A DATA QUALITY FRAMEWORK FOR THE SECONDARY USE OF ELECTRONIC HEALTH INFORMATION

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Steven G. Johnson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Advisor: Bonnie Westra, PhD, RN, FAAN, FACMI

April 2016

Acknowledgements

Completing a PhD has been a life-long goal, but it's taken me longer to get here than I anticipated. A number of people had an important role in helping me achieve this and I want to acknowledge their contribution.

First, I want to sincerely thank my advisor Bonnie Westra. She provided the opportunities to work with EHR data and her guidance made this research much better than it otherwise would have been. Her collaborative approach, passion and ability to get things done have given me a model for what a great informaticist should strive for.

My committee members have played a critical role in guiding my research. Stuart Speedie, Gyorgy Simon and Vipin Kumar provided significant feedback and advice that challenged me to deliver better research and writing.

I also want to thank the UMN CTSI flowsheet project team especially Lisiane Pruinelli, Jung In Park, Beverly Christie, Anne LaFlamme, Matt Byrne and Suzan Sherman. You helped me understand care delivery workflows and patiently answered my clinical questions.

My parents, Roger and Mary Jo, instilled in me that if I did well in school and worked hard, I could achieve my goals. That advice has served me well. My family, MaryAnn, Garret, Diana and Maya, have been patient and understanding as I tried to balance home, work and school. Home was sometimes neglected in that balancing act.

And last but not least, I want to thank my "PhD Support Group" – Tamara Winden, Sandy Long, Matt Breitenstein, Aaron Berdofe, and Lisiane Pruinelli. Without their friendship, support and encouragement I still would not be finished.

Dedication

This dissertation is dedicated to my family – MaryAnn, Garret, Diana and Maya. Their ongoing support, encouragement and understanding made this possible.

Abstract

A DATA QUALITY FRAMEWORK FOR THE SECONDARY USE OF ELECTRONIC HEALTH INFORMATION

Electronic health record (EHR) systems are designed to replace paper charts and facilitate the delivery of care. Since EHR data is now readily available in electronic form, it is increasingly used for other purposes such as clinical effectiveness research, predictive modeling, population health management and healthcare quality improvement. Use of EHR data is expected to improve health outcomes for patients; however, the benefits will only be realized if the data that is captured in the EHR is of sufficient quality to support these secondary uses.

This research demonstrated that a healthcare data quality framework can be developed that produces metrics that characterize underlying EHR data quality and it can be used to quantify the impact of data quality issues on the correctness of the secondary use of the data. The framework described in this research defined a Data Quality (DQ) Ontology and implemented an assessment method. The usefulness of this approach was illustrated by characterizing the data quality of EHR data and then quantifying the impact of data quality issues on the correctness of the CMS178 eMeasure.

The DQ Ontology was developed by mining the healthcare data quality literature for important terms used to discuss data quality concepts and these terms were harmonized into an ontology. Four high-level data quality dimensions (CorrectnessMeasure, ConsistencyMeasure, CompletenessMeasure and CurrencyMeasure) categorized 19 lower level Measures. The ontology serves as an unambiguous vocabulary and allows more precision when discussing healthcare data

quality. The DQ Ontology is expressed with sufficient rigor that it can be used for logical inference and computation.

The data quality framework was used to characterize data quality of an EHR for 10 data quality Measures. The results demonstrate that data quality can be quantified and metrics can track data quality trends over time and for specific domain concepts. The quantities produced by the ontology based assessment method are easier to use and understand than some of the existing, rule based, approaches to data quality assessment. The DQ framework produces scalar quantities which can be computed on individual domain concepts and can be meaningfully aggregated at different levels of an information model.

The data quality assessment process was also used to quantify the impact of data quality issues on a task. The EHR data was systematically degraded and a measure of the impact on the correctness of the CMS178 eMeasure (Urinary Catheter Removal after Surgery) was computed. A linear regression model that uses domain concept data quality measures as independent variables quantified the relative impact on CMS178. The un-degraded EHR data was used as a baseline to measure correctness and quantify how data quality issues affect secondary use. This information can help healthcare organizations prioritize data quality improvement efforts to focus on the areas that are most important and determine if the data can support its intended use.

Table of Contents

List of Tables

List of Figures

Abbreviations

APN – Advanced Practice Nurse

CAUTI - Catheter-Associated Urinary Tract Infection

CDC – Centers for Disease Control

CDR - Clinical Data Repository

CER - Clinical Effectiveness Research

CIHI – Canadian Institute for Health Information

CIMI – Clinical Information Modeling Initiative

CMS – Centers for Medicare and Medicaid Service

CQM - Clinical Quality Measures

CTSA - Clinical and Translational Science Award

CTSI - Clinical and Translational Science Institute

DQ – Data Quality

EDM - Electronic Data Methods

EHR – Electronic Health Record

FDA - Food and Drug Administration

FHIR - Fast Healthcare Interoperability Resources

FOPL- First Order Predicate Logic

ISO – International Standards Organization

MAR - Missing at Random

MCAR - Missing Completely at Random

MNAR - Missing Not at Random

NCRR - National Center for Research Resources

NHAMCS - National Hospital Ambulatory Medical Care Survey database

NIH - National Institutes of Health

NLP – Natural Language Processing

NQF - National Quality Forum

OCL – Object Constraint Language

OMOP - Observational Medical Outcomes Partnership

OWL – Web Ontology Language

PA – Physician Assistant

PPV – Positive Predictive Value

QDM – Quality Data Model

UML – Unified Modeling Language

Chapter 1: Introduction

1.1 Background and Significance

The healthcare system in the United States has adopted the use of electronic health records (EHR) at a rapid pace. As of the end of 2014, fully 83% of physicians and 93% of hospitals have an EHR[1,2]. An EHR system is designed to replace a paper chart and to document and facilitate the delivery of care. Since EHR data is now readily available in electronic form, it is increasingly used for other purposes such as clinical effectiveness research, predictive modeling, population health management and healthcare quality improvement. Use of EHR data is expected to improve health outcomes for patients; however, the benefits will only be realized if the data that is captured in the EHR is of sufficient quality to support these secondary uses[3].

Studies have shown that EHR data often contains errors that can impact research results[4]. Researchers need to understand the quality of the data that they use. In a recent literature review only 24% of studies had a data validation section[5]. Errors in EHR data not only have an economic cost, they can lead to patient safety issues and negative patient outcomes. Decisions based on inaccurate information can adversely affect patient care and any downstream use of the data[6–10]. This dissertation research is focused on the impact of data quality issues on the secondary use of EHR data.

While there have been data quality frameworks explored in computer and information science settings, there has been limited work applying them to healthcare data[11]. This dissertation develops a healthcare focused data quality framework that can be

used to quantify the quality of a set of data and provides a method to assess the correctness of results given the data's intended use. The framework relates measurements of data quality along different dimensions to a research question of interest and quantifies the impact of various data quality issues on the correctness of the result. The framework consists of a data quality ontology that defines data quality measures and provides an assessment method and software that can execute against an EHR dataset to produce data quality metrics.

1.2 Data Quality Dimensions

There is no quantitative definition of data quality; most authors define data quality in the context for how the data will be used. Juran defines high quality data as data that are fit for use in their intended operational, decision-making, planning, and strategic roles[12]. The International Standards Organization (ISO) defines data quality as the totality of features and characteristics of an entity that bears on its ability to satisfy stated and implied needs[13]. In order to assess the quality of data there needs to be an understanding of how the data will be used[14,15]. This "fitness-for-use" measure of data quality is an approach taken in information science. Kahn has proposed a simplified healthcare specific framework using a "fit-for-use" data quality assessment model[16]. It as an important approach since it is the ultimate consumer of the data that determines whether the data has met its purpose[12,17].

Different researchers have proposed multiple dimensions of data quality, but there is no overall consensus on the most important dimensions or even consistency in their definitions[17–20]. Wang and Strong proposed a framework that consolidates 118 different

data quality attributes found in the literature into 20 dimensions[11]. This framework is often cited and represents a useful set of data quality categories, but this framework is not healthcare focused.

Healthcare data is complex and diverse and there have been a few attempts to define EHR data quality[21]. The Canadian Institute for Health Information (CIHI) published a framework that describes a data quality model that includes five dimensions (Accuracy, Timeliness, Comparability, Usability and Relevance)[22]. Kahn also proposes five different dimensions (Accuracy, Objectivity, Believability, Timeliness, Appropriate amount of data)[16]. Liaw performed a literature review looking for commonalities on data quality dimensions and he also found five dimensions (Accuracy, Completeness, Consistency, Correctness and Timeliness)[23]. Based on a literature review Weiskopf proposed five similar but different categories of data quality (Completeness, Correctness, Concordance, Plausibility and Currency) and she also listed synonyms for the dimensions (Table 1.1)[24]. These attempts to define healthcare data quality have overlapping dimensions, but it is difficult to understand how similar they are without a better way to specify each data quality concept.

Dimension	Description
Completeness	Accessibility, Accuracy, Availability, Missingness, Omission, Presence, Quality, Rate of recording, Sensitivity, Validity
Correctness	Accuracy, Corrections made, Errors, Misleading, Positive predictive value, Quality, Validity
Concordance	Agreement, Consistency, Reliability, Variation
Plausibility	Accuracy, Believability, Trustworthiness, Validity
Currency	Recency, Timeliness

Table 1.1. Data Quality Dimensions

All of the aforementioned researchers define these dimensions using textual narrative and synonyms. As noted in Weiskopf's descriptions, the same terms may be used multiple times (i.e. "Accuracy" occurs 3 times) to mean different things and it is not

clear what aspect of data quality is described. A formal method for defining data quality characteristics is needed.

The "fit-for-use" view of data quality implies that different uses for a particular data set may require different levels of data quality. When health information was recorded in paper charts, it was difficult and time consuming to abstract the information of interest for a particular research project. Today, with every patient record stored in a database, it is easy to obtain a large volume of health information for research projects or studies. If the underlying data quality is not assessed for the intended purpose, errors and inappropriate results may occur. Clinical quality measures (CQMs) and clinical effectiveness research (CER) will be used as two examples of how data quality issues impact the secondary use of EHR data.

1.3 Clinical Quality Measures

The first example of data quality impact is illustrated with the calculation of Clinical Quality Measures (CQMs), sometimes called eMeasures, from EHR data. The creation of eMeasures from EHR data is an initiative developed to quantify how well patient care is meeting best practices[25,26]. Current EHR systems compute eMeasures[27,28], but studies show that eMeasures calculated from EHR data can vary substantially from measures calculated through manual chart review[29,30]. These results indicate that the current state of EHR data may not be of sufficient quality to correctly calculate eMeasures.

In order to ensure that eMeasures measure the same phenomenon across organizations, the National Quality Forum (NQF) recommends that all eMeasures be

empirically tested for validity and reliability[31], however current eMeasure validation does not take into account data quality. The validity of an eMeasure is critically dependent on identifying the correct patient population for the numerator or denominator of the measure[30]. One study compared a manual, paper based computation of a quality measure with the same one derived from EHR data[32]. Ten percent of the patients had missing information in the EHR. The eMeasures that could be calculated differed significantly from the manually computed ones. The results underestimated the quality of care due to incomplete or incorrect data items in the EHR data. Another study showed that when patients were allowed to review their EHR records, 25%-30% of patients found inaccuracies[33].

These examples demonstrate that poor data quality impacts the correctness of an eMeasure and that it would be useful to quantify the level of data quality needed to ensure the eMeasure correctly represents what it was intended to measure.

1.4 Clinical Effectiveness Research

As more healthcare data is recorded in electronic databases, there is a desire to use the data for clinical research. In order to understand the validity of the research results, the underlying data quality should be taken into account. Information in the electronic record is not necessarily recorded with clinical research in mind and the quality may not be what is needed for a specific research project[34]. Secondary use of electronic health information can be justified as long as data quality assessment metrics are compared against research requirements[24]. Clinical effectiveness research benefits from using aggregated data from multiple EHR systems in order to get larger sample sizes to

detect smaller effects, but the research must differentiate between true differences in treatment outcomes and differences due solely to variations in the quality of the aggregated data[16].

One example of the impact of not assessing data quality when conducting clinical research is illustrated by Green[35]. He describes data quality issues with the National Hospital Ambulatory Medical Care Survey database (NHAMCS), which is maintained by the Centers for Disease Control (CDC). Green shows that the CDC database had significant errors related to data for intubated patients (25% of the emergency department patients that were reviewed). This is important because NHAMCS is widely used for medical research. Shuur showed that there were similar issues with pregnancy data in NHAMCS[36], but these errors might have been discoverable with relatively straightforward data quality measures that may have uncovered issues before the data were used by researchers.

Another example is a study conducted in 2013 to evaluate the comparative effectiveness of antihypertensive medications on blood pressure control[37]. The study uncovered multiple issues with data that was extracted from four healthcare organization's EHRs used for the study. Data was missing, it contained errors (i.e. blood pressures with recording errors), it was inconsistent (i.e. comorbid conditions were not consistently listed) and some data was uninterpretable (i.e. blood pressure measures where the measurement technique was not captured). The researchers recommended giving clinicians periodic feedback on data quality issues to improve EHR documentation and ensure clinical research is not seen as a separate activity from clinical care.

Data quality issues for eMeasures and comparative effectiveness research may impact the ability to give correct results and demonstrates the importance of understanding the quality of data used in research.

1.5 Purpose and Specific Aims

The overall purpose for this dissertation is to demonstrate that a data quality framework can be developed to characterize underlying EHR data quality. The framework can be used to quantify the impact of data quality issues on the correctness of the data for an intended use. The specific research aims to support this purpose are:

1. Define a data quality framework that consists of a Data Quality (DQ) Ontology and a method for quantifying data quality for the secondary use of EHR data. The framework defines specific measures of data quality and describes an approach for quantifying these measures for a set of data.
2. Develop a software program that uses actual EHR data and the DQ Ontology to produce metrics that characterize that data. An example of computing the CMS178 eMeasure (Urinary Catheter Removal after Surgery) is developed to illustrate the usefulness of the approach.
3. Quantify the impact of two data quality issues, missing data and domain conformance, on the correctness of an eMeasure (CMS178) by systematically degrading the underlying quality of EHR data. A linear model that describes the change in the correctness of the eMeasure quantifies the impact of each data quality issue.

The research resulting from these aims provides a useful healthcare focused data quality assessment framework. The DQ Ontology concepts serve as an unambiguous vocabulary for discussing healthcare data quality. The ontology precisely defines data quality concepts better than using textual descriptions and is sufficiently rigorous to be used directly by software to quantify data quality. The DQ Ontology can be reused for different clinical domains and intended purposes to make validating data quality more common and reproducible. The metrics produced by this approach characterize data quality along a number of dimensions and the impact that data quality issues have on the intended use of the data can be quantified. Automating the data quality assessment process using this approach can enable sharing of data quality metrics that may aid in making research results that use EHR data more transparent and reproducible.

Chapter 2: A Data Quality Ontology for the Secondary Use of EHR Data

Steven G. Johnson, MS[a], Stuart Speedie, PhD, FACMI[a], Gyorgy Simon, PhD[a], Vipin Kumar, PhD[b], Bonnie L. Westra, PhD, RN, FAAN, FACMI[a,c]

[a]*University of Minnesota, Institute for Health Informatics*

[b]*University of Minnesota, Department of Computer Science*

[c]*University of Minnesota, School of Nursing*

Published in:

AMIA 2015 Annual Symposium Proceedings. American Medical Informatics Association; 2015:1937-1946.

2.1 Summary

The secondary use of EHR data for research is expected to improve health outcomes for patients, but the benefits will only be realized if the data in the EHR is of sufficient quality to support these uses. A data quality (DQ) ontology was developed to rigorously define concepts and enable automated computation of data quality measures. The healthcare data quality literature was mined for the important terms used to describe data quality concepts and harmonized into an ontology. Four high-level data quality dimensions ("correctness", "consistency", "completeness" and "currency") categorize 19 lower level measures. The ontology serves as an unambiguous vocabulary, which defines concepts more precisely than natural language; it provides a mechanism to automatically compute data quality measures; and is reusable across domains and use cases. A detailed example is presented to demonstrate its utility. The DQ ontology can make data validation more common and reproducible.

2.2 Introduction and Background

The healthcare system in the United States continues to adopt electronic health records (EHR) at a rapid pace.[38] The EHR is designed to replace a paper chart and to document and facilitate the delivery of care. Since this electronic data is now much more easily accessed than abstracting from paper charts, it is frequently used for other purposes such as clinical effectiveness research, predictive modeling, population health management and healthcare quality improvement. Secondary use of EHR data is expected to improve health outcomes for patients, but the benefits will only be realized if the data that is captured in the EHR is of sufficient quality to support these secondary

uses.[3] Investigators have shown that EHR data often contain errors that can impact research results, yet only 24% of clinical studies that use EHR data had a data validation section.[5] In order to measure the quality of data there must be an understanding of how the data will be used.[15]

There is no generally accepted quantitative measure of data quality, but Juran gives an often cited qualitative definition as "…high-quality data are data that are *fit for use* in their intended operational, decision-making, planning, and strategic roles."[12(p.34-8)] Data quality may be adequate when used for one task, but not for another. For example, a higher level of data quality is needed to count the number of diabetic patients with controlled HgA1C than to just count the number of patients. A *task* refers to concepts in a clinical *domain* and those concepts are represented by the data. For each task, a set of data quality measures must be developed that determine if the data are adequate to perform the task. The healthcare data quality literature provides terminology and definitions and attempts to organize data quality measures, but there is no general agreement on what these measures should be.[23] This terminology-based approach defines measures using natural language, which does not adequately represent the relationships between concepts and is too loosely defined to yield a quantifiable measure of data quality. A better approach is to use an ontology which provides a sufficiently rigorous foundation for concept definitions that enable automated methods for calculating data quality measures.

An ontology is a formal, explicit specification of a shared conceptualization.[39] Each concept (also called a "class") in the ontology has a name, attributes, properties (relations to other concepts) and constraints that must always be true for a concept. The

key benefits of defining data quality measures in terms of an ontology are that an ontology is: 1) a specification, written in a formal language and able to represent semantics, 2) a shared vocabulary that everyone can use to precisely refer to an aspect of the world, and 3) a sufficiently rigorous specification that can be used for logical inference and computation.[40] An ontology is a logical theory about a part of the world and it defines interrelationships between concepts and axioms that should be true about that world. Automated reasoning can be applied to check internal consistency and make inferences beyond what was explicitly stated in the ontology.[41] This automation eliminates the need for redefining the data quality measures for every task in every domain.

No formal healthcare data quality ontology currently exists, but there is research that examines core data quality concepts. Wang and Strong[11] proposed a framework that consolidates 118 different general data quality characteristics into 20 categories. Kahn[16] proposed a healthcare specific framework using a "fit-for-use" data quality model in which he proposes five high-level dimensions. Liaw[23] performed an extensive literature review looking for commonalities on data quality dimensions. He found consensus on the five most common occurring dimensions were "accuracy", "completeness", "consistency", "correctness" and "timeliness". While there is some agreement among investigators on these high-level dimensions, there is little agreement or consistency in definitions of more granular data quality concepts such as "validity", "reliability" and "believability".[18] In a 2012 paper, Weiskopf[24] defined five high-level dimensions of data quality and listed synonyms for each (Table 2.1).

Dimension	Synonyms
Completeness	Accessibility, Accuracy, Availability, Missingness, Omission, Presence, Quality, Rate of recording, Sensitivity, Validity
Correctness	Accuracy, Corrections made, Errors, Misleading, Positive predictive value, Quality, Validity
Concordance	Agreement, Consistency, Reliability, Variation
Plausibility	Accuracy, Believability, Trustworthiness, Validity
Currency	Recency, Timeliness

Table 2.1. Weiskopf Five Dimensions of Data Quality with Synonyms

While these dimensions capture orthogonal aspects of data quality, they are defined using natural language descriptions and synonyms. As can be seen from Weiskopf's descriptions, the same terms may be used multiple times to mean different things (i.e. "Accuracy" occurs 3 times), introducing confusion regarding what aspect of data quality is being described. To provide better conceptual clarity and precision, an ontology is needed.

This paper describes the development of a healthcare data quality ontology (DQ ontology) which provides rigorous definitions and can automate the computation of data quality measures. Given formal ontologies for a clinical domain and for a task, the DQ ontology enables measures to be reused without having to reinvent new data quality assessments for every research project. Ontologies for some clinical domains[42] and tasks[43] already exist and researchers can focus on creating additional ontologies that can be used by the DQ ontology to yield quantified measures. This can make it easier to incorporate data quality validation as a standard component of research results. The DQ ontology was developed from a comprehensive list of data quality terms present in the literature. The terms were organized into an ontology and constraints were defined that precisely describe a data quality measure better than natural language and enable quantification of the measure. It makes explicit which data quality concepts depend on the use of the data and which depend on the clinical domain. A detailed example

demonstrates the utility of this ontology for quantifying measures and for discussing aspects of data quality.

2.3 Materials and Methods

There are a number of methodologies for developing an ontology,[40] but the method described by Noy and McGuiness[44] was selected due to its simplicity and effectiveness. This methodology advocates a seven-step process that takes a list of terms and definitions and turns them into a formal ontology. The first step is to define the scope of the ontology. For this study, the scope is a shared vocabulary of data quality concepts with formal definitions that are automatically computable to quantify data quality. The software development community has had success adopting the approach of a common vocabulary to allow researchers to spend less time defining concepts and more time applying it in research.[45] Next, the reuse of existing ontologies was considered. No formal healthcare data quality ontology exists; but ontologies that describe clinical domains and tasks do exist and will be reused and referenced by the DQ ontology.[42,43]

In order to enumerate the important terms in the ontology, an extensive PubMed search for articles published between January 1995 and January 2015 was performed to obtain a comprehensive list of terms and definitions that are used to describe healthcare data quality. The goal was to find literature reviews and meta-analyses of papers about healthcare data quality to identify as many core concepts as possible. Also, all articles about informal healthcare data quality frameworks or ontologies were examined for key terms and definitions. Keywords included in the query were: ("data quality") and

("health" or EHR) and (“literature review” or framework or ontology or assessment or model) and (dimensions or accuracy or consistency or completeness or correctness).

There were 181 articles identified, which were manually reviewed by the first author and narrowed to five meta-analyses from Liaw[23], Weiskopf[24], Kahn[16], Chen[46], and Lima[47]. These papers were either reviews of other papers about healthcare data quality or they proposed an informal data quality framework. They all attempted to categorize data quality concepts into semi-orthogonal dimensions. The references from these papers were also reviewed, which yielded an additional five sources: Wang[11], Wand[20], Chan[6], CIHI[22], Stvilia[48]. Collectively, these 10 meta-analyses reviewed 412 papers looking for common aspects of healthcare data quality. There was similarity on high-level concepts such as “correctness”, “consistency” and “completeness”, but there were limited definitions for important terms such as “dataset”, “data”, “measurement”, “metric” and “measure”. Additional papers from the information science literature were found to further define these important concepts[49–51].

Ontologies can be specified using a number of methods including OWL[52], first order logic, or as UML[53]. For this paper, the ontology is documented using a UML diagram and a table that lists constraints. A bottom-up approach was taken in which terms and definitions from the meta-analyses were matched and harmonized into equivalent concepts and these concepts where grouped into higher-level categories. Each concept has properties and relationships with other concepts that were discerned from reading the description in the articles. The cardinality of relationships was also defined. Cardinality indicates whether an associated concept is optional, must always occur, or can occur multiple times. For example, a patient must always have a gender, but a blood

pressure reading is an optional observation. Constraints were also defined for each concept, describing what should always be true for a concept. The constraints evaluate to a Boolean (true/false) result and can be written in a number of languages including, Object Constraint Language (OCL), first order predicate logic (FOPL), pseudo-code or openEHR constraint language.[40,54] For this study, pseudo-code was chosen because it succinctly captures the important aspects of the constraint without introducing a specific, complex syntax.

2.4 Results

There were 96 terms and definitions extracted from the literature as a basis for the data quality measures of the ontology. Terms that described the same concept were matched based on their definition and use within the articles. Concepts that appeared in less than three of the articles were deemed non-core and were left out of this version of the DQ ontology. The resulting data quality ontology is shown in Figure 2.1 as a UML diagram depicting the relationships, attributes, and cardinality of the concepts. For readability, the 19 lower-level **Measures** were not included in the diagram and are listed in Table 2.2, which also provides a definition of the measure and a reference to equivalent terms from the meta-analyses. A **bold** font is used to indicate that a term refers to a concept from an ontology.

The meta-analyses articles make pervasive reference to concepts such as “data”, “information” and “value”. In the DQ ontology, a more precise concept, **Representation,** defines the lowest level, atomic piece of information that exists in the data being assessed (synonyms for this concept are data field, observation, value, etc). **Representations** have

a **DataValue** (the part that is stored somewhere) as well as a **DataValueType** that specifies a format to which the **DataValue** must conform (i.e. numeric quantity, string, choice field, etc). **DataValueTypes** put constraints on the **DataValue** of the **Representation**, and can only refer to intrinsic information about the value itself and not to relationships with other **Representations**. Formal semantics about concepts represented in the data are defined in a separate **Domain** ontology. **Representations** have an attribute, **DomainConcept,** which maps data to a concept in the clinical **Domain** ontology. There can be multiple **Representations** for each concept in the **Domain**. For example, a systolic blood pressure value can be represented as a single number (i.e. 123) or it can be encoded as the first part of a string (i.e. "123/92"). **DomainConcepts** can also have multiple synonyms in the **Domain** ontology (i.e. "BP" and "Blood Pressure"), but for the purpose of assessing data quality, they can all be mapped to a single, primary **DomainConcept** (i.e. "Blood Pressure"). The **Task** designates the context or the specific use of the data and is necessary for assessing fitness-for-purpose. The **Domain** and **Task** are separate, formal ontologies to which the DQ ontology refers. A **Dataset** is an arbitrary grouping of **Representations** of interest. For example, a **Dataset** can be all of the **Representations** in the entire EHR.

One of the key concepts in the DQ ontology is the **Measure,** which is defined as "a quantity that characterizes a quality of the data". Other possible terms considered were "dimension", "aspect", "measurement", "metric". **Measure** was chosen because it captured the notion of quantifying an aspect of interest. The word is used as a noun, not a verb. A **Measure** is quantified using a **MeasurementMethod.** A **Measurement** is a process that performs a **MeasurementMethod** on a specific **Representation** (or

Dataset) at a point in time that yields a **MeasurementResult** which is a quantity, usually numeric (but possibly a boolean or text value). A **Metric** is a statistic about a series of **MeasurementResults** along a dimension such as time or across patients. For example, a **MeasurementResult** could indicate that there were 72 data format errors in a **Dataset**. But a **Metric** for that situation would be that there were an average of 5.5 data format errors per day or per patient. This part of the DQ ontology was based in part on core concepts from the Ontology for Software Measurement[49].

Four high-level data quality dimensions (**CorrectnessMeasure**, **ConsistencyMeasure**, **CompletenessMeasure** and **CurrencyMeasure**) categorize 19 lower level **Measures**. "Accuracy" is one of the terms that had many definitions in the literature. In Weiskopf[24], she lists at least 3 different ways that the term is used. It sometimes means only correctness but it is also used to represent completeness or plausibility. For that reason, the term "accuracy" has been avoided in the DQ ontology because it is too overloaded. Instead, the term "correctness" was selected to represent this core concept.

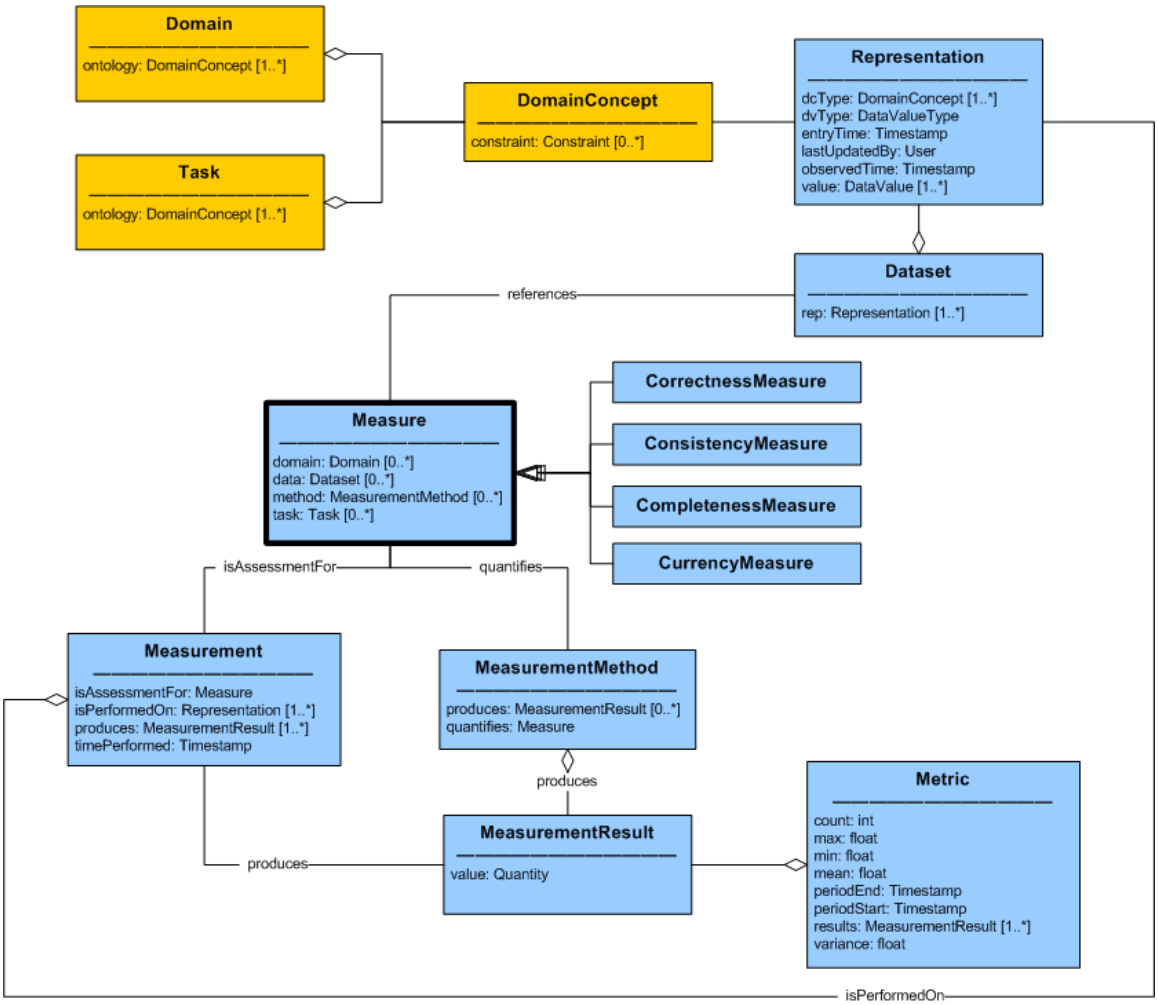

Figure 2.1. Data Quality Ontology

Concept	Definition	References / Synonyms
CorrectnessMeasure		
RepresentationIntegrity	Aspects of the Representation that reassure that data was not corrupted or subject to data entry errors.	Correctness: Credibility of source[23], Accuracy: …free of error[16], Integrity[46], Repeatability[46], Structural Consistency[48]
RelativeCorrectness	Assesses the quality of a Representation by comparing it to its counterpart in another Dataset which is a "relative standard", computed as PPV.	Accuracy: ...conformity with actual value[23], Correctness[24], Believability[16], Validity[24,47], Comparability[6,20], Accuracy[11,24,46,48], Corrections made[24], Errors[24], Misleading[24], PPV[24], Quality[24]
RepresentationCorrectness	A correct Representation has high accuracy and is complete.	Correctness: …accuracy and completeness[23], Accuracy[6,20]
Reliability	The data is correct and suitable for the Task.	Reliability[20,23,46,47], Accuracy: Measurement Error[22]
ConsistencyMeasure		
RepresentationConsistency	The data is a valid value and format for its DataValueType and all of the Representations for the same information have the same values.	Consistency: …values and physical representation of data[23], Concordance[24], Format[16], Internal Consistency[46], Consistency[24], Precision[20], Format[16,20], Reliability[24], Variation[24], Accuracy: Edit and Imputation[22], Representational Consistency[11]
DomainConsistency	Concepts in the Domain are represented in the data and the data satisfies syntactic and semantic rules. Constraints for the Domain are satisfied.	Accuracy: Refers to values and representation[23], Correctness: …format and types are valid[23], Plausibility[24], Believability[11,24], Relational Integrity Rules[16], Consistency[20,46,47], Measure validity[6], Accuracy[24], Trustworthiness[24], Validity[24,48]
CodingConsistency	Representations that are of coded text data type must be correctly mapped to an enumerated list or a terminology.	Consistency: …codes/terms…mapped to a reference terminology[23], Valid values[16], Comparability: Equivalency[22], Semantic Consistency[48]
DomainMetadata	Meta-data exists to describe the Domain and it is logically consistent.	Methodological Clarity[47], Metadata Documentation[46], Comparability: Data dictionary standards[22], Interpretability[11]

CompletenessMeasure		
RepresentationComplete	Domain independent extent to which data is not missing.	Completeness: …information is not missing[23], Completion[47], Completeness[6,46], Accuracy: Item Non-Response[22]
DomainComplete	The extent to which information is present or absent as expected.	Appropriate amount of data: Data are present or absent as expected[24], Optionality[16], Content[20]
RelativeCompleteness	The extent to which a truth about the world is represented in the data. This is computed as sensitivity relative to another Dataset.	Completeness: Is a truth…in the EHR?[24], Accessibility[11,24,47], Accuracy[24], Availability[24], Missingness[24], Omission[24], Presence[24], Quality[24], Rate of Recording[24], Sensitivity[24], Validity[24]
Sufficiency	The data has sufficient Representations along a given dimension (i.e. time, patient, encounter) to perform the Task.	Completeness: …sufficient breadth and depth for the task[23], Appropriate amount of data[16], Representativeness[46], Sufficiency[20], Accuracy: Coverage[22], Granularity[16,46], Continuity[16], Level of Detail[20], Completeness[11,48], Precision[48]
DomainCoverage	The data can represent the values and concepts required by the Domain.	Completeness: …represent every meaningful state of the […] real world[23], Completeness: All values for a variable are recorded[23], Coverage[47], Completeness[20]
TaskCoverage	The data contains all of the information required by the Task.	Completeness: …depict every possible state of the task[23], Usableness[20,46], Usability[46], Utility[46], Importance[20], Usefulness[20], Value-added[11]
Flexibility	The extent to which the data is sufficient to be used by many Tasks.	Consistency: information appl[ies] to different tasks[23], Flexibility[11,20], Relevance: Adaptability[22]
Relevance	Data is sufficient for the Task and conforms to the Domain.	Relevance[20,23,46,48], Relevance: Value[22], Relevancy[11]
CurrencyMeasure		
RepresentationCurrent	Calculation for time difference between when an observation was made and when it was entered into the system.	Timeliness: delay between a change of the real-world state and…the information system[23], Currency[24,46,48], Timeliness[20,24,46], Up-datedness[46], Recency[24]
DatasetCurrent	Time difference between when a Dataset was updated and when it was made available. For example, periodic repository updates.	Timeliness: …availability of output is on time[23], Opportunity[47], Periodicity[46], Currency[16,20], Timeliness: Data currency[22], Timeliness[11]
TaskCurrency	The Data is sufficiently up-to-date for the requirements of the Task.	Timeliness: …information is up to date for task[23], Timeliness: …age of the data is appropriate for the task[16], Timeliness (external)[20]

Table 2.2. Data Quality Ontology - Measure Detail

Illustrative Example of Using the DQ Ontology

In what follows, an example is provided to illustrate the utility of the DQ ontology concepts. Table 2.3 lists constraints (using pseudo-code) for some of the **Measures**. These will be used to show how data quality measures can be computed for a sample **Dataset** (Table 2.4) with respect to the task of calculating an eMeasure. An eMeasure[25] is a ratio for a health outcome of interest. For example, NQF 0018, "Controlling High Blood Pressure", is defined to be "The percentage of patients 18-85 years of age who had a diagnosis of hypertension and whose blood pressure was adequately controlled (<140/90mmHg) during the measurement period."

Measure	Constraint
RepresentationConsistency	Representation is valid format
DomainConsistency	RepresentationConsistency and (Representation DomainConcepts are in Domain) and DomainComplete and Representation's DomainConcept Constraints are satisfied
CodingConsistency	if Representation is coded text then Representation should have valid code
DomainMetadata	Domain ontology is consistent
RepresentationComplete	Representation value is not empty
DomainComplete	RepresentationComplete or Representation's DomainConcept cardinality is satisfied
Sufficiency	Task SufficiencyConstraint is satisfied
DomainCoverage	Domain's DomainConcepts are subset of Dataset's DomainConcepts
TaskCoverage	DomainCoverage and (Task's DomainConcepts are subset of Dataset's DomainConcepts)

Table 2.3. Examples of Data Quality Measure Constraints

For the DQ ontology to be applicable, a **Domain** and a **Task** need to be defined. In this case, the **Task** is to calculate the eMeasure defined above and the **Domain** consists of concepts related to blood pressure as well as some information about the patient and the encounter. To make the example more concrete, a minimalist (and incomplete) **Domain** and **Task** ontology will be defined. A portion of a blood pressure

(**Domain**) ontology is shown below (patterned after the openEHR blood pressure clinical model[42]):

BloodPressureDomain (portion) is an instance of a **Domain** ontology consisting of:

Patient is a Structure and has 1 MRN, [0 or more] Encounter, 1 Age

Age is a Quantity with a constraint of "Age > 0 and < 120"

Encounter is a Structure with [0 or more] Diagnosis, [0 or more] BloodPressureObservation

BloodPressureObservation has [0 or 1] Systolic, [0 or 1] Diastolic

Systolic is a Quantity with a constraint of "value > 0 and < 1000, Systolic > Diastolic"

Diastolic is a Quantity with a constraint of "value > 0 and < 1000, Systolic > Diastolic"

The **Task** usually has a formal ontology, but for simplicity's sake a task definition serves to illustrate how concepts in the **Domain** are referenced to specify the criteria for the patient population of interest. It defines the semantics of "diagnosis of hypertension" which, in this example, is a value set of codes from the ICD9 terminology. A portion of an example **Task** instance, TaskNQF0018 is shown below. It is patterned after the eMeasure Quality Data Model (QDM)[43].

TaskNQF0018 (portion) is an instance of a **Task** ontology consisting of:

PatientPopulation refers to Patients Age and Diagnosis:

InclusionCriteria: Diagnosis in {401.0, 401.1, 401.9} and Age ≥ 18 and Age ≤ 85

SufficiencyConstraint: At least 1 BloodPressureObservation per Encounter

Numerator refers to the most recent BloodPressureObservation: Formula is count(

BloodPressureObservation.Systolic > 140 and

BloodPressureObservation.Diastolic > 90)

Denominator refers to PatientPopulation: Formula is count(PatientPopulation)

Sample patient data is shown in Table 2.4. Each of the cells in the table shows the value of an instance of a **Representation**. The topmost column headers indicate the **DomainConcept** to which each of the cells map. The lower column headers show the **DataValueType** for the cells in the column. For brevity, other **Representation** information (entryTime, observedTime, etc.) is not shown.

Domain Concept	**Patient**				
	MRN	**Age**	**Encounter**		
			Diagnosis	**BloodPressureObservation**	
				Systolic	**Diastolic**
Data Value Type	*numeric*	*numeric*	*coded text*	*numeric*	*numeric*
Data Value	1	72	"ICD9:401.0"	147	92
	2	81	"ICD9:401.0"	142	"High"
	3	77	"ICD9:401.1"	140	
	4	60	"ICD9:xxx"	92	100
	5	44	"ICD9:401.9"		

Table 2.4. Example Patient Data

To assess the quality of the sample data, Measurements that quantify some of the Measures were performed. For this example, the MeasurementMethod evaluates the class constraint of a Measure for all of the Representations in a Dataset and produces a MeasurementResult, which is the proportion of constraints that were satisfied. These results are shown in Table 2.5. The quantity in the table cell is a fraction where the numerator is the number of constraints that are satisfied and the denominator is the

number of Representations for each concept. The cell also shows the decimal equivalent for the fraction. As an example, to compute RepresentationConsistency for the Diastolic DomainConcept, the three Representations in the last column of Table 2.4 are examined. It can be seen that these Representations have a DataValueType of numeric. But the value for Patient2 is not valid. Therefore, only two of the three Representations have RepresentationConsistency. The rest of the MeasurementResults are shown in the table.

Measure	Measurement Process Summary	MeasurementResult				
		Systolic	Diastolic	BloodPressure Observation	Encounter	Patient
Measures that involve only the Representation						
RepresentationConsistency	Satisfied if all **Representations** conform to their **DataValueTypes**. Patient2. Encounter.BloodPressureObservation.Diastolic is an invalid value.	4/4 1.0	2/3 .67	3/4 .75	4/5 .80	4/5 .80
RepresentationComplete	Patient3.Encounter.BloodPressureObservation.Diastolic has a missing value so it is not **RepresentationComplete**.	4/4 1.0	2/3 .67	3/4 .75	4/5 .80	4/5 .80
Measures that involve the Representation and Domain						
DomainConsistency	**DomainConsistency** is satisfied if all of the concepts in the **Domain** exist in the data (true for this example). Also, the data must have **RepresentationConsistency** (Patient2 does not) and all of the constraints for all of the **Domain** concepts must be satisfied. Patient4 has a diastolic blood pressure value that is higher than the systolic value, so the constraint is not satisfied. But Patient5's missing BloodPressureObservation is allowed by the **Domain**.	3/4 .75	1/3 .33	1/4 .25	2/5 .40	2/5 .40
CodingConsistency	True if all coded text **Representations** have valid values. Patient4.Encounter.Diagnosis is invalid in the ICD9 terminology.				4/5 .80	4/5 .80
DomainMetadata	The **Domain** ontology is defined and contains no logical inconsistencies. It would be considered inconsistent if it contained another rule that stated patient age was optional (i.e. "Patient has [0 or 1] Age").	4/4 1.0	3/3 1.0	4/4 1.0	5/5 1.0	5/5 1.0

DomainComplete	Even though Patient5.Encounter. BloodPressureObservation.Diastolic is missing, the **Domain** ontology indicates that it is optional, so the constraint is satisfied.	4/4 1.0	4/4 1.0	4/4 1.0	5/5 1.0	5/5 1.0
DomainCoverage	Satisfied since all of the **Domain** concepts are represented in the data.	4/4 1.0	3/3 1.0	4/4 1.0	5/5 1.0	5/5 1.0
Measures that involve the Representation, Domain and Task						
Sufficiency	The **Task** specifies a SufficiencyContraint that requires at least 1 **BloodPressureObservation** must exist during the assessment period. Patient5 and Patient3 don't have valid blood pressure observations recorded.				3/5 .60	3/5 .60
TaskCoverage	**TaskCoverage** is satisfied if the **Task** concepts are a subset of the concepts represented in the **Dataset**. In this case, only the data at the Patient level has all of the **Task** concepts represented. Therefore, the eMeasure can only be calculated when all the data from the Patient level and below is available.	0/4 0.0	0/3 0.0	0/4 0.0	0/5 0.0	5/5 1.0

Table 2.5. Measurement Process Summary for Some Measures

This example shows how the DQ ontology enables a meaningful discussion of data quality characteristics required for computing an eMeasure. It also illustrates a method for quantifying each **Measure** by evaluating the proportion of constraints satisfied by the **Representations**.

2.5 Discussion

The DQ ontology presented in this study harmonized data quality concepts from the literature and provides a practical framework to evaluate data quality in health care through explicit definitions using constraints and relationships between concepts. The ontological approach provides more precise definitions of concepts than simply relying on natural language, it enables computation of a quantity for a **Measure** (**MeasurementResult)** and it makes explicit the relationship between the DQ ontology and the **Task** and **Domain** ontologies. This allows the DQ ontology to be reused for

different **Domains** and for different **Tasks** without having to devise new **Measures**. The benefit of specifying these as separate ontologies was demonstrated in the previous section. For example, when calculating the **DomainConsistency Measure**, constraints from the **Domain** ontology (i.e. "Systolic > Diastolic") can be referenced when computing **MeasurementResults** without having to change the definition of the **MeasurementMethod** (or the computer program that implements it). The same benefit is true when calculating the **Sufficiency Measure**. A **SufficiencyConstraint** can be evaluated for different **Task** ontologies to yield a **MeasurementResult** without having to change how **Measures** are defined. Not having to invent a new data quality framework for every research project should make validating data quality more common and reproducible.

Precisely defining both the **Domain** and **Task** ontology are very important in accurately describing what each data quality **Measure** means. Some of the **Measures** have constraints that reference the **Task**; these are clearly context dependent. Other **Measures** reference only the **Representation** or the **Domain** and are task independent. The constraints make clear exactly how aspects of each are related and help sharpen definitions. An example will illustrate this. **DomainConsistency** and **RepresentationConsistency** often get intertwined in definitions found in the literature. Liaw[23] listed a number of sub-meanings under his "Consistency" dimension. One sub-definition ("Consistency: Representation of data values is same in all cases") is equivalent to **RepresentationConsistency**, but he did not list an exact equivalent to the concept of **DomainConsistency**. The closest mapping is "Accuracy: Refers to values and representation of output data". On the other hand, Weiskopf[24] separated and clearly

defined these differences. The concept of **RepresentationConsistency** is embodied as "Concordance: Is there agreement between elements in the EHR, or between the EHR and another data source?" and the concept of **DomainConsistency** is well defined as "Plausibility: Does an element in the EHR makes sense in light of other knowledge about what that element is measuring?" But there is an issue in the "Concordance" definition in that the last part of her definition "…or between the EHR and another data source" includes reference to another **Measure** (**RelativeCorrectness**). A **Representation** can have **RepresentationConsistency** without having **DomainConsistency**, but the reverse is not true. This is reflected in the constraint for **DomainConsistency** by explicitly referring to **RepresentationConsistency** as part of the definition. This also highlights the usefulness of a shared vocabulary for data quality. It makes it possible to discuss nuances of data quality characteristics.

Another issue that occurs frequently in the literature is the term "accuracy;" there is an assumption that it is possible to know what is absolutely true about the world. For EHR data, there are no true gold standards for comparison. There are only other sets of data whose "accuracy" is unknown which can be referred to as relative gold standards.[55] Comparing one dataset to another to yield a positive predictive value (PPV) and sensitivity measure are a useful way to characterize the data.[56] The concept of **RelativeCorrectness** measures whether data is likely correct by matching a **Representation** to its counterpart in another **Dataset**. The matches are considered true positives and are divided by the number of **Representations** in the **Dataset** to yield a PPV as a **CorrectnessMeasure**. Similarly, **RelativeCompleteness** looks to see which "truths" of the world are captured in the EHR data. If a **Representation** is present in one

Dataset and is also present in the other "relative gold standard", then these true positives are divided by the number of **Representations** in the other **Dataset** to yield sensitivity as a measure of how complete the first **Dataset** is.

There are a number of limitations to the current research. Data quality concepts described in the meta-analyses were harmonized and mapped to concepts in the DQ ontology. Care was taken to map based on meaning or context of use, but since the meaning was from an interpretation of a definition (or sometimes, a single term), the mapping might not represent what the author of the meta-analyses intended. This research depended heavily on the core data quality concepts contained in the meta-analyses. The literature search may not have been exhaustive in finding all of the meta-analyses or there may be important data quality concepts that were not discussed in those papers. Since many data quality concepts are repeated amongst the papers, it is likely that the most important ones were captured. It is expected that additional data quality concepts will be added to the DQ ontology as the need for having a formal definition for the concept arises. Concepts that did not appear in at least three of the papers were not included in the DQ ontology. This includes concepts such as objectivity, non-duplication, security and privacy. Future work is needed to incorporate these into the DQ ontology. The concept of **DomainComplete** is currently too simplistic. It will need to be expanded to better define types of missing data as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

The DQ ontology is applicable to structured EHR data. Additional research is needed to extend the DQ ontology to notes and other unstructured data present in EHRs. Natural language processing (NLP) techniques may be used to parse relevant

DomainConcepts from the unstructured information. In that case, the DQ assessment techniques described in this paper could be used to characterize that portion of the data.

The next phase of this research is to use the DQ ontology to perform data quality **Measurements** on actual EHR data. A **Domain** ontology for a clinical area will be developed in full and mapped through **Representations** to EHR **DataValues**. Similarly, a formal **Task** ontology will be created and referenced by the data quality **Measures**. The constraints for the DQ ontology **Measures** will be written in a formal language, which can then directly be used to compute **MeasurementResults** and **Metrics** for a real-world **Dataset**.

2.6 Conclusion

The healthcare data quality literature was mined for the important terms used to describe data quality concepts. These terms were harmonized into a DQ ontology that represents core data quality concepts. Four high-level data quality dimensions (**CorrectnessMeasure**, **ConsistencyMeasure**, **CompletenessMeasure** and **CurrencyMeasure**) categorize 19 lower level **Measures**. These concepts serve as an unambiguous vocabulary when discussing healthcare data quality. The class constraints precisely define concepts better than using natural language and provide a mechanism to automatically compute **MeasurementResults** to quantify data quality. The DQ ontology can be reused with different clinical **Domain** and **Task** ontologies to make validating data quality more common and reproducible.

Chapter 3: Application of an Ontology for Characterizing Data Quality for a Secondary Use of EHR Data

Steven G. Johnson, MS[a], Stuart Speedie, PhD, FACMI[a], Gyorgy Simon, PhD[a], Vipin Kumar, PhD[b], Bonnie L. Westra, PhD, RN, FAAN, FACMI[a,c]

[a]*University of Minnesota, Institute for Health Informatics*
[b]*University of Minnesota, Department of Computer Science*
[c]*University of Minnesota, School of Nursing*

Published in:

Journal of Applied Clinical Informatics. 2016;7:69-88.

3.1 Summary

Objective: The goal of this study is to apply an ontology based assessment process to electronic health record (EHR) data and determine its usefulness in characterizing data quality for calculating an example eMeasure (CMS178).

Methods: The process uses a data quality ontology that references separate data quality, domain and task ontologies to compute measures based on proportions of constraints that are satisfied. These quantities indicate how well the data conforms to the domain and how well it fits the task.

Results: The process was performed on a de-identified 200,000 encounter sample from a hospital EHR. CodingConsistency was poor (44%) but DomainConsistency (97%) and TaskRelevance (95%) were very good. Improvements in the data quality Measures correlated with improvements in the eMeasure.

Conclusion: This approach can encourage the development of new detailed Domain ontologies that can be reused for data quality purposes across different organizations' EHR data. Automating the data quality assessment process using this method can enable sharing of data quality metrics that may aid in making research results that use EHR data more transparent and reproducible.

3.2 Introduction

Big data is an overused buzzword that seems to be applied to any application where large amounts of data are being used to solve a problem; even so, it has yielded great successes in areas as diverse as web searches, product recommendations, and natural language translations. Nowhere is the promise of big data more anticipated than in healthcare [57]. The United States (US) healthcare system is going through a transformation and rapidly adopting electronic health records (EHR) which capture patient health information in structured, semi-structured and free-form notes to document care delivery [38,58]. Because the data is now available in electronic form, it is increasingly used in applications such as clinical effectiveness research, quality improvement, and clinical decision support [59,60]. The hope is that big data analytics can find patterns in large amounts of health data to reveal the best treatment practices for different patient populations, understand which medications work best for an individual, and precisely target interventions that are most beneficial for each patient [61]. But the promised benefits can only be achieved if the quality of the data in the EHR is sufficient to support these continuing (secondary) uses. A number of studies have shown that EHR data contain errors that can affect research results [62–64]. What is needed is a way to quantify the data quality for a data set and determine if that quality is sufficient for a specific purpose.

A few healthcare data quality frameworks exist to address specific purposes, but there are no generally accepted definitions of healthcare data quality, methods to best characterize the data, nor generalized processes for quantifying data quality [16,23,24]. The Canadian Institute for Health Information (CIHI) defined aspects of data quality and provided a process for assessing data based on those definitions [22]. The process consists

of a questionnaire and relies on answers provided by data stewards to assess quality. It is a manual process and does not result in measures of data quality that are easily comparable across different data sets. The Observational Medical Outcomes Partnership (OMOP) was established to develop best practices for using observational health data to monitor the safety of prescription medications in the US [65]. Part of their approach is to ensure that all reported data meets certain data quality standards and is amenable to the analytic methods they employ. OMOP has defined a common data model and a series of data quality rules that all data contributions must pass. The current rules evolved over time to meet the specific mission of OMOP; however, the rules are not easily transferable for assessing the quality of other data sets that do not conform to the OMOP data model. But an advantage of the OMOP data quality process is that it can be automatically applied to datasets from multiple parties and can scale [66]. Similarly, the MiniSentinel project grew out of a need for the Food and Drug Administration (FDA) to monitor the safety of medical products regulated by the agency [67]. A number of industry participants contribute data to faciliate medical product surveillance. There is a common data model and a set of data quality checks that must be adhered to by all contributors. While OMOP has approximately 35 data quality rules, MiniSentinel has a checklist of over 2,000 rules that must be satisfied for data to be acceptable. These sets of rules have evolved through multiple iterations to ensure that data are of sufficient quality, but the data quality rules are limited to medical product and safety surveillance.

While these frameworks produce useful information about how data satisfies quality rules along a number of dimensions, the rules are tailored to meet the goals of their respective organizations. The Electronic Data Methods (EDM) Data Quality

Collaborative proposed that there be a standard approach for reporting data quality that would ensure transparency and consistency in data quality assessments [68]. They recommend that data quality be reported for the data in general as well as how well the data are fit for a specific purpose.

The adoption of a standardized approach will lead to improved trust in research results and the ability to share data quality information across projects. Our recent work to define data quality as an ontology provides a good framework for characterizing aspects of the data[69]. The Data Quality Ontology (DQ Ontology) provides a vocabulary for discussing aspects of data quality and also defines a process to quantify it.

An ontology is a formal specification of a shared conceptualization [39]. Every concept in the ontology has a unique name, properties, relationships to other concepts and constraints that are always true for that concept. The benefits of using an ontology to describe data quality are that an ontology is written in a formal language, it is able to represent semantics, it provides a shared vocabulary for discussing data quality and it is sufficiently rigorous to be used directly in algorithms and computer programs [40]. Key concepts and their definitions from the DQ Ontology are listed in Table 3.1 and the relationships between them are shown in Figure 3.1[69(p1940)]. This ontology precisely defines data quality concepts in terms of relationships and constraints with other DQ concepts (shown in blue in Figure 3.1). Also included in Figure 3.1 is a link to 2 other ontologies described later in the paper – Task (shown in Figure 3.2 in orange) and Domain (shown in Figure 3.3 in green). The DQ Ontology is a meta-ontology that defines data quality concepts with respect to these two other ontologies.

Critically, a separate Domain ontology defines the formal semantics (using properties and constraints) of concepts represented in the data. The Task ontology is a specification for the concepts necessary to carry out a particular use of the data. The DomainConcepts link the Representations in the Dataset to the Domain and Task ontologies. Measures are further refined into ConsistencyMeasures and CompletenessMeasures. These are described in more detail in Table 3.2.

Table 3.1. Data Quality Ontology – Key Concepts

Concept	Definition
Measure	An aspect of data quality that quantifies a characteristic of the data.
CorrectnessMeasure	Measures that assess whether the data that exists in the Dataset is true.
ConsistencyMeasure	Measures that assess data conformance to constraints, rules and restrictions of the Domain.
CompletenessMeasure	Measures that assess whether a truth about the world is contained in the data.
CurrencyMeasure	Measures that assess timeliness of the data to represent the Domain and Task.
MeasurementMethod	A series of steps used to quantify an aspect of data quality for a Measure.
Measurement	The process of performing a MeasurementMethod to produce a MeasurementResult
MeasurementResult	The quantity produced by a MeasurementMethod.
Metric	Statistics for how a MeasurementResult varies over time or other dimensions.
Dataset	The entire set of Representations that are being assessed.
Representation	The lowest level, atomic piece of information that exists in the data being assessed (also known as a data field, observation, value).
DomainConcept	Concepts in the clinical Domain and Task of interest that map to Representations in the set of data being assessed.
Domain	A separate ontology describing the clinical domain of interest.
Task	A separate ontology describing the specific purpose of using the data.

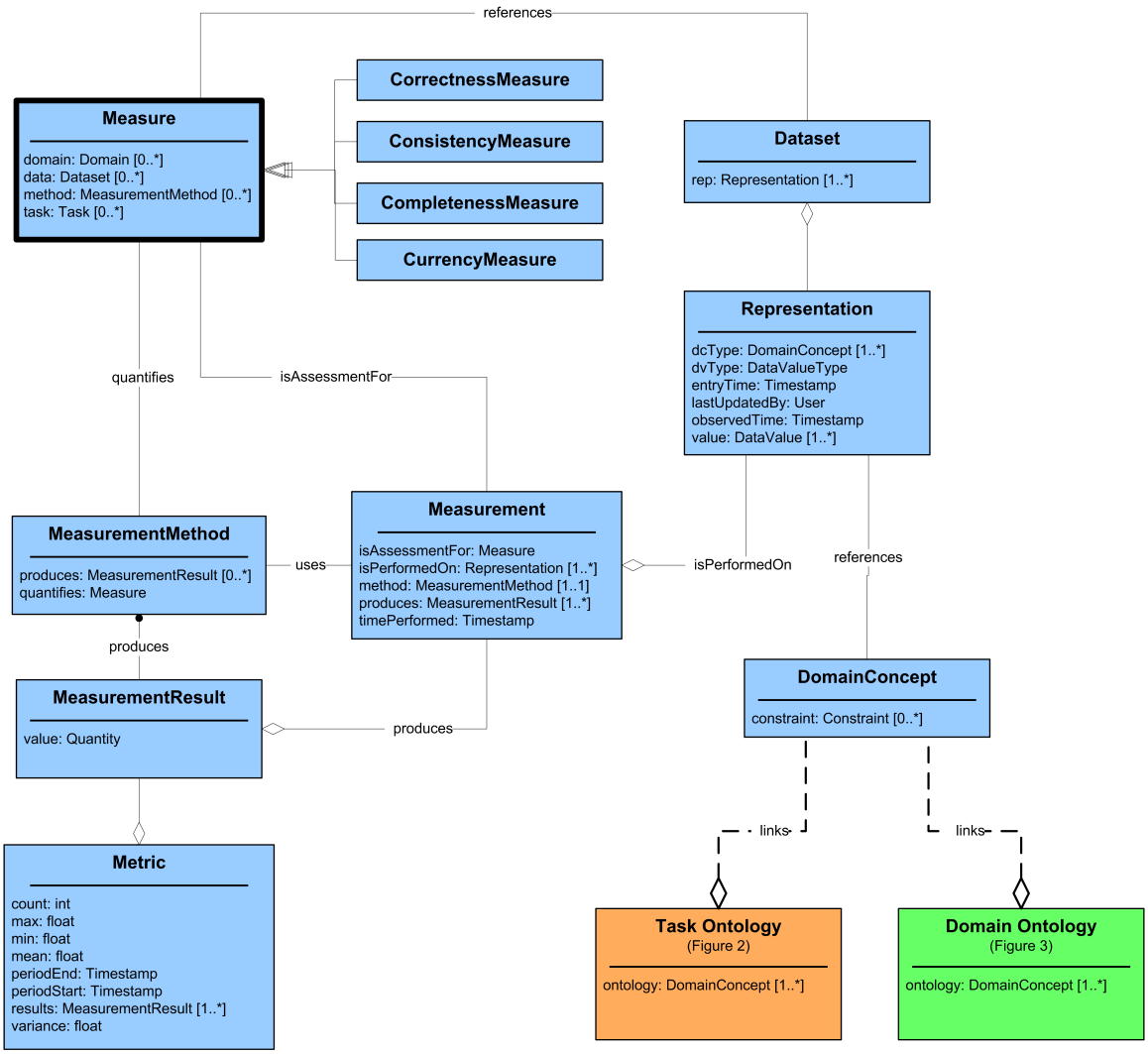

Figure 3.1. Data Quality Ontology Diagram

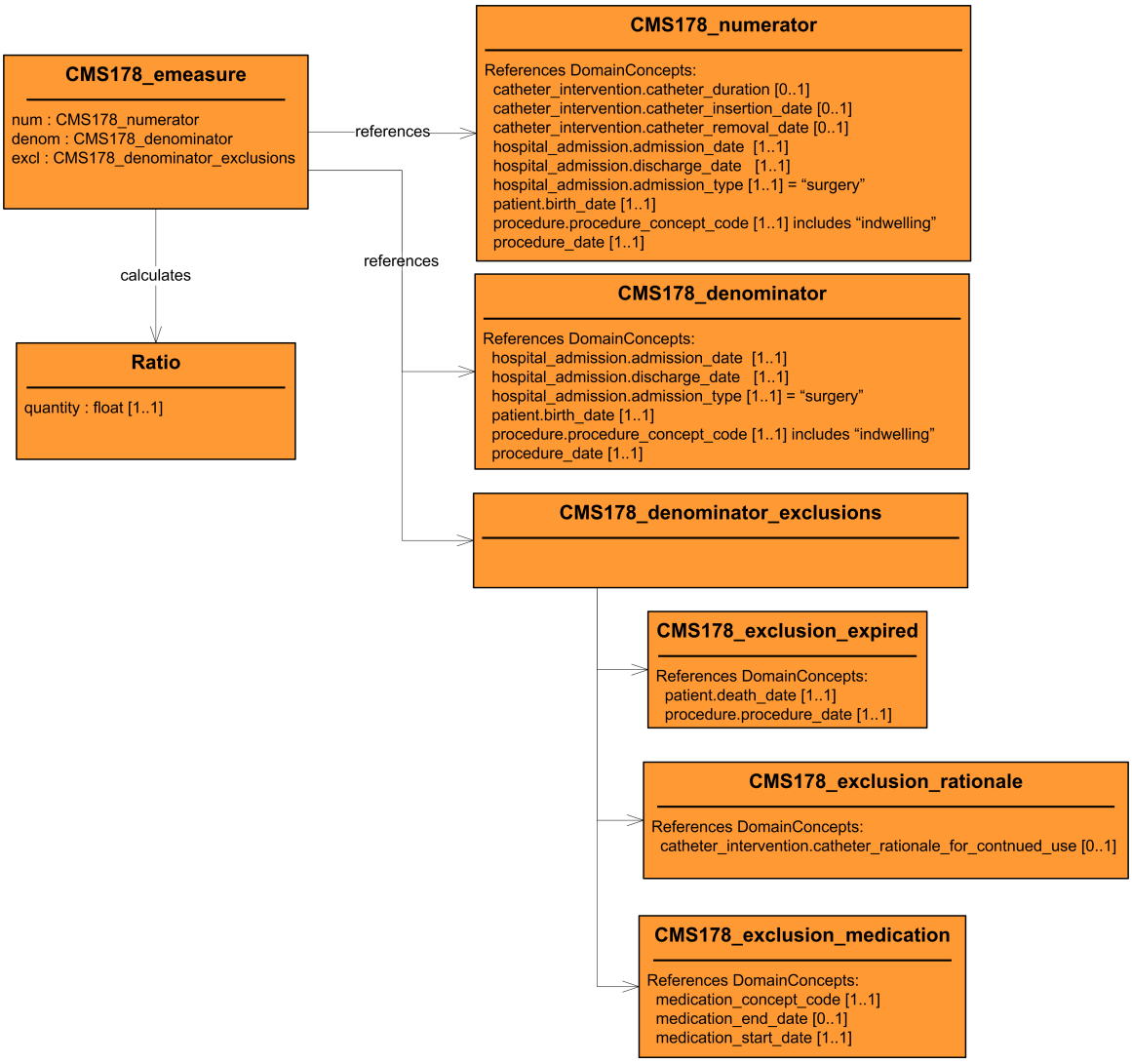

Figure 3.2. Task Ontology (CMS178)

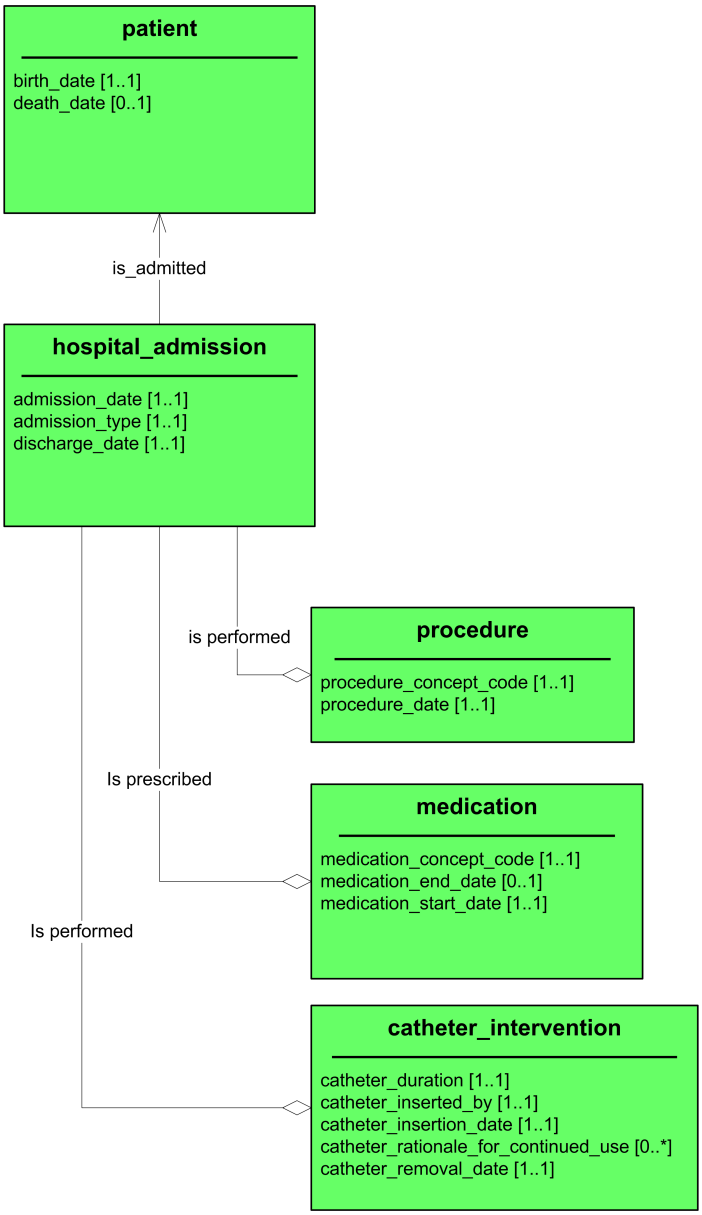

Figure 3.3. Domain Ontology

Table 3.2. Data Quality Ontology – Measure Detail with Constraints

Measure	Definition	Constraint
ConsistencyMeasure		
RepresentationConsistency	The data is a valid value and format for its DataValueType and all of the Representations for the same information have the same values.	value.isValidFormat()
DomainConsistency	Concepts in the Domain are represented in the data and the data satisfies syntactic and semantic rules. Constraints for the Domain are satisfied.	RepresentationConsistency and RepresentationComplete and CodingConsistency and DomainConstraints
DomainConstraints	All of the constraints defined for the DomainConcept are satisfied.	for each constraint in value.DomainConcept.constraints: constraint is True
CodingConsistency	Representations that are of coded text data type must be correctly mapped to an enumerated list or a terminology.	value.dataValueType() == 'coded' and value.isValidCode()
CompletenessMeasure		
RepresentationComplete	Domain independent extent to which data is not missing.	value is not null
DomainComplete	The extent to which information is present or absent as expected.	RepresentationComplete or (Cardinality == 'optional')
TaskSufficiency	The data has sufficient Representations along a given dimension (i.e. time, patient, encounter) to perform the Task.	if all(concept.DomainComplete > THRESHOLD) then average all concept.DomainComplete
TaskRelevance	The data is sufficient for the Task and conforms to the Domain.	TaskSufficiency and (for all concepts in Task.DomainConcepts: average all concept.DomainConsistency)
DomainCoverage	The data can represent the values and concepts required by the Domain.	For each concept in Domain.DomainConcepts: isMapped(concept)
TaskCoverage	The data contains all of the information required by the Task.	For each concept in Task.DomainConcepts: isMapped(concept)

Defining data quality as an ontology also provides a process for computing quantities that characterize data quality[69]. The data quality assessment process evaluates constraints defined for each Measure to compute a proportion of constraints that are satisfied. This MeasurementResult is a fraction where the denominator is the number of

Representations for each concept and the numerator is the number of Representations with all constraints satisfied. An example is RepresentationConsistency. The process (MeasurementMethod) counts the number of Representations that conform to its DataValueType (i.e. numeric fields only consist of numbers, decimal points or signs and dates have a valid format, etc). The RepresentationConsistency MeasurementResult is a fraction with the denominator being all Representations and the numerator being the number of Representations that satisfy the DataValueType formatting rules. A more complex example is CodingConsistency which assesses how well a coded Representation maps to standard terminologies. For example, medications should be mapped to valid RxNorm values. CodingConsistency is computed as the ratio of the number of Representations with valid codes to the total number of Representations. DomainConstraints are the proportion of constraints defined for the Domain that are satisfied by each Representation. If there are multiple constraints for a Representation, then all of them must be satisfied.

Measures such as DomainConsistency are based on other Measures. DomainConsistency requires that the combination of RepresentationConsistency, DomainComplete, CodingConsistency and DomainConstraints are all satisfied. The MeasurementResults for every DomainConcept are computed and then saved in a data quality database as meta-data about the Dataset.

The purpose of this study was to apply this DQ assessment process and determine its usefulness in characterizing data quality for data that is used in calculating an example eMeasure. To accomplish this goal, software was developed that implements the process and uses Domain and Task ontologies to produce Metrics for specific Measures of data

quality. The value of this approach is demonstrated by examining how these quantities characterize an EHR dataset for conformance in representing a Domain and for its fitness to be used for a particular Task.

3.3 Methods

Fairview Health Services and the University of Minnesota collaborated to create and maintain a clinical data repository (CDR) with over 2 million patients from seven hospitals and 40 clinics. Approval from the IRB (#1412E57982) was obtained to use the data for this study. A 200,000 encounter random sample was de-identified and used as the dataset for this research.

The process for characterizing data quality required the development of three ontologies and a software program that implements the data quality assessment process. The DQ Ontology defines Measures of interest and includes the constraints and interrelationships between data quality concepts. The computation of an eMeasure will be used as an example Task for this research. An eMeasure computes the proportion of a population conforming to a specific health outcome of interest[70]; CMS178 will be used as example eMeasure. It is defined as "Urinary catheter removed on Postoperative Day 1 (POD 1) or Postoperative Day 2 (POD 2) with day of surgery being day zero"[71,72]. Patients that have indwelling catheterization for long periods of time are at higher risk of developing catheter-associated urinary tract infection (CAUTI). This eMeasure quantifies the proportion of patients that receive the evidence based best practice of removing the catheter within 48 hours post-surgery[73]. It provides a real-world secondary use for EHR data that can be compared to underlying data quality for this research.

Constraints were defined for 10 Measures and are listed in Table 3.2. The full DQ Ontology describes 19 measures that characterize data quality[69]. Nine of these were selected for this research to illustrate how Measures in the ontology quantify data quality. The other Measures were not included as they either required another organization's data or relied on meta-data that is not captured by the EHR used for this study. An additional Measure, DomainConstraints, was included for this paper to better illustrate an intermediate aspect of DomainConsistency.

A Task ontology for the CMS178 eMeasure was developed. The eMeasure is a proportion that consists of a "Denominator", which is the entire patient population to which the eMeasure applies and a "Numerator", which is the subset of patients that conform to the characteristic of interest. The denominator also specifies "Denominator Exclusions" for patients that should not be counted in the eMeasure. The instructions for computing CMS178 is 64 pages long, but for this paper, CMS178 will be simplified by eliminating some of the denominator exclusions and including in the denominator all surgeries instead of just major surgeries. The simplified definition for CMS178 is:

Denominator:

- All hospital patients (age 18 and older) that had surgery during the measurement period with a catheter in place postoperatively.

Denominator Exclusions:

- Patients who expired perioperatively (CMS178_exclusion_expired).
- Patients who had physician/APN/PA documentation of a reason for not removing the urinary catheter postoperatively (CMS178_exclusion_rationale).

- Patients who had medications administered within 2 days of surgery that were Diuretics, IV Positive Inotropic and Vasopressor Agents or Paralytic Agents (CMS178_exclusion_medication).

Numerator:

- Number of surgical patients whose urinary catheter is removed on postoperative day (POD) 1 or postoperative day (POD) 2 with day of surgery being day zero.

The eMeasure is computed as:

$$CMS178_{simple} = \frac{CMS178\_numerator}{CMS178\_denominator - CMS178\_denominator\_exclusions}$$

These statements are specified in the CMS178 implementation guide and were mapped to concepts in the Domain ontology. An encounter was considered a surgery when the `admission_type` field was coded as "SURGERY". Patients who had catheters inserted during a procedure were indicated by the `procedure_concept_code` equalling "NUR380". The Task ontology, shown in Figure 3.2, specifies the relationship between aspects of the Task and the DomainConcepts that are required to calculate CMS178.

Ideally, the Domain ontology should represent all of the data that is in the EHR or CDR. A complete Domain ontology does not yet exist, but a Domain ontology was created for this research in order to illustrate the data quality assessment process. It includes all of the DomainConcepts referenced by the Task and which are required to compute the CMS178 eMeasure. For this paper, the Domain ontology is documented using a UML diagram (Figure 3.3) and a table that lists constraints (Table 3.3).

Table 3.3. Domain Ontology with Constraints

DomainConcept	Domain Complete (Cardinality)	Representation Consistency (DataValueType)	DomainConstraint
dataset			
patient			
birth_date	required	date	birth_date <= today
death_date	optional	date	if death_date is not null then death_date >= birth_date
hospital_admission			
admission_date	required	date	discharge_date - admission_date < 1000
admission_type	required	code:CHOICE	
discharge_date	required	date	admission_date <= discharge_date
procedure			
procedure_concept_code	required	code:CPT4	
procedure_date	required	date	procedure_date >= admission_date
medication			
medication_concept_code	required	code:RXNORM	
medication_end_date	optional	date	medication_start_date < medication_end_date
medication_start_date	required	date	medication_start_date >= admission_date
catheter_intervention			
catheter_duration	optional	numeric	catheter_duration >= 0 catheter_duration < 1000
catheter_insertion_date	optional	date	if catheter_insertion_date is not null then catheter_inserted_by is not null if catheter_insertion_date is not null and catheter_removal_date is null then catheter_rationale_for_continued_use is not null
catheter_removal_date	optional	date	if catheter_removal_date is not null then catheter_insertion_date is not null
catheter_rationale_for_continued_use	optional	string	if catheter_rationale_for_continued_use is not null then catheter_insertion_date is not null
catheter_inserted_by	optional	string	if catheter_inserted_by is not null then catheter_insertion_date is not null

Table 3.4. MeasurementResults for DomainConcepts

DomainConcept	Representation Consistency	Representation Complete	Domain Complete	Coding Consistency	Domain Constraints	Domain Consistency
dataset	100%	96%	98%	44%	97%	97%
patient	100%	55%	100%		100%	100%
birth_date	100%	100%	100%		100%	100%
death_date	100%	10%	100%		100%	100%
hospital_admission	100%	100%	100%	100%	100%	100%
admission_date	100%	100%	100%		100%	100%
admission_type	100%	100%	100%	100%		100%
discharge_date	100%	100%	100%		100%	100%
procedure	100%	99%	99%	29%	97%	63%
procedure_concept_code	100%	100%	100%	29%		29%
procedure_date	100%	97%	97%		97%	97%
medication	100%	92%	96%	92%	96%	96%
medication_concept_code	100%	92%	92%	92%		92%
medication_end_date	100%	90%	100%		97%	97%
medication_start_date	100%	95%	95%		95%	95%
catheter_intervention	100%	88%	100%		92%	92%
catheter_duration	100%	83%	100%		99%	99%
catheter_insertion_date	100%	92%	100%		78%	78%
catheter_removal_date	100%	85%	100%		98%	98%
catheter_rationale_for_continued_use	100%	99%	100%		89%	89%
catheter_inserted_by	100%	73%	100%		99%	99%

Domain constraints, including relationship cardinality (i.e. whether the data is optional or required) and data types for all of the fields are listed in Table 3.3. These constraints represent aspects of the data and its interrelationships that should always be true if the data accurately represents the clinical concepts of the Domain. For example, hospital discharge date should always occur after the hospital admission date. These were implemented as computer executable SQL but, for brevity, are shown as pseudo code in the table. For example, the first constraint for `catheter_insertion_date` is "`if catheter_insertion_date is not null then catheter_insertion_by is not null`" which can be paraphrased as ""if there is a catheter insertion documented, then the name of the clinician who inserted it should also be documented".

Concepts in the Domain ontology form a hierarchy and the parent concepts in the hierarchy can also have data quality Measures computed. There are MeasurementResults for parent concepts such as `medication, hospital_admission`, and `patient`. The denominator for the parent concept MeasurementResult is a count of all of the Representations for all of its sub-concepts. The numerator is a count of all of the Representations for all of the sub-concepts that satisfy the Measure. In this way MeasurementResults can be aggregated up the hierarchy, including aggregating Measures that apply to the Dataset as a whole.

Some Measures such as TaskRelevance and DomainConsistency combine other Measures. Pipino[74] discusses a number of methods for aggregating multiple data quality indicators that include min, max and average of the Measure quantities. This study used the simple approach of treating each Measure equally and averaging the MeasureResults.

3.4 Results

The data quality assessment process was performed on the de-identified 200,000 encounter sample from the Fairview Health EHR data. Table 3.4 shows the MeasurementResults (expressed as percentages) for DomainConcepts, parent concepts and the Dataset as a whole. DomainCoverage and TaskCoverage were 100% for the Dataset and are not listed. TaskSufficiency (99%) and TaskRelevance (95%) could only be calculated at the patient level since that was the only level in the Domain hierarchy that contained all of the DomainConcepts referenced by the Task.

RepresentationConsistency was 100%. RepresentationCompleteness assesses how many Representations have a data value that is not missing. It varied from 10% for `death_date` to 100% for `birth_date` and `procedure_concept_code`. DomainCompleteness indicates whether the Domain permits a value to be missing (i.e. it is optional). For example, `death_date` only has a value for 10% of the patients, but since it is an optional DomainConcept in the Domain model, its DomainCompleteness was 100%. CodingConsistency assesses how well the coded Representations conform to the standard terminology that is specified in the Domain ontology. This ranged from a low of 29% conformance with CPT4 procedure codes (`procedure_concept_code`) to a high of 100% for `admission_type`.

DomainConstraints were satisfied overall 97% of the time, but constraints for some concepts were much lower (`catheter_insertion_date` was 78%). DomainConsistency is the combination of RepresentationConsistency, DomainComplete, CodingConsistency and DomainConstraints and it is the best overall Measure to indicate

a Dataset's conformance to a Domain. Overall, this Dataset had a DomainConsistency of 97%.

The TaskSufficiency and TaskRelevance Measures were also computed. TaskSufficiency assesses whether a Dataset has enough data to be used to perform a Task. TaskSufficiency is calculated by examining DomainComplete for each of the referenced DomainConcepts and ensuring they are above a certain threshold. And if they are, the result is the average of all of the DomainComplete ratios. In this example, a threshold of 80% was used. This means that 80% of the Representations must be DomainComplete in order to be considered sufficient to carry out the calculation of the eMeasure. In this case, all of the DomainCompleteness' were above 80% and the DomainCompleteness ratios for all of the referenced DomainConcepts are averaged to produce an overall TaskSufficiency of 99%.

TaskRelevance not only assesses whether data is sufficient for a task but that it also conforms to the Domain (DomainConsistency). The DomainConsistency of each of the concepts referenced by a Task are averaged to produce an overall DomainConsistency which is then combined (averaged) with the TaskSufficiency value to yield a TaskRelevance value. For this example, the TaskRelevance of the Dataset for calculating the CMS178 eMeasure was 95%.

Measures can also be calculated for the Dataset at particular points in time. Using data for each month from April 2011 to July 2013, MeasurementResults were calculated. The graph in Figure 3.4 shows DomainConsistency Metrics for a few concepts of interest (`catheter_duration`, `catheter_insertion_date`, `catheter_removal_date` and `catheter_rationale_for_continued_use`). Figure 3.5 shows how TaskRelevance

changes over time. Figure 3.6 shows the value for the CMS178 eMeasure over the same time period. And Figure 3.7 displays the monthly trend of DomainConsistency for the entire Dataset. The Pearson correlation between DomainConsistency and the CMS178 eMeasure was 0.78.

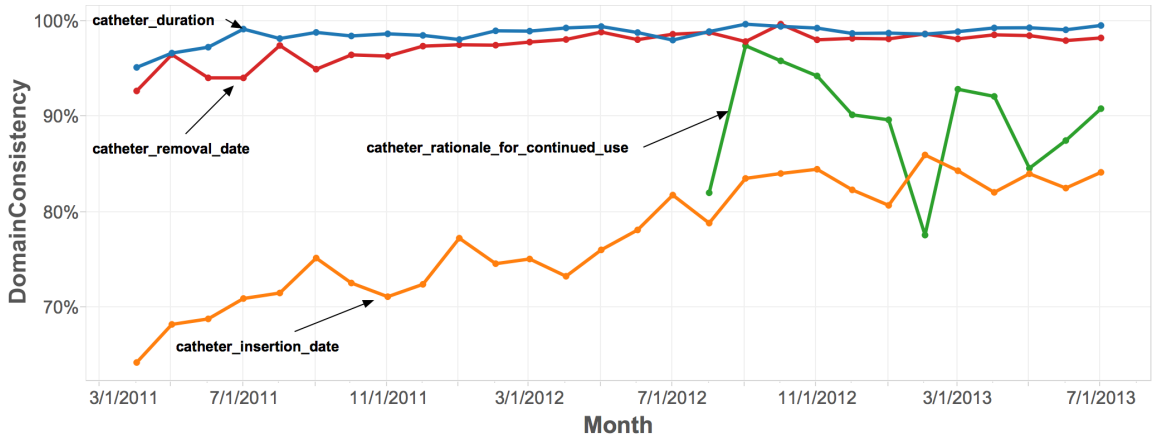

Figure 3.4. DomainConsistency for Selected DomainConcepts, by Month

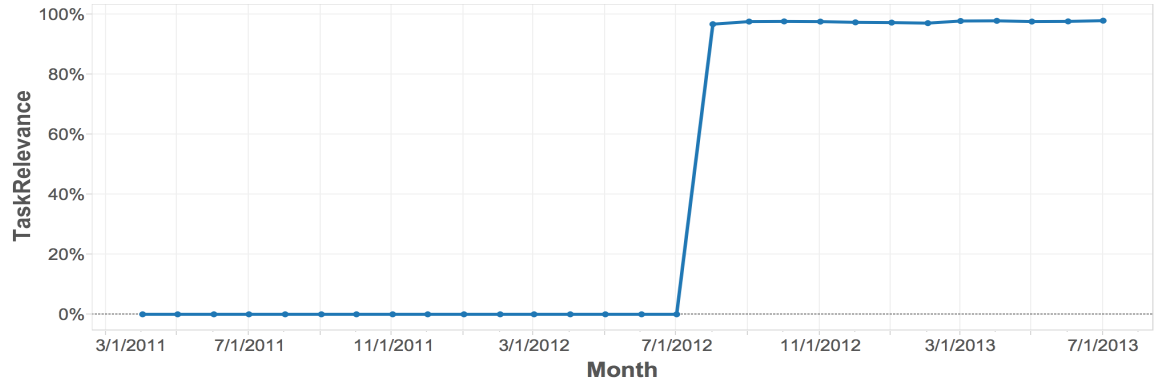

Figure 3.5. TaskRelevance by Month

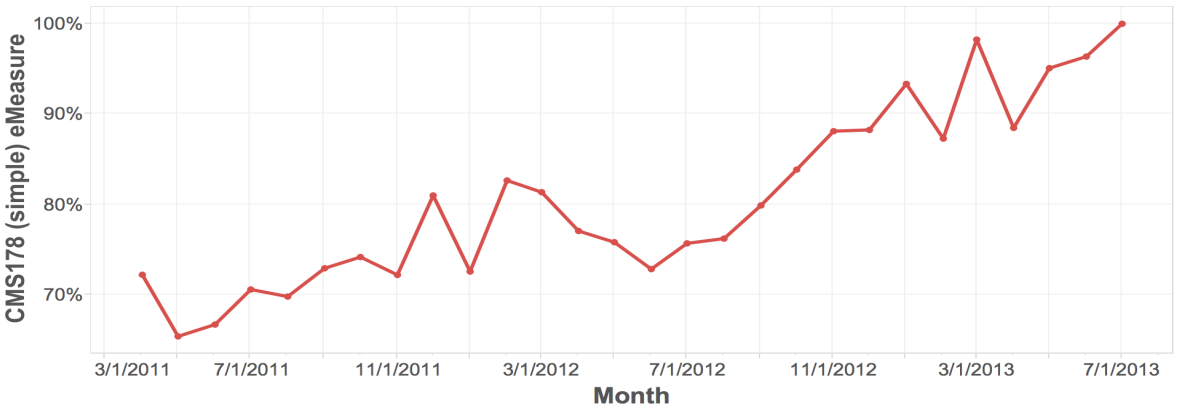

Figure 3.6. CMS178 (simple) eMeasure by Month

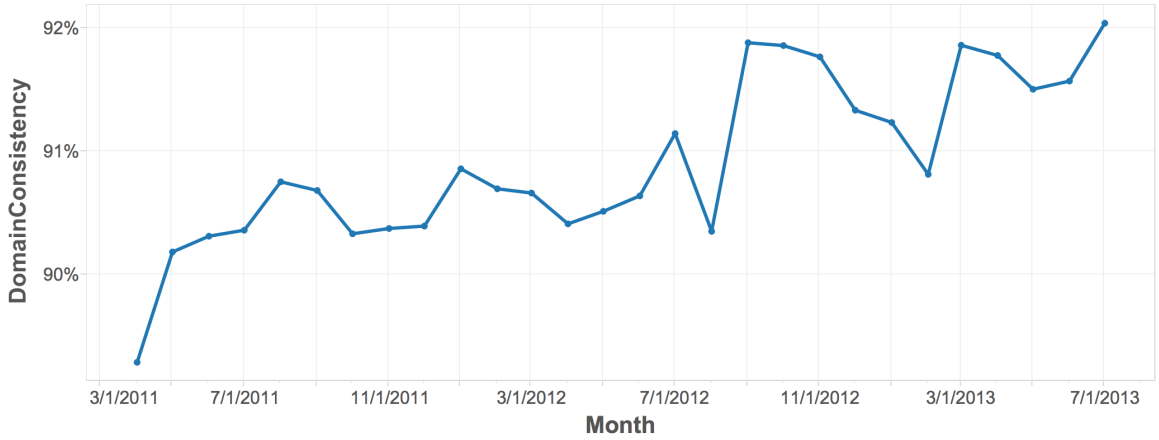

Figure 3.7. DomainConsistency for the Dataset

3.5 Discussion

The DQ assessment process described in this paper characterizes the data quality aspects of EHR data. The process requires correctly defined Task and Domain ontologies and yields specific quantities that indicate data quality. For this set of EHR data, RepresentationConsistency was very good. All Representations matched their data

formats 100% of the time. This high conformance is due to the data entry rules for the EHR that strictly enforce the correct data formats.

DomainCompleteness was also very good, with an overall Dataset conformance of 98%. Again, this is likely indicative of the EHR data entry rules ensuring that when a data value is required to exist that the clinician is guided to enter a value. For example, an important field like `birth_date` has a value for all patients (both DomainComplete and RepresentationComplete were 100%). CodingConsistency was very high except for `procedure_concept_code`, which was only 29%. When the data was further examined, it was revealed that the procedures were coded using valid CPT4 code or codes which only had meaning to the hospital (i.e. "NUR380") or which were variations of valid CPT4 codes (i.e. "82962.001").

The DomainConstraint results, for the set of constraints defined in this research, revealed an overall conformance to the Domain of 97%. But `catheter_insertion_date` had a relatively low DomainConstraint value. The constraint requires that if the patient has a catheter, then the name of the clinician who performed the insertion must be documented in `catheter_inserted_by` and that if a catheter is inserted but no removal date is documented, then there should be a `catheter_rationale_for_continued_use` documented by a clinician. These constraints were only satisfied 78% of the time.

Figure 3.4 shows that DomainConsistency was improving for `catheter_insertion_date` and `catheter_removal_date` over the measurement period. This parallels an improvement in the CMS178 eMeasure over that same time period (Figure 3.6). In fact, Fairview had undertaken a quality improvement initiative starting in November 2011 to better document catheter insertions and then in the summer

of 2012 to focus on reducing CAUTI. This initiative required improvement in indwelling catheter documentation, including documenting the rationale for not removing a catheter. The increasing DomainConsistency reflects the improved data quality as the initiative progressed. The correlation between DomainConsistency and the CMS178 eMeasure was 0.78, which is a moderately positive correlation. This suggests that as the data's conformance to the Domain improves, the computed value of CMS178 should converge on the true value.

DomainConsistency is the Measure that best reflects the Dataset's conformance to the Domain since it incorporates the other Measures. The DomainConsistency ratio continued to improve over time for the Dataset as a whole. Figure 3.6 shows that it improves from 89% to over 92% during the two years of the measurement period.

TaskRelevance is the Measure that best indicates that a Dataset can be used for a specific purpose. For this data, `catheter_rationale_for_continued_use` was not entered into the EHR before July 2012, so TaskRelevance was 0 prior to that date. If these data quality Measures had been in use by this healthcare organization, they might have decided not to compute the eMeasure before that date based on the low TaskRelevance.

OMOP and MiniSentinel have developed data quality rules that provide detailed information about specific pieces of data that don't conform to data quality expectations. The process described in this paper provides a data quality assessment approach that has several advantages over those methods. First, MeasurementResults are scalar quantities instead of lists of rules that failed. Scalar quantities are simpler to use and can be more easily compared across Datasets and across time. Heinrich[75] has proposed a set of

requirements that all data quality quantities should possess. They should be normalized, interval scaled, interpretable, aggregatable, adaptable and feasible. The quantities for the data quality assessment method described in this paper meet these requirements. Since the quantities are proportions, they are both normalized (range from 0 to 1), interval scaled (the difference between 20% and 30% is the same as the difference between 70% and 80%) and easily interpreted (researchers are familiar with using proportions). The quantities can be aggregated to parent concepts and to the entire Dataset. These quantities are also adaptable in that they can be used with different Tasks, and they are computationally feasible. The OMOP and MiniSentinel data quality rules are similar to DomainConstraints and they could be turned into a core set of constraints in a Domain ontology.

Secondly, this approach can be used to assess existing EHR data. The OMOP and MiniSentinel approaches assess the quality of incoming data feeds in order to filter out bad data from a central repository. Most healthcare data is already in an existing repository and the data quality assessment method described in this paper can be used to evaluate that pre-existing data.

Finally, the MeasurementResults can be used for different Tasks focusing on the same time periods without having to recompute them for the Domain. Once a Domain ontology has been defined, some Measures (such as RepresentationConsistency, RepresentationComplete, DomainComplete, CodingConsistency, DomainConstraints and DomainConsistency) will characterize the data regardless of how the data is to be used. This promotes reuse and sharing of the Metrics. If another Task is to be performed using the data, the already computed Domain Measures for each referenced DomainConcept

can be reused. In addition, these MeasurementResults are comparable across multiple Datasets if they use the same Domain ontology.

One potential limitation of this research is the choice of the 80% threshold for TaskSufficiency. The selection of this value is reasonable but arbitrary. It is possible that different Tasks will require different thresholds for the amount of data necessary for a result to be valid. More research is needed to quantify the impact of TaskSufficiency on the validity of results for different Tasks.

More research is also needed to determine the best way to combine multiple Measures. It is useful to be able to combine Measures to create a small number of quantities that can be used as a convenient score for the quality of a Dataset. The approach presented in this paper used a straightforward method of averaging the component Measures. For example, to compute TaskRelevance, the DomainConsistency of each DomainConcept referenced in the Task is averaged and then combined (averaged) with TaskSufficiency. This method may be appropriate if each DomainConcept is equally important to the overall Measure and there are a sufficient number of DomainConcepts to make an average with its implied normal distribution meaningful. However, it may be the case that some DomainConcepts are more important in a particular Task than others. In the example used in this paper, the DomainConsistency of `catheter_duration` is probably more important than the patient's `birth_date` (for determining age) when computing the CMS178 eMeasure. Further research is needed to determine if there is a better way to calculate Measures that combine other Measures that takes into account the data quality impact of each DomainConcept on the result. There are also additional Measures that should be defined

for aspects of data quality not addressed in this paper. For example, duplication of data and records is an important concept and should be included as an additional Measure in the DQ Ontology. The original ontology left it out because it didn't meet its inclusion criteria of being referenced in at least 3 data quality meta-analyses papers.

As more medical data is aggregated and organized, healthcare is able to benefit from big data analytic techniques. Future research should examine how the data quality assessment method described in this paper can be used for Tasks such as comparative effectiveness research and predictive modeling. In addition, this framework can be used to assess data quality in observational research. Measures of data quality could be computed on a timely basis (possibly nightly) so that researchers can quickly identify and mitigate data quality issues before they get too large.

3.6 Conclusions

This paper presented the results of a data quality assessment method that characterizes some aspects of the quality of EHR data. The method uses a DQ Ontology that references separate Domain and Task ontologies to compute Measures which quantify how well the data conforms to the Domain and how well it fits the Task. Metrics that show trends over time and for specific concepts in the data can be used to show changes in data quality and the Metrics can be compared to other Datasets that use the same Domain ontology. Different Tasks can reuse the Metrics without having to recompute them. These quantities may be easier to use and understand than some of the existing approaches to data quality assessment. This approach can encourage the use of existing or development of new detailed Domain ontologies that can be reused across different organizations'

EHR data. Automating the data quality assessment process using this method can enable sharing of data quality Metrics that may aid in making research results that use EHR data more transparent and reproducible.

Clinical Relevance Statement

The assessment process uses a Data Quality Ontology that references separate Domain and Task ontologies to compute Measures which quantify how well EHR data conforms to a Domain and how well it fits a specific Task. Automating the data quality assessment process using this approach can enable sharing of data quality Metrics that may aid in making research results that use EHR data more transparent and reproducible.

Conflict of Interest

The authors declare that they have no conflicts of interest in the research.

Human Subjects Protections

De-identified EHR data was used for this research and proper precautions were taken to minimize privacy risk. Patients were allowed to opt out of having their medical data used for research. IRB approval was obtained (University of Minnesota IRB #1412E57982).

Acknowledgments

This research was supported by Grant Number 1UL1RR033183 from the National Center for Research Resources (NCRR) of the National Institutes of Health (NIH) to the University of Minnesota Clinical and Translational Science Institute (CTSI). Its contents

are solely the responsibility of the authors and do not necessarily represent the official views of the CTSI or the NIH. The University of Minnesota CTSI is part of a national Clinical and Translational Science Award (CTSA) consortium created to accelerate laboratory discoveries into treatments for patients.

Chapter 4: Quantifying the Effect of Data Quality on the Correctness of an eMeasure

Steven G. Johnson, MS[a], Stuart Speedie, PhD, FACMI[a], Gyorgy Simon, PhD[a], Vipin Kumar, PhD[b], Bonnie L. Westra, PhD, RN, FAAN, FACMI[a,c]

[a]*University of Minnesota, Institute for Health Informatics*

[b]*University of Minnesota, Department of Computer Science*

[c]*University of Minnesota, School of Nursing*

4.1 Summary

Objective

The purpose of this study is to quantify the impact of two data quality issues, missing data (RepresentationCompleteness) and domain conformance (DomainConstraints), on the correctness of an eMeasure (CMS178 - Urinary Catheter Removal After Surgery).

Materials and Methods

Data quality issues were artificially created by systematically degrading the underlying quality of sample EHR data using two methods: independent and correlated degradation. A linear model that describes the change in the correctness of the eMeasure quantifies the impact of each data quality issue on the eMeasure.

Results

Birth date and admission type had the most impact on the CMS178 eMeasure for missing data quality issues; death date and medication end date had the most impact for domain conformance issues.

Discussion

The impact of data quality issues can be quantified using a generalized process. A 1% improvement in data quality of birth date or admission type yield 1% less missed catheter removal events whereas other variables did not have as great an impact. The correlated degradation method is the most robust approach, but independent degradation was most efficient.

Conclusion

Secondary use of EHR data is only warranted if the data is of sufficient quality. The assessment approach described in this study demonstrated how the impact of data quality

issues on an eMeasure can be quantified and the approach can be generalized for other data analysis tasks. Healthcare organizations can prioritize data quality improvement efforts to focus on the areas that will have the most impact on correctness and assess whether the values that are being reported should be trusted.

4.2 Background

The United States (US) healthcare system continues to invest in information technology to improve health outcomes[76]. This not only includes infrastructure such as electronic health record (EHR) systems and interoperability standards, but also initiatives for quickly translating clinical research into best practices[77]. Now that health information is in electronic form, it is made more available for research. This increasing secondary use of EHR data to improve health outcomes is promising, but it depends on clinical information being of sufficiently high quality to support the research[78].

One of the secondary uses of EHR data is evaluating care quality and outcomes. eMeasures are standardized performance measures based on data extracted and aggregated from EHRs to quantify how well patient care is meeting best practices[26]. eMeasures are just now becoming computable within EHR systems[27,28]. There are 297 active eMeasures listed in the US Department of Health & Human Services Measures Inventory[79] and many of these (93) are required to be reported by providers in order to meet the requirements of Meaningful Use[58,80,81].

Correctly computing an eMeasure depends on how well the data is recorded in the EHR[82]. But EHR vendors have not always ensured that data is captured at a quality sufficient to correctly compute the eMeasure[83]. Data may be adequate to document care delivery but may be insufficient to support the computation of an eMeasure[6]. Data may be missing, incorrect, out of range or inappropriate for the field. In these situations, the patient's record should not be used in the calculation of the eMeasure and some ability to quantify the best practice that the eMeasure was intended to assess will be lost. In a recent study, the data quality of a clinical data repository (CDR) was measured[84]. The

CDR was used to compute the CMS178 eMeasure (Urinary Catheter Removal After Surgery) but one data element, `catheter_rationale_for_continued_use`, was found to be of low quality. Technically, the eMeasure could be computed, but it did not reflect all of the care given to the population for which the eMeasure was intended to represent. These secondary uses of EHR data could be better trusted if the impact of the underlying data quality was assessed[85].

Recent work to define concepts as a data quality ontology (DQ Ontology) has improved the ability to discuss data quality issues and has lead to an assessment method that allows data sets to be characterized along a number of data quality dimensions[84]. This DQ Ontology is shown in Figure 4.1.

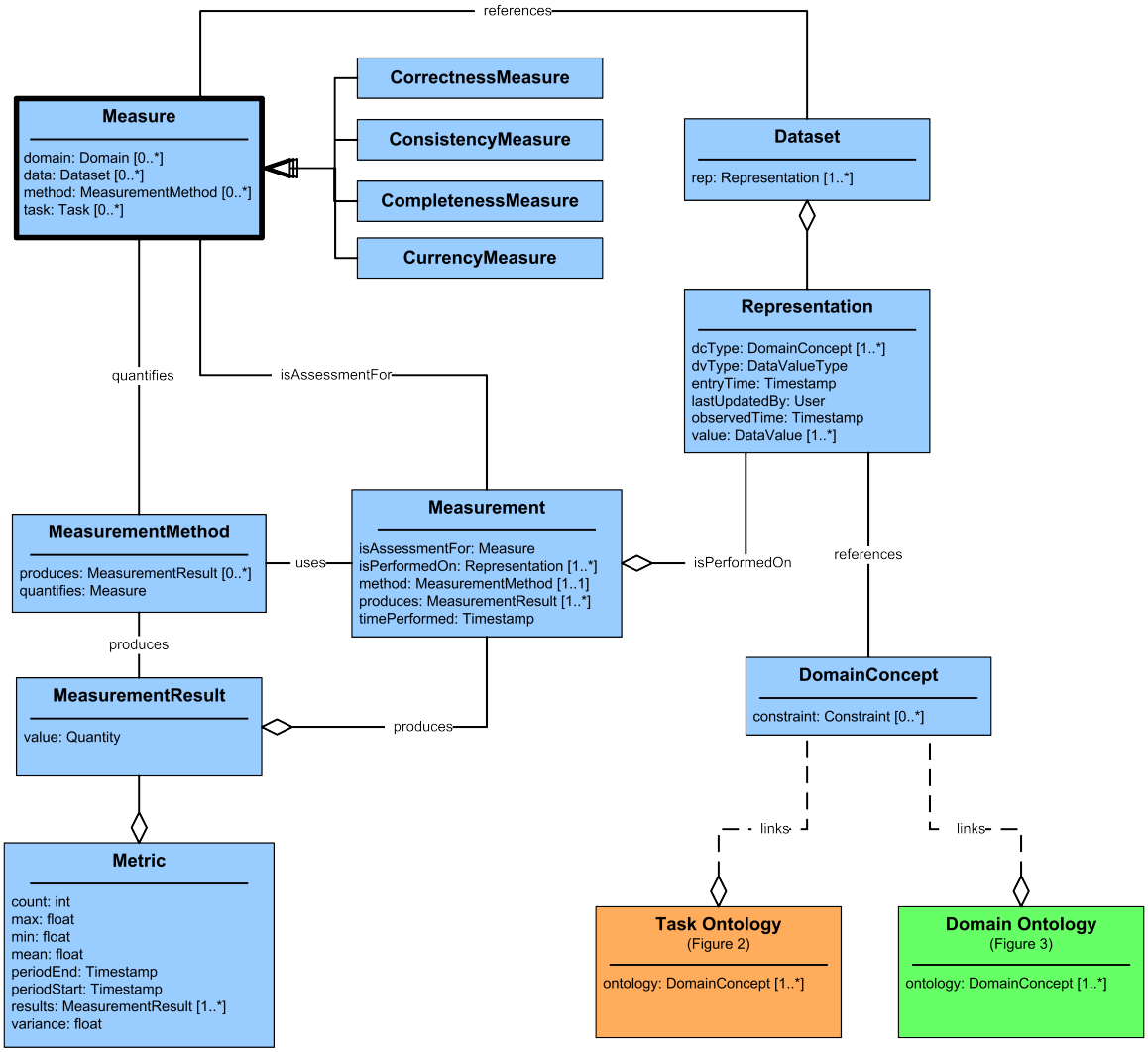

Figure 4.1. Data Quality Ontology

Data quality assessment using an ontology based framework has a number of benefits[69]. The DQ Ontology is specified in a formal language, is able to describe semantics, uses a shared vocabulary for data quality concepts, and is sufficiently well-defined to be used by computer software[40]. Concepts in the DQ Ontology are linked to two other ontologies: a Task ontology that describes the concepts necessary to carry out a particular use of the data and a Domain ontology that describes the semantics of data by specifying constraints (rules) that the data should satisfy if it accurately represents a

clinical area. From the DQ Ontology, the term DomainConcept, refers to concepts that both the Domain and Task share. The Domain ontology for this study is described in Table 4.1. An example of a DomainConcept is `admission_date` with its associated constraint rule that the `admission_date` must be earlier than the `discharge_date`. The concept of Representation is defined as the lowest level, atomic piece of information that exists in a set of data (Dataset). Synonyms for Representation are data field, observation, and value. Aspects of data quality are called Measures. The research described in this paper looks at two important Measures defined in the DQ Ontology: RepresentationComplete and DomainConstraints. RepresentationComplete measures the degree to which data in a Dataset is not missing. DomainConstraints assesses how well the data conforms to the Domain ontology.

A process for computing quantities that characterize data quality has been developed[69]. The process evaluates the constraints that are defined for each Measure for each DomainConcept to compute a proportion. The denominator is the number of Representations for each DomainConcept for a population and the numerator is the number of Representations which have all constraints satisfied. The quantities that are produced are called MeasureResults. The DQ Ontology defines correctness as Measures that assess whether data represent the real world. Data are deemed correct by comparison to another Dataset to determine whether a Representation in the first Dataset is equal to its counterpart Representation in the other Dataset[56]. Ideally, the other Dataset would be a "gold standard" for comparison, but the best that is usually available is a relative gold standard[55].

Objective

The purpose of this study is to quantify the impact of two data quality issues, RepresentationCompleteness and DomainConstraints, on the correctness of an eMeasure (CMS178). Data quality issues were artificially created by systematically degrading the underlying quality of sample EHR data using two methods: independent and correlated degradation. A linear model that describes the change in the correctness of the eMeasure was developed to quantify the impact of data quality issues for each DomainConcept on the eMeasure.

4.3 Materials and Methods

Data were obtained from a clinical data repository (CDR) at the University of Minnesota. IRB approval was received to extract a de-identified 72,127 encounter random sample to be used as the data source for this study. A Domain ontology and Task ontology were developed and the underlying data quality of the sample was assessed using the DQ Ontology.

For this research, the CMS178 eMeasure was used as an example Task to illustrate the assessment process. The definition of CMS178 is "Urinary catheter removed on Postoperative Day 1 (POD 1) or Postoperative Day 2 (POD 2) with day of surgery being day zero"[71]. Patients that are catheterized for long periods of time are at greater risk for developing catheter-associated urinary tract infection (CAUTI). The best practice is to remove the catheter within 48 hours after surgery[73]. CMS178 calculates the proportion of patient encounters that satisfy this best practice. Previous research defined a Domain

ontology (Table 4.1) and a simplified CMS178 eMeasure that was also used for the current research[84].

CMS178 was simplified for this study by eliminating some of the denominator exclusions and including in the denominator all surgeries instead of just major surgeries. The denominator criteria includes all hospital patients (age 18 and older) that had surgery during the measurement period with a catheter in place postoperatively. The denominator exclusions are 1) patients who expired perioperatively, or 2) patients who had physician documentation of a reason for not removing the urinary catheter postoperatively, or 3) patients who had medications administered within 2 days of surgery that were diuretics, IV positive inotropic and vasopressor agents or paralytic agents. The numerator is the number of denominator surgical patients whose urinary catheter was removed within 48 hours.

The eMeasure is computed as:

$$CMS178_{simple} = \frac{Numerator}{Denominator - Denominator\_exclusions}$$

The steps to compute the CMS178 numerator and denominator for the baseline (undegraded) data are shown in Figure 4.2.

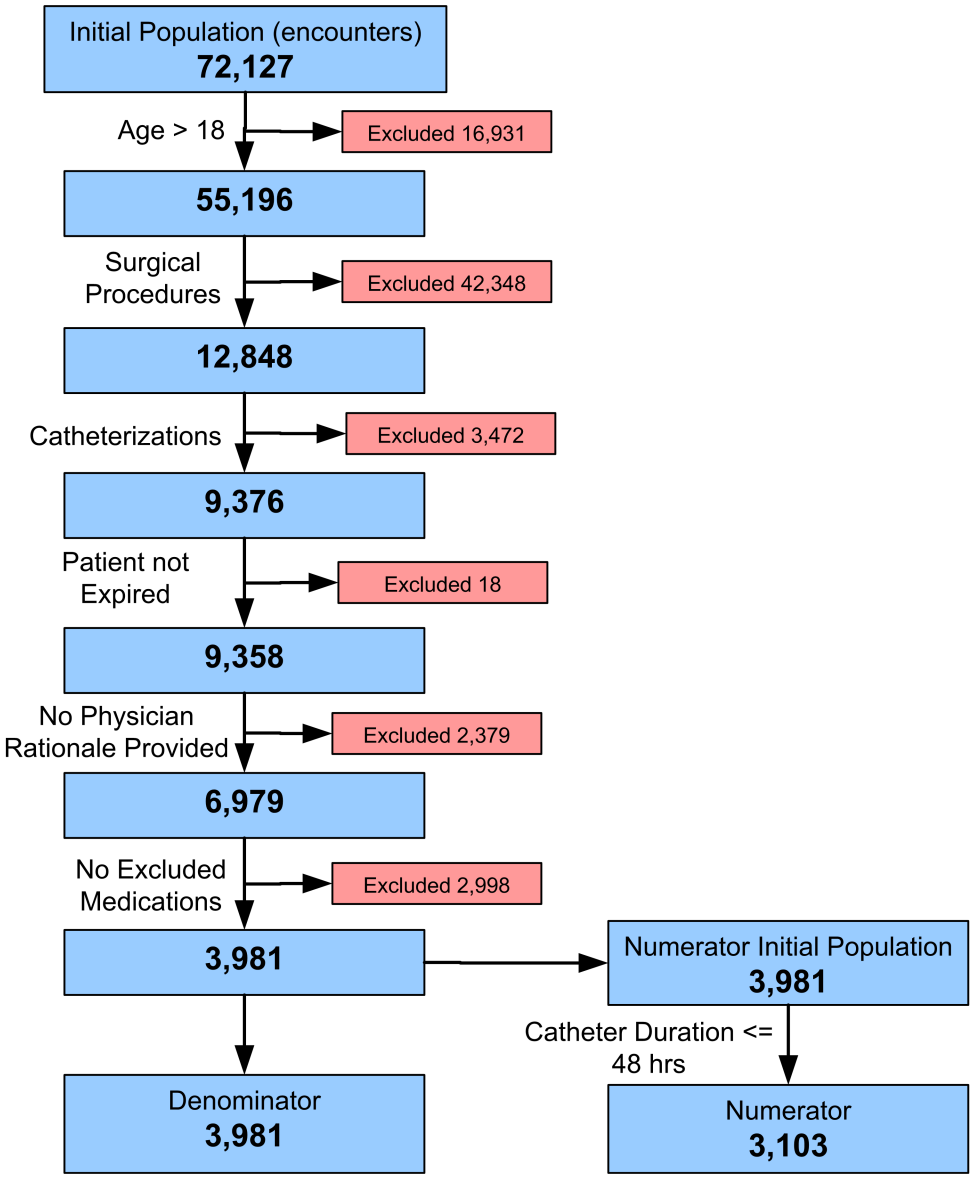

Figure 4.2. Computation of CMS178 Numerator and Denominator for Baseline Data (Undegraded)

Table 4.1. Domain Ontology with Constraints

DomainConcept	Type	DomainConstraint
Dataset		
Patient		
birth_date	date	birth_date <= today
death_date	date	if death_date is not null then death_date >= birth_date
Hospital Admission		
admission_date	date	discharge_date - admission_date < 1000
admission_type	code:choice_list	
discharge_date	date	admission_date <= discharge_date
Procedure		
procedure_concept_code	code:CPT4	
procedure_date	date	procedure_date >= admission_date
Medication		
medication_concept_code	code:RXNORM	
medication_end_date	date	medication_start_date < medication_end_date
medication_start_date	date	medication_start_date >= birth_date
Catheter Intervention		
catheter_duration	numeric	catheter_duration >= 0 catheter_duration < 1000
catheter_insertion_date	date	if catheter_insertion_date is not null then catheter_inserted_by is not null if catheter_insertion_date is not null and catheter_removal_date is null then catheter_rationale_for_continued_use is not null
catheter_removal_date	date	if catheter_removal_date is not null then catheter_insertion_date is not null
catheter_rationale_for_continued_use	string	if catheter_rationale_for_continued_use is not null then catheter_insertion_date is not null
catheter_inserted_by	string	if catheter_inserted_by is not null then catheter_insertion_date is not null

For this research, two data quality Measures, RepresentationComplete and DomainConstraints, were studied. These Measures were selected because they assess important aspects of data quality and they are fundamental components of other Measures. RepresentationComplete quantifies the extent of missing data. It is the proportion of Representations that have data values divided by the total number of Representations in the data. The DomainConstraints Measure quantifies how well the data satisfies all of the rules (constraints) defined in the Domain ontology. An example is that `death_date` must be after a patient's `birth_date` and the `death_date` DomainConstraint is the proportion of encounters in a population where that is true.

A data quality assessment process was developed to compute these Measures on EHR data repositories[84]. The research described in this paper extends that approach by quantifying the degree that data quality issues for each DomainConcept impact a Task. This was accomplished by deliberately changing the underlying EHR data in a systematic way and observing how those changes affected the use of the data. For this research, the Task was to compute the CMS178 eMeasure after each change. Data quality changes typically caused encounters to be removed from both the numerator and denominator of CMS178 and, since it is a ratio, CMS178 remained essentially unchanged.

In this study, correctness is a relative measure and will be operationalized by comparison to a relative gold standard. The baseline, unmodified sample CDR data will be used as the relative gold standard. A data element will be considered correct if it matches the baseline data. A new variable, `missing_events`, will be computed that quantifies the correctness of the CMS178 eMeasure after the data is modified. This variable represents the number of patients that had a catheter removed within 48 hours in

the baseline data but, after the data was degraded, were subsequently not counted as satisfying the CMS178 numerator criteria. These are events of interest that were therefore missing due to the data quality issues. Therefore, the `missing_events` variable is an indicator of correctness and quantifies the impact that data quality issues have on the ability to detect when catheters were removed within 48 hours of surgery.

Each type of data quality Measure has a specific method for degrading the data. For the RepresentationComplete Measure, missing data is simulated by removing Representations. For the DomainConstraint Measure, the data is degraded by changing the values of Representations to no longer satisfy the Domain constraints. For example, for `discharge_date`, the underlying data was changed to occur before the `admission_date` by a random number of days.

The full degradation process consists of iteratively applying the degradation method to the data for each of the DomainConcepts listed in Table 4.1. The Task is performed (in this case, computing CMS178 and `missing_events`) and the RepresentationComplete and DomainConstraint Measures were recomputed on the degraded data. The MeasureResults for RepresentationComplete and DomainConstraints for every DomainConcept, the CMS178 eMeasure, and `missing_events` were recorded in an analysis database.

Two approaches to degrading data were examined: 1) independent and 2) correlated. Each process was performed to yield 1,200 observations to build each of the linear models. To independently degrade each DomainConcept, a random set of 0% to 10% of records in the underlying data for each variable was degraded in a succession of 1% increments leaving the data for all other variables unchanged. The degradation

procedure either replaced a data value for a DomainConcept with a null value (to assess RepresentationComplete) or changed the value to something that would ensure the Domain constraints for that DomainConcept would be violated (to assess DomainConstraints).

The correlated approach to degrading data ensured that highly correlated DomainConcepts remain correlated. If each DomainConcept was arbitrarily degraded, it would not necessarily reflect how data quality impairments for related DomainConcepts would likely occur in the real world. For example, `catheter_insertion_date` and `catheter_inserted_by` are often missing together. If the reason they are missing is correlated (i.e. a clinician forgot or didn't have time to record the information before discharge), they would often be missing at the same time. Each DomainConcept was degraded from 0% to 10% leaving all other data unchanged unless the DomainConcept was part of a highly correlated cluster. In that case, the other DomainConcepts in the cluster would also be degraded by the same percentage.

The pairwise association between the presence of each of the DomainConcepts was computed at the encounter level. An encounter could have multiple instances of medication or catheter data associated with it. Data for each encounter was aggregated to indicate whether there was at least one data value for each of the DomainConcepts for the encounter. For example, consider the association between `admission_date` and `medication_start_date`. An encounter may have multiple `medications` (and therefore, multiple `medication_start_dates`). The association was computed between the presence of an `admission_date` and the presence of a `medication_start_date` for at least one of the medications associated with a

particular encounter. The Pearson correlation coefficient and a chi-square were calculated in a similar manner for the presence of data for each pair of DomainConcepts. Variables were considered highly correlated if they showed a significant chi-squared association and had a Pearson correlation coefficient above 0.90. The degradation process for correlated variables ensured that when one of the variables in a correlated cluster was degraded by a specific percent, the other variables in that cluster were also degraded by the same percentage.

A linear regression model was fit to `missing_events` as the dependent variable, with the data quality MeasureResults for RepresentationComplete and DomainComplete for each DomainConcept as the predictor variables. Feature selection was performed by computing an initial linear model using all of the predictor variables. Variables with a p-value ≤ 0.1 were then selected as predictors for a second linear model. The resulting coefficients from the second linear model are the final results. The regression model quantifies the effect of each DomainConcept. Negative changes (degradation) to the data increase `missing_events` and can be used to quantify what would happen if instead, data quality improved. If data in an EHR is of low quality (i.e. the degraded data) and a method existed to somehow improve it by fixing the data (assuming the incorrect data could be identified) then `missing_events` would be reduced.

4.4 Results

The data quality degradation process was performed first to assess how the RepresentationCompleteness data quality issue can affect `missing_events` and then

to assess the impact that the DomainConstraint issues had on `missing_events`. Each data quality issue was evaluated using the independent and correlated degradation methods. In order to perform the correlated degradation process, the pairwise Pearson correlation and chi-square association between all 12 DomainConcepts was computed. Three clusters of highly correlated variables were found:

Cluster 1: `admission_date, discharge_date`

Cluster 2: `medication_concept_code, medication_start_date, medication_end_date`

Cluster 3: `catheter_duration, catheter_insertion_date, catheter_removal_date`

4.4.1 Results for RepresentationComplete

The linear regression models for `missing_events` when degrading DomainConcepts by the independent and correlated degradation methods are shown in Table 4.2.

Predictor	Coefficient	SE Coefficient	t-value	p-value
Independent Degradation				
birth_date	-0.9996	0.0029	-345.66	< 0.0001
admission_type	-1.0039	0.0029	-347.13	< 0.0001
medication_start_date	-0.3765	0.0030	-123.84	< 0.0001
catheter_duration	-0.3281	0.0035	-93.87	< 0.0001
catheter_rationale_for_ continued_use	-0.1102	0.0029	-37.7	< 0.0001
Correlated Degradation				
birth_date	-0.9916	0.0037	-267.92	< 0.0001
admission_type	-0.9978	0.0037	-269.59	< 0.0001
medication_start_date	-0.3990	0.0026	-151.7	< 0.0001
catheter_removal_date	-0.3270	0.0029	-114.16	< 0.0001
catheter_rationale_for_ continued_use	-0.1070	0.0037	-28.59	< 0.0001

Table 4.2. Linear Models for Independent and Correlated Degrading of RepresentationComplete Measure

Degrading the data for RepresentationComplete removes data for a variable and causes the numerator or the denominator to change for the CMS178 eMeasure. The impact that each variable has on the value of the CMS178 eMeasure is proportional to the amount of relevant data removed. Table 4.3 shows the baseline number of encounters for the numerator and denominator and what those values are when 10% of the data is degraded for two example variables, `catheter_duration` and `birth_date`.

		10% Degrade	
	Baseline	**birth_date**	**catheter_duration**
Numerator	3103	2841	3208
Denominator	3981	3623	3981
missing_events	0	262	-105
CMS178	78%	78%	81%

Table 4.3. Example Impact of 10% Degradation vs. Baseline

4.4.2 Results for DomainConstraints

The linear regression model for `missing_events` when degrading DomainConstraints for the two methods are shown in Table 4.4.

Predictor	Coefficient	SE Coefficient	t-value	p-value
Independent Degradation				
death_date	-0.0294	0.0017	-17.568	< 0.0001
medication_end_date	-0.3546	0.0018	-195.244	< 0.0001
medication_start_date	-0.0223	0.0026	-8.481	< 0.0001
Correlated Degradation				
death_date	-0.0306	0.0019	-16.21	< 0.0001
medication_end_date	-0.3729	0.0012	-310	< 0.0001

Table 4.4. Linear Model for Independent and Correlated Degrading of DomainConstraint Measure

The CMS178 eMeasure was also computed as data quality was being degraded in order to show how it changed as the number of `missing_events` increased. A graph of CMS178 compared to the RepresentationComplete data quality Measure for the entire Dataset is shown in Figure 4.3.

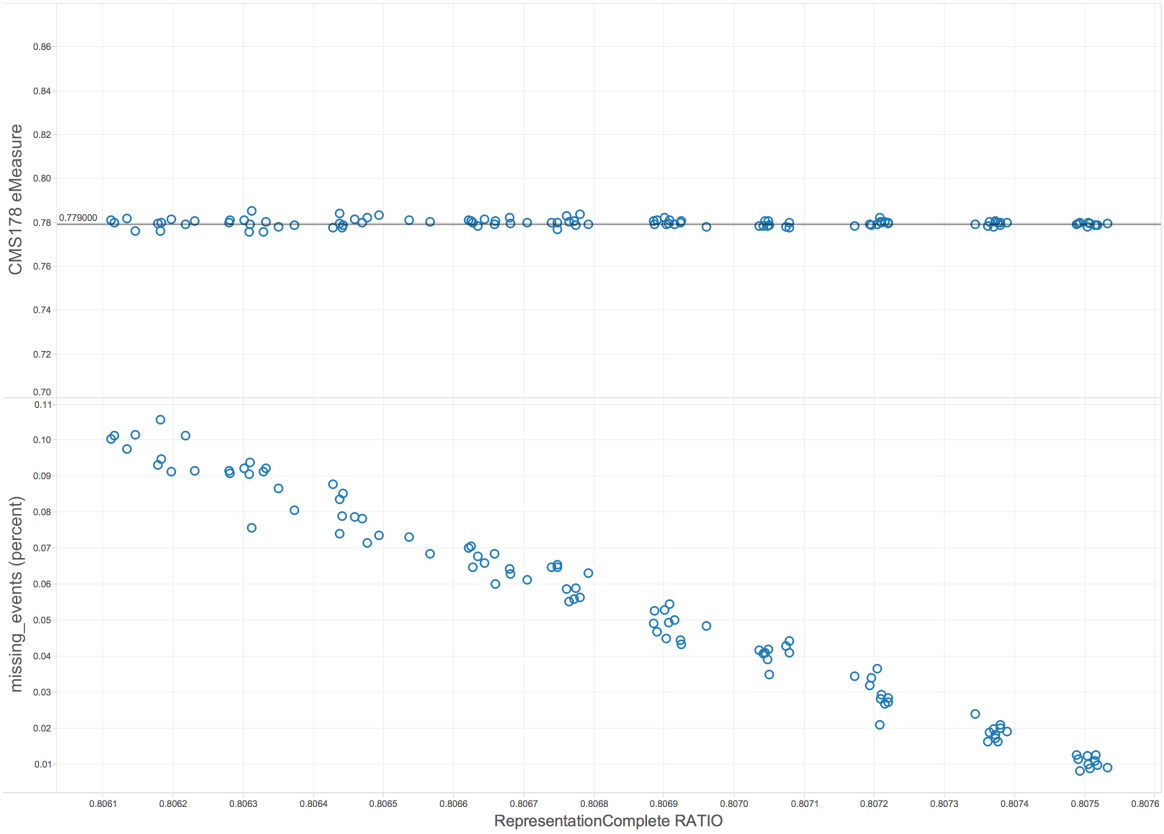

Figure 4.3. CMS178 eMeasure and `missing_events` vs. Dataset RepresentationComplete Ratio

CMS178 remains relatively constant when data quality improves, whereas missing events decreases as data quality improves.

4.5 Discussion

The purpose of this research was to demonstrate how two data quality issues, RepresentationCompleteness and DomainConstraints, can impact the correctness of an eMeasure (CMS178). Data quality issues were artificially introduced into the underlying data using two methods: independent and correlated degradation. A linear regression model quantified the effect that the data quality issues have on `missing_events` and a comparison to the CMS178 eMeasure was also shown. The results of this study clearly show that: 1) data quality issues (i.e. RepresentationCompleteness and DomainCompleteness) for different variables (i.e. `birth_date`, `catheter_duration`, etc.) impact the correctness of an eMeasure and the impact can be quantified and 2) the CMS178 eMeasure, as currently defined, may not measure how well an organization is meeting the best practice goal of removing catheters within 48 hours of surgery.

In support of the first finding, the impact of the data quality of each DomainConcept on the correctness of the eMeasure is reflected in the results of linear regression models. The coefficients in the models can be interpreted to quantify the magnitude of the impact on `missing_events` for a 1 unit improvement in a Measure. For example, for the independently degraded RepresentationComplete Measure in Table 4.2, for every 1% reduction in RepresentationComplete data quality for `admission_type` there were 1.0039% more events missed and the eMeasure is further from its correct value. But a 1% reduction in data quality for `catheter_rationale_for_continued_use` only results in 0.11% cases being missed.

Table 4.3 helps illustrate how different DomainConcepts have different impact on missing_events. The table shows denominator and numerator values for two example DomainConcepts: birth_date and catheter_duration. It shows that there were 262 missing_events when the birth_date field was degraded by 10%. CMS178 excludes patients who are younger than 18, so when the birth_date is removed, the encounter is no longer included in either the denominator or the numerator. In the study data, 100% of the encounters have a birth_date. Every encounter that is removed from the denominator (due to missing birth_date) will also be removed from the numerator so there is a one-for-one impact on the denominator. This is reflected in the coefficient of the linear model which is approximately equal to 1.0 for birth_date.

In the case of catheter_duration, degrading the data only affects the numerator (since it is not part of the inclusion or exclusion criteria) and a 10% degradation of the data causes 105 additional events. Additional events are added to the numerator because these are encounters that had durations of more than 48 hours that, after degradation, had a catheter_duration of 0. Therefore, there are 105 encounters that then meet the CMS178 criteria. For the study data, approximately 32% of the encounters were catheterized (22,744 of 72,127). Removing 10% of the catheter data (due to degradation) resulted in removing catheter_duration from approximately 3.2% of the encounters, or 0.32% for every 1% degradation of the data. The coefficient for catheter_duration in the Independent Degradation linear model is 0.33, which supports this interpretation.

It is surprising that degrading the data would cause encounters to be added to the numerator. This issue arises because it is known that a catheter was used on a patient (from other information like `catheter_inserted_by`), but if `insertion_date` is missing, then the `catheter_duration` can't be computed. There are two ways to handle the situation: 1) exclude the encounter because not all of the data needed to compute the measure is available or 2) set the `catheter_duration` to 0. The definition of CMS178 does not explicitly state that all data needed to compute the numerator must be present in order to include an encounter in the denominator, so it is left to the EMR vendors to decide how to implement the computation. For this research, the second approach was used, which is why the numerator increases. If the first approach had been chosen, then the numerator would not change because the encounter would be excluded from both the numerator and denominator. But that would mask the fact that there are catheterized patients where it is unknown if CMS178 is satisfied or not. In the DQ Ontology, there is a Measure called Sufficiency[69] that quantifies whether the data is sufficient to perform the Task. It may be beneficial to add to the inclusion criteria of CMS178 that the Sufficiency Measure of an encounter be 100% (i.e. that all data elements exist) but also report the overall Sufficiency Measure of the Dataset in order to quantify the number of encounters that were excluded due to data quality issues.

In the Independent model, `admission_type` and `birth_date` were the most impactful variables. For every 1% decrease in the RepresentationCompleteness of these variables (i.e. more missing data), there is approximately a 1% increase in the number of `missing_events`. Since `age` is used in the denominator (and numerator) inclusion criteria, when `age` can't be calculated because `birth_date` is missing the encounter is

removed from both the numerator and denominator. The eMeasure proportion stays the same, but `missing_events` increases. The same is true for `admission_type` since a non-surgical case is removed from both the numerator and denominator and the CMS178 eMeasure stays the same. On the other hand, when `catheter_duration` is degraded, the encounter is added to the numerator. In this case, only 0.33% of the events would be missed, but the CMS178 eMeasure would change by a small amount to reflect the change in the numerator. Since `admission_type` and `birth_date` have the largest impact on the number of missed encounters, any data quality initiatives should focus on improving the data quality of those items.

Degrading each DomainConcept independently compared to degrading in a correlated manner produced approximately the same set of variables that were most impactful. In the independent model, five variables were found to be significant in the model. These were `birth_date, admission_type, medication_start_date, catheter_duration` and `catheter_rationale_for_continued_use`. For the correlated data, four of the variables were the same and the fifth was `medication_end_date`, which is in the same cluster of highly correlated variables as the `medication_start_date`. Those variables are essentially equal from a RepresentationComplete point of view since they are likely missing (or present) at the same time. All of these variables, except `catheter_duration`, are part of the inclusion or exclusion criteria of the eMeasure and if they change, both the numerator and denominator are changed by the same amount. If only `catheter_duration` changes, that would only affect the numerator. But the effect was not that large; only 0.33% of the events would be missed for every 1%

change in RepresentationCompleteness of `catheter_duration`. Since Results of this study demonstrate that degrading each variable independently or in a correlated way made no difference to which variables where found to be significant and the magnitude of the impact (the coefficients). This is useful information in that it is less computationally expensive to degrade each variable independently versus having to degrade a variable and maintain all of its correlations.

Degrading the Domain constraints yielded a different set of variables that were impactful compared to RepresentationComplete. For independent degradation, the DomainConcepts of `death_date, medication_start_date` and `medication_end_date` were significant. But only `medication_end_date` had an appreciable impact. A 1% improvement in the `medication_end_date` reduces the missed cases by 0.38%, but the other DomainConcepts only change the difference by less than 0.1%. This has a smaller impact than RepresentationCompleteness. This is likely because this data is only altered to violate the constraints, it is not missing, so the calculation of CMS178 can still occur. But since it is using data that doesn't make clinical sense (i.e. `discharge_date` is before the `admission_date`), the computation, as currently defined, will blindly use the variable and compute a result. But as the low impact shows, most of the time, it doesn't change the count for which patients are included in the numerator. For example, since the eMeasure just looks at the `admit_date` to determine inclusion in the reporting period, even if the `discharge_date` is before the `admit_date`, the encounter will still be included.

The second finding is that the CMS178 eMeasure may not adequately measure catheter removal within 48 hours of surgery. As seen in Figure 4.3, even though

`missing_events` increases as the underlying data is degraded, CMS178 itself does not appreciably change. This is due to the fact that the eMeasure is a proportion. As the data is changed, it generally causes patient encounters to be removed from both the denominator and numerator. But the absolute number of missed events increases significantly over the range of the degradation. This highlights a potential problem with using CMS178 to assess catheterization best practices. The eMeasure is not affected by significant changes in data quality that generate missed events as it does not seem to be defined in such a way as to capture all of the best practice catheter removal events. The way the eMeasure is currently defined does not give CMS an accurate quantification of how well an organization is removing catheters within 48 hours of surgery for all such patients.

Understanding how data quality for each DomainConcept impacts the Task can be used to prioritize data quality improvement efforts. A healthcare organization can target data quality issues for DomainConcepts that have the most chance of improving eMeasure correctness. If data quality Measures are too low in a particular area, it may be advisable to not report the eMeasure or at least indicate the level of data quality when the eMeasure is reported.

4.5.1 Limitations and Future Work

There are some limitations to this research. The current research showed that degrading each DomainConcept independently produced about the same results as degrading the DomainConcepts in a correlated manner. This may not always be the case with other, more complex, Tasks. Only pairwise associations between DomainConcepts were examined. It is likely that degrading in a correlated manner is the best, most robust

approach. But degrading each DomainConcept independently has the fastest execution time. Further research is needed with additional Tasks to understand which degradation technique is most effective. Missing data in the real world is likely more complex than what can just be represented by pairwise associations of DomainConcepts. Future research should build complete correlation networks between all of the DomainConcepts so that the correlated degradation process can precisely maintain the correlations between all of the variables as the data quality is reduced.

Another limitation is that the RepresentationComplete Measure should be expanded to encompass different types of missing data. The definition of completeness is contextual and dependent on how data will be used[86]. RepresentationComplete should differentiate between data that is missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

This research did not attempt to quantify every type of data quality issue and only looked at two types of problems: RepresentationComplete and DomainConstraints. There are other types of data quality issues that should be explored. For example, an error in a date variable can occur in many ways. This research examined errors in dates that were large enough to cause a Domain constraint to be violated. But an error could occur that just affects the day of the month, which would not necessarily violate the Domain constraint. Other Measures are needed to quantify those types of errors. The impact of a Measure is also dependent on the specific Domain model that is defined as well as the amount of data that is degraded. This research modified up to 10% of the data, but further research is needed to determine the typical proportion of data errors in a CDR.

The results from this research showed that missing data (i.e. RepresentationCompleteness) has a greater impact on an eMeasure than the Domain constraint errors that were introduced into the data. Further research is needed to determine if this is the case for other Domains and Tasks. This study used the CMS178 eMeasure as an example Task to study in detail the process for assessing the impact data quality has on Task correctness. The technique can be generalized for other data analysis Tasks that depend on secondary use of EHR data such as predictive modeling and comparative effectiveness research. It is necessary to define a Task and Domain ontology with constraints, but the same data quality assessment approach can be used. Future research should evaluate this approach for other secondary uses.

4.6 Conclusion

Access to a significant amount of structured electronic health data allows researchers to identify evidence based best practices that improve patient outcomes. This secondary use of data is only warranted if the data is of sufficient quality to support the secondary use. eMeasures have been introduced as a method to assess how well evidence based practices are being followed at a healthcare organization. The research described in this paper quantified the impact of RepresentationComplete and DomainConstraint data quality issues on the correctness of an eMeasure and the assessment approach can be generalized for other data analysis Tasks. The research also showed that the CMS178 eMeasure, as it is currently defined, may not adequately assess how well an organization is removing catheters within 48 hours of surgery due to cases that are improperly excluded. The usefulness of characterizing data quality using these methods enables

healthcare organizations to prioritize data quality improvement efforts to focus on the areas that will have the most impact on correctness and assess whether the values that are being reported should be trusted.

Chapter 5: Conclusions

The purpose of this study was to demonstrate that a healthcare data quality framework can be developed that produces metrics that characterize underlying EHR data quality and it can be used to quantify the impact of data quality issues on the correctness of the intended use of the data. The framework described in this research successfully defined a Data Quality (DQ) Ontology and implemented an assessment method. The usefulness of this approach was illustrated by characterizing the data quality of EHR data and then quantifying the impact of data quality issues on the correctness of the CMS178 eMeasure.

Research detailed in the first paper produced a DQ Ontology that serves as a foundation for describing aspects of data quality. The DQ Ontology was developed by mining the healthcare data quality literature for important terms used to discuss data quality concepts and these terms were harmonized into an ontology. Four high-level data quality dimensions (CorrectnessMeasure, ConsistencyMeasure, CompletenessMeasure and CurrencyMeasure) categorized 19 lower level Measures. The ontology serves as an unambiguous vocabulary and allows more precision when discussing healthcare data quality. The terms from the DQ Ontology were used throughout this research to more easily refer to very specific aspects of data quality. The DQ Ontology is expressed with sufficient rigor that it can be used for logical inference and computation.

Current literature on healthcare data quality defines terms for data quality concepts using textual descriptions or sometimes doesn't define them and leaves it up to the reader to infer which aspect of data quality to which a term refers[69]. Different authors often use the same term to refer to different concepts. As an extreme example, the term

"accuracy" is used in the literature to refer to different concepts including how correct the data is (RelativeCorrectness)[23], whether it is missing or not (RelativeCompleteness)[24], how consistent it is with a domain model (DomainConsistency)[23] and whether or not it is corrupted (RepresentationIntegrity)[16]. To help resolve this ambiguity, the DQ Ontology gives each concept a name, defines key attributes for each concept and specifies how concepts are related to one another (see Figure 2.1). If researchers agree to use a common terminology like the DQ Ontology when discussing data quality, it will make it easier to describe and understand issues. As an example, the field of software development adopted the Design Patterns ontology for discussing software architecture and it improved developers' ability to discuss software design issues[45]. At first, developers weren't necessarily familiar with the terms used for each software concept, but over time the terms were adopted and they now stand for specific aspects of good software architecture[87]. The same can happen for discussions of healthcare data quality.

Each concept in an ontology is defined with a textual description but, more importantly, it has relationships to other concepts, attributes and constraints that more precisely define it. The DQ Ontology specifies concepts concerning data quality and it refers to a separate Domain and Task ontology. A Domain ontology defines the formal semantics (using attributes and constraints) of concepts represented in the data. A Task ontology is a specification for the concepts necessary to carry out a particular use of the data. The use of the constraints defined in the DQ, Domain and Task ontologies is one of the key contributions of this research. The constraints precisely define concepts better than using textual definitions and can be used directly by software to quantify data quality. The assessment method described in this research computes Measures which

quantify how well the data conforms to a Domain and how well it fits a Task. The method produces MeasureResults that quantify the proportion of data that satisfy the constraints for a Measure or DomainConcept.

The DQ Ontology and assessment method were used in the second paper to characterize data quality from an EHR. A 72,127 encounter de-identified random sample of EHR data was assessed for 10 data quality Measures. Domain and Task ontologies were developed which included constraint definitions. MeasureResults for the assessment were detailed in Table 3.4. The results demonstrate that data quality can be quantified and Metrics can track data quality trends over time and for specific DomainConcepts.

Another key contribution of this research is that the quantities (MeasureResults) produced by the constraint based assessment method are easier to use and understand than some of the existing, rule based, approaches to data quality assessment[84]. The DQ framework produces scalar quantities which can be computed on individual DomainConcepts and can be meaningfully aggregated at different levels of an information model. The DomainConcepts in a Domain model usually form a hierarchy of concepts. For example, `Patients` have `Encounters` which have associated `Observations`. The research detailed how MeasureResults can be computed for the lowest level DomainConcepts in the hierarchy (`Observations`), at higher levels (like the `Encounter` or `Patient` level) and also for the Dataset as a whole.

The third paper described enhancements to the data quality assessment process to systematically degrade the EHR data and record the impact of data quality issues on a Task (CMS178 – Urinary Catheter Removal after Surgery). A linear regression model that used the DomainConcept Measures as independent variables quantified the relative

impact of DomainConcepts on a Task (in this case, the CMS178 eMeasure). The un-degraded EHR data was used as a relative gold standard and served as a baseline to measure correctness and quantify how data quality issues affect a Task. This information can help healthcare organizations prioritize data quality improvement efforts to focus on the areas that are most important and determine if the data can support its intended use.

This data degradation technique was also helpful in demonstrating when a Task might not measure what it was intended to measure. Results of this research demonstrate that the CMS178 eMeasure, as it is currently defined, may not adequately assess how well an organization is removing catheters within 48 hours of surgery. Different choices for how to handle data quality issues may bias the eMeasure. The assessment showed that even when the data had significant data quality issues that caused missing catheterization events, the CMS178 eMeasure didn't reflect the change. The current definition of the eMeasure may not necessarily give CMS an unbiased measure of how well an organization is removing catheters within 48 hours of surgery for all such patients.

The data degradation and impact assessment technique is generalizable for assessing other secondary uses of the data. For example, assessing the impact of data quality on a predictive model or for comparative effectiveness research can quantify the relative importance of variables used in the model or research. As the underlying data is degraded, the impact of the data quality issues can be determined using a linear regression model just as it was in the case of the eMeasure. Decisions for how trustworthy and correct the results are can be based on Measurements of data quality for the Datasets used for the models or in the clinical research. Thresholds can be established for what level of data quality is required for these Tasks.

5.1 Limitations

There were a number of limitations in this research. The first is that the DQ Ontology only incorporated concepts that occurred in at least three of the meta-analyses papers. Important data quality Measures such as non-duplication of data, accessibility and privacy need to be incorporated into the ontology. Also, the Measures for missing data (RepresentationComplete and DomainComplete) are too simplistic. They need to account for data missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR)[86].

The aggregation and combining of multiple Measures is a useful aspect of this assessment method, but the approach presented in this research uses a simple method of averaging multiple Measures to create a single quantity. For example, TaskRelevance is the average of DomainConsistency and TaskSufficiency. Further research is needed to determine if each Measure should carry equal weight or if some of the Measures should have a larger impact in the combined Measure.

Another limitation of this research is that the method for degrading data only simulated two types of data quality issues. To assess the impact of the data quality of a DomainConcept, data was degraded to simulate missing data or data that doesn't satisfy a constraint. But there are other types of data quality issues that weren't simulated. Additional research should explore other types of issues such as data transformation errors and typos during data entry. Also, in the data degradation simulations, up to 10% of the data was degraded. More research is needed to determine what a typical error rate is for each type of Measure.

This research was carried out on a single health organization's EHR data. A relatively simple Task of computing the CMS178 eMeasure was used to illustrate data quality characterization and impact analysis. Further research is needed to confirm that this data quality assessment approach holds for other organization's EHR data and for more complex Tasks such as predictive modeling and comparative effectiveness research.

5.2 Future Directions

This research demonstrated a method for computing data quality Metrics. Metrics for one Dataset can be compared to Metrics from other Datasets that use the same Domain ontology and different Tasks can reuse those Metrics without having to re-compute them. This could encourage the development of Domain ontologies that can be reused across different organizations' EHR data. Organizations that have expertise and interest in a clinical area could take responsibility for developing and maintaining a Domain ontology that would be useful not only for data quality assessment but also for other purposes such as data exchange and clinical quality research. Recent workgroups such as Fast Healthcare Interoperability Resources (FHIR)[88] and the Clinical Information Modeling Initiative (CIMI)[89] have started to develop domain specific information models. These could become Domain ontologies by explicitly defining constraints and relationships between concepts. As these initiatives progress, the information models could include definitions of acceptable data quality for important Measures. The DQ Ontology and assessment framework can help model aspects of data quality and be incorporated into the information models.

Another future direction to explore is to use the DQ Ontology to assess data as it is added to the EHR. Data quality Measures and Metrics could be computed and delivered in real-time to interested stakeholders so that data quality is continually assessed and monitored. Data quality issues and trends may be more quickly identified and remediated.

The DQ Ontology and assessment framework provides a practical mechanism for assessing EHR data quality and for determining the impact of data quality issues on the intended use of a healthcare Dataset. Data quality assessment can be automated to encourage the sharing of data quality Metrics which could help make research results more transparent and reproducible. Disclosure of data quality Measures and Metrics could become a standard practice in research. Standards should be developed for specifying data quality Measures. Standards for saving data quality Metrics as metadata along with the EHR data are starting to be developed[90]. As data is moved from place to place or transformed through various processes within healthcare systems, data quality Metrics should be recomputed and saved (with history) so that data consumers are aware of the underlying data quality. And when EHR data is used in research or for other secondary uses, the data quality Metrics could be summarized in a report (possibly as a standard appendix) to reassure readers that the data used to reach a conclusion was fit for that purpose.

Bibliography

1. Office of the National Coordinator for Health Information Technology. Office-based Physician Electronic Health Record Adoption: 2004-2014, Health IT Quick-Stat #50. http://dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php. Accessed January 15, 2016.

2. Hsiao C, Hing E. *Use and Characteristics of Electronic Health Record Systems Among Office-Based Physician Practices: United States, 2001-2013*.; 2014.

3. Ancker JS, Shih S, Singh MP, et al. Root Causes Underlying Challenges to Secondary Use of Data. *AMIA Symp Proc*. 2011:57-62.

4. Data Quality Collaborative. *DQC White Paper Draft 1: A Consensus-Based Data Quality Reporting Framework for Observational Healthcare Data*.; 2013.

5. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. *Med Care Res Rev*. 2009;66(6):611-638.

6. Chan KS, Fowles JB, Weiner JP. Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature. *Med Care Res Rev*. 2010;67(5):503-527. doi:10.1177/1077558709359007.

7. Floor-Schreudering A, De Smet P, Buurma H, Egberts A, Bouvy M. Documentation quality in community pharmacy: completeness of electronic patient records after patients' first visits. *Ann Pharmacother*. 2009;43(11):1787-1794. doi:10.1345/aph.1M242.

8. Kohn LT, Corrigan JM, Donaldson MS, eds. *To Err Is Human: Building a Safer Health System*. Washington, DC: National Academy Press; 2000.

9. Cornish PL, Knowles SR, Marchesano R, et al. Unintended medication discrepancies at the time of hospital admission. *Arch Intern Med*. 2005;165(4):424-429. doi:10.1001/archinte.165.4.424.

10. Parsons A, McCullough C, Wang J, Shih S. Validity of electronic health record-derived quality measurement for performance monitoring. *J Am Med Inform*

Assoc. 2012;19(4):604-609. doi:10.1136/amiajnl-2011-000557.

11. Wang RY, Strong DM. Beyond Accuracy : What Data Quality Means to Data Consumers. *J Manag Inf Syst*. 1996;12(4):5-33.

12. Juran JM, Godfrey AB. *Juran's Quality Control Handbook*. 5th ed. New York: McGraw-Hill; 1999.

13. The International Standards Organization (ISO). 8402-1986 Quality Vocabulary. 1986.

14. Orfanidis L, Bamidis PD, Eaglestone B. Data Quality Issues in Electronic Health Records: An Adaptation Framework for the Greek Health System. *Health Informatics J*. 2004;10(1):23-36. doi:10.1177/1460458204040665.

15. Orr K. Data Quality and Systems Theory. *Commun ACM*. 1998;41(2):66-71.

16. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. ANALYTIC METHODS: A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health. *Med Care*. 2012;50(7):21-29.

17. Strong DM, Lee YW, Wang RY. Data quality in context. *Commun ACM*. 1997;40(5):103-110.

18. Almutiry O, Wills G, Crowder R. *Toward a Framework for Data Quality in Electronic Health Record*.; 2013.

19. Tayi G, Ballou D. Examining data quality. *Commun ACM*. 1998;41(2):54-57.

20. Wand Y, Wang RY. Anchoring Data Quality Dimensions in Ontological Foundations. *Commun ACM*. 1996;39(11):86-95.

21. Leitheiser RL. Data Quality in Health Care Data Warehouse Environments. *Proc 34th Hawaii Int Conf Syst Sci*. 2001:1-10.

22. Canadian Institute of Health. *The CIHI Data Quality Framework*.; 2009.

23. Liaw S, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. *Int J Med Inform*. 2013;82(1):10-24.

24. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Informatics Assoc*. 2012:2-8.

25. Centers for Medicare & Medicaid Services. The CMS EHR Incentive Programs : Small-Practice Providers and Clinical Quality Measures. 2011. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/CQM_Webinar_10-25-2011.pdf.

26. Conway PH, Mostashari F, Clancy C. The Future of Quality Measurement for Improvement and Accountability. *J Am Med Assoc*. 2013;309(21):2215-2216.

27. Torda P, Tinoco A. Achieving the Promise of Electronic Health Record-enabled Quality Measurement: a Measure Developer's Perspective. *eGEMs (Generating Evidence and Methods to Improve patient outcomes)*. 2013;1(2). doi:10.13063/2327-9214.1031.

28. Amster a., Jentzsch J, Pasupuleti H, Subramanian KG. Completeness, accuracy, and computability of National Quality Forum-specified eMeasures. *J Am Med Informatics Assoc*. 2014:409-416. doi:10.1136/amiajnl-2014-002865.

29. Kern L, Malhotra S, Barron Y, et al. Accuracy of Electronically Reported "Meaningful Use" Clinical Quality Measures. *Ann Intern Med*. 2013;158:77-83.

30. Fowles JB, Kind EA, Awwad S, et al. PERFORMANCE MEASURES USING ELECTRONIC HEALTH RECORDS : FIVE CASE STUDIES. *Commonw Fund*. 2008;(1132).

31. National Quality Forum. *Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties*.; 2011.

32. Dentler K, Cornet R, ten Teije A, et al. Influence of data quality on computed Dutch hospital quality indicators: a case study in colorectal cancer surgery. *BMC Med Inform Decis Mak*. 2014;14:32. doi:10.1186/1472-6947-14-32.

33. Roth CP, Lim Y-W, Pevnick JM, Asch SM, McGlynn E. The Challenge of Measuring Quality of Care From the Electronic Health Record. *Am J Med Qual*. 2009;24(5):385-394. doi:10.1177/1062860609336627.

34. Weiner M, Embi P. Toward Reuse of Clinical Data for Research and Quality Improvement : The End of the Beginning ? *Ann Intern Med*. 2009;151(5):359-360.

35. Green SM. Congruence of Disposition After Emergency Department Intubation in the National Hospital Ambulatory Medical Care Survey. *Ann Emerg Med*. 2012.

36. Schuur J, Tibbetts S, Pines J. Pregnancy testing in women of reproductive age in US emergency departments, 2002 to 2006: assessment of a national quality measure. *Ann Emerg Med*. 2010;55(5):449-457.e2. doi:10.1016/j.annemergmed.2009.08.017.

37. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. *Med Care*. 2013;51(8 Suppl 3):80-86. doi:10.1097/MLR.0b013e31829b1d48.

38. King J, Patel V, Furukawa MF. *Physician Adoption of Electronic Health Record Technology to Meet Meaningful Use Objectives: 2009-2012*.; 2012.

39. Studer R, Benjamins R, Fensel D. Knowledge engineering: Principles and methods. *Data Knowl Eng*. 1998;25(1-2):161-198.

40. Staab S, Studer R. *Handbook on Ontologies*. Springer; 2010.

41. Horrocks I. What Are Ontologies Good For ? *Evol Semant Syst*. 2013:175-188.

42. Heard S. openEHR Clinical Knowledge Manager. *openEHR*. 2015. http://openehr.org/ckm/. Accessed January 1, 2015.

43. National Quality Forum. *Quality Data Model December 2013*.; 2013.

44. Noy NF, McGuinness DL. *Ontology Development 101 : A Guide to Creating Your First Ontology*.; 2001.

45. Gamma E, Helm R, Johnson R, Vlissides J. *Design Patterns: Elements of Reusable Object-Oriented Software*. Pearson Education; 1994.

46. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health*. 2014;11:5170-5207. doi:10.3390/ijerph110505170.

47. Lima CRDA, Schramm JMDA, Coeli CM, Silva MEM Da. Revisão das dimensões de qualidade dos dados e métodos aplicados na avaliação dos sistemas de informação em saúde. *Cad Saúde Públ*. 2009;25(10):2095-2109.

48. Stvilia B, Gasser L, Twidale MB, Smith LC. A Framework for Information Quality Assessment. *J Am Soc Info Sci Tech*. 2007;58(12):1720-1733. doi:10.1002/asi.

49. Bertoa M, Vallecillo A. An Ontology for Software Measurement. In: Calero C, Ruiz F, Piattini M, eds. *Ontologies for Software Engineering and Software Technology*. Heidelberg: Springer; 2006:175-196.

50. Pipino LL, Lee YW, Wang RY. Data quality assessment. *Commun ACM*. 2002;45(4):211. doi:10.1145/505248.506010.

51. Fox C, Levitin A, Redman T. The notion of data and its quality dimensions. *Inf Process Manag*. 1994;30(1):9-19.

52. W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview. 2012. http://www.w3.org/TR/owl2-overview/.

53. Object Management Group. Ontology Definition Metamodel. 2014. http://www.omg.org/spec/ODM/1.1/.

54. Beale ET, Heard S. Architecture Overview. 2008:1-79.

55. Kahn MG, Eliason BB, Bathurst J. Quantifying clinical data quality using relative gold standards. *AMIA Annu Symp Proc*. 2010;2010:356-360.

56. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. *J Am Med Inform Assoc*. 1997;4(5):342-355.

57. Kayyali B, Knott D, Kuiken S Van. The big-data revolution in US health care : Accelerating value and innovation. 2013;(April):1-6.

58. Blumenthal D, Tavenner M. The "Meaningful Use" Regulation for Electronic Health Records. *N Engl J Med*. 2010;363(6):501-504. doi:10.1056/NEJMp1006114.

59. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *J Am Med Informatics Assoc*. 2007;14(1):1. doi:10.1197/jamia.M2273.Introduction.

60. Holve E, Segal C, Hamilton Lopez M. Opportunities and Challenges for Comparative Effectiveness Research (CER) With Electronic Clinical Data. *Med Care*. 2012;50(7):S11-S18. doi:10.1097/MLR.0b013e318258530f.

61. Mirnezami R, Nicholson J, Darzi A. Preparing for Precision Medicine. *N Engl J Med*. 2012;366(6):489-491. doi:10.1056/NEJMp1114866.

62. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51(8):S30-S37. doi:10.1097/MLR.0b013e31829b1dbd.

63. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *AMIA Proc from Summit Transl Sci*. 2010;2010:1-5.

64. Arts D, Keizer N, Scheffer G-J. Defining and Improving Data Quality in Medical Registries : A Literature Review , Case Study , and Generic Framework. *J Am Med Inf Assoc*. 2002;9:600-611. doi:10.1197/jamia.M1087.

65. Observational Medical Outcomes Partnership (OMOP). http://omop.org/. Accessed July 15, 2015.

66. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Informatics Assoc*. 2012;19:54-60. doi:10.1136/amiajnl-2011-000376.

67. Platt R, Carnahan R, Brown J, et al. The U.S. Food and Drug Administration's

Mini-Sentinel Program. *Pharmacoepidemiol Drug Saf*. 2012;21(1):1-8.

68. Kahn MG, Brown JS, Davidson BN, et al. Transparent Reporting of Data Quality in Distributed Data Networks. *eGEMs (Generating Evidence and Methods to Improve patient outcomes)*. 2015;3(1).

69. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. A Data Quality Ontology for the Secondary Use of EHR Data. In: *AMIA 2015 Annual Symposium Proceedings*. American Medical Informatics Association; 2015:1937-1946.

70. Centers for Medicare & Medicaid Services. Proposed Clinical Quality Measures for 2014 CMS EHR Incentive Programs for Eligible Hospitals & CAHs. 2014:1-20. http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Eligible_Hospital_Information.html. Accessed August 1, 2015.

71. CMS Clinical Quality eMeasure Logic and Implementation Guidance v1.3. 2014. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/2014_eCQM_Measure_Logic_Guidancev13_April2013.pdf. Accessed August 1, 2015.

72. Centers for Medicare & Medicaid Services (CMS). Urinary catheter removed on Postoperative Day 1 (POD 1) or Postoperative Day 2 (POD 2) with day of surgery being day zero. 2014. https://ecqi.healthit.gov/system/files/ecqm/2014/EH/measures/CMS178v5_1.html#toc. Accessed August 1, 2015.

73. Stéphan F, Sax H, Wachsmuth M, Hoffmeyer P, Clergue F, Pittet D. Reduction of urinary tract infection and antibiotic use after surgery: a controlled, prospective, before-after intervention study. *Clin Infect Dis*. 2006;42(11):1544-1551. doi:10.1086/503837.

74. Pipino LL, Lee YW, Wang RY. Data quality assessment. *Commun ACM*. 2002;45(4):211. doi:10.1145/505248.506010.

75. Heinrich B, Kaiser M, Klier M. How to Measure Data Quality? A Metric Based Approach. 2007:1-15. http://epub.uni-regensburg.de/23633/1/heinrich.pdf. Accessed August 1, 2015.

76. Blumenthal D. Launching HITECH. *N Engl J Med*. 2010;362(5):382-385.

doi:10.1056/NEJMp0912825.

77. Zerhouni E. Translational and Clinical Science — Time for a New Vision. *N Engl J Med*. 2005;353(15):1621-1623. doi:10.1056/NEJMsb053723.

78. Richesson RL, Rusincovitch SA, Simon GE. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research. (919):1-26.

79. (AHRQ) Agency for Healthcare Research and Quality. Measures Inventory. 2015. http://www.qualitymeasures.ahrq.gov/hhs/matrix.aspx. Accessed November 3, 2015.

80. (AHRQ) Agency for Healthcare Research and Quality. Clinical Quality Measures. 2015. https://ushik.ahrq.gov/QualityMeasuresListing?&system=mu&filterLetter=&resultsPerPage=50&filterPage=2&sortField=570&sortDirection=ascending&stage=Stage 2&filter590=April 2014 EH&filter590=July 2014 EP&enableAsynchronousLoading=true. Accessed November 3, 2015.

81. Centers for Medicare & Medicaid Services (CMS). EHR Incentive Programs: 2015 through 2017 Overview. 2015. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/Stage3Overview2015_2017.pdf. Accessed November 8, 2015.

82. Persell S, Wright J, Kmetik K, Baker D, Thompson J. Assessing the validity of national quality measures for coronary artery disease using an electronic health record. *Arch Intern Med*. 2006;166(20):2272-2277. doi:10.1001/archinte.166.20.2272.

83. HealthCatalyst. The Unintended Consequences of Electronic Clinical Quality Measures. 2015. https://www.healthcatalyst.com/electronic-clinical-quality-measures-impact-data-quality. Accessed November 8, 2015.

84. Johnson SG, Speedie SM, Simon G, Kumar V, Westra BL. Application of an Ontology for Characterizing Data Quality for a Secondary Use of EHR Data. *J Appl Clin Informatics*. 2016;7:69-88.

85. Hasan S, Padman R. Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach. *AMIA Annu Symp Proc*. 2006:324-328.

86. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46(5):830-836. doi:10.1016/j.jbi.2013.06.010.

87. Beck K, Crocker R, Meszaros G, et al. Industrial experience with design patterns. *Softw Eng 1996 Proc 18th Int Conf*. 1996:103-114. doi:10.1109/ICSE.1996.493406.

88. Health Level Seven (HL7). Welcome to FHIR. 2015. https://www.hl7.org/fhir/. Accessed February 15, 2016.

89. Health Level Seven (HL7). Clinical Information Modeling Initiative. 2016. http://www.hl7.org/Special/Committees/cimi/index.cfm. Accessed January 2, 2016.

90. Data Quality Collaborative. Data Quality Common Data Model. 2015;(March). http://repository.academyhealth.org/dqc/3.