*, 1259–1264. https://doi.org/10.1109/WAINA.2013.218

Andry, F., Ridolfo, R., & Huffman, J. (2015). Migrating healthcare applications to the cloud through containerization and service brokering. *HEALTHINF 2015 - 8th International Conference on Health Informatics, Proceedings; Part of 8th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2015*, 164–171. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84938849051&partnerID=tZOtx3y1

Aziz, B., Arenas, A., & Wilson, M. (2011). SecPAL4DSA: A policy language for specifying data sharing agreements. *Communications in Computer and Information Science, 186 CCIS*, 29–36. https://doi.org/10.1007/978-3-642-22339-6_4

Bahga, A., & Madisetti, V. K. (2013). A cloud-based approach for interoperable electronic health records (EHRs). *IEEE Journal of Biomedical and Health Informatics*, *17*(5), 894–906. https://doi.org/10.1109/JBHI.2013.2257818

Benson, T. (2012). Why Interoperability is Hard. In *Health Information Technology Standards*. *Principles of Health Interoperability HL7 and SNOMED SE - 2* (pp. 21–32). https://doi.org/10.1007/978-1-4471-2801-4_2

Berler, A., & Apostolakis, I. (2014). Research Perspectives on the Role of Informatics in Health Policy and Management. In C. El Morr (Ed.), *Research Perspectives on the Role of Informatics in Health Policy and Management* (pp. 168–207). https://doi.org/10.4018/978-1-4666-4321-5

Berwick, D. M., Nolan, T. W., & Whittington, J. (2008). The triple aim: care, health, and cost. *Health Affairs*, *27*(3), 759–769.

Bhaskaran, S., Suryanarayana, G., Basu, A., & Joseph, R. (2013). Cloud-Enabled Search for Disparate Healthcare Data: A Case Study. *2013 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, 1–8. https://doi.org/10.1109/CCEM.2013.6684431

Biswas, S., Akhter, T., Kaiser, M. S., & Mamun, S. A. (2014). Cloud based healthcare application architecture and electronic medical record mining: An integrated approach to improve healthcare system. *2014 17th International Conference on Computer and Information Technology (ICCIT)*, 286–291. https://doi.org/10.1109/ICCITechn.2014.7073139

Bohmer, R. M. J. (2016). The hard work of health care transformation. *New England Journal of*

Medicine, *375*(8), 709–711.

Boissonnealth, P., & Lafreniere, N. (2014). Deploying Information Systems throughout the Community Care Sector of the Champlain Region. *E-Health 2014 Conference, Vancouver, BC*. Retrieved from http://www.e-healthconference.com/pastpresentations/2015/201462198156/CS122.pdf

Caimi, C., Gambardella, C., Manea, M., Petrocchi, M., & Stella, D. (2016). Legal and technical perspectives in data sharing agreements definition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9484*, 178–192. https://doi.org/10.1007/978-3-319-31456-3_10

Cao, J., Carminati, B., Ferrari, E., & Tan, K. L. (2011). CASTLE: Continuously anonymizing data streams. *IEEE Transactions on Dependable and Secure Computing*, *8*(3), 337–352. https://doi.org/10.1109/TDSC.2009.47

Cavoukian, A., & Emam, K. El. (2011). Dispelling the Myths Surrounding Anonymization Remains a Strong Tool for Protecting Privacy. *Information and Privacy Commissioner, Ontario, Canada*, (June). Retrieved from http://www.ipc.on.ca/images/Resources/anonymization.pdf

CDC. (2018). What is Population Health. Retrieved September 26, 2018, from https://www.cdc.gov/pophealthtraining/whatis.html

Chalasani, S., Jain, P., Dhumal, P., Moghimi, H., & Wickramasinghe, N. (2014). Content architecture applications in healthcare. *Health and Technology*, *4*(1), 11–19. https://doi.org/10.1007/s12553-014-0075-x

CIHR. (2017). Community-Based Primary Health Care. Retrieved July 15, 2017, from Canadian Institutes of Health Research website: http://www.cihr-irsc.gc.ca/e/43626.html

Coats, B., & Acharya, S. (2013). The forecast for electronic health record access. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 937–942. https://doi.org/10.1145/2492517.2500329

Consolidate CDA Overview. (2015). Retrieved December 12, 2015, from https://www.healthit.gov/policy-researchers-implementers/consolidated-cda-overview

CPCSSN Data for Research. (n.d.). Retrieved May 26, 2019, from https://cpcssn.ca/research-resources/cpcssn-data-for-research/

De la Rosa Algarín, A., Demurjian, S. A., Ziminski, T. B., Rivera Sánchez, Y. K., & Kuykendall, R. (2014). Architectures and Protocols for Secure Information Technology Infrastructures. In A. Ruiz-Martinez, R. Marin-Lopez, & F. Pereniguez-Garcia (Eds.), *Architectures and Protocols for Secure Information Technology Infrastructures* (pp. 334–363). https://doi.org/10.4018/978-1-4666-4514-1

de la Torre-Díez, I., González, S., & López-Coronado, M. (2013). EHR Systems in the Spanish Public Health National System: The Lack of Interoperability between Primary and Specialty Care. *Journal of Medical Systems*, *37*(1), 9914. https://doi.org/10.1007/s10916-012-9914-3

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Demster, B., Dooling, J., Kadlec, L., Torzewski, S., Walker, R., Warner, D., & Weidemann, L. A. (2011). Limiting the Use of the Social Security Number in Healthcare. *Journal of AHIMA*, *82*(6), 52–56. Retrieved from http://library.ahima.org/doc?oid=104465#.WHJx3FMrJhE

Dixon, B. E., Vreeman, D. J., & Grannis, S. J. (2014a). The long road to semantic interoperability in support of public health: Experiences from two states. *Journal of Biomedical Informatics*, *49*, 3–8. https://doi.org/10.1016/j.jbi.2014.03.011

Dixon, B. E., Vreeman, D. J., & Grannis, S. J. (2014b). The long road to semantic interoperability in support of public health: Experiences from two states. *Journal of Biomedical Informatics*, *49*, 3–8. https://doi.org/10.1016/j.jbi.2014.03.011

Donnelly, N., Irving, K., & Roantree, M. (2014). Cooperation across Multiple Healthcare Clinics on the Cloud. In K. Magoutis & P. Pietzuch (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-662-43352-2

Dwork, C. (2006). Differential privacy. *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*, 1–12. https://doi.org/10.1007/11787006_1

El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., … Bottomley, J. (2009). A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. *Journal of the American Medical Informatics Association*, *16*(5), 670–682. https://doi.org/10.1197/jamia.M3144

Eze, B., Kuziemsky, C., Lakhani, R., & Peyton, L. (2016). Leveraging Cloud Computing for Systematic Performance Management of Quality of Care. *Procedia Computer Science*, *58*. https://doi.org/10.1016/j.procs.2016.09.048

Eze, B., Kuziemsky, C., & Peyton, L. (2017). Cloud-based performance management of community care services. *Journal of Software: Evolution and Process*. https://doi.org/10.1002/smr.1897

Eze, B., & Peyton, L. (2015). Systematic literature review on the anonymization of high dimensional streaming datasets for health data sharing. *Procedia Computer Science*, *63*. https://doi.org/10.1016/j.procs.2015.08.353

Eze, Benjamin, Kuziemsky, C., Lakhani, R., & Peyton, L. (2016). Leveraging Cloud Computing for Systematic Performance Management of Quality of Care. *Procedia Computer Science*, *98*, 316–323.

Eze, Benjamin, Kuziemsky, C., & Peyton, L. (2017). A Patient Identity Matching Service for Cloud-based Performance Management of Community Healthcare. *Procedia Computer Science*, *113*, 287–294. https://doi.org/10.1016/j.procs.2017.08.321

Farmanova, E., Kirvan, C., Verma, J., Mukerji, G., Akunov, N., Phillips, K., & Samis, S. (2016). Triple Aim in Canada: developing capacity to lead to better health, care and cost. *International Journal for Quality in Health Care*, *28*(6), 830.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Frontoni, E., Baldi, M., Zingaretti, P., Landro, V., & Misericordia, P. (2014). Security issues for data sharing and service interoperability in eHealth systems: The Nu.Sa. test bed. *2014*

International Carnahan Conference on Security Technology (ICCST), 2014-Octob(October), 1–6. https://doi.org/10.1109/CCST.2014.6986999

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010a). Privacy-preserving data publishing: A Survey of Recent Developments. *CSUR*, *42*(4), 1–53. https://doi.org/10.1145/1749603.1749605

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010b). Privacy-preserving data publishing. *ACM Computing Surveys*, *42*(4), 1–53. https://doi.org/10.1145/1749603.1749605

Furht, B., & Escalante, A. (2010). A Handbook of Cloud Computing. In *A Handbook of Cloud Computing* (pp. 3–19). Springer US.

Gatner. (2015). Big Data. Retrieved December 12, 2015, from Gartner website: http://www.gartner.com/it-glossary/big-data

Gaynor, M., Yu, F., Andrus, C. H., Bradner, S., & Rawn, J. (2014). A general framework for interoperability with applications to healthcare. *Health Policy and Technology*, *3*(1), 3–12. https://doi.org/10.1016/j.hlpt.2013.09.004

Gazzarata, G., Gazzarata, R., & Giacomini, M. (2015). A Standardized SOA Based Solution to Guarantee the Secure Access to EHR. *Procedia Computer Science*, *64*, 1124–1129. https://doi.org/10.1016/j.procs.2015.08.582

Gomatam, S., Carter, R., Ariet, M., & Mitchell, G. (2002). An empirical comparison of record linkage procedures. *Statistics in Medicine*, *21*(10), 1485–1496.

Grant, R. W., Ashburner, J. M., Hong, C. C., Chang, Y., Barry, M. J., & Atlas, S. J. (2011). Defining patient complexity from the primary care physician's perspective: a cohort study. *Annals of Internal Medicine*, *155*(12).

Gregor, S., & Hevner, A. R. (2017). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*. https://doi.org/10.25300/misq/2013/37.2.01

Gu, L., Baxter, R., Vickers, D., & Rainsford, C. (2003). Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report*, *3*, 83.

Guarrera, T. K., McGeorge, N. M., Ancker, J. S., Hegde, S., Zhou, Y., Lin, L., … Bisantz, A. M. (2014). Characterising the effect of interoperability on healthcare work: a novel framework. *Theoretical Issues in Ergonomics Science*, *15*(6), 578–594. https://doi.org/10.1080/1463922X.2013.838318

Hincapie, A., & Warholak, T. (2011). The impact of health information exchange on health outcomes. *Applied Clinical Informatics*, *2*(4), 499–507. https://doi.org/10.4338/ACI-2011-05-R-0027

Hsieh, G., & Chen, R.-J. (2012). Design for a secure interoperable cloud-based Personal Health Record service. *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, 472–479. https://doi.org/10.1109/CloudCom.2012.6427582

Hu, J. (2011). *Privacy-Preserving Data Integration in Public Health Surveillance*. University of Ottawa.

Hu, Y., & Bai, G. (2014). A cloud model for interoperable Home-based Chronic Diseases Healthcare. *2014 World Symposium on Computer Applications & Research (WSCAR)*, 1–6. https://doi.org/10.1109/WSCAR.2014.6916770

IBM. (2017). Understanding the Common Data Model. https://doi.org/2017-07-19

IHI. (2016). *Better Health and Lower Cost for Patients with Complex Needs. An IHI Triple Aim Collaborative*. Retrieved from http://www.ihi.org/Engage/collaboratives/BetterHealthLowerCostsPatientswithComplexNeeds/Documents/2015_BetterHealthLowerCosts_Collaborative_Prospectus.pdf

International Organization For Standardization. (2005). *Health informatics - Electronic health record - Definition, scope and context*. https://doi.org/ISO/TR 20514:2005(E)

Introducing HL7 FHIR. (n.d.). Retrieved May 26, 2019, from https://www.hl7.org/fhir/summary.html

Introduction to HL7 Standards. (2016). Retrieved December 4, 2016, from http://www.hl7.org/implement/standards/

Just, B. H., Fabian, D. P., Webb, L. L., & Hjort, B. M. (2009). Managing the integrity of patient identity in health information exchange. *Journal of the American Health Information Management Association*, *80*(7), 62–69.

Katsis, Y., & Papakonstantinou, Y. (2017). View-based Data Integration. In *Encyclopedia of Database Systems*. https://doi.org/10.1007/978-1-4899-7993-3_1072-2

Kemper, H. G., Rausch, P., & Baars, H. (2013). Business Intelligence and Performance Management: Introduction. *Business Intelligence and Performance Management*, 3–10. https://doi.org/10.1007/978-1-4471-4866-1

Klann, J. G., Buck, M. D., Brown, J., Hadley, M., Elmore, R., Weber, G. M., & Murphy, S. N. (2014). Query Health: standards-based, cross-platform population health surveillance. *Journal of the American Medical Informatics Association : JAMIA*, *21*(4), 650–656. https://doi.org/10.1136/amiajnl-2014-002707

Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., & Kuhn, K. A. (2012). Flash: Efficient, stable and optimal k-anonymity. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 708–717. https://doi.org/10.1109/SocialCom-PASSAT.2012.52

Kohlmayer, F., Prasser, F., Eckert, C., & Kuhn, K. a. (2013). A flexible approach to distributed data anonymization. *Journal of Biomedical Informatics*, *50*, 62–76. https://doi.org/10.1016/j.jbi.2013.12.002

Koshy, E., Valsa, K., & Waterman, H. (2010). What is action research? *Action Research in Healthcare*.

Kuziemsky, C. (2013). A Multi-Tiered Perspective on Healthcare Interoperability. In M. Á. Sicilia & P. Balazote (Eds.), *Interoperability in Healthcare Information Systems: Standards, Management, and Technology* (pp. 1–18). https://doi.org/10.4018/978-1-4666-3000-0

Kuziemsky, C., Liu, X., & Peyton, L. (2010). Leveraging goal models and performance indicators to assess health care information systems. *Quality of Information and Communications Technology (QUATIC), 2010 Seventh International Conference on The*, 222–227.

Kuziemsky, C E, Randell, R., & Borycki, E. M. (2016). Understanding Unintended

Consequences and Health Information Technology. *IMIA Yearbook*, *20*(01), 53–60.

Kuziemsky, Craig E., & Peyton, L. (2016). A framework for understanding process interoperability and health information technology. *Health Policy and Technology*, *5*(2), 196–203. https://doi.org/10.1016/j.hlpt.2016.02.007

LeFevre, K., DeWitt, D. J. D. J., & Ramakrishnan, R. (2005). Incognito: efficient full-domain K-anonymity. *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, 49–60. https://doi.org/http://doi.acm.org/10.1145/1066157.1066164

LHIN. (2018). Local Health Integration Network. Retrieved April 3, 2018, from http://www.lhins.on.ca/

Li, J., Wong, R. C. W., Fu, A. W. C., & Pei, J. (2008). Anonymization by local recoding in data with attribute hierarchical taxonomies. *IEEE Transactions on Knowledge and Data Engineering*, *20*(9), 1181–1194. https://doi.org/10.1109/TKDE.2008.52

Li, Y., & Guo, Y. (2015). Wiki-Health: From Quantified Self to Self-Understanding. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2015.08.008

Liu, J., Erdal, S., Silvey, S. A., Ding, J., Riedel, J. D., Marsh, C. B., & Kamal, J. (2009). Toward a fully de-identified biomedical information warehouse. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2009*, 370–374.

Ma, J., Peng, C., & Chen, Q. (2014). Health Information Exchange for Home-Based Chronic Disease Self-Management -- A Hybrid Cloud Approach. *2014 5th International Conference on Digital Home*, 246–251. https://doi.org/10.1109/ICDH.2014.54

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). ℓ-Diversity: Privacy beyond k-anonymity. *Proceedings - International Conference on Data Engineering*, *2006*, 24. https://doi.org/10.1109/ICDE.2006.1

Martin, K. (CPCSSN). (2018). CPCSSN and Canada's National Health Data Platform. Retrieved May 26, 2019, from https://cpcssn.ca/cpcssn-and-rapid-hc/

Mathew, P. S., & Pillai, A. S. (2015). Big Data solutions in Healthcare: Problems and perspectives. *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference On*, pp. 1–6. https://doi.org/10.1109/ICIIECS.2015.7193211

Matteucci, I., Petrocchi, M., & Sbodio, M. L. (2010). CNL4DSA: a controlled natural language for data sharing agreements. *Proceedings of the 2010 ACM Symposium on Applied Computing*, 616–620.

Matteucci, I., Petrocchi, M., Sbodio, M. L., & Wiegand, L. (2012). A design phase for data sharing agreements. In *Data Privacy Management and Autonomous Spontaneus Security* (pp. 25–41). Springer.

Mcgregor, J., Mercer, S. W., & Harris, F. M. (2016). Health benefits of primary care social work for adults with complex health and social needs: A systematic review. *Health and Social Care in the Community*, *26*(1), 1–13. https://doi.org/10.1111/hsc.12337

McLaughlin, C. P., & Kaluzny, A. D. (2004). *Continuous Quality Improvement in Health Care: Theory, Implementation and Applications* (2nd Editio). Jones and Bartlett Publishers.

Mendelson, D. S., Erickson, B. J., & Choy, G. (2014). Image Sharing: Evolving Solutions in the Age of Interoperability. *Journal of the American College of Radiology*, *11*(12), 1260–1269. https://doi.org/10.1016/j.jacr.2014.09.013

Méray, N., Reitsma, J. B., Ravelli, A. C. J., & Bonsel, G. J. (2007). Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *Journal of Clinical Epidemiology*, *60*(9). https://doi.org/10.1016/j.jclinepi.2006.11.021

Miguel-Angel, S., & Pablo, S. B. (2013). *Interoperability in Healthcare Information Systems: Standards, Management, and Technology*. IGI Global.

Mills, M. E. (2006). Linkage of patient records to support continuity of care: Issues and future directions. *Studies in Health Technology and Informatics*, *122*, 320.

Minutolo, A., Esposito, A., Ciampi, M., Esposito, M., & Cassetti, G. (2014). An Automatic Method for Deriving OWL Ontologies from XML Documents. *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2014 Ninth International Conference On*, pp. 426–431. https://doi.org/10.1109/3PGCIC.2014.88

Moumtzoglou, A. (2014). Cloud Computing Applications for Quality Health Care Delivery. In A. Moumtzoglou & A. N. Kastania (Eds.), *Cloud Computing Applications for Quality Health Care Delivery* (pp. 284–301). https://doi.org/10.4018/978-1-4666-6118-9

Mouttham, A., Kuziemsky, C., Langayan, D., Peyton, L., & Pereira, J. (2011). Interoperable support for collaborative, mobile, and accessible health care. *Information Systems Frontiers*, *14*(1), 73–85. https://doi.org/10.1007/s10796-011-9296-y

Mouttham, A., Peyton, L., Eze, B., & El Saddik, A. (2009). Event-driven data integration for personal health monitoring. *Journal of Emerging Technologies in Web Intelligence*, *1*(2), 110–118.

Navarro, R. (2008). An ethical framework for sharing patient data without consent. *Informatics in Primary Care*, *16*(4), 257–262.

Nergiz, M. E., Atzori, M., & Clifton, C. (2007). Hiding the presence of individuals from shared databases. *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, 665–676. https://doi.org/10.1145/1247480.1247554

Ninghui, L., Tiancheng, L., & Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. *Proceedings - International Conference on Data Engineering*, (2), 106–115. https://doi.org/10.1109/ICDE.2007.367856

Noumeir, R. (2012). Requirements for Interoperability in Healthcare Information Systems. *Journal of Healthcare Engineering*, *3*(2), 323–346. https://doi.org/10.1260/2040-2295.3.2.323

Ochian, A., Suciu, G., Fratu, O., Voicu, C., & Suciu, V. (2014). An overview of cloud middleware services for interconnection of healthcare platforms. *2014 10th International Conference on Communications (COMM)*, 1–4. https://doi.org/10.1109/ICComm.2014.6866753

openEHR Architecture. (2015). Retrieved December 3, 2016, from http://www.openehr.org/what_is_openehr

OPM.GOV. (2017). Performance Management Overview. Retrieved July 15, 2017, from

https://www.opm.gov/policy-data-oversight/performance-management/overview-history/

Orit, L. (Microsoft). (2013). *Big Data Ecosystem Reference Architecture*. Retrieved from http://bigdatawg.nist.gov/_uploadfiles/M0015_v1_1596737703.docx

Overview of privacy legislation in Canada. (2014). Retrieved December 10, 2016, from Office of the Privacy Commissioner of Canada website: https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/02_05_d_15/

Pang, Z., Zheng, L., Tian, J., Kao-Walter, S., Dubrova, E., & Chen, Q. (2013). Design of a terminal solution for integration of in-home health care devices and services towards the Internet-of-Things. *Enterprise Information Systems*, *9*(1), 86–116. https://doi.org/10.1080/17517575.2013.776118

Pathak, J., Kiefer, R. C., & Chute, C. G. (2012). Applying linked data principles to represent patient's electronic health records at Mayo clinic. *Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics - IHI '12*, 455. https://doi.org/10.1145/2110363.2110415

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006). The Design Science Research Process: A Model for Producing and Presenting Information Systems Research. *Proceedings of Design Research in Information Systems and Technology DESRIST'06*. https://doi.org/10.2753/MIS0742-1222240302

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*. https://doi.org/10.2753/mis0742-1222240302

Perakis, K., Bouras, T., Ntalaperas, D., Hasapis, P., Georgousopoulos, C., Sahay, R., … Usurelu, D. (2013). Advancing Patient Record Safety and EHR Semantic Interoperability. *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 3251–3257. https://doi.org/10.1109/SMC.2013.554

Piyare, R. (2013). Integrating Wireless Sensor Network into Cloud services for real-time data collection. *2013 International Conference on ICT Convergence (ICTC)*, 752–756. https://doi.org/10.1109/ICTC.2013.6675470

Poulymenopoulou, M., Papakonstantinou, D., Malamateniou, F., & Vassilacopoulos, G. (2015). A health analytics semantic ETL service for obesity surveillance. *Studies in Health Technology and Informatics*, *210*, 840–844. https://doi.org/10.3233/978-1-61499-512-8-840

Prasser, F., Bild, R., Eicher, J., Spengler, H., Kohlmayer, F., & Kuhn, K. A. (2016). Lightning: Utility-driven anonymization of high-dimensional data. *Transactions on Data Privacy*, *9*(2), 161–185.

Prasser, F., Kohlmayer, F., Lautenschläger, R., & Kuhn, K. A. (2014). ARX - A Comprehensive Tool for Anonymizing Biomedical Data. *AMIA Annual Symposium Proceedings*, 984–993.

Ribeiro, L. S., Viana-Ferreira, C., Oliveira, J. L., & Costa, C. (2014). XDS-I Outsourcing Proxy: Ensuring Confidentiality While Preserving Interoperability. *IEEE Journal of Biomedical and Health Informatics*, *18*(4), 1404–1412. https://doi.org/10.1109/JBHI.2013.2292776

Ringard, Å., Sagan, A., Sperre Saunes, I., & Lindahl, A. K. (2013). Norway: health system review. *Health Systems in Transition*, *15*(8), 1–162. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24434287

Roughead, E. E., Kalisch, L. M., Ramsay, E. N., Ryan, P., & Gilbert, A. L. (2011). Continuity of care: When do patients visit community healthcare providers after leaving hospital? *Internal Medicine Journal*, *41*(9), 662–667. https://doi.org/10.1111/j.1445-5994.2009.02105.x

Ruiz, J. F., Petrocchi, M., Matteucci, I., Costantino, G., Gambardella, C., Manea, M., & Ozdeniz, A. (2016). A lifecycle for data sharing agreements: How it works out. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9857 LNCS*, 3–20. https://doi.org/10.1007/978-3-319-44760-5_1

Sabooniha, N., Toohey, D., & Lee, K. (2012). An evaluation of hospital information systems integration approaches. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics - ICACCI '12*, 498. https://doi.org/10.1145/2345396.2345479

Sachs, P., Gall, W., Marksteiner, A., & Dorda, W. (2000). Unambiguous identification of hospital patients: Case study at the university departments of the General Hospital, Vienna. *International Journal of Medical Informatics*, *57*(2–3). https://doi.org/10.1016/S1386-5056(00)00063-0

Samarati, P., & Sweeney, L. (1998). Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion. *Proc of the IEEE Symposium on Research in Security and Privacy*, 384–393.

Sauleau, E. A., Paumier, J.-P., & Buemi, A. (2005). Medical record linkage in health information systems by approximate string matching and clustering. *BMC Medical Informatics and Decision Making*, *5*. https://doi.org/10.1186/1472-6947-5-32

Sheikh, A., Sood, H. S., & Bates, D. W. (2015). Leveraging health information technology to achieve the "triple aim" of healthcare reform. *Journal of the American Medical Informatics Association : JAMIA*, *22*(4), 849–856. https://doi.org/10.1093/jamia/ocv022

Sheikh, Aziz, Sood, H. S., & Bates, D. W. (2015). Leveraging health information technology to achieve the "triple aim" of healthcare reform. *Journal of the American Medical Informatics Association*, *22*(4), 849–856.

Sinaci, A. A., & Laleci Erturkmen, G. B. (2013). A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. *Journal of Biomedical Informatics*, *46*(5), 784–794. https://doi.org/10.1016/j.jbi.2013.05.009

Spruijt-Metz, D., Hekler, E., Saranummi, N., Intille, S., Korhonen, I., Nilsen, W., … Pavel, M. (2015). Building new computational models to support health behavior change and maintenance: new opportunities in behavioral research. *Translational Behavioral Medicine*, *5*(3), 335–346. https://doi.org/10.1007/s13142-015-0324-1

Summary of the HIPAA Privacy Rule. (2015). Retrieved December 12, 2015, from http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/index.html

Sun, J., & Reddy, C. (2013). Big data analytics for healthcare. *19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013*, 1525–1525. https://doi.org/10.1145/2487575.2506178

Swarup, V., Seligman, L., & Rosenthal, A. (2006). A Data Sharing Agreement Framework. *Information Systems Security*, 22–36. Retrieved from http://www.springerlink.com/index/x5100184p5x6u871.pdf

SWEENEY, L. (2002a). ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, pp. 571–588. https://doi.org/10.1142/S021848850200165X

SWEENEY, L. (2002b). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, pp. 557–570. https://doi.org/10.1142/S0218488502001648

Sweeney, L. (2000). Uniqueness of Simple Demographics in the U.S. Population, LIDAP-WP4. In *Forthcoming book entitled, The Identifiability of Data*.

Sweeney, Latanya. (2002). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Tapuria, A., Kalra, D., & Kobayashi, S. (2013). Contribution of Clinical Archetypes, and the Challenges, towards Achieving Semantic Interoperability for EHRs. *Healthcare Informatics Research*, *19*(4), 286–292. https://doi.org/10.4258/hir.2013.19.4.286

The European Parliament, & The European Council. (2016). General Data Protection Regulation. *Official Journal of the European Union*. https://doi.org/http://eur-lex.europa.eu/pri/en/oj/dat/2003/l_285/l_28520031101en00330037.pdf

Vanhaecht, K., De Witte, K., Depreitere, R., Van Zelm, R., De Bleser, L., Proost, K., & Sermeus, W. (2007). Development and validation of a care process self-evaluation tool. *Health Services Management Research*, *20*(3), 189–202.

Vatsalan, D., Christen, P., & Verykios, V. S. (2013). A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, *38*(6), 946–969. https://doi.org/10.1016/j.is.2012.11.005

Verma, A., & Bhatia, S. (2016). A policy framework for health systems to promote triple aim innovation. *Healthcare Papers*, *15*(3), 9–23.

Vida, M., Lupse, O., & Stoicu-Tivadar, L. (2012). Improving the interoperability of healthcare information systems through HL7 CDA and CCD standards. *SACI 2012 - 7th IEEE International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, 157–161. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84866792295&partnerID=tZOtx3y1

Vie, L. L., Griffith, K. N., Scheier, L. M., Lester, P. B., & Seligman, M. E. P. (2013). The Person-Event Data Environment: leveraging big data for studies of psychological strengths in soldiers. *Frontiers in Psychology*, *4*(DEC), 934. https://doi.org/10.3389/fpsyg.2013.00934

Waterloo, U. of. (2017). Elements of a data sharing agreement: An example. Retrieved August 1, 2017, from https://uwaterloo.ca/research/office-research-ethics/research-human-participants/pre-submission-and-training/human-research-guidelines-and-policies-alphabetical-list/data-sharing-or-transfer-agreements-what-are-they-and-when/elements-data-sharing-agreemen

Weber-Jahnke, J. H., Price, M., & Williams, J. (2013). Software engineering in health care: Is it really different? and how to gain impact. *2013 5th International Workshop on Software*

Engineering in Health Care, SEHC 2013 - Proceedings, 1–4. https://doi.org/10.1109/SEHC.2013.6602469

WHO. (2018). Public Health Surveillance. Retrieved September 7, 2018, from http://www.who.int/topics/public_health_surveillance/en/

Wilk, S., Michalowski, M., Michalowski, W., Rosu, D., Carrier, M., & Kezadri-Hamiaz, M. (2017). Comprehensive mitigation framework for concurrent application of multiple clinical practice guidelines. *Journal of Biomedical Informatics*, *66*. https://doi.org/10.1016/j.jbi.2016.12.002

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W.-C. (2006). Utility-based anonymization using local recoding. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785. https://doi.org/10.1145/1150402.1150504

Zhu, Y., Matsuyama, Y., Ohashi, Y., & Setoguchi, S. (2015). When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *Journal of Biomedical Informatics*, *56*, 80–86. https://doi.org/10.1016/j.jbi.2015.05.012

Appendix A

Pilot Project - Sample operationalized Privacy Compliance Definition Document for Community Care organizations

- This is a redacted version showing only 3 Community care organizations, consent definitions, and definitions for 3 tables. The actual definitions has over 4,000 lines of table entries and 54 CCO organization consent definitions.

```
<?xml version="1.0" encoding="UTF-8"?>
<PrivacyComplianceDefinition>
   <RunSettings
    job_admin="etljobadmins@rha.ca"
        smtp_server="191.15.0.3"
        filter_clients="false"
        send_notification="onerror"
        set_defaults="false" />
   <AnonymizationSettings apply_to="all-or-nothing" />
   <PatientConsent
                reference="C3CLSTAT.CONSENT_TO_SHARE"
                consent_values="A,W,H"
                denied_consent_values="D"
                consent_date="C3CLINT.CONSENT_DATE" />
   <OrganizationConsent>
      <Databases>
         <Database
                server="sso-champ-cluster"
                target_database="CCO-ADOPTION-INSTANCE"
                org_code="100730"
                active="true"
                excluded_programs="INTK,CIA,MOW" />
         <Database
                server="sso-champ-cluster"
                target_database="CCO-MEALS-ON-WHEELS"
                org_code="100726" active="true"
                excluded_programs="ADDK,INTK,CIA,MOW" />
         <Database
                server="sso-champ-cluster"
                target_database="CCO-PHYIO-HELP"
                org_code="100786"
                active="true"
                excluded_programs="INTK" />
      </Databases>
      <Tables>
         <Table name="C3CLINT" ignore="false">
            <Description>Client Master table.</Description>
            <Exceptions>
               <!-- Option for meta: exclude, null, anonymize, mask-->
               <Field name="ADD_SOURCE" meta="exclude" />
               <Field name="ADD_TIME" meta="exclude" />
               <Field name="ADD_USER" meta="exclude" />
               <Field name="ALPHA_KEY" meta="exclude" />
               <Field name="BP_CALL_SAFETY" meta="exclude" />
               <Field name="BP_CDISP_YN" meta="exclude" />
               <Field name="BP_UNLISTED_YN" meta="exclude" />
               <Field name="C5TRNZON_ZONE" meta="exclude" />
               <Field name="CALL_DISPLAY_YN" meta="exclude" />
               <Field name="CAR_REQD" meta="exclude" />
               <Field name="CLOSE_USER" meta="exclude" />
               <Field name="CP_CALL_SAFETY" meta="exclude" />
               <Field name="CP_CDISP_YN" meta="exclude" />
               <Field name="CP_MSG_SAFETY" meta="exclude" />
               <Field name="CP_UNLISTED_YN" meta="exclude" />
               <Field name="CURRENT_CLASS" meta="exclude" />
               <Field name="DECEASED_USER" meta="exclude" />
```

```
            <Field name="DEFAULT_POS" meta="exclude" />
            <Field name="DEFAULT_ROUTE" meta="exclude" />
            <Field name="DEL_COMMENTS1" meta="exclude" />
            <Field name="DEL_COMMENTS2" meta="exclude" />
            <Field name="DEL_COMMENTS3" meta="exclude" />
            <Field name="DIET_NUMBER" meta="exclude" />
            <Field name="DIRECT_LINE1" meta="exclude" />
            <Field name="DIRECT_LINE2" meta="exclude" />
            <Field name="DIRECT_LINE3" meta="exclude" />
            <Field name="DIRECT_LINE4" meta="exclude" />
            <Field name="DONATION_COUNT" meta="exclude" />
            <Field name="ENTRANCE_CODE" meta="exclude" />
            <Field name="ENTRY_CODE" meta="exclude" />
            <Field name="EXPORT_TO_MS" meta="exclude" />
            <Field name="EXPORT_TO_NOVUS" meta="exclude" />
            <Field name="EXPORT_TO_TRACT" meta="exclude" />
            <Field name="FILE_STARTED_BY" meta="exclude" />
            <Field name="FILEAS" meta="exclude" />
            <Field name="FIRST_DONATION_AMOUNT" meta="exclude" />
            <Field name="FIRST_DONATION_DATE" meta="exclude" />
            <Field name="FP_CALL_SAFETY" meta="exclude" />
            <Field name="FP_CDISP_YN" meta="exclude" />
            <Field name="FP_MSG_SAFETY" meta="exclude" />
            <Field name="FP_UNLISTED_YN" meta="exclude" />
            <Field name="FUNDING_INFO" meta="exclude" />
            <Field name="FUTURE_USE1" meta="exclude" />
            <Field name="FUTURE_USE2" meta="exclude" />
            <Field name="FUTURE_USE3" meta="exclude" />
            <Field name="HEIGHT_KEY" meta="exclude" />
            <Field name="HP_CALL_SAFETY" meta="exclude" />
            <Field name="HP_MSG_SAFETY" meta="exclude" />
            <Field name="INTAKE_WORKER" meta="exclude" />
            <Field name="IVR_FLAGS" meta="exclude" />
            <Field name="L_DONOR_ID_NUMBER" meta="exclude" />
            <Field name="L_WORKER_NUMBER" meta="exclude" />
            <Field name="LARGEST_DONATION_AMOUNT" meta="exclude" />
            <Field name="LARGEST_DONATION_DATE" meta="exclude" />
            <Field name="LAST_DONATION_AMOUNT" meta="exclude" />
            <Field name="LAST_DONATION_DATE" meta="exclude" />
            <Field name="LAST_EDIT_TIME" meta="exclude" />
            <Field name="LAST_EDIT_USER" meta="exclude" />
            <Field name="OL_ADD_DATE" meta="exclude" />
            <Field name="OL_ADD_TIME" meta="exclude" />
            <Field name="OL_ADD_USER" meta="exclude" />
            <Field name="OL_EMAIL_DATE" meta="exclude" />
            <Field name="OL_EMAIL_TIME" meta="exclude" />
            <Field name="OL_IMPORT_DATE" meta="exclude" />
            <Field name="OL_IMPORT_TIME" meta="exclude" />
            <Field name="OL_PURGE_DATE" meta="exclude" />
            <Field name="OL_PURGE_TIME" meta="exclude" />
            <Field name="OL_RECORDS_ADDED" meta="exclude" />
            <Field name="REBOOK_WORKER" meta="exclude" />
            <Field name="REOPEN_USER" meta="exclude" />
            <Field name="SEE_ALSO" meta="exclude" />
            <Field name="SEND_TO_MAIN" meta="exclude" />
            <Field name="SEND_TO_OFFLINE" meta="exclude" />
            <Field name="SIGN_TSHEET" meta="exclude" />
            <Field name="SITE_PREFIX" meta="exclude" />
            <Field name="TOTAL_DONATION_AMOUNT" meta="exclude" />
            <Field name="UNLISTED_YN" meta="exclude" />
            <Field name="UOM_HEIGHT" meta="exclude" />
            <Field name="UOM_WEIGHT" meta="exclude" />
            <Field name="USE_CBI" meta="exclude" />
            <Field name="USER_SORT1" meta="exclude" />
            <Field name="USER_SORT2" meta="exclude" />
            <Field name="USER_SORT3" meta="exclude" />
            <Field name="WEIGHT_KEY" meta="exclude" />
            <Field name="WEIGHT_UPDATED" meta="exclude" />
            <Field name="vwBiz_code" meta="exclude" />
            <Field name="vwCon_conID" meta="exclude" />
            <Field name="vwBizOff_code" meta="exclude" />
            <Field name="SUPERVISOR_SUMMARY" meta="exclude" />
            <Field name="HCN_EXPIRY_DATE" meta="exclude" />
            <Field name="BRN_LIST" meta="exclude" />
         </Exceptions>
      </Table>
      <Table name="C3CLASCD" ignore="true">
         <Description>Diet instruction classifications</Description>
      </Table>
      <Table name="C3CLASS" ignore="false">
         <Description>
```

```
                    Client Assessment table. Please do not
                    remove without contacting Jim Brophy
                    </Description>
      <!-- This table is a child table to the client table-->
      <Relationship parent="C3CLINT">
         <!-- ignore the target field name if it is the same as the source -->
         <Map field="CLIENT_NUMBER" />
      </Relationship>
      <Exceptions>
         <!-- Option for meta: exclude, null, mask-->
         <Field name="ADD_SOURCE" meta="exclude" />
         <Field name="ADD_USER" meta="exclude" />
         <Field name="CHECK_IN_DATE" meta="exclude" />
         <Field name="CHECK_IN_TIME" meta="exclude" />
         <Field name="CHECK_IN_USER" meta="exclude" />
         <Field name="CHECK_OUT_DATE" meta="exclude" />
         <Field name="CHECK_OUT_TIME" meta="exclude" />
         <Field name="CHECK_OUT_USER" meta="exclude" />
         <Field name="CLOSE_BY" meta="exclude" />
         <Field name="CLOSE_USER" meta="exclude" />
         <Field name="LAST_EDIT_USER" meta="exclude" />
         <Field name="OL_ADD_DATE" meta="exclude" />
         <Field name="OL_ADD_TIME" meta="exclude" />
         <Field name="OL_ADD_USER" meta="exclude" />
         <Field name="OL_EMAIL_DATE" meta="exclude" />
         <Field name="OL_EMAIL_TIME" meta="exclude" />
         <Field name="REOPEN_USER" meta="exclude" />
         <Field name="TIME_SLOT" meta="exclude" />
         <Field name="TRAVEL" meta="exclude" />
         <Field name="UW_SURVEY_ID" meta="exclude" />
      </Exceptions>
   </Table>
   <Table name="C3CLSTAT" ignore="false">
      <Description>
                     Client status table.
                     Note: A client can have multiple status records for the same department.
                     The most recent status record is the current status record.
                     Extract based on Client Number.
          </Description>
      <Relationship parent="C3CLINT">
         <!-- ignore the target field name if it is the same as the source -->
         <Map field="CLIENT_NUMBER" />
      </Relationship>
      <Exceptions>
         <!-- Option for meta: exclude, null, mask-->
         <Field name="ADD_SOURCE" meta="exclude" />
         <Field name="ADD_USER" meta="exclude" />
         <Field name="C9CBIHD_ID" meta="exclude" />
         <Field name="LAST_EDIT_TIME" meta="exclude" />
         <Field name="LAST_EDIT_USER" meta="exclude" />
         <Field name="OL_ADD_DATE" meta="exclude" />
         <Field name="OL_ADD_TIME" meta="exclude" />
         <Field name="OL_ADD_USER" meta="exclude" />
         <Field name="OL_EMAIL_DATE" meta="exclude" />
         <Field name="OL_EMAIL_TIME" meta="exclude" />
         <Field name="STATUS_SOURCE" meta="exclude" />
      </Exceptions>
   </Table>
   <Table name="C3CLCONT" ignore="false">
      <Description>Client Contacts Master table</Description>
      <!-- This table is a child table to the client table-->
      <Relationship parent="C3CLINT">
         <!-- ignore the target field name if it is the same as the source -->
         <Map field="CLIENT_NUMBER" />
      </Relationship>
      <Exceptions>
         <!-- Option for meta: exclude, null, mask-->
         <Field name="ADD_SOURCE" meta="exclude" />
         <Field name="ADD_TIME" meta="exclude" />
         <Field name="ADD_USER" meta="exclude" />
         <Field name="FP_AREA_CODE" meta="exclude" />
         <Field name="FP_EXCHANGE" meta="exclude" />
         <Field name="FP_NOTES" meta="exclude" />
         <Field name="FP_NUMBER" meta="exclude" />
         <Field name="LAST_EDIT_TIME" meta="exclude" />
         <Field name="LAST_EDIT_USER" meta="exclude" />
         <Field name="OL_ADD_DATE" meta="exclude" />
         <Field name="OL_ADD_TIME" meta="exclude" />
         <Field name="OL_ADD_USER" meta="exclude" />
         <Field name="OL_EMAIL_DATE" meta="exclude" />
         <Field name="OL_EMAIL_TIME" meta="exclude" />
```

```
            <Field name="USER1" meta="exclude" />
            <Field name="USER2" meta="exclude" />
            <Field name="USER3" meta="exclude" />
          </Exceptions>
        </Table>
      </Tables>
   </OrganizationConsent>
</PrivacyComplianceDefinition>
```

Appendix B

Sample ER Data Systematic Data Collection Service mapping definition

```
<RunSettings
  use_bulk_import="true"
  job_admin="etljobadmins@rha.ca"
  smtp_server="191.15.0.3"
  filter_clients ="false"
  send_notification="onerror"
  set_defaults="false"
  mode="configurableccp"
  test_mode="false"
  >
  <Servers>
    <Server name="{sourceServer}" default_database="IDS " default_schema="dbo" role="source" />
    <Server name="{targetServer}" default_database="Coordinated_Care_Plans" default_schema="dbo" role="target"
/>
  </Servers>
  <DocumentMap>
    <VersionControl path="//clientCarePlan:DocumentControl/clientCarePlan:TemplateVersionNumber"
match="v1.0|v0.62|v0.57" schema="v1">
      <Namespace name="clientCarePlan" url="http://schema.ccac-ont.ca/CHRIS/ClientCarePlan/" />
    </VersionControl>
    <!--
    Document Control Section
    Creates and populates the Ccp_Document_Control table, and audit tables.
  -->
    <Table name="CcpDocument_Control" path="//clientCarePlan:DocumentControl">
      <Field name="ClientId" type="string:small" value="{ClientId}" isKey="true" />
      <Field name="ClientCarePlanId" type="uuid" path="clientCarePlan:ClientCarePlanId" isKey="true" />
      <Field name="CarePlanId" type="bigint" value="{CarePlanId}" />
      <Field name="CarePlanDocumentId" type="bigint" value="{CarePlanDocumentId}" />
      <Field name="DocumentNumber" type="string:small" path="clientCarePlan:DocumentNumber" />
      <Field name="Status" type="string:small" path="clientCarePlan:Status/clientCarePlan:Name" />
      <Field name="LastUpdate" type="date" value="{LastUpdate}" />
      <Field name="TemplateVersionNumber" type="string:small" path="clientCarePlan:TemplateVersionNumber" />
      <Table name="Audit" path="clientCarePlan:DocumentCreationAudit|clientCarePlan:DocumentUpdateAudit">
        <Field name="Type" type="string:small" value="{node-name}" />
        <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
        <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
        <Field name="UpdaterAccount" type="string:medium" path="clientCarePlan:UpdaterAccount" />
        <Field name="UpdaterQualifications" type="string:medium" path="clientCarePlan:UpdaterQualifications" />
        <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
        <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
      </Table>
    </Table>
    <!--Clients section-->
    <Table name="Clients" path="//clientCarePlan:Client" parent="CcpDocument_Control">
      <Field name="ClientNumberCcp" type="string:small" path="clientCarePlan:ChrisClientNumber" isKey="true" />
      <Field name="Surname" type="string:small" path="clientCarePlan:Surname" />
      <Field name="GivenNames" type="string:small" path="clientCarePlan:GivenNames" />
      <Field name="PreferredName" type="string:small" path="clientCarePlan:PreferredName" />
      <Field name="PreferredPronoun" type="string:small" path="clientCarePlan:PreferredPronoun" />
      <Field name="DateOfBirth" type="date" path="clientCarePlan:DateOfBirth" />
      <Field name="Hcn" type="string:small" path="clientCarePlan:HealthCard/clientCarePlan:Hcn" />
      <Field name="Hcnverion" type="string:small" path="clientCarePlan:HealthCard/clientCarePlan:HcnVersion" />
      <Field name="FirstLanguage" type="string:small" path="clientCarePlan:FirstLanguage/clientCarePlan:Name"
/>
      <Field name="PreferredLanguage" type="string:small" path="clientCarePlan:PreferredLanguage/clientCare-
Plan:Name" />
      <Field name="MaritalStatus" type="string:small" path="clientCarePlan:MaritalStatus/clientCarePlan:Name"
/>
      <Field name="LivingArrangement" type="string:small" path="clientCarePlan:LivingArrangement/clientCare-
Plan:Name" />
      <Field name="ResidenceType" type="string:small" path="clientCarePlan:ResidenceType/clientCarePlan:Name"
/>
      <Table name="Audit" path="clientCarePlan:SectionUpdateAudit">
        <Field name="Type" type="string:small" value="SectionUpdateAudit-Client" />
        <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
        <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
        <Field name="UpdaterAccount" type="string:small" path="clientCarePlan:UpdaterAccount" />
        <Field name="UpdaterQualifications" type="string:small" path="clientCarePlan:UpdaterQualifications" />
```

```
        <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
        <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
      </Table>
      <Table name="ContactNumbers" path="clientCarePlan:PrimaryPhone|clientCarePlan:AlternatePhone">
        <Field name="Target" type="string:small" value="Client-{node-name}" />
        <Field name="Type" type="string:small" path="clientCarePlan:Type/clientCarePlan:Name" />
        <Field name="Number" type="string:small" path="clientCarePlan:Number" />
        <Field name="Comment" type="string:medium" path="clientCarePlan:Comment" />
      </Table>
      <Table name="EmergencyContact" path="clientCarePlan:EmergencyContact|clientCarePlan:PrimaryContact">
        <Field name="Surname" type="string:small" path="clientCarePlan:Surname" />
        <Field name="FirstName" type="string:small" path="clientCarePlan:FirstName" />
        <Field name="PhoneType" type="string:small" path="clientCarePlan:Type/clientCarePlan:Name" />
        <Field name="Number" type="string:small" path="clientCarePlan:Number" />
        <Field name="Comment" type="string:medium" path="clientCarePlan:Comment" />
      </Table>
      <Table name="ClientAddress" path="clientCarePlan:Address">
        <Field name="Address" type="string:small" path="clientCarePlan:AddressFreeLine" />
        <Field name="City" type="string:small" path="clientCarePlan:City" />
        <Field name="Province" type="string:medium" path="clientCarePlan:Province/clientCarePlan:Name" />
        <Field name="PostalCode" type="string:medium" path="clientCarePlan:PostalCode" />
      </Table>
    </Table>
    <Section name="CareTeams" path="//clientCarePlan:CareTeam" parent="CcpDocument_Control">
      <Table name="Audit" path="clientCarePlan:SectionUpdateAudit">
        <Field name="Type" type="string:small" value="SectionUpdateAudit-CareTeams" />
        <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
        <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
        <Field name="UpdaterAccount" type="string:small" path="clientCarePlan:UpdaterAccount" />
        <Field name="UpdaterQualifications" type="string:small" path="clientCarePlan:UpdaterQualifications" />
        <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
        <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
      </Table>
      <Table name="CareTeamMembers" path="clientCarePlan:CareTeamMember">
        <Field name="TeamId" type="int" value="{repeat-index}" />
        <Field name="Name" type="string:medium" path="clientCarePlan:Name" />
        <Field name="Role" type="string:medium" path="clientCarePlan:Role" />
        <Field name="OrganizationName" type="string:medium" path="clientCarePlan:OrganizationName" />
        <Field name="PrimaryPhone" type="string:medium" path="clientCarePlan:PrimaryPhone" />
        <Field name="SecondaryPhone" type="string:medium" path="clientCarePlan:SecondaryPhone" />
        <Field name="IsLeadCoordinator" type="boolean" path="clientCarePlan:IsLeadCoordinator" />
        <Field name="IsHomeCaregiver" type="boolean" path="clientCarePlan:IsHomeCaregiver" />
        <Field name="IsRegularCareTeamMember" type="boolean" path="clientCarePlan:IsRegularCareTeamMember" />
      </Table>
      <Table name="HomeCaregiverStatus" path="clientCarePlan:HomeCaregiverStatus">
        <Field name="TeamId" type="int" value="{repeat-index}" />
        <Field name="Name" type="string:medium" path="clientCarePlan:Name" />
      </Table>
    </Section>
    <Table name="Goals" path="//clientCarePlan:Goals" parent="CcpDocument_Control">
      <Field name="MostImportant" type="string:large" path="clientCarePlan:MostImportant" />
      <Field name="Concerns" type="string:large" path="clientCarePlan:Concerns" />
      <Field name="ContributingCareTeamMembers" type="string:medium" path="clientCarePlan:ContributingCareTeam-
Members" />
      <Table name="Audit" path="clientCarePlan:SectionUpdateAudit">
        <Field name="Type" type="string:small" value="SectionUpdateAudit-Goals" />
        <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
        <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
        <Field name="UpdaterAccount" type="string:small" path="clientCarePlan:UpdaterAccount" />
        <Field name="UpdaterQualifications" type="string:small" path="clientCarePlan:UpdaterQualifications" />
        <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
        <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
      </Table>
      <Table name="SpecificGoal" path="clientCarePlan:SpecificGoal">
        <Field name="SpecificGoalId" type="int" value="{repeat-index}" isKey="true" />
        <Field name="Description" type="string:medium" path="clientCarePlan:Description" />
        <Field name="ChallengesDescription" type="string:medium" path="clientCarePlan:ChallengesDescription" />
        <Field name="SuggestedBy" type="string:medium" path="clientCarePlan:SuggestedBy/clientCarePlan:Name" />
        <Field name="ExpectedOutcome" type="string:medium" path="clientCarePlan:ExpectedOutcome" />
        <Field name="AchievedResults" type="string:large" path="clientCarePlan:AchievedResults" />
        <Field name="ReviewDate" type="date" path="clientCarePlan:ReviewDate" />
        <Field name="ContributingCareTeamMembers" type="string:medium" path="clientCarePlan:ContributingCare-
TeamMembers" />
        <Table name="SpecificGoalAction" path="clientCarePlan:Actions">
          <Field name="ActionId" type="int" value="{repeat-index}" />
          <Field name="Description" type="string:large" path="clientCarePlan:Description" />
          <Field name="ResponsibleParty" type="string:small" path="clientCarePlan:ResponsibleParty" />
        </Table>
```

```
      </Table>
      <Table name="FutureSituation" path="clientCarePlan:FutureSituation">
          <Field name="Id" type="int" value="{repeat-index}" />
          <Field name="Description" type="string:large" path="clientCarePlan:Description" />
          <Field name="WhatWillDo" type="string:medium" path="clientCarePlan:WhatWillDo" />
          <Field name="WhatWillNotDo" type="string:medium" path="clientCarePlan:WhatWillNotDo" />
          <Field name="Telephone" type="string:small" path="clientCarePlan:Telephone" />
          <Field name="ReviewDate" type="datetime" path="clientCarePlan:ReviewDate" />
        </Table>
      <Table name="Plan" path="clientCarePlan:Plan">
        <Field name="Id" type="int" value="{repeat-index}" />
        <Field name="PoaCompleted" type="string:small" path="clientCarePlan:PoaCompleted/clientCarePlan:Name"
/>
        <Field name="PoaLocation" type="string:medium" path="clientCarePlan:PoaLocation" />
        <Field name="PoaContactSurname" type="string:small" path="clientCarePlan:PaoContact/clientCarePlan:Sur-
name" />
        <Field name="PoaContactFirstName" type="string:small" path="clientCarePlan:PaoContact/clientCare-
Plan:FirstName" />
        <Field name="PoaContactRelationship" type="string:small" path="clientCarePlan:PaoContact/clientCare-
Plan:RelationshipCodedValue/clientCarePlan:Name" />
        <Field name="PoaContactPhoneType" type="string:small" path="clientCarePlan:PaoContact/clientCare-
Plan:Phone/clientCarePlan:Type/clientCarePlan:Code" />
        <Field name="PoaContactPhone" type="string:small" path="clientCarePlan:PaoContact/clientCare-
Plan:Phone/clientCarePlan:Number" />
      </Table>
    </Table>
    <Section name="HealthConditions" path="//clientCarePlan:HealthConditions" parent="CcpDocument_Control">
      <Table name="Audit" path="clientCarePlan:SectionUpdateAudit">
        <Field name="Type" type="string:small" value="SectionUpdateAudit-HospitalVisits" />
        <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
        <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
        <Field name="UpdaterAccount" type="string:small" path="clientCarePlan:UpdaterAccount" />
        <Field name="UpdaterQualifications" type="string:small" path="clientCarePlan:UpdaterQualifications" />
        <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
        <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
      </Table>
      <Table name="HealthIssue" path="clientCarePlan:HealthIssue">
        <Field name="IssueType" type="string:medium" path="clientCarePlan:IssueType/clientCarePlan:Name" />
        <Field name="HealthCondition" type="string:medium" path="clientCarePlan:HealthCondition/clientCare-
Plan:Condition" />
        <Field name="Notes" type="string:large" path="clientCarePlan:HealthCondition/clientCarePlan:Notes" />
      </Table>
    </Section>
    <Section name="HospitalVisits" path="//clientCarePlan:HospitalVisits" parent="CcpDocument_Control">
      <Table name="Audit" path="clientCarePlan:SectionUpdateAudit">
        <Field name="Type" type="string:small" value="SectionUpdateAudit-HospitalVisits" />
        <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
        <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
        <Field name="UpdaterAccount" type="string:small" path="clientCarePlan:UpdaterAccount" />
        <Field name="UpdaterQualifications" type="string:small" path="clientCarePlan:UpdaterQualifications" />
        <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
        <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
      </Table>
      <Table name="HospitalVisits" path="clientCarePlan:HospitalVisit">
        <Field name="Hospital" type="string:medium" path="clientCarePlan:Hospital/clientCarePlan:Name" />
        <Field name="VisitType" type="string:medium" path="clientCarePlan:VisitType/clientCarePlan:Name" />
        <Field name="VisitDate" type="string:medium" path="clientCarePlan:VisitDate" />
        <Field name="VisitReason" type="string:medium" path="clientCarePlan:VisitReason" />
        <Field name="HospitalAdviceNote" type="string:large" path="clientCarePlan:HospitalAdviceNote" />
        <Field name="Complications" type="string:large" path="clientCarePlan:Complications" />
      </Table>
    </Section>
    <Section name="Situation" path="//clientCarePlan:Situation" parent="CcpDocument_Control">
      <Table name="Audit" path="clientCarePlan:SectionUpdateAudit">
        <Field name="Type" type="string:small" value="SectionUpdateAudit-Situation" />
        <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
        <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
        <Field name="UpdaterAccount" type="string:small" path="clientCarePlan:UpdaterAccount" />
        <Field name="UpdaterQualifications" type="string:small" path="clientCarePlan:UpdaterQualifications" />
        <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
        <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
      </Table>
      <Section name="Situations" path="//clientCarePlan:Situation" parent="CcpDocument_Control">
        <Table name="Situation" path="clientCarePlan:Employment|clientCarePlan:IncomeAdequacy|clientCare-
Plan:SupplementaryBenefit|clientCarePlan:SmokesTobacco">
          <Field name="Type" type="string:medium" value="{section-node-name}" />
          <Field name="Name" type="string:medium" path="clientCarePlan:Name" />
          <Field name="Code" type="string:medium" path="clientCarePlan:Code" />
```

```
      </Table>
    </Section>
    <Table name="HospitalVisits" path="clientCarePlan:HospitalVisit">
      <Field name="Hospital" type="string:medium" path="clientCarePlan:Hospital/clientCarePlan:Name" />
      <Field name="VisitType" type="string:medium" path="clientCarePlan:VisitType/clientCarePlan:Name" />
      <Field name="VisitDate" type="string:medium" path="clientCarePlan:VisitDate" />
      <Field name="VisitReason" type="string:medium" path="clientCarePlan:VisitReason" />
      <Field name="HospitalAdviceNote" type="string:large" path="clientCarePlan:HospitalAdviceNote" />
      <Field name="Complications" type="string:large" path="clientCarePlan:Complications" />
    </Table>
  </Section>
  <Section name="Supports" path="//clientCarePlan:Supports" parent="CcpDocument_Control">
    <Table name="Audit" path="clientCarePlan:SectionUpdateAudit">
      <Field name="Type" type="string:small" value="SectionUpdateAudit-Supports" />
      <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
      <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
      <Field name="UpdaterAccount" type="string:small" path="clientCarePlan:UpdaterAccount" />
      <Field name="UpdaterQualifications" type="string:small" path="clientCarePlan:UpdaterQualifications" />
      <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
      <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
    </Table>
    <Table name="CommunitySupport" path="clientCarePlan:CommunitySupport">
      <Field name="Organization" type="string:medium" path="clientCarePlan:Organization/clientCarePlan:Name"
/>
      <Field name="ContactName" type="string:medium" path="clientCarePlan:ContactName" />
      <Field name="ProvidedServices" type="string:medium" path="clientCarePlan:ProvidedServices" />
      <Field name="Phone" type="string:small" path="clientCarePlan:Phone" />
    </Table>
  </Section>
  <Section name="Treatments" path="//clientCarePlan:Treatments" parent="CcpDocument_Control">
    <Table name="Audit" path="clientCarePlan:SectionUpdateAudit">
      <Field name="Type" type="string:small" value="SectionUpdateAudit-Treatments" />
      <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
      <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
      <Field name="UpdaterAccount" type="string:small" path="clientCarePlan:UpdaterAccount" />
      <Field name="UpdaterQualifications" type="string:small" path="clientCarePlan:UpdaterQualifications" />
      <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
      <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
    </Table>
    <Table name="Prescription" path="clientCarePlan:Prescription">
      <Field name="Drug" type="string:small" path="clientCarePlan:Drug/clientCarePlan:Name" />
      <Field name="Strength" type="string:small" path="clientCarePlan:Strength" />
      <Field name="Frequency" type="string:small" path="clientCarePlan:Frequency" />
      <Field name="RouteCoded" type="string:small" path="clientCarePlan:RouteCoded/clientCarePlan:Name" />
    </Table>
    <Table name="Allergy" path="clientCarePlan:Allergy">
      <Field name="Substance" type="string:small" path="clientCarePlan:Substance" />
      <Field name="AllergyCategory" type="string:small" path="clientCarePlan:AllergyCategory/clientCare-
Plan:Name" />
    </Table>
  </Section>
  <Section name="HealthAssessments" path="//clientCarePlan:HealthAssessments" parent="CcpDocument_Control">
    <Table name="Audit" path="clientCarePlan:SectionUpdateAudit">
      <Field name="Type" type="string:small" value="SectionUpdateAudit-HealthAssessments" />
      <Field name="LastUpdateDateTime" type="datetime" path="clientCarePlan:LastUpdateDateTime" />
      <Field name="UpdaterName" type="string:small" path="clientCarePlan:UpdaterName" />
      <Field name="UpdaterAccount" type="string:small" path="clientCarePlan:UpdaterAccount" />
      <Field name="UpdaterQualifications" type="string:small" path="clientCarePlan:UpdaterQualifications" />
      <Field name="UpdatedSinceLastTime" type="string:small" path="clientCarePlan:UpdatedSinceLastTime" />
      <Field name="UpdaterOrganizationName" type="string:small" path="clientCarePlan:UpdaterOrganizationName"
/>
    </Table>
    <Table name="HealthAssessment" path="clientCarePlan:HealthAssessment">
      <Field name="ActionsTaken" type="string:small" path="clientCarePlan:ActionsTaken" />
    </Table>
  </Section>
 </DocumentMap>
</RunSettings>
```

Mapping Process

This process that executes the data translation mapping generates a relational database schema that automatically persists the data from semi-structured data sources to a structured table. The mapping must adhere to the following rules:

1. Each CDM table maps to one or more paths in the source data stream nodes. For example, "*patient/personalDetails*" is a node with patient personal and contact details. The assumption is that each mapped path is a collection of records. This results in the automatic creation and populating of the table(s) that match the defined schema.
2. For each field in the CDM table with an applicable field, create a field definition that identifies the node or attribute from the CDM mapped node. The path for each field definition must be a different node or attribute of the parent node. Where a field is mapped to a direct child node, the path is set with the name of the child node. If the field maps to an embedded node within the tree of the parent, then an XPath that points to the node must be set. If the field maps to an attribute, the attribute name is set with an "@" prefix.
3. Each field has a type value that identifies the data type for each field. Supported types are listed in Table A2-1.

Table A2-1 Support Types and matching SQL types

Type	SQL Type
Uuid	Guid
Int	Int
Bigint	bigint
String:n	nvarchar(n)
Float	float
Double	double
Currency	currency
Dates	date/datetime/time

4. Table Fields can be set with a constant value in the definition. These constants override the defined path expressions where they exist.
5. Primary key constraints are defined using the isKey=”true” flag. When a field is set as a key, it automatically creates these field(s) as Primary Key within the table. If the table definition includes a parent table, all the keys from the parent are automatically inherited by the child table. Fields can also be set to take a repeat-index from the XPath query, automatically seeding an integer index for each record returned from executing the XPath query.