SemDQ: A Semantic Framework for Data Quality Assessment

by

Lingkai Zhu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Health Studies and Gerontology

Waterloo, Ontario, Canada, 2014

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Objective: Access to, and reliance upon, high quality data is an enabling cornerstone of modern health delivery systems. Sadly, health systems are often awash with poor quality data which contributes both to adverse outcomes and can compromise the search for new knowledge. Traditional approaches to purging poor data from health information systems often require manual, laborious and time-consuming procedures at the collection, sanitizing and processing stages of the information life cycle with results that often remain sub-optimal. A promising solution may lie with semantic technologies — a family of computational standards and algorithms capable of expressing and deriving the meaning of data elements. Semantic approaches purport to offer the ability to represent clinical knowledge in ways that can support complex searching and reasoning tasks. It is argued that this ability offers exciting promise as a novel approach to assessing and improving data quality. This study examines the effectiveness of semantic web technologies as a mechanism by which high quality data can be collected and assessed in health settings. To make this assessment, key study objectives include determining the ability to construct of valid semantic data model that sufficiently expresses the complexity present in the data as well as the development of a comprehensive set of validation rules that can be applied semantically to test the effectiveness of the proposed semantic framework.

Methods: The Semantic Framework for Data Quality Assessment (SemDQ) was designed. A core component of the framework is an ontology representing data elements and their relationships in a given domain. In this study, the ontology was developed using openEHR standards with extensions to capture data elements used in for patient care and research purposes in a large organ transplant program. Data quality dimensions were defined and corresponding criteria for assessing data quality were developed for each dimension. These criteria were then applied using semantic technology to an anonymized research dataset containing medical data on transplant patients. Results were validated by clinical researchers. Another test was performed on a simulated dataset with the same attributes as the research dataset to confirm the computational accuracy and effectiveness of the framework.

Results: A prototype of SemDQ was successfully implemented, consisting of an ontological model integrating the openEHR reference model, a vocabulary of transplant variables and a set of data quality dimensions. Thirteen criteria in three data quality dimensions were transformed into computational constructs using semantic web standards. Reasoning and logic inconsistency checking were first performed on the simulated dataset, which contains carefully constructed test cases to ensure the correctness and completeness of logical computation. The same quality checking algorithms were applied to an established research database. Data quality defects were successfully identified in the dataset which was manually cleansed and validated periodically. Among the 103,505 data entries, application of two criteria did not return any error, while eleven of the criteria detected erroneous or missing data, with the error rates ranging from 0.05% to 79.9%. Multiple review sessions were held with clinical researchers to verify the results. The SemDQ framework was refined to reflect the intricate clinical knowledge. Data corrections were implemented in the source dataset as well as in the clinical system used in the transplant program resulting in improved quality of data for both clinical and research purposes.

Implications: This study demonstrates the feasibility and benefits of using semantic technologies in data quality assessment processes. SemDQ is based on semantic web standards which allows easy reuse of rules and leverages generic reasoning engines for computation purposes. This mechanism avoids the shortcomings that come with proprietary rule engines which often make ruleset and knowledge developed for one dataset difficult to reuse in different datasets, even in a similar clinical domain. SemDQ can implement rules that have shown to have a greater capacity of detect complex cross-reference logic inconsistencies. In addition, the framework allows easy extension of knowledge base to cooperate more data types and validation criteria. It has the potential to be incorporated into current workflow in clinical care setting to reduce data errors during the process of data capture.

Acknowledgements

My first and most sincere gratitude and appreciation goes to Dr. Helen H. Chen for her guidance, patience, encouragement and knowledge throughout the long journey of my study. I would have given up at the hardest time without her encouraging words and confidence in me. I would also like to thank Dr. Ian McKillop for his invaluable guidance and encouragement.

I wish to thank Dr. John P. Hirdes and Dr. Christopher M. Perlman for their guidance and patience. I would also like to thank Julie Koreck, Tracy Taves, Carol West-Seebeck, Tracie Wilkinson and the rest of the Applied Health Science Faculty for their assistance during my program of study at University of Waterloo.

I would like to express my deepest gratitude to Cesar Leos-Toro and Kevin Quach, who provided assistance in understanding health concepts and my academic writing. I would also like to thank Dr. Joseph S. Kim and Segun Famure at the Multiple Organ Transplant institute in the Toronto General Hospital for providing the dataset and valuable feedbacks. This project could not be accomplished without their help.

I received many supports from, and spent pleasant and memorable times with my classmates and colleagues: Dr. Josephine McMurray, Dr. Huang Hong, André Carrington, Chenyu Liu, Lana Vanderlee, Louise Li, Tim Wadman, Chong Li, Zhonghan Li at the Waterloo Health Information Systems and Technology Lab. My best wishes are always with you.

My parents, Jiazheng Zhu and Yuhui Hou, my sister and brother-in-law, Lingke Zhu and Xin Fang, cared about me while I worked towards my degree. The completion of this

thesis means as much to me as to them. So I dedicate this thesis to my loving family, without their love, affection and encouragement this work would not have been possible.

A special thank you goes to Xiaogeng Deng, Zhixia Xi and Bo Hu in Italy, who provided necessary documents for my application in this program.

This work was conducted using the Protégé resource, which is supported by grant LM007885 from the United States National Library of Medicine. I am also grateful to the open-source communities of Linux, Apache Jena, Weka and R for their invaluable tools for my work.

Table of Contents

List of Tables

List of Figures

Glossary

Cohen's Kappa A statistical measure of inter-rater agreement for categorical items.

Cronbach's Alpha A measure of internal consistency of a test or scale, widely used in psychological test.

Data Quality Assessment Data quality assessment is the process of exposing technical and business data issues in order to plan data cleansing and data enrichment strategies.

Data Quality Assurance Data quality assurance is the process of profiling the data to discover inconsistencies, and other anomalies in the data and performing data cleansing activities (e.g. removing outliers, missing data interpolation) to improve the data quality.

Data Quality Dimension A data quality dimension is an aspect or feature of information and a way to classify information and data quality needs.

Electronic Health Record (EHR) An Electronic Health Record (EHR) is an official health record for an individual that is shared among multiple facilities and agencies.

Health Care The maintenance and improvement of physical and mental health, esp. through the provision of medical services. The word “Healthcare” (without the space) refers to the industry that provide health care actions.

ICD The International Classification of Diseases (ICD) is the standard diagnostic tool for epidemiology, health management and clinical purposes. The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) was the official system used in the United States to classify and assign codes to health conditions and related information until the use of the Tenth Revision (ICD-10) started. ICD-10 was endorsed by the Forty-third World Health Assembly in May 1990 and came into use in WHO Member States as from 1994.

Ontology A formal description of the concepts and their relationships about a knowledge domain

OpenEHR A non-profit organization which aims to provide systems and tools for semantically manipulating health data and has developed a wide collection of “archetypes” (a.k.a, clinical models) that has received formal acceptance as an International Organization for Standardization (ISO) standard (ISO 13606-2)

OWL The Web Ontology Language (OWL), a W3C standard, is a family of knowledge representation languages or ontology languages for authoring ontologies or knowledge bases.

RDF Resource Description Framework (RDF) is a data model specification according to which data are stored and linked in triple statements

Semantic Technology The technologies that express the meanings of resources and their relationships in a way that machine can compute them during program execution

SNOMED The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting.

Software Framework A software framework is a generic and reusable software platform to develop applications, products and solutions

SPARQL A recursive acronym for “SPARQL Protocol and RDF Query Language”, a query language for RDF datasets

Chapter 1

Introduction

Modern health care involves multiple medical services and care providers and heavily relies on massive data generated during the process (Groves, Kayyali, Knott, & Van Kuiken, 2013). Multidisciplinary care teams in different care settings can enter and access patient data. Transcription errors, misinterpretations, inaccurate or missing records may occur during the process of busy, demanding care. Erroneous data can have a serious impact on patient safety, care quality and the output of research (Hickey et al., 2013). If there are flaws in the source of data, any further analysis on them could only result in incomplete, inaccurate or sometimes fatally wrong results. To further complicate the problem of data quality, patient data are often managed by multiple systems in a mixture of formats ranging from digital images, structure reports, free-text clinical notes to letters on paper. As the demand for more data in healthcare is increasing rapidly in order to provide more efficient and better care, the quality of data becomes more important than ever.

Ensuring a level of high quality in health data is a daunting task. Traditional forms of data quality checking include the use of data entry forms with defined ranges and limits of a data value (Chen, Chen, Conway, Hellerstein, & Parikh, 2011). However, not every information system currently used in hospitals is capable of supporting standardized data entry forms. In addition, not all types of errors can be detected by using simple value restrictions, especially in the case of detecting errors in multiple data fields that requires complex clinical knowledge. To achieve better data quality, data are re-extracted from databases and complex validation rules are imposed on them (Barrett et al., 2011). Sometimes, many man-hours are spent on data aggregation and cleansing for such purposes. However, such data validation rules are not transferable since each program has its own data collection tool, such as macros in Microsoft Excel or arithmetic calculations and functions in SAS. In addition, validation rules typically fit one specific setting but become inapplicable in a different setting. For example, a set of rules may be hard-coded for several drugs which are eligible for treating a disease, but later when a new drug is approved for this treatment, the rules have to be reviewed and rewritten.

With this challenge of maintaining high data quality at hand, an alternative data quality assessment that combines new ways of data organization and rule description is to be explored in this thesis. The proposed semantic framework separates data quality validation rules from a specific dataset and is expected to achieve wider applicability and better reusability than the traditional approach that relies on accumulating rules. For instance, considering the previous drug eligibility example, an alternative way to approach this problem is to design a rule describing some of the drugs that are eligible for a treatment while maintaining the drug list in another place, then any newly-added eligible drug would only results in a minor update in the list. However, challenges arise on how to properly, explicitly and interoperably[1] represent the corresponding knowledge. One could express rules in any proprietary format but it would be hard to exchange and share them with another person or computing entity.

Semantic Web technologies are believed to be a promising solution for data quality management (Fürber & Hepp, 2013). By adopting Semantic Web technologies, everyone can share a common understanding on how to describe data and the meaning of data will not be misinterpreted during reuse. Thus, in this thesis, a semantic framework is designed based on semantic web technologies. The framework formally defines the whole process of conceptualizing, organizing, importing and evaluating data. Data validation rules are developed explicitly and stand-alone from any platforms, preventing knowledge from being proprietary and buried in programs. Data will be imported into the framework and will be checked against the aforementioned validation rules. Data quality errors will be highlighted and the cause of errors annotated for users to review. After the corrections are made, the data can be exported for data analysis. The whole process is reproducible and repeatable. This thesis will demonstrate how a semantic framework could accelerate the data quality assessing process. Data cleansing is a related but separate research topic and is out of the scope of this thesis.

In order to provide a more detailed context for this topic, a literature review examining existing data quality assessing approaches in a healthcare context will be presented in Chapter 2. In Chapter 3, the methodology of the study is discussed. Next, the process

[1] Mead et al. (2006) defined interoperability as “the ability of two parties, either human or machine, to exchange data or information.”

of establishing and applying a semantic framework for health data quality assessment is described in Chapter 4. Two datasets were assessed using the framework and the results are also presented in this chapter. Finally, the discussion and future work is presented in Chapter 5 and 6, respectively.

Chapter 2

Literature Review

2.1 Importance of Data Quality in Health Care and Health Research

Medical error is a major source of injury and death in North America. A report from the Institute of Medicine, U.S. estimated that preventable medical errors cause between 44,000 and 98,000 deaths every year (Kohn, Corrigan, Donaldson, et al., 2000). Poor data quality is a major cause leading to medical errors. For example, a follow-up to Kohn et al. (2000)'s report estimated that at least 1.5 million people are injured by preventable medication errors, which yielded an extra annual cost of $3.5 billion in 2006 dollars (Aspden, Wolcott, Bootman, Cronenwett, et al., 2006). According to the report, possible causes include improper representation of drug information and unstandardized terms, which are directly related to the poor quality of data. Furthermore, incorrect information access is a significant cause leading to wrong site / wrong patient surgery, which is the most prevalent sentinel event reported by The Joint Commission, U.S. (Spath, 2011). In addition, poor data quality can hinder the acquisition of business intelligence about healthcare operation, thus resulting in sub-optimal or wasteful resource management. A business report estimated that due to the inability to render needed information from collected data, the thirty North American health providers that were surveyed lost an average of $70.2 million or 15% of additional avenue annually. Although poor data quality was not given as an explicit reason, 47% of the executives stated they could not translate the captured data into meaningful interpretations and 63% said they needed greater ability of data analytics in order to achieve this (Oracle, 2012). Given that total health expenditures in Canada have been constantly growing over the last thirty years, reaching $193.1 billion or 11.9% of Canada's GDP in 2010 (CIHI, 2012), it could be estimated, although no exact numbers have been seen, that a considerable amount of loss due to poor data quality occurs each year. Attaining clean and high quality data is a major challenge facing every sector of the healthcare industry, and the importance of improving health data quality cannot be overstated.

It is believed that health information technology can help promote patient-focused care (Goldschmidt, 2005), reduce medical errors (Hoffman & Podgurski, 2008), improve care outcomes and manage costs (Bughin, Livingston, & Marwaha, 2011). However, technology alone will not achieve a better patient outcome if the wrong data are captured in the system or if some crucial information is missing. For example, the data quality problem was emphasized during an ongoing effort to establish a surveillance system for chronic disease in primary care across Canada (Birtwhistle et al., 2009). In the case presented by Birtwhistle, when massive data were collected into the surveillance system, various issues of data quality were immediately found, e.g. missing drug doses or dates of the onset, referral to an unlisted doctor and including identifiable data like names. In a national health administrative database in Portugal containing over 9 million episodes between 2000 and 2007, 26.5% were found missing the "type of care" variable, which was essential to group patients and split other variables (Freitas, Silva-Costa, Marques, & Costa-Pereira, 2010). Users may also have unrealistic expectations of the Electronic Health Record (EHR) system which may not be met due to poor data quality. Campbell, Sittig, Guappone, Dykstra, and Ash (2007) found that clinicians believe that all patient-related data will exist in the EHR system or trust EHR data despite possible inaccuracies. Although the tendency to overly rely on computers should be corrected, improving the data quality is vital in both health care and research activities that depend on patient data.

2.2 Data Quality Dimensions

The most prevalent definition of data quality can be succinctly summed up as "fitness to use", i.e., how well do the data serve the data consumer's purposes (Wang, Strong, & Guarascio, 1996; Orfanidis, Bamidis, & Eaglestone, 2004; Watts, Shankaranarayanan, & Even, 2009). In order to further understand data quality, many efforts were devoted to divide the concept into data quality dimensions. Generally, two research communities aid in the development of data quality dimensions: management science and information science. It is helpful to distinguish them since they share a common vocabulary but have different viewpoints. Data quality dimensions in information science often solely describe the data, whereas management science describes the whole workflow from acquisition to utilization (Oliveira, Rodrigues, & Henriques, 2005). This study focuses on the information side of data quality dimensions. The three studies that will be reviewed discussed data quality in general (Wand & Wang, 1996; Pipino, Lee, & Wang, 2002; Sebastian-Coleman, 2012) while two studies looked specifically in a health context (CIHI, 2009; Liaw et al., 2012). Commonly observed dimensions include accuracy, completeness, consistency and timeliness. Although there is no definitive agreement on each definition, similarities can be found. The operationalization of each dimension used for this study is listed below.

- **Accuracy** (as known as "Free-of-Error") refers to the condition in which a recorded value is unbiased from the actual value (Wand & Wang, 1996; Pipino et al., 2002; CIHI, 2009; Liaw et al., 2012). For example, for a record which shows a patient having no history of smoking, accuracy is satisfied only when the patient indeed has never smoked.

- **Completeness** refers to the state in which data are not missing or when all necessary data have been included (Pipino et al., 2002; Sebastian-Coleman, 2012; Liaw et al., 2012). For example, every patient record must have a gender value. Missing such a value would be considered a violation of the data validation rule.

- **Consistency** (also seen as comparability) refers to the condition in which data are represented in conformity with other sources or at different times, or use standard formats (Wand & Wang, 1996; Pipino et al., 2002; CIHI, 2009; Sebastian-Coleman, 2012; Liaw et al., 2012); for example, lists or diagnoses recorded in EHR are mapped to a standardized terminology such as the International Classification of Disease (ICD), so their meanings and categorization can be consistent across different information systems and in different contexts.

- **Timeliness** refers to the state in which data are up-to-date, or delivered on time (Wand & Wang, 1996; Pipino et al., 2002; CIHI, 2009; Sebastian-Coleman, 2012; Liaw et al., 2012); e.g., flu surveillance requires up-to-date epidemiological data.

These data quality dimensions are generic descriptors of data quality dimensions which are applied to assess data quality in any business domain, i.e. banking, manufacture, etc. Two additional quality dimensions are common and of particular interests in health studies: namely, reliability and validity.

- **Reliability** refers to the conformity of recorded data when collection practices are repeatedly performed on the same data source (Mor et al., 2003; Greiver, Barnsley, Glazier, Harvey, & Moineddin, 2012). The difference between reliability and consistency is that reliability describes data conformity across collection practices, while consistency describes data conformity during storage and transfer. Reliability receives particular attention in health research because healthcare organizations often collect data from multiple sources (e.g., subject assessments such as diagnosis and self-reported questionnaire, objective measurements such as lab tests and data generated from monitoring instruments. Reliability is measured by Cronbach's alpha value. A value of 0.7 is a measurement of good reliability (Streiner & Norman, 2008). High reliability in data quality is of particular importance in health studies involving psychometric properties of a patient (Hirdes et al., 2013; Naus & Hirdes, 2013).

- **Validity** is defined as "the degree to which a test measures what it claims, or purports, to be measuring" (Brown, 1996) and refers to the degree to which the data conforms to a defined business rules [1]. In health studies, validity also refers to the proportion of cases that truly reflects the actual values (Bray & Parkin, 2009). Validity is not synonymous to accuracy as data can have a high degree of validity but not be accurate. For example, a set of data representing patient age can have a high degree of validity if the values are within the range of 0-100, but the accuracy of a patient's age cannot be guaranteed since a wrong value can be assigned to him/her. Criterion, content, and construct are three basic types of validity (Kaplan, Bush, & Berry, 1976; Mokkink et al., 2010). Criterion validity is met when a proposed measure reflects an accurate observation of the interested value. For example, predictive validity, a subcategory of criterion validity, refers to the correlation between a predicted value and the later obtained actual value (Mor, Intrator, Unruh, & Cai, 2011). Content validity is achieved when the test covers all the items from the domain to be observed (Kaplan et al., 1976). Construct validity refers to whether a test adequately measures the "construct", that is, the theoretical concept that is being intended to measure. One sub-type of construct validity is "convergent validity", which measures the degree to which two similar concepts that are both theoretically related to the construct are in fact related in the data collected from the test (Rabinowitz, Pérez, Nancy Curtin Telegdi RN, & Prendergast, 2002; Bray & Parkin, 2009).

[1] Please see definition on http://iaidq.org/main/glossary.shtml

The assessment of reliability and validity of health data often requires in-depth clinical and healthcare operational knowledge. In the computer and information science domain, reliability and validity are expressed via rigorous and enforceable logic constraints on the data model. The term "logic consistency" can be viewed as another dimension of data quality. In computing terms, it is defined as the logical concordance among data values (Bryan & George, 2003). Further expanding the concept, this dimension requires that obtained data values follow certain logic which should be translated from domain knowledge (medical or clinical knowledge in this study) or common sense. For example, it is illogical that a male patient would receive a diagnosis of a female-specific disease like ovarian cancer. Logic consistency can be checked for one attribute, e.g., a person's body mass index (BMI) is unlikely to exceed 45, or cross-referenced involving multiple attributes, e.g., a patient cannot receive a nephrectomy procedure (removal of kidney) if the kidneys have already been removed. In practice, rules have been written to examine logical errors like whether a height measurement is unusually high, or a re-admission date is before the first admission date (Hirdes et al., 2013).

2.3 Current Frameworks for Data Quality Assurance

The total data quality management (TDQM) cycle, which originated from MIT, is a popularly adapted data quality assurance framework in the management science field (Wang, 1998; Baskarada, Koronios, & Gao, 2006). TDQM includes four steps: (1) Define - identify information quality dimensions (which can be regarded as the same as data quality in this study's context); (2) Measure - develop evaluable metrics under dimensions and measure the quality of data; (3) Analyze - based on the results from the last step, find the root causes of observed data quality problems, and check whether the established dimensions are suitable; and (4) Improve - provide suggestions on improving current work-flow and data quality dimensions.

Aimed at providing the highest quality information to Canadian health systems, the Canadian Institute for Health Information (CIHI) has developed a comprehensive data quality framework. The latest version (year 2009) of the framework stated that data quality is a responsibility of all staff members and are a part of the roles and responsibilities for each position. It defined a data quality work cycle with three phases (CIHI, 2009), as illustrated in Figure 2.1. First, the planning phase refers to the preparation and prioritization of a data activity to improve data quality before data collection. Secondly, the implementing phase refers to the actual implementation of the data activities. Lastly, the assessing phase refers to evaluation of the obtained data. Feedback about previous phases is provided as well during the assessing phase so that the methods used in previous phases could be continuously improved.

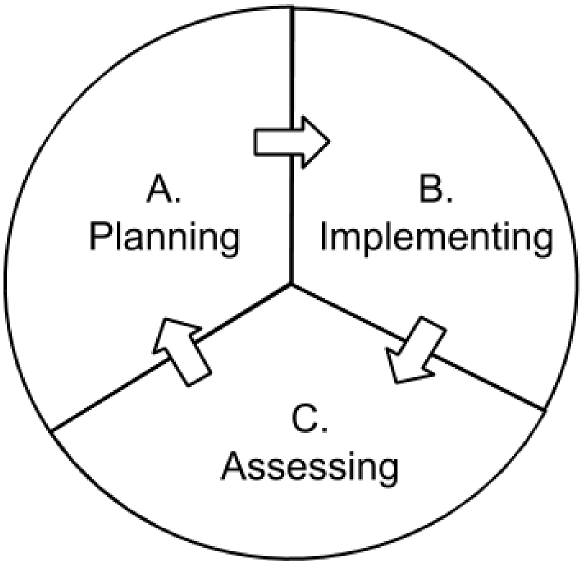

***Figure 2.1:* CIHI's data quality work cycle, reproduced from CIHI (2009)**

Compared to the CIHI Data Quality Work Cycle, the TDQM cycle makes the "improvement" explicit in the process, while the CIHI cycle implies the continuous feedback and improvement mechanism. In addition to the aformentioned TDQM and CIHI's framework, numerous data quality frameworks have been proposed, such as Total Information Quality Management (TIQM) (English, 2003), Data QUality In Cooperative Information Systems (DaQuinCIS) (Scannapieco, Virgillito, Marchetti, Mecella, & Baldoni, 2004), Complete Data Quality (CDQ) (Batini, Cabitza, Cappiello, & Francalanci, 2008), etc. They are designed for different interests but share a similar logic structure of "define - assess - improve". The work of this thesis focuses on the "Assessing Phase" of the CIHI framework. CIHI is used herein because it is the current standard of data quality assurance in the Canadian healthcare industry; however, our implementation should be able to fit into the assessing phase of other frameworks as well.

2.4 Data Quality Assessing Approaches

To ensure data quality, various approaches have been taken based on the understanding of data quality dimensions. Simply matching data with another source can examine accuracy, e.g., matching electrical records vs. paper file records (Mor et al., 2011), or comparing patient-answered questionnaires with general practitioner records (Mant, Murphy, Rose, & Vessey, 2000). Data reliability can be tested using statistics. For example, the level of inter-rater agreement (i.e., how similar two different data interpreters complete a task) can be measured using Cohen's Kappa (Mor et al., 2003; Hessol, Missett, & Fuentes-Afflick, 2004). Internal consistency, another form of reliability, is measured by Cronbach's alpha (Hirdes et al., 2013). Consistency can be improved through a manual effort, e.g., hiring a data clerk to input structured data (Greiver et al., 2011). Incompleteness is usually easy to detect when appearing as null values or missing (Batini, Cappiello, Francalanci, & Maurino, 2009). To accelerate the data cleansing process, rules and scripts are often employed. However, they are usually programmed specifically to a particular dataset and thus lack reusability.

No matter which approach is used, an assessment system is essential. Pipino et al. (2002) summarized two common types of assessments. Subjective assessments examine a stakeholder's subjective perception about data quality, typically in a questionnaire form. A questionnaire is useful when collecting general perception over the whole dataset, but cannot reflect details in large datasets since subjective perceptions are not processable by machines. Three forms of objective assessment are: (1) “Simple Ratio”, which measures the functional proportion of valid records out of total records. Free-of-error (a.k.a. accuracy), completeness and consistency can be measured using this form, e.g., the ratio of records without missing data; (2) “Min or Max Operation”, which is used to measure dimensions related to aggregated data like timeliness, e.g., specifying a maximum delay of data delivery; and (3) “Weighted Average”, which is used in a situation where multiple variables are involved and an overall parameter needs to be set (Pipino et al., 2002). The data assessment methodology used in this study is “objective assessment” and the “Simple Ratio” forms is used.

2.5 Semantic Technology

Semantic technology is defined as the technology “for expressing the meaning of resources and their relationships in machine processable ways and for drawing conclusions (reasoning) based on this meaning with mechanisms that are independent of meaning” (Tiropanis, Davis, Millard, & Weal, 2009). Compared to the traditional programming approaches, semantic technology separates the vague concept of “data” into the values, the format, the semantics and the reasoning mechanism (i.e., how a rule is executed on a given computing platform) of data. For example, a string of digits “20110110” could be interpreted as a number, an ID or a date and each interpretation leads to different methods to handle the information. A semantic statement could specify the string type and enables a computer to automatically select appropriate methods to deal with the string, e.g., if the string “20110110” is specified as a date, it could be transformed into a human-friendly format like “Jan 10th, 2011”. A semantics-featured database could answer complex queries that requires concept grouping and inference, e.g., a query for all patients with a diagnosis of “diabetes” regardless of the sub-type of diabetes with which the patient is diagnosed. In this case, the application advances beyond simple string matching by “understanding” the question and automatically expanding the query to include all strings representing sub-types and variations of diabetes.

Semantic Web technologies are a set of standard solutions for semantic technologies proposed by the World Wide Web Consortium (W3C) that aims at storing, describing and handling data over the web as well as at local repositories. The term “Semantic Web” refers to the W3C’s vision of “the Web of linked data” (W3C, 2013). Core components of Semantic Web standards include (1) Resource Description Framework (RDF), a specification for data modeling; (2) SPARQL (a recursive acronym for “SPARQL Protocol and RDF Query Language”), a query language for RDF datasets; and (3) the Web Ontology Language (OWL), a specification for ontology construction. Originally, the word “ontology” refers to the study of “the nature and structure of ‘reality”’ (i.e., the existence, beings and their categorization) (Guarino, Oberle, & Staab, 2009). In computer science, an ontology is defined as “a specification of a conceptualization”, i.e., a formal description of the concepts and their relationships about a knowledge domain (Gruber, 1995). The OWL specification contains a set of vocabulary describing its components: individuals, classes, properties and operators (McGuinness, Van Harmelen, et al., 2004). Individuals, such as a book, a flower or a man, are concrete entities. Classes are abstract containers of individuals that share the same properties, e.g., the Ocean class represents the common characteristics of all oceans in the world. Properties are relationships between different classes or entities that define their characteristics, e.g., the color in the statement, "the ocean's color is blue". Operators allow computations on classes such as union and intersection, as well as cardinality restrictions, e.g., a man can have a maximum of two hands. The main vocabulary of OWL is listed in Table 2.1.

Table 2.1: **The main vocabulary of OWL, version 2, summarized from Motik et al. (2009)**

Construct	Description
owl:NamedIndividual	Declare a named individual
owl:Class	Declare a named class
rdf:Property	Declare a named property
rdfs:subClassOf	A property stating that all the individual of one class are also instances of another class (the parent class)
owl:equivalentClass	A property stating that two class share the exactly same individuals
owl:disjointWith	A property stating that individuals in one class do not belong to another class
owl:ObjectProperty	Properties relating individuals to other individuals
owl:DatatypeProperty	Properties relating individuals to datatype values, e.g., an integer, a string or a date
owl:FunctionalProperty	A functional property can have only one value in maximum at the same time
owl:allValuesFrom	Specifying a class that in a triple, all objects must come from this class
owl:someValuesFrom	Specifying a class that in a triple, at least one object must come from this class

In addition to the ability to express an ontology, OWL also enables reasoning, i.e., the ability to use given facts to infer new information or find inconsistencies within them.

For example, from inputs such as “Socrates is a man” and “all men are mortal”, an OWL reasoner will infer that “Socrates is mortal”.[2] Due to the limited scope of this thesis, the technical details of OWL reasoning and other Semantic Web components will not be elaborated on.

2.6 Semantic Approaches for Improving Data Quality

In recent years, semantic technologies have received increasing attention among the informatics community because they can be a powerful enabler of interoperability between information systems. While many organizations face the challenge of processing large scale, heterogeneous and dynamic data, semantic technology holds the promise of facilitating better data integration and deriving relations by the application of interfacing rules (Sheth & Ramakrishnan, 2003). For example, Sonntag, Setz, Ahmed-Baker, and Zillner (2012) have applied semantic annotations on medical imaging. With their solution, radiologists can compare medical images with previous diagnoses, not just by simply string matching, but by using semantically similar concepts and easily browse related medication and treatment plans. This is achieved by linking disease, drug and clinical trial databases via semantic technologies. Data quality assessment in the health domain inevitably requires the expression of domain knowledge and the implicit and explicit relations between concepts from multiple sources. Semantic technology is well equipped to meet this need.

Several studies reported the application of semantic technology on data quality assurance. Brüggemann and Gruening (2008) demonstrated how a domain ontology

[2] Modeled and tested using HermiT reasoner v1.3.8 from http://hermit-reasoner.com/

can help finding errors in data. Their data contained two variables: one was disease classification and the other was condition. Given one condition, only part of the disease classifications can form a valid combination with the condition. An ontology was constructed to describe all valid combinations of two variables, as shown in Figure 2.2. For each entry of data containing a combination, the program could refer to the ontology and judge whether the combination was valid. Furthermore, if an invalid combination was detected, the program could provide correction suggestions by assuming either value was correct and listing possible valid combinations. For the same problem, traditional programming approaches would have to enumerate every valid combinations; the semantic approach mentioned above only requires a generic rule referring to the ontological knowledge, which could be constructed using an ontology editor.

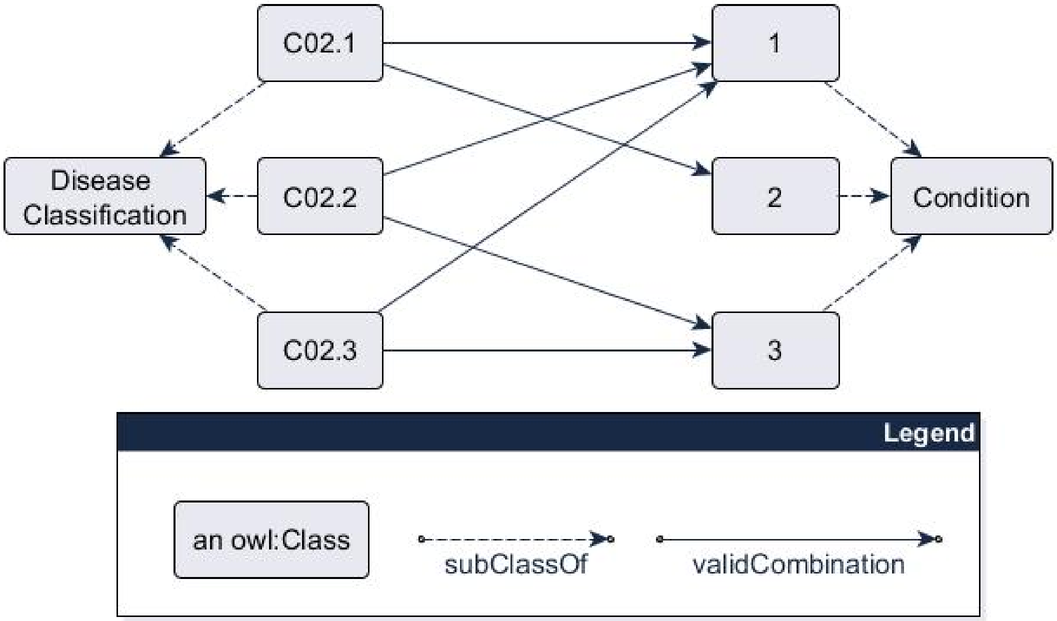

Figure 2.2: **An example of ontology-based consistency checking**

Fürber and Hepp (2010, 2013) pursued a semantic approach that handles missing value, illegal value, and functional dependency data quality problems. Straightforward SPARQL queries were constructed to implement rules detecting data deficiencies. For instance, a "missing value check" query searches for any variable without a supplied value in data; in the same way, an "illegal value check" query searches for pre-defined illegal values. Their approach was similar to Brüggemann and Gruening's error checking method but the difference was that the knowledge was hard-coded in queries instead of an ontology. Their query constructs were generalized, but they also provided an option of importing a trusted knowledge base to specify a query. For example, when checking illegal values, a list of valid values could be provided by binding local knowledge bases like a constructed ontology.

A dearth of literature is available about applying semantic technologies on data quality assurance for healthcare data registries. Liaw et al. (2012) observed this gap in their review: "There is an increasing amount of work on ontology of chronic disease, but little on ontological approaches to DQ (Data Quality) in CDM (Chronic Disease Management) specifically or in health generally." The aim of this study is to investigate the feasibility and effectiveness of semantic technology in improving the data quality in a specific clinical domain (organ transplant).

Chapter 3

Methodology

3.1 Stages of Research

The proposed study encompasses five stages, namely (1) data acquisition; (2) knowledge representation; (3) framework design; (4) selection of data quality dimensions; (5) system implementation and (6) evaluation. At the data collection stage, an existing dataset to be analyzed was acquired and a simulated dataset was prepared. Then, the structure of the proposed framework was established and the technologies to be adopted were explained during the design stage. Subsequently, knowledge identified in the analysis was represented in an OWL ontology. The dimensions to assess data quality were then selected (to be discussed in Section 3.5). Afterwards, data quality assessment queries were implemented according to chosen dimensions and scripts (which are used to transfer data and information) were imported to perform the queries. Finally, datasets were assessed using the framework and the results were validated by other researchers during the evaluation stage.

3.2 Data Acquisition

The Multi-Organ Transplant (MOT) program at the University Health Network is the first and largest transplant program in Canada (UHN, 2013). MOT provides health care services covering heart, lung, liver, kidney, pancreas and small bowel transplantation programs. This is a retrospective study that uses a research dataset containing clinical data of approximately 2,000 kidney transplant patients. The dataset was extracted from the Comprehensive Renal Transplant Research Information System (CoReTRIS), which collects data from EHR systems, labs, paper-based documents and other sources to serve kidney transplant research purposes. Patients have consented to this data release, and data anonymization has been performed. The dataset remains at the facility and is kept confidential when being analyzed. All data in CoReTRIS are checked against field restrictions when inserted into the database and are manually cleaned on a three-six data validation cycle. Therefore, high data quality is expected. The proposed framework was applied on this dataset to discover additional data quality issues and their underlying causes.

3.3 Knowledge Representation

It is common that a healthcare setting defines its own proprietary standard for representing health information which may be incompatible with another setting. To overcome this obstacle and achieve high interoperability, our framework needs a commonly-accepted health information model. The *open*EHR Reference Model (Garde, Knaup, Hovenga, & Heard, 2007) was chosen as a basis to develop the health information model, namely, MOT EHR Ontology, in this study. OpenEHR is a non-profit organization which aims to provide electronic medical records standards and tools for semantically manipulating health data. The standardization of a large collection of "archetypes" (a.k.a, clinical models) in the openEHR model has received formal acceptance as an International Organization for Standardization (ISO) standard (ISO 13606-2).

The openEHR Reference Model is publicly available as an eXtensible Markup Language (XML) document. A script was written to convert it into OWL classes which formed the top level classes of the ontology model. The "Patient" archetype and "Entry" archetypes were of particular interest. Each archetype is represented as a "Class" in the MOT EHR Ontology. The "Patient" class represents all patients in the program and is a main class in the ontology. However, clinicians can assume a subclass such that data can be entered to associate with patients. An entry archetype represents a self-contained piece of medical information that exists in the EHR and is further divided into five subtypes in openEHR (Beale, 2011):

- **Observation**: A measurement of the patient that has not been interpreted (e.g., a weight measurement, a lab report, a self-reported event)

- **Evaluation**: A subjective opinion or statement, often based on observations (e.g., a diagnosis, a risk assessment)
- **Instruction**: An order leading to an intervention (e.g., a prescription, an order to conduct a therapy)
- **Action**: An implementation of the ordered intervention (e.g., an operation, a drug intake)
- **Administrative**: A record of administrative events (e.g., an admission, a discharge)

In the meantime, a series of informal interviews were conducted with an experienced researcher at MOT to understand their data quality problems and clinical knowledge related to data.

3.4 SemDQ Framework Design

In computer science, a software framework is "a reusable design and building blocks for a software system and/or subsystem" (Shan & Hua, 2006). Reusability and extensibility are the key features of a framework and are desired in this study. As illustrated in Figure 3.1, the designed semantic framework is divided into four components: data sources, run-time system, knowledge platform and output.

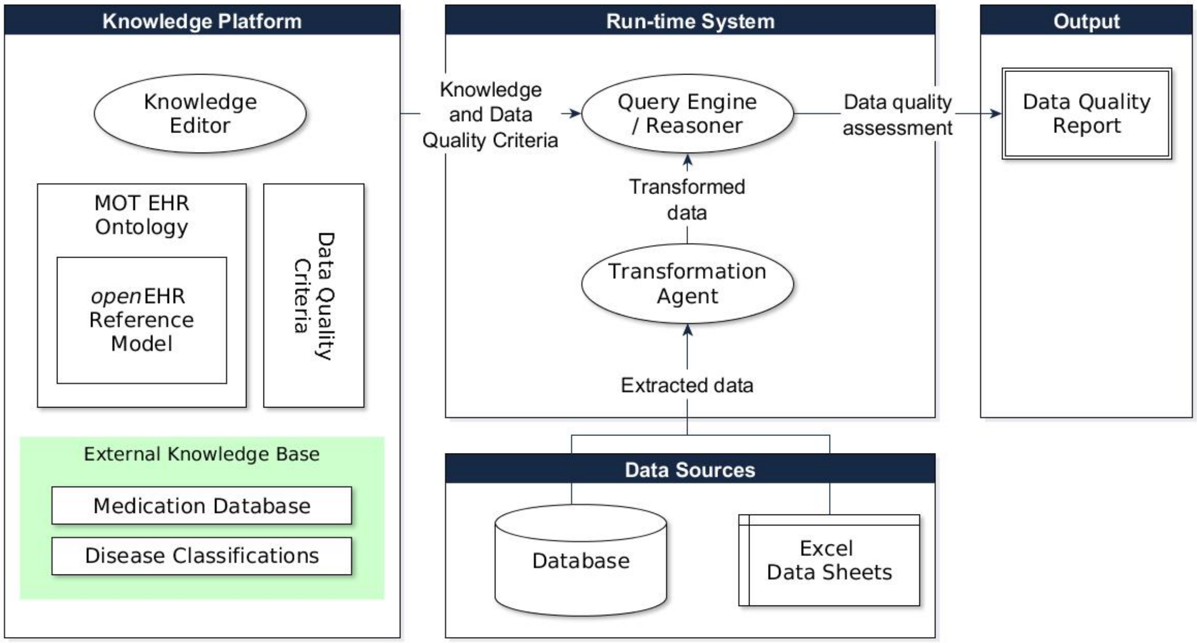

Figure 3.1: **The architecture of the semantic framework**

Data sources refer to the raw data to be analyzed, which usually need to be transformed to be compatible with the designed framework. A data source could come from an existing database, an excel data sheet or other data repositories. Depending on the type of data source, different transformation methods are required to extract the data and necessary information about data into the run-time system. In our framework, the transformation agent provides a collection of such methods. If the data come from a database, the agent will prompt proper SQL statements to query for data tables and use a script to transform the results into RDF datasets. If the data come from an excel data sheet, it will be converted into a comma-separated values (CSV) file and a script will then transform the data into RDF syntax and insert into the MOT EHR Ontology. Although not illustrated in

Figure 3.1, the data transformation process also includes translating variable information (labels, field definitions, cardinality relationships, etc.) into ontology classes, which is done either manually by using a knowledge editor or automatically by executing a script which analyzes a database schema and translates the information. It this study, the outcome of this translation process is a set of variable classes in the MOT EHR Ontology, which is one of the core components of the knowledge platform. The knowledge platform consists of the MOT EHR ontology, a data quality criteria ontology, external knowledge bases and a knowledge editor. The MOT EHR Ontology contains MOT-specific domain knowledge including data elements of all MOT programs, clinical entries of MOT programs (each is composed from some data elements) and logical restraints about the elements and entries. The entry classes are developed based on the openEHR Reference Model which has become an international standard of clinical entry specifications. The data quality criteria ontology defines data quality dimensions and includes SPARQL implementations. The data quality criteria include general and domain-specific criteria. General data quality criteria are independent of the domain and can be applied to any system (e.g., no missing value). Domain-specific data quality criteria are customized to suit domain-specific needs (e.g., a normal range of BMI for kidney transplant patients is different from the general population's range of BMI). The external knowledge bases contain a medication database ontology and a disease classifications ontology; some data quality criteria require inputs from them. In addition, a knowledge editor is utilized to input the knowledge into the ontologies and edits them. The data quality assessment process is performed at the query engine/reasoner module. A reasoner checks the ontologies in the knowledge platform regarding their consistency. Then, the query engine applies SPARQL implementations of the data quality criteria on the transformed datasets with the aid of knowledge inputted from the knowledge platform. Violations in data against each criterion are found after the queries, and a data quality report describing the data quality of the original dataset is produced as output.

3.5 Selection of Data Quality Dimensions

When determining data quality dimensions to measure, it is important to determine the view point. Wand and Wang (1996) provided two views of data quality dimensions. Internal view refers to dimensions selected by developers during the design and implementation stage of information systems. For example, accuracy, met when the recorded value is a true record of the reality, is associated with the internal view. External view refers to dimensions that are perceived by users when using the system. For example, timeliness, when defined by how recently the data were acquired, is relevant to the time point when the user uses the data and is associated with the external view. This study leaned towards an external view for the data quality dimensions because it mainly reused a readily collected and cleaned dataset. Practical needs of researchers at the collaboration site were also taken into consideration. As a result, completeness, consistency and logic consistency were chosen to guide the examination of data. The logic consistency dimension was further divided into three sub-dimensions. Definitions of the dimensions are:

- **Completeness**: Data are not missing or all necessary data have been included
- **Consistency**: Data are represented in conformity with other sources or at different

times, preferably in standard formats

- **Logic Consistency**: The logical concordance among data values
  - **Value in Range**: Inputted values should fit in predefined absolute and conditional ranges (e.g., aspirin could be prescribed in many circumstances but its use is contraindicated in hemophilia patients)
  - **Correct Temporal Sequence**: Recorded events should follow a reasonable temporal sequence (e.g., the first discharge date of a patient at a facility cannot occur ahead of the first admission date at the same facility)
  - **Correct Events According to Clinical Knowledge**: Recorded event occurrences are required to be in logical correspondence of other attributes (e.g., if a dialysis event is recorded immediately after kidney transplant, that indicates a graft failure and the corresponding diagnosis should be recorded)

3.6 System Implementation

The SemDQ ontology was constructed using the Protégé 4.3 software developed by Stanford University.[1] Knowledge was captured, represented in OWL language, and inserted into the ontology. Chosen data quality dimensions were also documented as annotated classes. Under each data quality dimension, one or more queries were implemented based on either dimension definitions or opinions from interviews with a researcher. The queries were written in SPARQL and referred to the variable vocabulary in the SemDQ knowledge

[1] This software is open source and available at http://protege.stanford.edu/download/registered.html

platform. A text editor was used to write the queries. Written in pseudo-code (informal programming language describing the algorithm, intended for human reading rather than machine reading), generic queries for each data quality sub-dimension are listed in Table 3.1.

Table 3.1: **Generic queries for data quality checking**

Data quality sub-dimension	**Query in plain words**	**SPARQL pseudo code***

Completeness	Query for missing or invalid values	SELECT ?element
		WHERE
		?element :hasValue [list of missing or invalid values]

Consistency	Query for informally represented values	SELECT ?element
		WHERE
		?element :hasValue [list of informal representations]

Value in Range	Query for values that fall out of a defined absolute or conditional range	SELECT ?entryA WHERE ?entryA :contains ?elementA. ?elementA :hasValue [list of out-of-range values]. OPTIONAL ?entryB :contains ?elementB. ?elementB :hasValue [list of values that defines out-of-range values of ?elementA].

Continued on Next Page...

Table 3.1 – Continued

Data quality sub-dimension	**Query in plain words**	**SPARQL pseudo code***
Correct Temporal Sequence	Query for a series of time points not following a reasonable order	SELECT ?entry WHERE ?entry :contains ?precedingElement. ?precedingElement :hasTime ?timepointA. ?entry :contains ?succeedingElement. ?succeedingElement :hasTime ?timepointB. FILTER (?timeValueA :laterThan ?timeValueB)**
Correct Events According to Clinical Knowledge	If an entry exists and must have some corresponding entries, query for records without one	SELECT ?EHR ?entry WHERE ?EHR :hasEntry ?entry. ?entry :hasCorrespondence [list of corresponding events] NOT EXIST :EHR :hasEntry [list of corresponding events]

* openEHR notions (EHR, entry, element, data value) are used as query variables

These SPARQL statements were executed on a SPARQL server which was established via Apache's Fuseki.[2] The server allowed any RDF dataset to be uploaded and queried by SPARQL statements. Each query was executed on the whole dataset. Problematic entries were found and a "Simple Ratio" assessment which measures the proportion of invalid records out of total records was made. After all queries were executed, the results were summarized.

[2] Fuseki is open-source and publicly available at http://jena.apache.org/download/index.html

3.7 Validation

Two parts of the proposed method need validation. One is the constructed semantic framework and the other is the generated data quality report. First, a simulated dataset was prepared by generating data values (in accordance of the data range and field restrictions of the variables found in CoReTRIS using the dataset codebook) which purposefully trigger all known types of data quality violations. The developed queries were applied on the simulated dataset to test the accuracy and effectiveness of the SemDQ framework. Deliberate errors were inserted into the simulated dataset for each criterion and were recorded, so that results from every query could be examined against the record to verify whether the framework is working correctly.

After the SemDQ framework passed the validation step, the CoReTRIS dataset was uploaded and examined. The SemDQ framework was applied to the dataset and problematic entries were reported. For all errors identified by the framework, a research assistant at MOT reviewed these errors. The results were then summarized and presented at a MOT seminar. Two senior MOT researchers reviewed and validated the errors identified by the SemDQ.

3.8 Ethics Approval

The institutional authorization to use the CoReTRIS dataset was granted by RQI (Research Quality Integration), the REB (Research Ethics Board) and the Department/Division Head of UHN on July 10th, 2013. The approval letter is attached as

Appendix A. The approval from the Ethics Office, University of Waterloo was obtained on November 25th, 2013. The approval letter is attached as Appendix B.

Chapter 4

Results

4.1 Description of the Data Sets

The CoReTRIS includes a data repository which continuously collects data of patients who have received care from the Kidney Transplant Program. The data collection is dated back to January 1, 2000. The data tables and variables in CoReTRIS and used in this study are listed in Table 4.1. All tables use the “Recipient research ID” variable as the primary key to protect patient identities.

Table 4.1: **Variables in CoReTRIS used in this study**

Data table	**Variables**
Demographic	**Recipient research ID** Recipient sex Recipient history of stroke
Transplant Admission Information	**Recipient research ID** Date of admission Date of discharge Date of transplant Whether a dialysis is needed within the 1st week after transplant First date of dialysis Last date of dialysis Height at admission
Transplant	**Recipient research ID** Date of transplant
Induction Therapy	**Recipient research ID** Type of therapy Total dose Total dose in ML Start date of induction therapy Stop date of induction therapy
CMV Prophylaxis	**Recipient research ID** Start date of prophylaxis Stop date of prophylaxis Prophylaxis drug type Prophylaxis drug dose Prophylaxis drug code

Continued on Next Page…

Table 4.1 – Continued

Data table	**Variables**
Recipient Diagnosis	**Recipient research ID** Date of transplant Whether is a pre-transplant diagnosis Date of diagnosis ICD code version ICD code Name of diagnosis
New-Onset Diabetes Information	**Recipient research ID** Date of transplant Method of diagnosis Date of diagnosis Date of last review
Recipient Weight	**Recipient research ID** Weight value Date of measure

A dataset containing data on 2,051 patients, whose transplant dates were between January 1, 1998 and June 1, 2013, was extracted from CoReTRIS and analyzed in this study. Also, a simulated dataset containing the same tables and variables as the CoReTRIS dataset was prepared. It included 29 patient instances with 465 entries. Data values in tables were randomly generated within defined ranges (e.g., dates were randomized between 1/1/1990 and 12/31/2012). Table 4.2 displays the distribution of entries of both datasets.

Table 4.2: **Entity distributions in the CoReTRIS and simulated datasets**

Data table	**Count of entries in the CoReTRIS dataset**	**Count of entries in the simulated dataset**
Demographic	2043	29
Transplant Admission Information	1749	58
Transplant	2400	29
Induction Therapy	2184	87
CMV Prophylaxis	4119	58
Recipient Diagnosis	32306	87
New-Onset Diabetes Information	367	*30
Recipient Weight	58379	87

* one additional entry in duplicate with an entry in this table was created for validation purposes

4.2 The Semantic Framework for Health Data Quality Assessment

4.2.1 The Health Data Quality Assurance Ontology

The constructed ontology consisted of five major classes, illustrated in Figure 4.1. The “Base Model” class contained basic definitions of variables from both the openEHR model and the CoReTRIS research database. All classes and datatypes of the openEHR model were completely retained from the original[1]; however, the entire hierarchy of openEHR was hard to navigate for visualization. For better visualization, the “Patient” class was isolated under the “Demographic” archetype class and various clinical episodes with data were put in the “Entry” archetype class. This arrangement is in accordance with openEHR online Clinical Knowledge Manager’s display of concepts (OpenEHR, 2012). The entry types in CoReTRIS (e.g., diagnosis, induction therapy) were modeled as extended subclasses of openEHR’s entry archetypes. The “Data Quality Criteria” class consisted of definitions of all data quality rules developed in this project. The “Disease Classification” class contained disease classification information. Details about each class are explained in subsequent sections.

[1] Resource available at http://www.openehr.org/wiki/display/spec/openEHR+1.0.2+UML+resources, retrieved on Oct 10, 2013

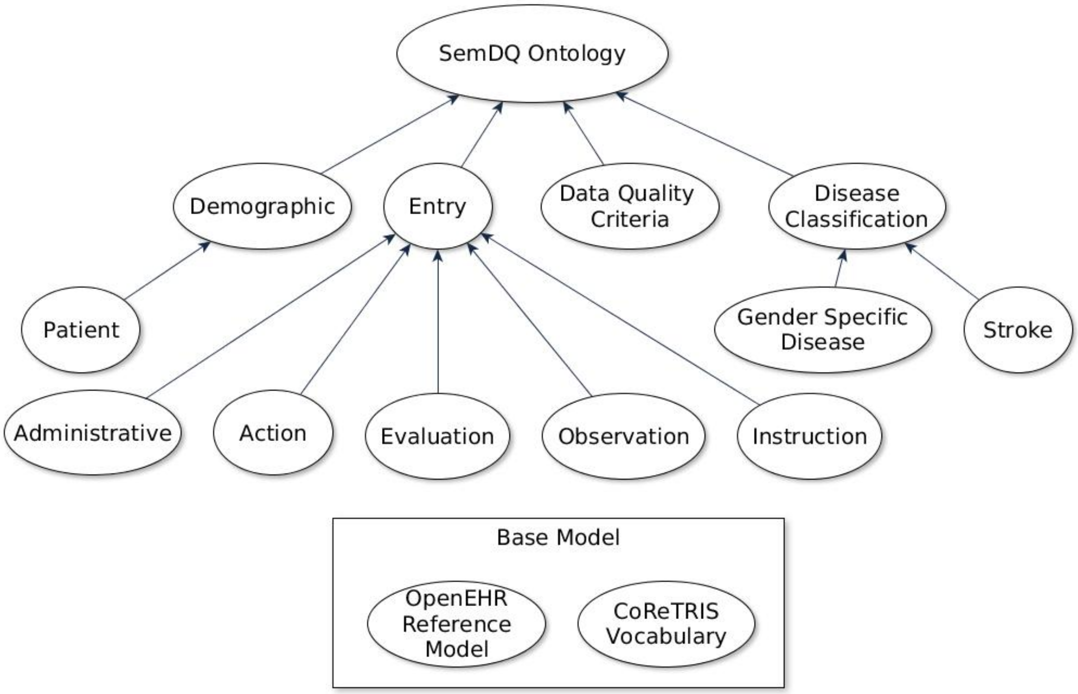

Figure 4.1: **The top-level classes of the SemDQ ontology**

4.2.2 Entry Mapping

Each data table from the CoReTRIS dataset except for the "Demographic" table was modeled as a subclass of one of the five openEHR "Entry" archetype classes. Table 4.3 lists the mapping relationship between the data tables and the modeled classes.

The variables in each table were modeled as datatype properties and each row of data was converted as an instance of the corresponding class. All instances were associated with a patient EHR instance, which was converted from data in the "Demographic" table if they shared an identical research ID. An object property named "has_entry" was constructed to

Table 4.3: **Mapping CoReTRIS data tables into entry subclasses**

Data table	**The corresponding parent class in the SemDQ ontology**
Transplant Admission Information	Admin Entry
Transplant	Action
Induction Therapy	Action
CMV Prophylaxis	Action
Recipient Diagnosis	Evaluation
New-Onset Diabetes Information	Evaluation
Recipient Weight	Observation

denote this association relationship. One patient EHR instance could be associated with multiple entries.

4.2.3 Standard Knowledge Base of Disease Classification

As a preliminary attempt, two categories of diseases were semantically expressed in the SemDQ ontology. One class labeled as “FemaleOnlyDisease” provided information about a number of diseases which are female-specific. Two breast-related problems and two uterine-related problems were modeled as its subclasses. Another class labeled as “Stroke” included a classification of different types of stroke, according to the American Heart Association (Kokotailo & Hill, 2005). Two datatype properties defined the disease classes. One was “hasICDCodeVersion” which indicated which version of ICD classification is in use and the other was “hasICDCode” which recorded the specific disease code. A restriction combining uses of both properties identified a disease. For example, “Cerebral infarction” is a subclass of Acute Ischemic Stroke (AIS), and it was defined through a OWL restriction statement written as “(has_ICD_code_version value "ICD 10") and (has_ICD_code value "I63.x")”. This statement is equivalent to the use of ICD-10 codes and associates I63 with cerebral infraction. The full hierarchy of expressed disease classifications with ICD codes is displayed in Table 4.4.

Table 4.4: **Expressed disease classifications**

Hierarchy of disease classification	**ICD code**	**ICD version**
Female only diseases		
Breast-related diseases		
Malignant neoplasm of central portion of breast, female	C50.11	ICD 10
Unspecified lump in breast	N63	ICD 10
Uterine Problem		
Leiomyoma of uterus, unspecified	D25.9	ICD 10
Other specified abnormal uterine and vaginal bleeding	N93.8	ICD 10
Stroke		
Stroke - Type AIS (Acute Ischemic Stroke)		
Acute, but ill-defined cerebrovascular disease	436	ICD 9
Central retina artery occlusion	H34.1	ICD 10
Cerebral infarction	I63.x	ICD 10
Occlusion and stenosis of precerebral arteries	433.x1	ICD 9
Occlusion of cerebral arteries	434.x1	ICD 9
Retinal vascular occlusion	362.3	ICD 9
Stroke, not specified as hemorrhage or infarction	I64.x	ICD 10
Stroke - Type ICH (Intracerebral Hemorrhage)		

Continued on Next Page...

Table 4.4 – Continued

Hierarchy of disease classification	ICD code	ICD version
Intracerebral hemorrhage*	431.x	ICD 9
	I61.x	ICD 10
Stroke - Type SAH (Subarachnoid Hemorrhage)		
Subarachnoid hemorrhage*	430.x	ICD 9
	I60.x	ICD 10
Stroke - Type TIA (Transient Ischemic Attack)		
Transient cerebral ischemia	435.x	ICD 9
Transient cerebral ischemic attacks and related syndromes	G45.x	ICD 10

* multiple codes describing the same disease are supported

4.2.4 Data Quality Criteria

A total of twelve data quality criteria were developed under the three data quality dimensions (for the list of dimensions please check Section 3.5). Each criterion is accompanied with a SPARQL-query implementation by which the query engine can traverse the whole dataset and identify questionable instances. The criteria are explained below, and the query implementations are presented in Appendix D. Also, whether the criterion is general or domain-specific is noted.

- Criterion under the Completeness dimension

  - **No missing value**: An input should be available for the variables that must have a value; general criterion (abbreviation: CMP_Missing)
  - **No out-of-range value**: Inputs of variables should fall into the valid range; general criterion (abbreviation: CMP_Out_Range)

- Criteria under the Consistency dimension

  - **Correct unit use in induction therapy entries**: The unit "milligram" should be used for all drug doses in induction therapy entries; other units such as "milliliter" are not allowed; domain-specific criterion (abbreviation: CST_Indu_Unit)
  - **No duplicate entries**: The existence of two or more identical entries is not allowed (by "identical", it means that all variables and all values are the same for both entries); general criterion (only examined new onset diabetes entries; abbreviation: CST_Duplicate)

- Criteria under the Value in Range sub-dimension, Logic Consistency dimension

  - **Body Mass Index (BMI) in expected range**: The patient's calculated BMI should be within an expected range; domain-specific criterion (set as $14<\text{BMI}<45$ according to feedback from MOT's researchers; abbreviation: RLC_BMI)
  - **Accordance between diagnoses of gender-specific disease and gender**: Gender-specific diseases should not appear in diagnoses of patients of the

other gender; general criterion (only male patients were checked against four female-specific diseasesf abbreviation: RLC_Gender)

  - **Correct dose in cytomegalovirus (CMV) drug records**: If either the recipient or the donor's CMV serology test result is positive, CMV prophylactic drug should be taken by the kidney recipient with a dose greater than zero; domain-specific criterion (abbreviation: RLC_CMV)

- Criteria under the Correct Temporal Sequence sub-dimension, Logic Consistency dimension

  - **Date of the Post-Transplant New-Onset Diabetes Mellitus (PTND) is in accordance with the date of transplant**: PTND is a research interest at MOT; any diabetes diagnosis that occurred before the transplant should not be marked as a PTND. Thus, the transplant date should be earlier than any PTND diagnosis dates; domain-specific criterion (abbreviation: TLC_New_Diabetes)
  - **Date of the pre-transplant diagnosis is in accordance with the date of transplant**: If one diagnosis is recorded as a pre-transplant diagnosis, the transplant date should be later than the diagnosis date; domain-specific criterion (abbreviation: TLC_PreTx_Diag)
  - **The first week dialysis record is in accordance with the date of transplant**: If there is a need for dialysis in the first week post-transplant, the first dialysis date should be within one week of the transplant date; domain-specific criterion (rule abbreviation: TLC_1stWeekDialysis)

  - **The start date is before the stop date**: For any paired dates indicating a period, the start date should be earlier than the stop date; general criterion (abbreviation: TLC_Paired_Dates)

- Criteria under the Correct Events According to Clinical Knowledge sub-dimension, Logic Consistency dimension

  - **Existence of diagnosis records when a disease history is recorded**: When a history of a disease category is recorded as true, there should be corresponding disease diagnosis entries; domain-specific criterion (only examined history of strokef abbreviation: ALC_History)
  - **Indication of delayed graft function (DGF) records when a first week dialysis is recorded**: When a kidney recipient undergoes dialysis within the first week after a kidney transplant, this is an indication of DGF and there should be a corresponding DGF diagnosis entry; domain-specific criterion (abbreviation: ALC_DGF)

4.3 Results of Data Quality Assessment

4.3.1 Overview

All data quality queries were executed on the simulated dataset first. The results were in accordance to the error records as expected, which confirmed the framework was functioning properly. A thorough data quality check was then performed on the CoReTRIS dataset, which took approximately 200 seconds to process 2,051 patient instances with 103,505 entries. A sum of 648 data quality rule violations were found. Detailed per-criterion results are listed in Table 4.5.

Table 4.5: **Overview of results of data quality checking on the CoReTRIS dataset**

Criterion	**Entities on which this criterion applies**	**Violation Count**	**Total Count**	**Percentage**
CMP_Missing	all data elements	48	103563	0.05%
CMP_Out_Range	all data elements	54	103563	0.05%
CST_Indu_Unit	all induction therapy entries	51	2184	2.34%
CST_Duplicate	all new onset diabetes entries	0	357	0.00%
RLC_BMI	all patient instances with both height and weight records available	93	1719	5.41%
RLC_Gender	all male patient instances	0	1275	0.00%
RLC_CMV	all patient instances necessary to be issued with CMV prophylactic drugs	83	1022	8.12%
TLC_New_Diabetes	all new onset diabetes entries	1	367	0.27%
TLC_PreTx_Diag	all diagnoses with the pre-transplant diagnosis variable recorded as "true"	15	5248	0.29%
TLC_1stWeekDialysis	all admission entries with the first week dialysis variable recorded as "true"	3	343	0.87%
TLC_Paired_Dates	all entries containing a pair of dates*	10	7657	0.13%

Continued on Next Page…

Table 4.5 – Continued

Criterion	Entities on which this criterion applies	Violation Count	Total Count	Percentage
ALC_History	all patient instances with the stroke history variable recorded as "true"	19	78	24.36%
ALC_DGF	all patient instances with the first week dialysis variable recorded as "true"	271	339	79.94%

* Paired dates include: admission date vs. discharge date; first dialysis date vs. last dialysis date; start date of prophylaxis vs. stop date of prophylaxis; start date of induction therapy vs. stop date of induction therapy

4.3.2 Explanation of Results by Each Criterion

Dimension: Completeness

The *"No missing value"* and *"No out-of-range value"* criteria were developed to denote the completeness of single data elements. When a data value was imported through the data import module, it was checked against the range definition of the variable it belonged to, and an output value with three possibilities would be generated: a regular value which is the standardized format of the original input, a "missing" mark, or an "out of range" mark. This process is illustrated by Figure 4.2. During the evaluation phase, the "missing" and "out of range" marks were extracted by the query module.

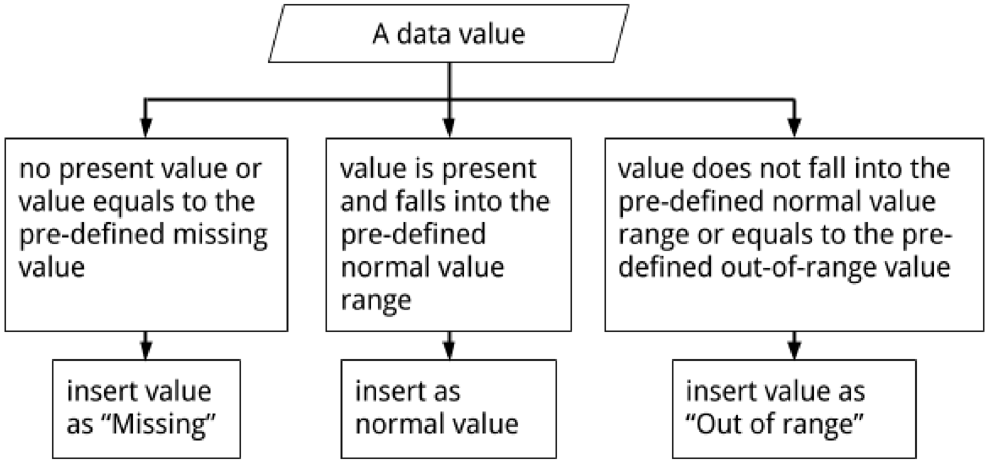
A data value
no present value or value equals to the pre-defined missing value
value is present and falls into the pre-defined normal value range
value does not fall into the pre-defined normal value range or equals to the pre-defined out-of-range value
insert value as "Missing"
insert as normal value
insert value as "Out of range"

Figure 4.2: **Mechanism of determining a missing, out-of-range or normal value**

Among the 103,563 entries from the CoReTRIS dataset, 48 (0.05%) entries were found containing at least one "missing" data mark. 54 (0.05%) entries were found containing at least one "out-of-range" data mark.

Dimension: Consistency

This category collected criteria that requires entries to be in standard forms. The *"Correct unit use in induction therapy entries"* criterion reviewed the unit of the drug dose in induction therapy entries, which was "milliliter (mL)" in the old database scheme but "milligram (mg)" in the new one. Any drug dose that recorded an mL unit was marked as a violation. Among the 2184 induction therapy entries from the CoReTRIS dataset, there were 51 (2.3%) entries that contained a drug dose in mL unit.

The *"No duplicate entries"* criterion examined whether there were duplicate entries. This was achieved by comparing two entries of the same type. If all variables in one entry shared the same values with the ones in another entry, the two entries were marked as duplicates. The total number of variables in the entry was needed for the comparison and it was interpreted from the ontology. The entries of the "New-Onset Diabetes Information" table were checked as a demonstration. Among the 357 new onset diabetes entries from the CoReTRIS dataset, no duplications were found.

Dimension: Logic Consistency, sub-dimension: Value in Range

This dimension contained three criteria checking whether a value fell in a pre-defined valid range. The ranges could be absolute; for example, *"Body Mass Index (BMI) in expected range"* examined whether the patient's BMI fell within a reasonable range, which was set between 14 and 45. The ranges could also vary under different conditions. The *"Accordance between diagnoses of gender-specific disease and gender"* criterion expressed that patients with a specific gender could not be diagnosed with a disease of the opposing gender. Four diseases were checked, as seen in Table 4.4. The *"Correct dose in cytomegalovirus (CMV) drug records"* checked whether a dose exists when a CMV prophylaxis drug has been prescribed to the recipient. The logic is illustrated in Figure 4.3.

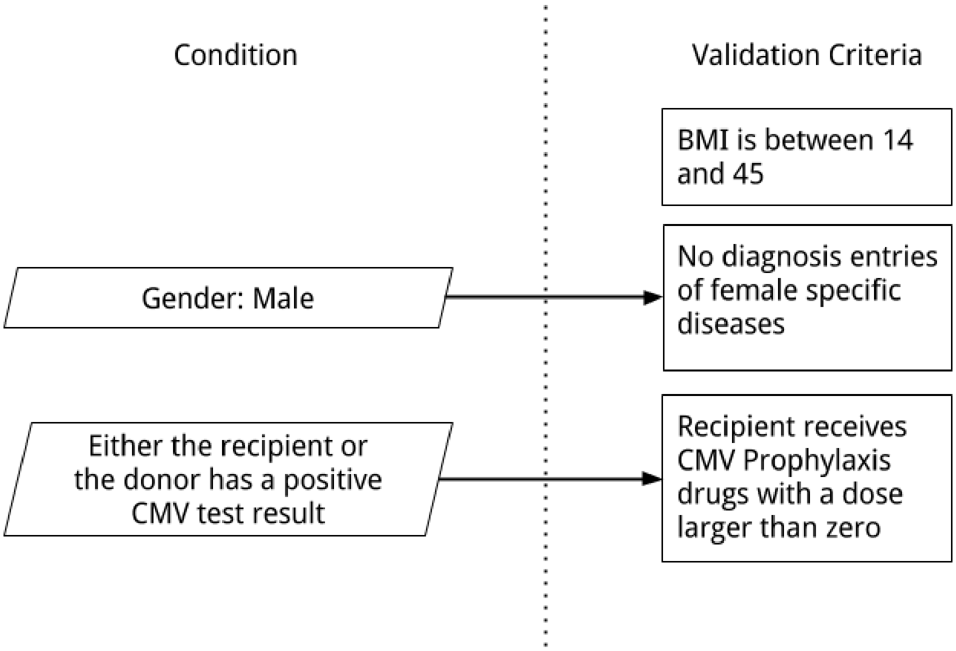
Condition
Validation Criteria
BMI is between 14 and 45
Gender: Male
No diagnosis entries of female specific diseases
Either the recipient or the donor has a positive CMV test result
Recipient receives CMV Prophylaxis drugs with a dose larger than zero

Figure 4.3: **Criteria in the Value in Range sub-dimension**

In the CoReTRIS dataset, among the 1,719 patient instances who had at least one weight and height measurement, BMI was computed for each pair of measurement, and 93 (5.4%) values were found having at least one abnormal BMI. Among 1275 male patient instances, none were found having a diagnosis of female-specific diseases. Among 1022 patient instances who had been issued with CMV prophylaxis drugs, 83 (8.1%) had at least one record where the dose was zero.

Dimension: Logic Consistency, sub-dimension: Correct Temporal Sequence

The criteria in this category compared dates and any incorrect sequence of events was noted. The *"Date of the post-transplant new-onset diabetes mellitus (PTND) is in accordance with the date of transplant"* criterion checked whether the date of any new onset diabetes diagnoses was earlier than the transplant date. If yes, then an inconsistency has occurred. Similarly, the *"Date of the pre-transplant diagnosis is in accordance with the date of transplant"* criterion checked if a recorded pre-transplant diagnosis had a date earlier than the transplant date. The *"The first week dialysis record is in accordance with the date of transplant"* criterion examined if the patient received dialysis in the first week after transplant, and if so, the first dialysis date should be within one week of the transplant date. Finally, the "The start date is before the stop date" criterion compared any group of paired dates containing a start date and a stop date whereby the start date should be earlier than the stop date. The logic is illustrated in Figure 4.4.

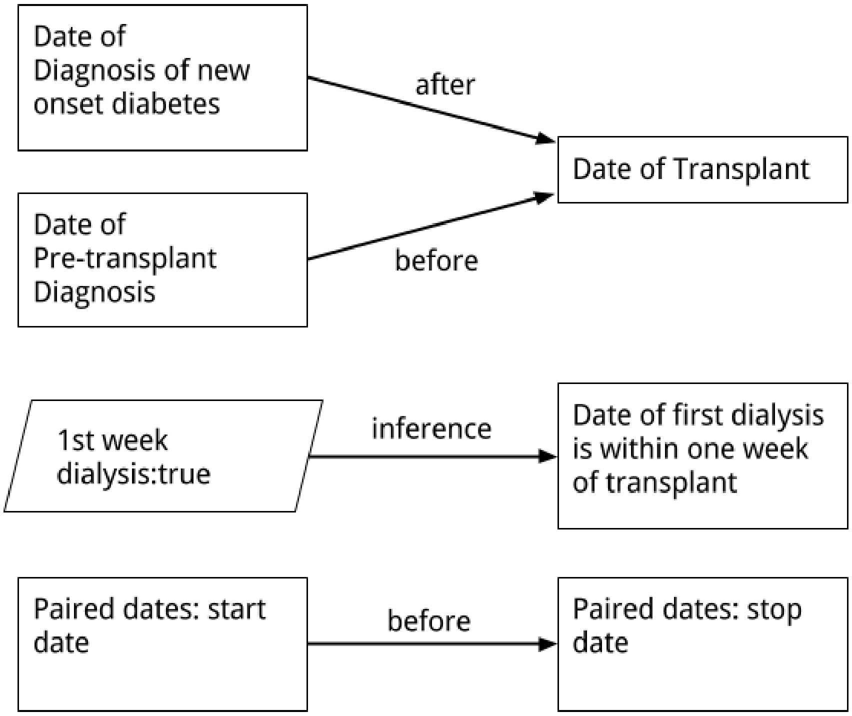
Date of Diagnosis of new onset diabetes
after
Date of Transplant
Date of Pre-transplant Diagnosis
before
1st week dialysis:true
inference
Date of first dialysis is within one week of transplant
Paired dates: start date
before
Paired dates: stop date

Figure 4.4: **Criteria in the Correct Temporal Sequence sub-dimension**

In the CoReTRIS dataset, among the 367 new onset diabetes entries, 1 (0.3%) entry was found having a diagnosis date earlier than the transplant date. Among the 5,248 pre-transplant entries, 15 (0.3%) entries were found having a transplant date earlier than the diagnosis date. Among the 343 transplant admission entries with a record of first week dialysis after transplant, 3 (0.9%) entries were found having the first dialysis date more than seven days later than the transplant date. Among 7,657 entries that included paired dates, there were 10 (0.1%) entries in which the start date occurred after the stop date of the pair.

Dimension: Logic Consistency, sub-dimension: Correct Events According to Clinical Knowledge

Criteria in this dimension examine the sequential integrity of a series of clinical events. Implemented criteria included *"Existence of diagnosis records when a disease history is recorded"* and *"Indication of delayed graft function (DGF) records when a first week dialysis is recorded"*. The former expressed that if one disease history variable has a true value, existences of diagnosis entries of that type of disease should be found. In the latter, if a patient was put on dialysis within the first week of his/her kidney transplant, some existences of diagnosis entries of delayed graft function should be found because immediate dialysis treatment indicates the DGF. The logic is illustrated in Figure 4.5.

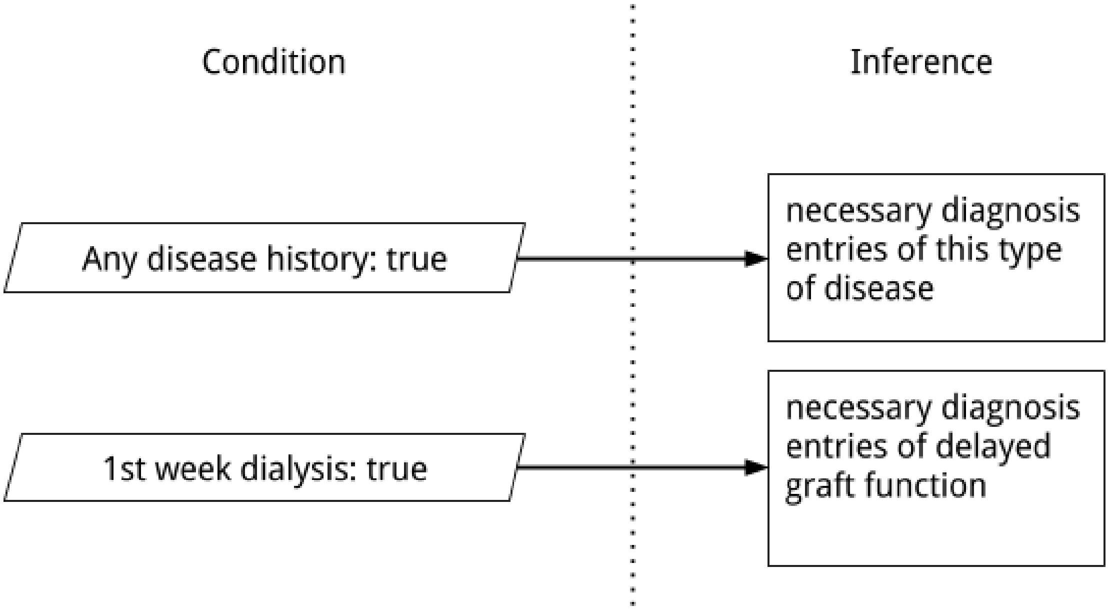
Condition
Inference
Any disease history: true
necessary diagnosis entries of this type of disease
1st week dialysis: true
necessary diagnosis entries of delayed graft function

Figure 4.5: **Mechanism of detecting disagreement in records**

In the CoReTRIS dataset, among the 78 patient instances that were marked as having a history of stroke, 19 (24.4%) were not associated with any diagnosis of stroke in the diagnosis table. Among the 339 patient instances that had a record of required dialysis within the first week after transplant, 271 (79.9%) were not related to any diagnosis of delayed graft function.

4.4 Result Validation

All queries successfully examined the simulated dataset. Results from assessing the CoReTRIS dataset were presented to the researchers at MOT and feedback were acquired. All queries and resulting violation cases were understood and discussed among the researchers. All violations found in the CoReTRIS dataset were confirmed, while some results were provided with interpretations. The validation by researchers is explained in more detail in the next chapter. So far, the pilot implementation of the SemDQ framework has been proven as accurate and effective.

Chapter 5

Discussion

5.1 Data Quality of the CoReTRIS Dataset

Our data analysis used an audited CoReTRIS dataset and therefore the rate of data violations was predicted to be rare. The rate of violations was generally low except for four criteria where it was above 5%:

- Body Mass Index (BMI) in expected range (RLC_BMI) violations: 5.41% of the patients had at least one BMI out of the expected range (between 14 and 45).
- Correct dose in cytomegalovirus (CMV) drug records (RLC_CMV) violations: 8.12% of the drug records recorded a drug dose of zero.
- Existence of diagnosis records when a disease history is recorded (ALC_History)

violations: 24.36% of the patients recorded with a history of stroke were not associated with any diagnoses of stroke.

- Indication of delayed graft function (DGF) records when a first week dialysis is recorded (ALC_DGF) violations: 79.94% of the patients recorded with a first week dialysis were not associated with any DGF records.

Among all data entries, 48 (0.05%) entries contained at least one missing value and 54 (0.05%) entries contained at least one out-of-range value. These two numbers were initially high (10,106 and 1,548, respectively), but after a thorough investigation, violations that are false positive or not clinically meaningful were excluded. For example, some medication or therapy records did not have a stop date, which means that they were still open at the time of data obtained instead of missing the stop date. Also, the query detected many missed diagnosis dates. This is because when the diagnoses database was collected, any pre-transplant diagnosis dates were not recorded because it was considered to not be relevant in transplant research. Hence, if the field “Is this a pre-transplant diagnosis” responds to “Yes”, then no date is assumed. Dates are only issued for post-transplant diagnoses. The semantic framework is capable of gradually adding such exclusion rules, so that the results will be more precise and meaningful.

The 48 entries detected with values missing included 34 admission entries missing a height measurement. Since the variable defines the height at the time of transplant, a missing record implies that the research assistant who searched for this value had exhausted all options and was unable to locate the value. If a research study is flexible enough to replace this height value with the most recent available height measurement, the investigator can choose to do so and request a research assistant to abstract this information. At the time of this study, this value is considered missing. In addition, 13 CMV prophylaxis entries were found without a drug dose recorded, which is critical for some research projects. Finally, 1 diagnosis record was not associated with an ICD code. There are two issues at hand when ICD codes are missing. The first most likely reason is that there is no such ICD code to represent the diagnosis in question. The second reason is that the data abstractor did not have sufficient clinical knowledge to link a diagnosis to another one with an available ICD code.

The 54 out-of-range violations were all abnormal diagnosis dates in diagnosis entries. Either wrong content was in that field (e.g., a note like "left leg" was typed there) or the date was not correctly inputted (e.g., "in 1996", "9/27/20210" or "23/7/9"). These out-of-range values cannot be used for research.

The *"Correct unit use in induction therapy entries"* criterion aimed to reduce confusion about drug dose units in induction therapy entries. The unit "milliliter" was used in the old database, but this was changed to "milligram" due to new treatments. The query engine found that 51 entries were still using the old milliliter unit, out of the 2184 entries with a record of drug units. One research assistant at MOT confirmed this as a defect left from one upgrade of the database.

The two criteria checking diagnosis records, stroke and delayed graft function, in the "event correspondence" dimension yielded results with violation rates higher than those of other rules. The rates were 24.36% and 79.94% of the applicable entries, respectively. There are in fact two diagnoses tables in CoReTRIS. One captures major categories of diagnoses (e.g., cardiovascular disease, stroke, diabetes mellitus) and the other table captures specific diagnoses. In practice, the major diagnosis category information might be enough for research purposes. This piece of information was introduced at the seminar and this study has not been updated with the new information. Another possible interpretation for the high rates is that the diagnosis table did not contain diagnoses for all patient instances occurring in other tables, which is likely to happen since the research dataset that was used stored a proportion of all transplant patients. Other than this, the results have shown few omissions of clinical events.

Results of most rules indicated that the CoReTRIS dataset is of high quality. Among 1,719 patients with at least one computed BMI record, 93 (5.41%) patients had at least one BMI out of the expected range (between 14 and 45). Manual review is needed for the researchers to determine whether the out-of-range entry indicates an actual state of patient health or a measurement error. The third criterion in the range accuracy dimension examined 1,022 CMV prophylaxis entries and 83 (8.12%) of them recorded a drug dose of zero. When the drug issued was on hold or out of inventory, a research assistant would record a dose of zero to indicate this anomaly. However, the codebook did not reflect this implicit knowledge. The MOT EHR Ontology has captured this new knowledge, as illustrated by Figure 5.1.

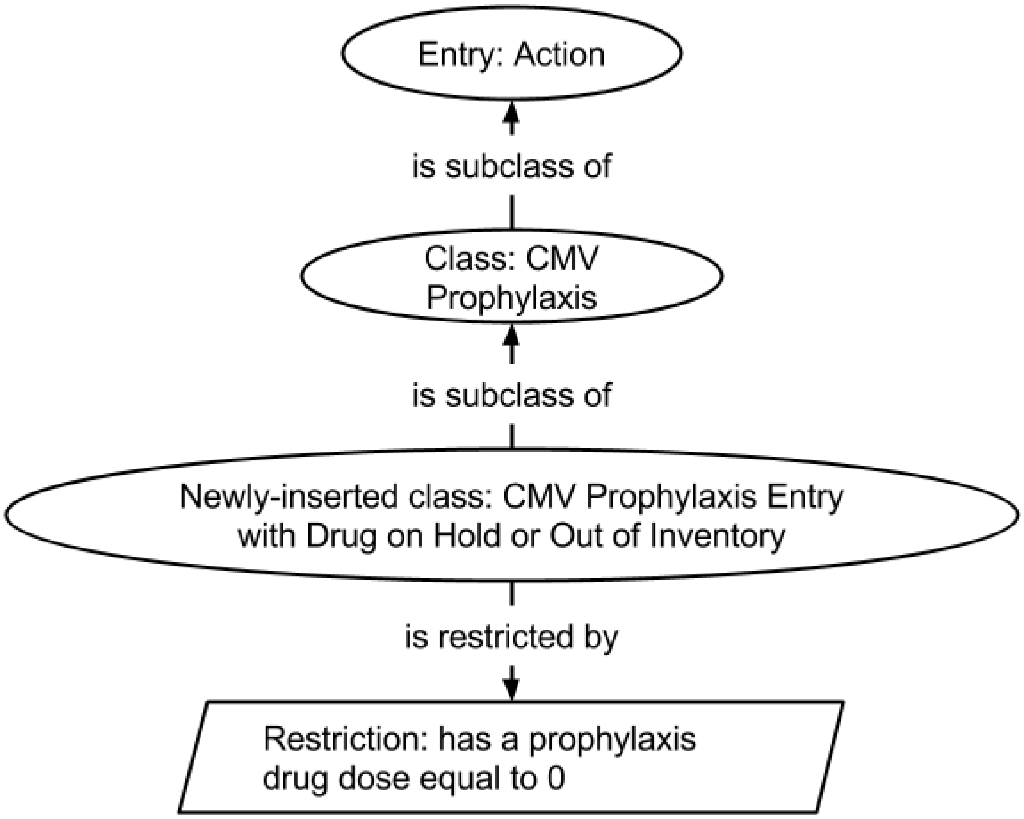
Entry: Action
is subclass of
Class: CMV Prophylaxis
is subclass of
Newly-inserted class: CMV Prophylaxis Entry with Drug on Hold or Out of Inventory
is restricted by
Restriction: has a prophylaxis drug dose equal to 0

Figure 5.1: **Updating the ontology with new knowledge**

Four criteria examined temporal accuracy about the CoReTRIS dataset. One out of 367 post-transplant new-onset diabetes mellitus entries was found to have a diagnosis date before transplant. Fifteen out of 5,248 pre-transplant diagnosis entries were found to have a diagnosis date after transplant. 343 admission entries were associated with a dialysis received within the first week after transplant; however, 3 of them were found to have the first dialysis date after one week of the transplant. Finally and interestingly, 10 of 7,657 pairs of start and stop dates had a reverse order, i.e., the start date occurred after the stop date. Although violations were found, only less than one percent of each set of related entries were affected, indicating an overall good status of temporal accuracy in the dataset.

Zero violations were found when checking gender-specific disease diagnoses and duplicate entries. Two possible reasons are that no violation indeed existed or that the rules did not apply correctly. To demonstrate that the rules are effective, we referred to the results by checking the simulated dataset. Two violations were found against the *"Accordance between diagnoses of gender-specific disease and gender"* criterion and one against the *"No duplicate entries"* criterion, thus showing effectiveness of the rules.

In this work, all criteria were treated equally so that any criterion was considered as important as the other. One of the reasons for doing so is that the importance of a given criteria (e.g. BMI $<$ 45) varies across different organ programs. The implemented framework did not expand on user preferences yet, though the SemDQ framework can be easily expanded to handle "weighting" on each criterion.

All criteria implementations are included in the Appendix D for review.

5.2 Discussion on the Ontology

The semantic framework and the established ontology served well for the purposes of this study. The ontology conceptualized knowledge about variables and entries, provided information of disease classifications and documented data quality criteria. All other components seamlessly connected to the ontology. The clinical knowledge base and its update are reflected in the error detection process instantaneously without reprogramming or re-factoring of the dataset. The separation of the knowledge base (what) and the execution (how) is one of the greatest advantages of the SemDQ framework, where the knowledge base and validation rules, especially the meta rules can be easily reused in a different dataset collected for a different study whenever appropriate.

Disadvantages of the established ontology existed as well. First, the ontology was peer-reviewed but no quantified evaluation was performed. The validity of class hierarchy organization and semantic definition could be disputed. However, this shortcoming is acceptable because the focus of this thesis is not about establishing a complete ontology of the domain knowledge, but introducing a new way to assess data quality. Our ontology successfully provided the semantic framework for the data quality assessing process. Secondly, although some validation rules were incorporated, the corresponding knowledge was not exhaustive or complete, e.g., information about gender specific diseases only listed four diseases for females. This is because the current work was more demonstrative than productive. A comprehensive disease classification ontology can be developed using SNOMED terminology and would be a natural next step of this work.

The strength of MOT EHR ontology included integrating the commonly adopted openEHR reference model and archetypes which features descriptive power and allows collaboration. Our ontological hierarchy of the clinical episodes was built based on the "Entry" archetypes of the openEHR reference model. The same model has been used across the world, allowing other organizations to share terminologies and definitions. Integrating the openEHR reference model ensures that the SemDQ framework possesses wide compatibility with other health care information systems and long-term extensibility. In addition, our ontology was written in the web ontology language, which is an international standard. The expressiveness of logic relations embedded in OWL language has been proven to make writing non-proprietary and platform-independent validation rules much easier as well as exchangeable in a range of computing environments. Database schemas focus on data storage and easily lose meaning during an upgrade, while ontologies maintain semantics and allow the framework to integrate multiple sources of data as long as the concepts are consistent. Finally, our ontology was built with complete open-source and standard technologies. Anyone willing to contribute will benefit from its transparent structure and syntax, instead of facing obscure and proprietary implementations.

5.3 Discussion on the Semantic Approach

5.3.1 Advantages

The greatest advantage of the semantic framework is the reusability of both the ontological model and the set of rules. The ontological model was developed as a collection of well defined concepts for the MOT kidney transplant program. Many variables in the kidney transplant program are identical to variables from other solid organ transplant programs in the heart, pancreas, etc. Thus, if there is a need for a data quality check on datasets from those domains, variable definitions and rules are ready to be reused.

Extensibility is another inherent advantage of the SemDQ framework. The open source ontology editor, Protégé, provides an easy-to-use graphical interface to view and update the ontology. Disease classifications, coding definitions and drug interaction information can be effortlessly added as subclasses into the current ontology, or imported from an existed ontology. The new information will be instantly introduced into the reasoning process without any recoding.

The efficiency of our framework was impressive. Following a structured programming approach, the whole process of importing the data, setting up the query engine and running the queries were semi-automatically run by scripts. It took less than half an hour to complete a cycle. However, experiments are needed to compare the total time with current Excel or SAS based solutions.

5.3.2 Technology Choice

There are two possible technologies we could employ when implementing the data assessment criteria. One is SPARQL Protocol and RDF Query Language (SPARQL), which is a RDF query language, and the other is Semantic Web Rule Language (SWRL),[1] a rule language. SWRL respects OWL restrictions while SPARQL does not. For example, SWRL can directly express the rule "a parent is a person having at least one child" using OWL's minimum cardinality restriction. To achieve the same in SPARQL, a sub-query must be established to count the number of children before the query defining a parent, which is tedious. The most appealing feature of SWRL for the SemDQ is its implementation of OWL inference rules while SPARQL has only a limited set of inference rules, limited mostly to the subsumption rule (i.e. the expression of type A may also be given type B if B is a subtype or a part of A). For example, if a patient is known as a kidney organ transplant patient, SPARQL is able to infer that he/she is also an "organ transplant" patient. In summary, SPARQL was selected over SWRL in the implementation of SemDQ despite SWRL's strength in expressing complex inference rules because: 1) SPARQL

[1] Documentation at http://www.w3.org/Submission/SWRL/, accessed on May 21, 2013

reflects current knowledge (i.e., limited inferences) thus a SPARQL query's efficiency is considerably higher than that of a SWRL rule's for several criteria. This is particularly true during the checking for duplicate entries. 2) SPARQL supports a negation query (e.g., query for a set of patients not associated with a specific diagnosis) but SWRL cannot because it respects OWL's open world assumption.[2] 3) Although SPARQL's capacity is limited compared to SWRL, it is sufficient for implementing all criteria identified in this study. Moving forward, the two technologies can be regarded as complementary, and should be employed according to the complexity of inference rules and the requirements for performance when checking large datasets.

5.3.3 Implementation Constraints

One concern is over- or under- generalization of criteria since the scope (i.e., how wide is the range of problems covered by a criterion) matters. Narrowing down to a specific problem will lead the criterion to be precise, but may lose extensibility. For example, the *"Correct unit use in induction therapy entries"* criterion only dealt with one variable - the drug dose unit of an induction therapy entry. It could have been extended to correct any unit errors if a group of unit classes was properly defined. The other extreme is to design a criterion with a broad compatibility, but this may cause ambiguous interpretation of a dataset. For instance, the *"No out-of-range value"* criterion examined out-of-range values, but it could have further clarified by denoting multiple types of out-of-range values such as lexical errors, wrong syntax of coding or erroneous variable types. A balance is worth

[2] Explained in <OWL Web Ontology Language Guide>, which is available at http://www.w3.org/TR/owl-guide/

pursuing when the rules are reviewed and updated in the future.

Although the current data quality ensuring approach at MOT contains flaws, it is smoothly functioning and well integrated with the clinical processes. Introducing a new system requires huge efforts on development, testing, learning and deployment. Time and funding will need to be provided. Furthermore, maintaining and updating the knowledge base and rules requires continuous inputs. Although the Protégé ontology editor is intuitive to learn and use, people trained with an understanding of computational logic are required to code for class restrictions. Knowledge translation requires intensive collaboration of experts from both the medical research and the information technology teams. A constructive collaboration has taken place during this study, and will be essential for the future success of the framework.

Chapter 6

Future Work

The constructed SemDQ framework has shown to be effective and has great potential in actual practice. Since the current validation criteria are illustrative and far from complete, the next immediate step is to expand the validation criteria for the CoReTRIS system with the aid of researchers at MOT. More medical knowledge translated into the framework is needed to suit the practical needs of research-grade data. Further, the use of SemDQ could be expanded over other transplant programs at MOT, such as heart, pancreas and liver transplant programs. An assessment on the reusability of criteria in other groups or the extent of modifications to current criteria would help us understand the scope of changes if the framework were to be applied elsewhere.

Patient profiling is another potential outcome of this study. With the computational abilities in data validation and organization, the SemDQ framework is able to detect outliers in data as well as abnormal changes, e.g., a sudden drop of BMI within a period.

The framework could also identify patient transplant status by applying logic consistency checks across different sub-domains (e.g., the correlation between lab results, diagnosis and medication).

There is also a plan to fully integrate the ICD-10 disease classification system, so that each diagnosis can be mapped to a corresponding ICD code. Once integrated, gender-specific rules and problem list rules could be extended to cover a wider range of diseases. Another possible extension of the knowledge base is to include a standardized drug database. With a comprehensive drug and diagnosis knowledge base, the detection of the adverse drug effect violation can be carried out at a deeper level. One such example would be the detection of the use of anticoagulant agents on a patient with an ulcer or open-wound-related condition.

The framework could be directly integrated into the data collection process at MOT. Data quality issues occur at the start of data entry. For example, due to insufficient training, a user may introduce mistakes when entering data. The SemDQ framework is able to perform continuous and real-time data quality checks when an entry is inputted into the database, and it can provide feedback to accelerate the data cleansing process.

Finally, the error reports generated from the SemDQ framework can serve as an indicator of a dataset's quality using the "Simple Ratio" score. By summarizing violation counts of each rule, categorizing the violations and determining scores in each category, a total score can be computed. People could manually allocate weights to each category, e.g., put heavy weights on the categories they are most concerned with, so that scoring customization could be achieved.

APPENDICES

Appendix A

Research Ethics Board Approval from the University Health Network

UHN Toronto General
Toronto Western
Princess Margaret
Toronto Rehab

University Health Network
Research Ethics Board
10th Floor, Room 1056
700 University Ave
Toronto, Ontario, M5G 1Z5
Phone: (416) 581-7849

Notification of REB Approval for Access to Retrospective Data for Research Purposes

Date: July 10th, 2013

To: Dr. S. Joseph Kim
11C, 1183, CSB, Toronto General Hospital, 200 Elizabeth St.
Toronto, Ontario, Canada M5G 2C4

Re: 13-6402-AE
Accelerating Error Detection in Health Data Using Semantic Technology

REB Review Type:	Expedited
REB Initial Approval Date:	July 10th, 2013
REB Expiry Date:	July 10th, 2014

Documents Approved:

Protocol v3.0	Version date: July 10th, 2013
Data Collection Form v2.0	Version date: June 29th, 2013

The UHN Research Ethics Board operates in compliance with the Tri-Council Policy Statement; ICH Guideline for Good Clinical Practice E6(R1); Ontario Personal Health Information Protection Act (2004); Part C Division 5 of the Food and Drug Regulations; Part 4 of the Natural Health Products Regulations and the Medical Devices Regulations of Health Canada. The approval and the views of the REB have been documented in writing.

Furthermore, members of the Research Ethics Board who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB.

Best wishes on the successful completion of your project.

Sincerely,

Anna Gagliardi, PhD
Co-Chair, University Health Network Research Ethics Board

Appendix B

Research Ethics Approval from the Office of Research Ethics, University of Waterloo

Lingkai Zhu <pig.junior@gmail.com>

Ethics Clearance (b) (ORE # 19180)

ORE Ethics Application System <OHRAC@uwaterloo.ca> Mon, Nov 25, 2013 at 8:15 AM
To: helen.chen@uwaterloo.ca
Cc: l49zhu@uwaterloo.ca

Dear Researcher:

This is to advise that the ethics review of your application to conduct research:

Title: A Semantic Framework for Data Quality Assurance in Medical Research
ORE #: 19180
Faculty Supervisor: Helen Chen (helen.chen@uwaterloo.ca)
Student Investigator: Lingkai Zhu (l49zhu@uwaterloo.ca)

has been completed through a University of Waterloo Research Ethics Committee. Based on the outcome of the ethics review process, I am pleased to advise you that your project has received ethics clearance.

Note 1: This ethics clearance from a University of Waterloo Research Ethics Committee is valid for one year from the date shown on the certificate and is renewable annually. Renewal is through completion and ethics clearance of the Annual Progress Report for Continuing Research (ORE Form 105).

Note 2: This project must be conducted according to the application description and revised materials for which ethics clearance has been granted. All subsequent modifications to the project also must receive prior ethics clearance (i.e., Request for Ethics Clearance of a Modification, ORE Form 104) through the Office of Research Ethics and must not begin until notification has been received by the investigators.

Note 3: Researchers must submit a Progress Report on Continuing Human Research Projects (ORE Form 105) annually for all ongoing research projects or on the completion of the project. The Office of Research Ethics sends the ORE Form 105 for a project to the Principal Investigator or Faculty Supervisor for completion. If ethics clearance of an ongoing project is not renewed and consequently expires, the Office of Research Ethics may be obliged to notify Research Finance for their action in accordance with university and funding agency regulations.

Note 4: Any unanticipated event involving a participant that adversely affected the participant(s) must be reported immediately (i.e., within 1 business day of becoming aware of the event) to the ORE using ORE Form 106. Any unanticipated or unintentional change which may impact the research protocol, information-consent document or other study materials, must be reported to the ORE within 7 days of the deviation usng ORE Form 107.

Best wishes for success with this study.

Susanne Santi, M. Math.,
Senior Manager
Office of Research Ethics
NH 1027
519.888.4567 x 37163
ssanti@uwaterloo.ca

Appendix C

Common Used RDFS and OWL Vocabularies

C.1 Common Used RDFS vocabulary

This vocabulary is summarized from http://www.w3.org/TR/rdf-schema/.

- Classes
  - **rdfs:Class**: classes are groups of resources. The members of a class are often referred as instances of the class
  - **rdfs:Resource**: all things described by RDF are instances of rdfs:Resource
  - **rdfs:Literal**: the class of literal values such as strings and integers

  - **rdfs:Datatype**: the class of RDF datatypes[1]
  - **rdf:Property**: the class of RDF properties, which describe relations between subject resources and object resources.

- Properties
  - **rdf:type**: a property stating that the subject resource is an instance of the object resource
  - **rdfs:subClassOf**: a property stating that the subject resource is a subclass of the object resource and they are both classes. All instances in the subject class are also the instances of the object class.
  - **rdfs:domain**: states that all subjects of one property are instances of one or more classes
  - **rdfs:range**: states that all objects of one property are instances of one or more classes
  - **rdfs:comment**: used to provide a human-readable description of a resource
  - **rdfs:label**: used to provide a human-readable version of a resource's name

C.2 Common Used OWL vocabulary

This vocabulary is summarized from http://www.w3.org/TR/owl-ref.

[1] The applicable datatypes are listed in http://www.w3.org/TR/2004/REC-rdf-mt-20040210/#DTYPEINTERP

- Classes
  - **owl:Thing**: the set of all individuals. (The word “individual” can be interchangeably used with “instance” except in some versions of OWL, where “individual” strictly refers to an instance of a class and cannot be a class or a property)
  - **owl:Class**: defines a class similar to rdfs:Class but with OWL features
  - **owl:Restriction**: a subclass of owl:Class that describes a value constraint, using the owl:onProperty property to link to a particular property
  - **owl:ObjectProperty**: in OWL, properties are distinguished into two main categories, object properties link individuals to individuals.
  - **owl:DatatypeProperty**: datatype properties link individuals to data values.
  - **owl:FunctionalProperty**: a functional property is a property that can have only one (unique) value for each individual
  - **owl:SymmetricProperty**: if a subject is pointed to an object by a symmetric property, the reverse will hold as well
  - **owl:TransitiveProperty**: if individual A is pointed to individual B by a transitive property P and individual B is pointed to individual C by P, the statement that A is pointed to individual C by P will hold
- Properties
  - **owl:allValuesFrom**: used to describe a class of all individuals for which all values of the property concerned are either members of the class extension of

the class description or are data values within the specified data range

- **owl:someValuesFrom**: describes a class of all individuals for which at least one value of the property concerned is an instance of the class description or a data value in the data range.
- **owl:hasValue**: describes a class of all individuals for which the property concerned has at least one value semantically equal to a particular value ("semantically equal" means mapping to the same URI)
- **owl:cardinality** / **owl:minCardinality** / **owl:maxCardinality**: describes a class of all individuals that have exactly / a minimum of / a maximum of N semantically distinct values (individuals or data values) for the property concerned, where N is the value of the cardinality constraint.
- **owl:equivalentClass**: when one class is linked to another class by the owl:equivalentClass property, both class extensions contain exactly the same set of individuals
- **owl:disjointWith**: when one class is linked to another class by the owl:disjointWith property, the class extensions of the two class descriptions involved have no individuals in common

Appendix D

Data Quality Rules

All data quality criteria were implemented in SPARQL, the codes are listed below.

Prefixes

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX k: <http://mot.uhn.ca/ottr/kidney#>

PREFIX o: <http://mot.uhn.ca/ottr/openEHR#>
```

Data Description - Count Patient Instances

```
SELECT (COUNT(distinct ?p) as ?patientCount)

WHERE {?p a k:Patient}
```

Data Description - Count Entries

```
SELECT (COUNT(distinct ?x) as ?entryCount)
?entryLabel #distinct entry count
WHERE { ?x a owl:NamedIndividual . ?x a ?c.
?c rdfs:label ?entryLabel.
?c rdfs:subClassOf* o:ENTRY . }
GROUP BY ?entryLabel
```

Completeness - No missing value

```
SELECT  (COUNT(DISTINCT ?entry) as ?missingEntryCount)
WHERE {?p k:has_entry ?entry .
{ ?entry ?v "Missing"} UNION
{ ?entry ?v ?miss. ?miss a k:MissingValue. } }
```

Completeness - No out-of-range value

```
SELECT  (COUNT(DISTINCT ?entry) as ?outofrangeEntryCount)
WHERE { ?p k:has_entry ?entry .
{ ?entry ?v "OutOfRange"} UNION
{ ?entry ?v ?out. ?out a k:OutOfRangeValue . } }
```

Consistency - Correct unit use in induction therapy entries

```
SELECT (COUNT(distinct ?indu) as

?inductionTherapyWithMLUnitentryCount)

WHERE {

?p a k:Patient. #?p a patient

?indu a k:class580. #?t a Induction Therapy Record

?p k:has_entry ?indu. #?p has entry of ?induRATS

?indu k:type_of_therapy ?type.

?indu k:total_dose_ML ?doseML.}
```

Consistency - No duplicate entries

```
SELECT (COUNT (?wC)/2  as ?duplicateEntryCount) # ?eC ?attrCount

{{

SELECT (COUNT (?w) /2  as ?wC) ?x ?y ?p ?attrCount

WHERE { ?p k:has_entry ?x .

?p k:has_entry ?y .

?x a ?Class. ?y a ?Class. ?x ?w ?z . ?y ?w ?z .

?Class rdfs:subClassOf ?r .

?r owl:onProperty o:hasAttributeCount ;

owl:hasValue ?attrCount .

FILTER ( ?Class != owl:NamedIndividual && ?w != rdf:type )

FILTER (?x != ?y)

} GROUP BY ?x ?y ?p ?attrCount }

FILTER (?wC*2 = ?attrCount)
} GROUP BY ?p
```

Logic Consistency - Value in Range - BMI in expected range

```
SELECT (COUNT (DISTINCT ?p) as
?patientWithAbnormalBMICount)
WHERE { ?p k:has_entry ?x .
?p k:has_entry ?y .
?x ?weight ?weightvalue.
?weight rdfs:subPropertyOf k:has_weight .
?y ?height ?heightvalue.
?height rdfs:subPropertyOf k:has_height .
FILTER ( ?weightvalue *10000 /
(?heightvalue * ?heightvalue) > 45 ||
?weightvalue *10000 /
(?heightvalue * ?heightvalue) <14 )
}
```

Logic Consistency - Value in Range - Accordance between diagnoses of gender specific disease and gender

```
SELECT (COUNT(DISTINCT ?malePatientHavingFemaleSpecificDisease)
as ?malePatientHavingFemaleSpecificDiseaseCount)
WHERE { ?malePatientHavingFemaleSpecificDisease
k:has_entry ?diagnosis , k:sex_Male .
?diagnosis k:has_ICD_version ?ICDver ;
k:has_ICD_Code ?ICDcode .
{
SELECT ?ICDver ?ICDcode
WHERE {
?cls rdfs:subClassOf+ o:FemaleOnlyDisease .
?cls rdfs:subClassOf/owl:intersectionOf/
(rdf:first|rdf:rest)+ ?ver , ?code.
?ver owl:onProperty k:has_ICD_version ;
owl:hasValue ?ICDver .
?code owl:onProperty k:has_ICD_Code ;
owl:hasValue ?ICDcode . }}}
```

Logic Consistency - Value in Range - Correct dose in cytomegalovirus (CMV) drug records

```
SELECT (COUNT(DISTINCT ?p) as
?patientReceivingCMVbutDoseIsZeroCount)
WHERE {
?p a k:Patient. #?p a patient
?CMV a k:CMVProphylaxis.
?p k:has_entry ?CMV.
{?p k:has_recipient_CMV true.} UNION
{?p k:has_donor_CMV true.}
?CMV k:Prophylaxis_drug_type ?type ;
k:Prophylaxis_drug_dose ?dose .
FILTER (?dose = 0) . }
```

Logic Consistency - Temporal Sequence - Date of the post-transplant new-onset diabetes mellitus (PTND) is in accordance with the date of transplant

```
SELECT (COUNT (DISTINCT ?x) as
?newDiabeteDiagBeforeTxEntryCount)
WHERE { ?p k:has_entry ?x .
?x k:new_diabetes_diagnosis_date ?newDiabetesDiagDate .
?x k:has_transplant_date ?txDate .
FILTER ( ?newDiabetesDiagDate < ?txDate ) }
```

Logic Consistency - Temporal Sequence - Date of the pre-transplant diagnosis is in accordance with the date of transplant

```
SELECT (COUNT (DISTINCT ?x) as ?preTxDiagLaterThanTxEntryCount)
WHERE { ?p k:has_entry ?x .
?x k:has_diagnosis_date ?diagDate .
?x k:is_pre-tx_diagnosis true .
?x k:has_transplant_date ?txDate .
FILTER ( ?diagDate > ?txDate ) }
```

Logic Consistency - Temporal Sequence - The first week dialysis record is in accordance with the date of transplant

```
SELECT (COUNT (DISTINCT ?e) as
?dialysisNotin1stWeekEntryCount)
WHERE {
?p k:has_entry ?e .
?e k:has_transplant_date ?txDate .
?e a k:TransplantAdmissionInformation .
?e k:has_admission_date ?admDate .
?e k:has_first_dialysis_date ?1stDiaDate.
?e k:is_1stweek_dialysis_needed true .
FILTER ((?1stDiaDate - ?txDate) >
"P0Y0M7DT0H0M0.000S"^^xsd:duration ) }
```

Logic Consistency - Temporal Sequence - The start date is before the stop date

```
SELECT (COUNT ( DISTINCT ?e) as
?startDateLaterThanStopDateEntryCount)
WHERE {
?p k:has_entry ?e .
?e ?x ?xv . ?e ?y ?yv.
?x rdfs:subPropertyOf o:startDate .
?y rdfs:subPropertyOf o:stopDate .
FILTER ( ?xv > ?yv ) }
```

Logic Consistency - Agreement in Records - Existence of diagnosis records when a disease history is recorded

```
SELECT (COUNT (DISTINCT ?patientWithStrokeHistoryButNoDiagnosis)
as ?patientWithStrokeHistoryButNoDiagnosisCount)
WHERE { ?patientWithStrokeHistoryButNoDiagnosis
k:has_entry k:has_stroke_history_yes ;
k:has_entry ?diagnosisALL .
?diagnosisALL a k:RecipientDiagnosis ;
k:has_ICD_version ?ICDversion ;
k:has_ICD_Code ?ICDcoding .
MINUS {
SELECT DISTINCT (?patient as
?patientWithStrokeHistoryButNoDiagnosis) ?ICDcodeA
WHERE { ?patient k:has_entry
?diagnosis , k:has_stroke_history_yes .
?diagnosis a k:RecipientDiagnosis .
?diagnosis k:has_ICD_version ?ICDver ;
k:has_ICD_Code ?ICDcodeA .
FILTER regex(?ICDcodeA, ?convertedCode, "i")
{ SELECT ?ICDver ?ICDcode ?convertedCode
WHERE {
?cls rdfs:subClassOf+ o:Stroke .
?cls rdfs:subClassOf/owl:intersectionOf/
(rdf:first|rdf:rest)+ ?ver , ?code.
?ver owl:onProperty k:has_ICD_version ;
owl:hasValue ?ICDver .
?code owl:onProperty k:has_ICD_Code ;
owl:hasValue ?ICDcode .
BIND(REPLACE(?ICDcode, ".x", ".?[0-9]*", "i") AS ?convertedCode)
}}}}}
```

Logic Consistency - Agreement in Records - Indication of delayed graft function (DGF) records when a first week dialysis is recorded

```
SELECT  (COUNT (DISTINCT ?p) as
?patientWithDGFButNoDGFRecordCount)
WHERE {
?p k:has_entry ?e .
?e a k:TransplantAdmissionInformation .
?e k:is_1stweek_dialysis_needed true .
MINUS {?p k:has_entry ?dEntry .
?dEntry k:has_diagnosis_name "delayed graft function" . }}
```

Appendix E

Designed Test Cases in the Simulated Dataset

Query ID and Abbreviation: 1-CMP_Missing

Reflecting Criterion: No missing value

Query Applied On: all entries

Violation Example:

```
D_rid_510614_2f97272e has_ICD_code "Missing"
(D: Diagnosis Entry)
```

Note: When transforming a raw dataset to RDF, the transformation script replaces null values by "Missing".

Query ID and Abbreviation: 2-CMP_Out_Range

Reflecting Criterion: No out-of-range value

Query Applied On: all entries

Violation Example:

```
D_rid_510614_2f977760 is_pre-tx_diagnosis "OutOfRange"
(D: Diagnosis Entry)  (tx: Transplant)
```

Note: When transforming a raw dataset to RDF, the transformation script replaces out-of-range values (defined by MOT's codebook) by "OutOfRange".

Query ID and Abbreviation: 3-CST_Indu_Unit

Reflecting Criterion: Correct unit use in induction therapy entries

Query Applied On: all induction therapy entries

Violation Example:

```
Indu_rid_312907_2f99090e type_of_therapy "Thymoglobulin"
Indu_rid_312907_2f99090e total_dose_ML 35
(Indu: Induction Therapy Entry)
```

Query ID and Abbreviation: 4-CST_Duplicate

Reflecting Criterion: No duplicate entries

Query Applied On: all new on-set diabetes entries

Violation Example:

```
P_rid_149664 has_entry NewD_rid_149664_1ecaae2e
P_rid_149664 has_entry NewD_rid_149664_2f99e4be
(P: Patient)

NewD_rid_149664_1ecaae2e method_of_diagnosis
"Fasting blood sugars (WHO/ADA/CDA criteria)"
NewD_rid_149664_1ecaae2e date_diagnosis "2002-11-04T00:00:00-05:00"
NewD_rid_149664_1ecaae2e has_tx_date "2005-02-06T00:00:00-05:00"
NewD_rid_149664_1ecaae2e last_review "2006-05-06T00:00:00-05:00"

NewD_rid_149664_2f99e4be method_of_diagnosis
"Fasting blood sugars (WHO/ADA/CDA criteria)"
NewD_rid_149664_2f99e4be date_diagnosis "2002-11-04T00:00:00-05:00"
NewD_rid_149664_2f99e4be has_tx_date "2005-02-06T00:00:00-05:00"
NewD_rid_149664_2f99e4be last_review "2006-05-06T00:00:00-05:00"
(NewD: New On-set Diabetes Entry)
```

Note: In this example, the patient (id=149664) is associated with two new on-set diabetes entries (id=149664_1ecaae2e and 149664_2f99e4be), which both have four attributes with identical values and the pair of these two entries is marked as a violation.

Query ID and Abbreviation: 5-RLC_BMI

Reflecting Criterion: Body Mass Index (BMI) in expected range

Query Applied On: all patients

Violation Example:

```
P_rid_164423 has_entry W_rid_164423_2f98b6de
(P: Patient)           (W: Weight Measurement Entry)
W_rid_164423_2f98b6de has_weight_value "40.0"^^xsd:float
                                       (Unit:kg)
P_rid_164423 has_entry A_rid_164423_2f98232c
(P: Patient)           (A: Admission Entry)
A_rid_164423_2f98232c has_height_at_admission 176
                                              (Unit:cm)
```

Note: In this example, the patient's height is 176cm and has a weight measurement of 40.0kg. The calculated BMI is 12.9, not in the (14, 45) range.

Query ID and Abbreviation: 6-RLC_Gender

Reflecting Criterion: Accordance between diagnoses of gender specific disease and gender

Query Applied On: all male patients

Violation Example:

```
P_rid_504486 has_entry D_rid_504486_2f97df5c
(P: Patient)           (D: Diagnosis Entry)
D_rid_504486_2f97df5c has_ICD_code "C50.11"
```

Note: “C50.11” refers to “Malignant neoplasm of central portion of breast, female” in ICD 10, which is not a valid diagnosis for a male patient.

Query ID and Abbreviation: 7-RLC_CMV

Reflecting Criterion: Correct dose in cytomegalovirus (CMV) drug records

Query Applied On: all patients

Violation Example:

```
P_rid_149664 has_entry CMVP_rid_149664_2f99a92c
(P: Patient)           (CMVP: CMV Prophylaxis)
CMVP_rid_149664_2f99a92c Prophylaxis_drug_dose 0
```

Note: Dose should be larger than zero.

Query ID and Abbreviation: 8-TLC_New_Diabetes

Reflecting Criterion: Date of the Post-Transplant New-Onset Diabetes Mellitus (PTND) is in accordance with the date of transplant

Query Applied On: all PTND entries

Violation Example:

```
NewD_rid_370506_2f99d258 date_diagnosis "2002-01-14T00:00:00-05:00"^^xsd:dateTime
NewD_rid_370506_2f99d258 has_tx_date "2012-06-03T00:00:00-05:00"^^xsd:dateTime
(NewD: New On-set Diabetes)
```

Note: In this example, the PTND was diagnosed in 2002 but the corresponding transplant date happened in 2012.

Query ID and Abbreviation: 9-TLC_PreTx_Diag

Reflecting Criterion: Date of the pre-transplant diagnosis is in accordance with the date of transplant

Query Applied On: all pre-transplant diagnosis entries

Violation Example:

```
D_rid_938980_2f97974a is_pre-tx_transplant true
D_rid_938980_2f97974a has_diagnosis_date "2005-09-23T00:00:00-05:00"^^xsd:dateTime
D_rid_938980_2f97974a has_tx_date "2003-06-25T00:00:00-05:00"^^xsd:dateTime
(D: Diagnosis Entry)
```

Note: In this example, the diagnosis is marked as a pre-transplant diagnosis but its diagnosis date is later than the transplant date.

Query ID and Abbreviation: 10-TLC_1stWeekDialysis

Reflecting Criterion: The first week dialysis record is in accordance with the date of transplant

Query Applied On: all admission information entries recording a first week dialysis

Violation Example:

```
A_rid_510614_2f987818 is_1st_week_dialysis_needed true
A_rid_510614_2f987818 has_tx_date "2005-05-29T00:00:00-05:00"^^xsd:dateTime
A_rid_510614_2f987818 has_first_dialysis_date
"2009-09-28T00:00:00-05:00"^^xsd:dateTime
(A: Transplant Admission Information Entry)
```

Note: In this example, the admission information recorded that 1st week dialysis (after the transplant) is needed. However, the date of the 1st dialysis is far behind one week since the transplant.

Query ID and Abbreviation: 11-TLC_Paired_Dates

Reflecting Criterion: The start date is before the stop date

Query Applied On: entries containing paired dates

Violation Example:

```
Indu_rid_504486_2f992c72 induction_start_date
"2007-03-08T00:00:00-05:00"^^xsd:dateTime
Indu_rid_504486_2f992c72 induction_stop_date
"2006-10-09T00:00:00-05:00"^^xsd:dateTime
(Indu: Induction Therapy Entry)
```

Query ID and Abbreviation: 12-ALC_History

Reflecting Criterion: Existence of diagnosis records when a disease history is recorded

Query Applied On: patients with a stroke history

Violation Example:

```
P_rid_199879 has_entry rstrk_yes
(P: Patient)           (rstrk: History of Stroke)
P_rid_199879 has_entry D_rid_199879_2f96d53a
P_rid_199879 has_entry D_rid_199879_2f973d9a
P_rid_199879 has_entry D_rid_199879_2f978d2c
                       (D: Diagnosis Entry)
D_rid_199879_2f96d53a has_diagnosis_name "abdominal abscess"
D_rid_199879_2f973d9a has_diagnosis_name "abdominal hernia"
D_rid_199879_2f978d2c has_diagnosis_name "abdominal pain"
```

Note: In this example, the patient is recorded with a history of stroke but none of his three diagnoses belongs to the stroke disease family.

Query ID and Abbreviation: 13-ALC_DGF

Reflecting Criterion: Indication of delayed graft function (DGF) records when a first week dialysis is recorded

Query Applied On: patients who needed a first week dialysis

Violation Example:

```
P_rid_230753 has_entry A_rid_230753_2f985ee6
(P: Patient)           (A: Transplant Admission Information Entry)
A_rid_230753_2f985ee6 is_1stweek_dialysis_needed true


P_rid_230753 has_entry D_rid_230753_2f9704ec
P_rid_230753 has_entry D_rid_230753_2f9756b8
P_rid_230753 has_entry D_rid_230753_2f97a7a8


D_rid_230753_2f9704ec has_diagnosis_name "abdominal aortic aneurysm"
D_rid_230753_2f9756b8 has_diagnosis_name "abdominal hernia"
D_rid_230753_2f97a7a8 has_diagnosis_name "abdominal pain"
```

Note: In this example, the patient is recorded with a need of first week dialysis, which indicates delayed graft function (DGF), but none of his three diagnoses is DGF.

References

Aspden, P., Wolcott, J., Bootman, J. L., Cronenwett, L. R., et al. (2006). *Preventing medication errors: quality chasm series*. National Academies Press.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., ... Sherman, P. M., et al. (2011). Ncbi geo: archive for functional genomics data sets - 10 years on. *Nucleic acids research*, *39*(suppl 1), D1005–D1010.

Baskarada, S., Koronios, A., & Gao, J. (2006). Towards a capability maturity model for information quality management: a tdqm approach. In *11th international conference on information quality (iciq-06), mit, cambridge, massachusetts, usa, november* (pp. 10–12).

Batini, C., Cabitza, F., Cappiello, C., & Francalanci, C. (2008). A comprehensive data quality methodology for web and structured data. *International Journal of Innovative Computing and Applications*, *1*(3), 205–218.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, *41*(3), 16.

Beale, T. (2011). Openehr entry types faqs. Retrieved October 31, 2013, from http://www.openehr.org/wiki/display/resources/openEHR+ENTRY+Types+FAQs

Birtwhistle, R., Keshavjee, K., Lambert-Lanning, A., Godwin, M., Greiver, M., Manca, D., & Lagacé, C. (2009). Building a pan-canadian primary care sentinel surveillance network: initial development and moving forward. *The Journal of the American Board of Family Medicine*, *22*(4), 412–422.

Bray, F. & Parkin, D. M. (2009). Evaluation of data quality in the cancer registry: principles and methods. part i: comparability, validity and timeliness. *European Journal of Cancer*, *45*(5), 747–755.

Brown, J. D. (1996). *Testing in language programs.* Prentice Hall Regents New Jersey.

Brüggemann, S. & Gruening, F. (2008). Using domain knowledge provided by ontologies for improving data quality management. *Proceedings of I-Know*, 251–258.

Bryan, K. N. & George, R. (2003). Geographic information systems. *The methods and materials of demography*, 733.

Bughin, J., Livingston, J., & Marwaha, S. (2011). Seizing the potential of 'big data'. *McKinsey Quarterly*, 103–109.

Campbell, E. M., Sittig, D. F., Guappone, K. P., Dykstra, R. H., & Ash, J. S. (2007). Overdependence on technology: an unintended adverse consequence of computerized provider order entry. In *Amia annual symposium proceedings* (Vol. 2007, p. 94). American Medical Informatics Association.

Chen, K., Chen, H., Conway, N., Hellerstein, J. M., & Parikh, T. S. (2011). Usher: improving data quality with dynamic forms. *Knowledge and Data Engineering, IEEE Transactions on*, *23*(8), 1138–1153.

CIHI. (2009). The CIHI data quality framework. Ottawa, Ont.: Canadian Institute for Health Information. Retrieved October 17, 2013, from http://www.cihi.ca/CIHI-ext-portal/pdf/internet/data_quality_framework_2009_en

CIHI. (2012). National health expenditure trends, 1975 to 2012. Retrieved October 8, 2013, from https://secure.cihi.ca/free_products/NHEXTrendsReport2012EN.pdf

English, L. P. (2003). Total information quality management: a complete methodology for iq management. *Dm Review*, *9*(03), 7320–1.

Freitas, J., Silva-Costa, T., Marques, B., & Costa-Pereira, A. (2010). Implications of data quality problems within hospital administrative databases. In *XII mediterranean conference on medical and biological engineering and computing 2010* (pp. 823–826). Springer.

Fürber, C. & Hepp, M. (2010). Using sparql and spin for data quality management on the semantic web. In *Business information systems* (pp. 35–46). Springer.

Fürber, C. & Hepp, M. (2013). Using semantic web technologies for data quality management. In *Handbook of data quality* (pp. 141–161). Springer.

Garde, S., Knaup, P., Hovenga, E. J., & Heard, S. (2007). Towards semantic interoperability for electronic health records–domain knowledge governance for open ehr archetypes. *Methods of information in medicine*, *46*(3), 332–343.

Goldschmidt, P. G. (2005). Hit and mis: implications of health information technology and medical information systems. *Communications of the ACM*, *48*(10), 68–74.

Greiver, M., Barnsley, J., Aliarzadeh, B., Krueger, P., Moineddin, R., Butt, D. A., … White, D., et al. (2011). Using a data entry clerk to improve data quality in primary care electronic medical records: a pilot study. *Informatics in Primary Care*, *19*(4), 241–250.

Greiver, M., Barnsley, J., Glazier, R. H., Harvey, B. J., & Moineddin, R. (2012). Measuring data reliability for preventive services in electronic medical records. *BMC Health Services Research*, *12*(1), 116.

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). The big data revolution in healthcare: accelerating value and innovation. *New York (NY): McKinsey Global Institute*, 1–3.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, *43*(5), 907–928.

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In *Handbook on ontologies* (pp. 1–17). Springer.

Hessol, N. A., Missett, B., & Fuentes-Afflick, E. (2004). Lower agreement on behavioral factors than on medical conditions in self-reported data among pregnant latina women. *Archives of medical research*, *35*(3), 241–245.

Hickey, G. L., Grant, S. W., Cosgriff, R., Dimarakis, I., Pagano, D., Kappetein, A. P., & Bridgewater, B. (2013). Clinical registries: governance, management, analysis and applications. *European Journal of Cardio-Thoracic Surgery*.

Hirdes, J. P., Poss, J. W., Caldarelli, H., Fries, B. E., Morris, J. N., Teare, G. F., … Jutan, N. (2013). An evaluation of data quality in canada's continuing care reporting system (ccrs): secondary analyses of ontario data submitted between 1996 and 2011. *BMC medical informatics and decision making*, *13*(1), 27.

Hoffman, S. & Podgurski, A. (2008). Finding a cure: the case for regulation and oversight of electronic health record systems, 105.

Kaplan, R. M., Bush, J. W., & Berry, C. C. (1976). Health status: types of validity and the index of well-being. *Health services research*, *11*(4), 478.

Kohn, L. T., Corrigan, J. M., Donaldson, M. S., et al. (2000). *To err is human: building a safer health system*. National Academies Press.

Kokotailo, R. A. & Hill, M. D. (2005). Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10. *Stroke*, *36*(8), 1776–1781.

Liaw, S., Rahimi, A., Ray, P., Taggart, J., Dennis, S., de Lusignan, S., … Talaei-Khoei, A. (2012). Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. *International journal of medical informatics*.

Mant, J., Murphy, M., Rose, P., & Vessey, M. (2000). The accuracy of general practitioner records of smoking and alcohol use: comparison with patient questionnaires. *Journal of Public Health*, *22*(2), 198–201.

McGuinness, D. L., Van Harmelen, F. et al. (2004). Owl web ontology language overview. *W3C recommendation*, *10*(2004-03), 10.

Mead, C. N. et al. (2006). Data interchange standards in healthcare it-computable semantic interoperability: now possible but still difficult. do we really need a better mousetrap? *Journal of Healthcare Information Management*, *20*(1), 71.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. (2010). The cosmin checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international delphi study. *Quality of Life Research*, *19*(4), 539–549.

Mor, V., Angelelli, J., Jones, R., Roy, J., Moore, T., & Morris, J. (2003). Inter-rater reliability of nursing home quality indicators in the US. *BMC Health Services Research*, *3*(1), 20.

Mor, V., Intrator, O., Unruh, M. A., & Cai, S. (2011). Temporal and geographic variation in the validity and internal consistency of the nursing home resident assessment minimum data set 2.0. *BMC health services research*, *11*(1), 78.

Motik, B., Patel-Schneider, P. F., Parsia, B., Bock, C., Fokoue, A., Haase, P., ... Sattler, U., et al. (2009). Owl 2 web ontology language: structural specification and functional-style syntax. *W3C recommendation*, *27*, 17.

Naus, T. E. & Hirdes, J. P. (2013). Psychometric properties of the interrai subjective quality of life instrument for mental health. *Health*, *5*, 637.

Oliveira, P., Rodrigues, F., & Henriques, P. (2005). A Formal Definition of Data Quality Problems. In *Proceedings of the 2005 International Conference on Information Quality (MIT IQ Conference), Cambridge, MA, USA*. MIT.

OpenEHR. (2012). Clinical knowledge manager. Retrieved November 12, 2013, from http://www.openehr.org/ckm/

Oracle. (2012). From overload to impact: an industry scorecard on big data business challenges. Retrieved October 7, 2013, from http://www.oracle.com/us/industries/oracle-industries-scorecard-1692968.pdf

Orfanidis, L., Bamidis, P. D., & Eaglestone, B. (2004). Data quality issues in electronic health records: an adaptation framework for the greek health system. *Health Informatics Journal*, *10*(1), 23–36.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218.

Rabinowitz, T., Pérez, E., Nancy Curtin Telegdi RN, M., & Prendergast, P. (2002). The resident assessment instrument-mental health (rai-mh): inter-rater reliability and convergent validity. *The journal of behavioral health services & research*, *29*(4), 419–432.

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. (2004). The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems. *Information Systems*, *29*(7), 551–582.

Sebastian-Coleman, L. (2012). *Measuring data quality for ongoing improvement: a data quality assessment framework*. Access Online via Elsevier. Retrieved November 17, 2013, from http://store.elsevier.com/product.jsp?lid=0&iid=73&sid=0&isbn=9780123977540

Shan, T. C. & Hua, W. W. (2006). Taxonomy of java web application frameworks. In *E-business engineering, 2006. icebe'06. ieee international conference on* (pp. 378–385). IEEE.

Sheth, A. P. & Ramakrishnan, C. (2003). Semantic (web) technology in action: ontology driven information systems for search, integration and analysis. *IEEE Data Eng. Bull.* *26*(4), 40–48.

Sonntag, D., Setz, J., Ahmed-Baker, M., & Zillner, S. (2012). Clinical trial and disease search with ad hoc interactive ontology alignments. In *The semantic web: research and applications* (pp. 674–686). Springer.

Spath, P. L. (2011). *Error reduction in health care: a systems approach to improving patient safety*. John Wiley & Sons.

Streiner, D. L. & Norman, G. R. (2008). *Health measurement scales: a practical guide to their development and use*. Oxford university press.

Tiropanis, T., Davis, H., Millard, D., & Weal, M. (2009). Semantic technologies for learning and teaching in the web 2.0 era. *Intelligent Systems, IEEE*, *24*(6), 49–53.

UHN. (2013). About the multi organ transplant program. Retrieved November 11, 2013, from http://www.uhn.ca/MOT/About

W3C. (2013). Semantic web. Retrieved October 11, 2013, from http://www.w3.org/standards/semanticweb/

Wand, Y. & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95.

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), 58–65.

Wang, R. Y., Strong, D. M., & Guarascio, L. M. (1996). Beyond accuracy: what data quality means to data consumers. *J. of Management Information Systems*, *12*(4), 5–33.

Watts, S., Shankaranarayanan, G., & Even, A. (2009). Data quality assessment in context: a cognitive perspective. *Decision Support Systems*, *48*(1), 202–211.