A Data Quality Ontology for the Secondary Use of EHR Data

Steven G. Johnson, MS[1]; Stuart Speedie, PhD, FACMI[1]; Gyorgy Simon, PhD[1]; Vipin Kumar, PhD[2]; Bonnie L. Westra, PhD, RN, FAAN, FACMI[1,3]

[1]University of Minnesota, Institute for Health Informatics; [2]University of Minnesota, Department of Computer Science; [3]University of Minnesota, School of Nursing

Abstract

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.

Introduction and Background

The healthcare system in the United States continues to adopt electronic health records (EHR) at a rapid pace.[1] 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.[2] 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.[3] In order to measure the quality of data there must be an understanding of how the data will be used.[4]

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."[5(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.[6] 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.[7] 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.[8] 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.[9] 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[10] proposed a framework that consolidates 118 different general data quality characteristics into 20 categories. Kahn[11] proposed a healthcare specific framework using a "fit-for-use" data quality model in which he proposes five high-level dimensions. Liaw[6] 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".[12] In a 2012 paper, Weiskopf[13] defined five high-level dimensions of data quality and listed synonyms for each (Table 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 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[14] and tasks[15] 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.

Materials and Methods

There are a number of methodologies for developing an ontology,[8] but the method described by Noy and McGuiness[16] 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.[17] 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.[14,15]

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[6], Weiskopf[13], Kahn[11], Chen[18], and Lima[19]. 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[10], Wand[20], Chan[21], CIHI[22], Stvilia[23]. 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[24–26].

Ontologies can be specified using a number of methods including OWL[27], first order logic, or as UML[28]. 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.[8,29] For this study, pseudo-code was chosen because it succinctly captures the important aspects of the constraint without introducing a specific, complex syntax.

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 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, 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[24].

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[13], 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.

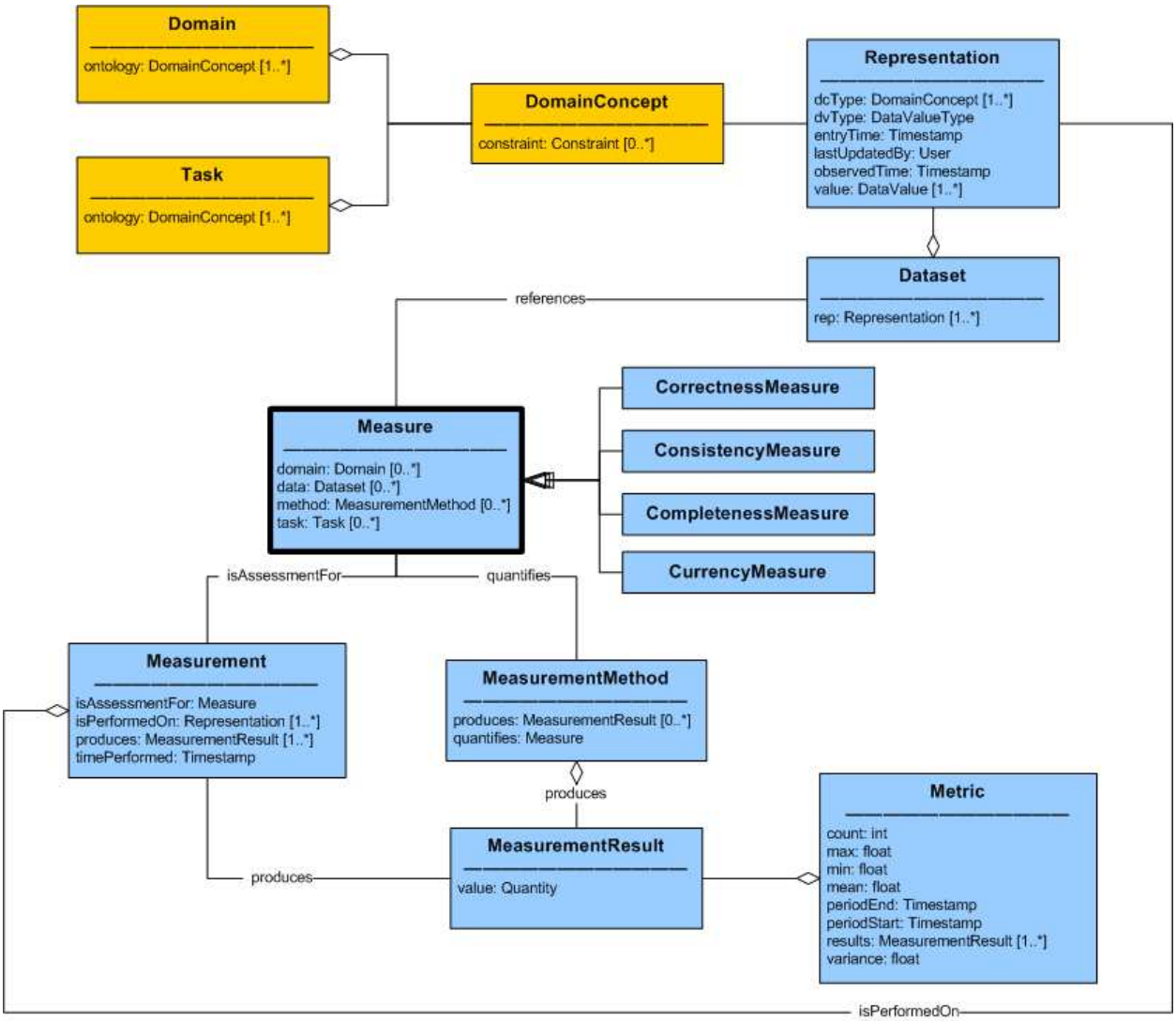

Figure 1: Data Quality Ontology

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[6], Accuracy: …free of error[11], Integrity[18], Repeatability[18], Structural Consistency[23]
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[6], Correctness[13], Believability[11], Validity[13,19], Comparability[20,21], Accuracy[10,13,18,23], Corrections made[13], Errors[13], Misleading[13], PPV[13], Quality[13]
RepresentationCorrectness	A correct Representation has high accuracy and is complete.	Correctness: …accuracy and completeness[6], Accuracy[20,21]
Reliability	The data is correct and suitable for the Task.	Reliability[6,18–20], 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[6], Concordance[13], Format[11], Internal Consistency[18], Consistency[13], Precision[20], Format[11,20], Reliability[13], Variation[13], Accuracy: Edit and Imputation[22], Representational Consistency[10]
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[6], Correctness: …format and types are valid[6], Plausibility[13], Believability[10,13], Relational Integrity Rules[11], Consistency[18–20], Measure validity[21], Accuracy[13], Trustworthiness[13], Validity[13,23]
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[6], Valid values[11], Comparability: Equivalency[22], Semantic Consistency[23]
DomainMetadata	Meta-data exists to describe the Domain and it is logically consistent.	Methodological Clarity[19], Metadata Documentation[18], Comparability: Data dictionary standards[22], Interpretability[10]
CompletenessMeasure		
RepresentationComplete	Domain independent extent to which data is not missing.	Completeness: …information is not missing[6], Completion[19], Completeness[18,21], 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[13], Optionality[11], 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?[13], Accessibility[10,13,19], Accuracy[13], Availability[13], Missingness[13], Omission[13], Presence[13], Quality[13], Rate of Recording[13], Sensitivity[13], Validity[13]
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[6], Appropriate amount of data[11], Representativeness[18], Sufficiency[20], Accuracy: Coverage[22], Granularity[11,18], Continuity[11], Level of Detail[20], Completeness[10,23], Precision[23]
DomainCoverage	The data can represent the values and concepts required by the Domain.	Completeness: …represent every meaningful state of the […] real world[6], Completeness: All values for a variable are recorded[6], Coverage[19], Completeness[20]
TaskCoverage	The data contains all of the information required by the Task.	Completeness: …depict every possible state of the task[6], Usableness[18,20], Usability[18], Utility[18], Importance[20], Usefulness[20], Value-added[10]
Flexibility	The extent to which the data is sufficient to be used by many Tasks.	Consistency: …information…appl[ies] to different tasks[6], Flexibility[10,20], Relevance: Adaptability[22]
Relevance	The data is sufficient for the Task and conforms to the Domain.	Relevance[6,18,20,23], Relevance: Value[22], Relevancy[10]
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[6], Currency[13,18,23], Timeliness[13,18,20], Up-datedness[18], Recency[13]
DatasetCurrent	Time difference between when a Dataset was updated and when it was made available. For example, periodic updates to a repository.	Timeliness: …availability of output is on time[6], Opportunity[19], Periodicity[18], Currency[11,20], Timeliness: Data currency[22], Timeliness[10]
TaskCurrency	The Data is sufficiently up-to-date for the requirements of the Task.	Timeliness: …information is up to date for task[6], Timeliness: …age of the data is appropriate for the task[11], Timeliness (external)[20]

Table 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 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 4) with respect to the task of calculating an eMeasure. An eMeasure[30] 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 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[14]):

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)[15].

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 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 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 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 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 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**.

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[6] 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[13] 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.[31] Comparing one dataset to another to yield a positive predictive value (PPV) and sensitivity measure are a useful way to characterize the data.[32] 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**.

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.

References

1. King J, Patel V, Furukawa MF. *Physician Adoption of Electronic Health Record Technology to Meet Meaningful Use Objectives: 2009-2012*.; 2012.
2. Ancker JS, Shih S, Singh MP, et al. Root Causes Underlying Challenges to Secondary Use of Data. *AMIA Symp Proc*. 2011:57-62.
3. 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.
4. Orr K. Data Quality and Systems Theory. *Commun ACM*. 1998;41(2):66-71.
5. Juran JM, Godfrey AB. *Juran's Quality Control Handbook*. 5th ed. New York: McGraw-Hill; 1999.
6. 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.
7. Studer R, Benjamins R, Fensel D. Knowledge engineering: Principles and methods. *Data Knowl Eng*. 1998;25(1-2):161-198.
8. Staab S, Studer R. *Handbook on Ontologies*. Springer; 2010.
9. Horrocks I. What Are Ontologies Good For ? *Evol Semant Syst*. 2013:175-188.
10. Wang RY, Strong DM. Beyond Accuracy : What Data Quality Means to Data Consumers. *J Manag Inf Syst*. 1996;12(4):5-33.
11. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Ascp MT, 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.
12. Almutiry O, Wills G, Crowder R. *Toward a Framework for Data Quality in Electronic Health Record*.; 2013.
13. 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.
14. Heard S. openEHR Clinical Knowledge Manager. *openEHR*. 2015. http://openehr.org/ckm/. Accessed January 1, 2015.
15. National Quality Forum. *Quality Data Model December 2013*.; 2013.

16. Noy NF, McGuinness DL. *Ontology Development 101 : A Guide to Creating Your First Ontology*.; 2001.
17. Gamma E, Helm R, Johnson R, Vlissides J. *Design Patterns: Elements of Reusable Object-Oriented Software*. Pearson Education; 1994.
18. 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.
19. 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.
20. Wand Y, Wang RY. Anchoring Data Quality Dimensions in Ontological Foundations. *Commun ACM*. 1996;39(11):86-95.
21. 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.
22. Canadian Institute of Health. *The CIHI Data Quality Framework*.; 2009.
23. 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.
24. 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.
25. Pipino LL, Lee YW, Wang RY. Data quality assessment. *Commun ACM*. 2002;45(4):211. doi:10.1145/505248.506010.
26. Fox C, Levitin A, Redman T. The notion of data and its quality dimensions. *Inf Process Manag*. 1994;30(1):9-19.
27. W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview. 2012. http://www.w3.org/TR/owl2-overview/.
28. Object Management Group. Ontology Definition Metamodel. 2014. http://www.omg.org/spec/ODM/1.1/.
29. Beale ET, Heard S. Architecture Overview. 2008:1-79.
30. 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.
31. Kahn MG, Eliason BB, Bathurst J. Quantifying clinical data quality using relative gold standards. *AMIA Annu Symp Proc*. 2010;2010:356-360.
32. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. *J Am Med Inform Assoc*. 1997;4(5):342-355.