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

Steven G. Johnson[1]; Stuart Speedie[1]; Gyorgy Simon[1]; Vipin Kumar[2]; Bonnie L. Westra[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

Keywords

Data quality, electronic health record, data validation and verification, ontology

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.

Correspondence to:

Steven G. Johnson
University of Minnesota,
Institute for Health Informatics
330 Diehl Hall
505 Essex Street SE
Minneapolis, MN 55455
E-mail: joh06288@umn.edu

Appl Clin Inform 2016; 7: 69–88
http://dx.doi.org/10.4338/ACI-2015-08-RA-0107
received: August 28, 2015
accepted: December 29, 2015
published: February 10, 2016
Citation: Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Application of an ontology for characterizing data quality for a secondary use of EHR data. Appl Clin Inform 2016; 7: 69–88
http://dx.doi.org/10.4338/ACI-2015-08-RA-0107

1. 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 [1]. 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 [2,3]. 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 [4,5]. 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 [6]. 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 [7–9]. 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 [10–12]. The Canadian Institute for Health Information (CIHI) defined aspects of data quality and provided a process for assessing data based on those definitions [13]. 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 [14]. 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 [15]. 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 [16]. 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 [17]. 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 [18]. 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 [19]. 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 [20]. Key concepts and their definitions from the DQ Ontology are listed in ▶Table 1 and the relationships between them are shown in ▶Figure 1 [18, p.1940]. This ontology precisely defines data quality concepts in terms of relationships and constraints with other DQ concepts (shown in blue in ▶Figure 1). Also included in ▶Figure 1 is a link to 2 other ontologies described later in the paper – Task (shown in ▶Figure 2 in orange) and Domain (shown in ▶Figure 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 2.

Defining data quality as an ontology also provides a process for computing quantities that characterize data quality [18]. 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.

2. 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 [21]; 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" [22][23]. 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 [24]. 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 2. The full DQ Ontology describes 19 measures that characterize data quality [18]. 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 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) and a table that lists constraints (▶Table 3).

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. 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 [25] 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. Results

The data quality assessment process was performed on the de-identified 200,000 encounter sample from the Fairview Health EHR data. ▶Table 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 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 5 shows how TaskRelevance changes over time. ▶Figure 6 shows the value for the CMS178 eMeasure over the same time period. And ▶Figure 7 displays the monthly trend of DomainConsistency for the entire Dataset. The Pearson correlation between DomainConsistency and the CMS178 eMeasure was 0.78.

4. 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 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 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 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 [26] 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.

5. 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.

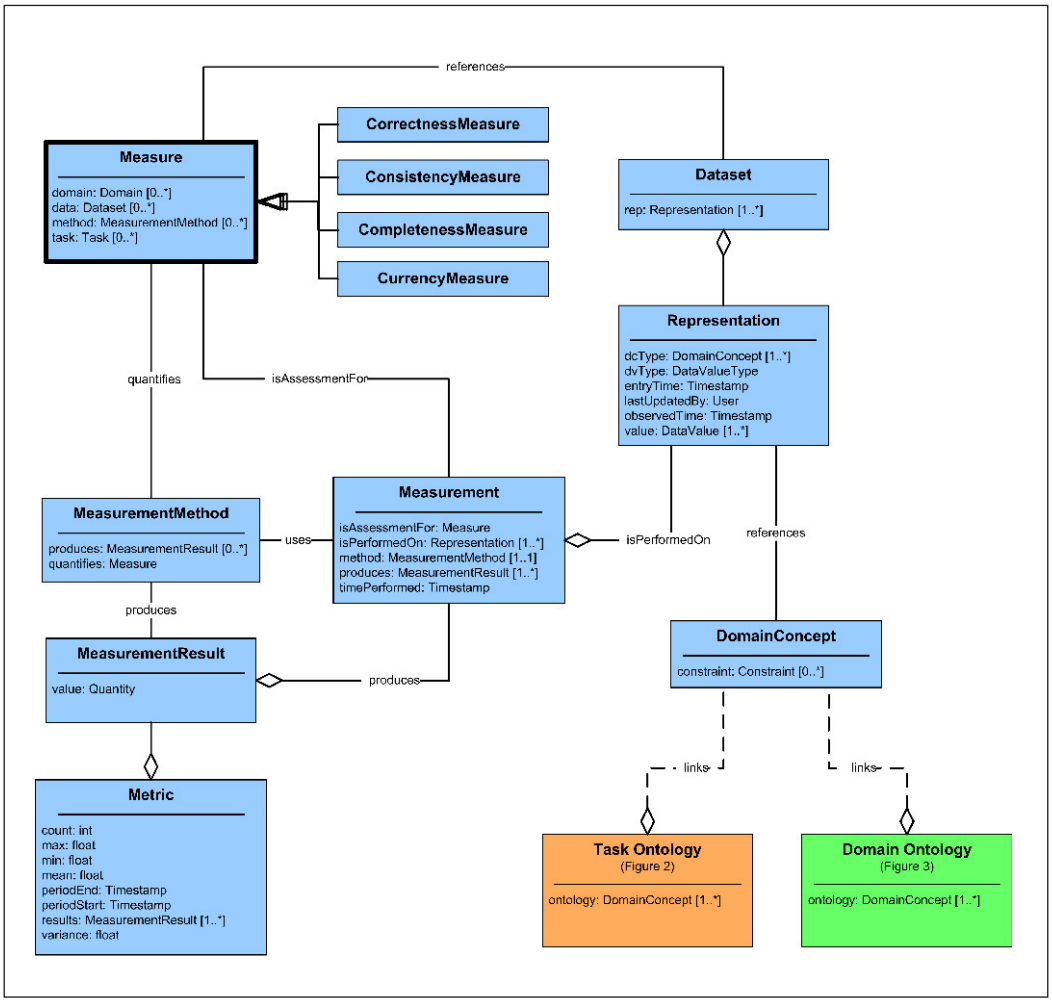

Fig. 1 Data Quality Ontology Diagram

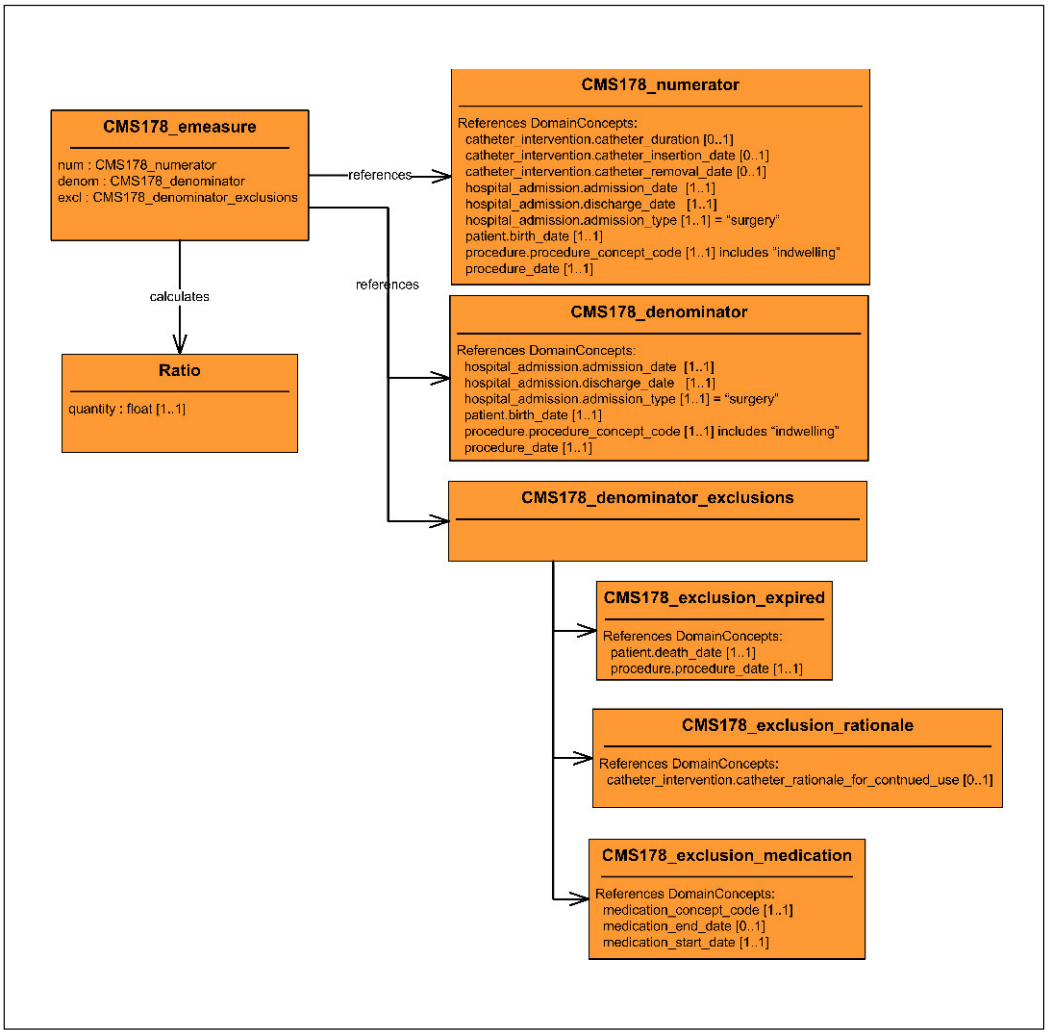

Fig. 2 Task Ontology (CMS178)

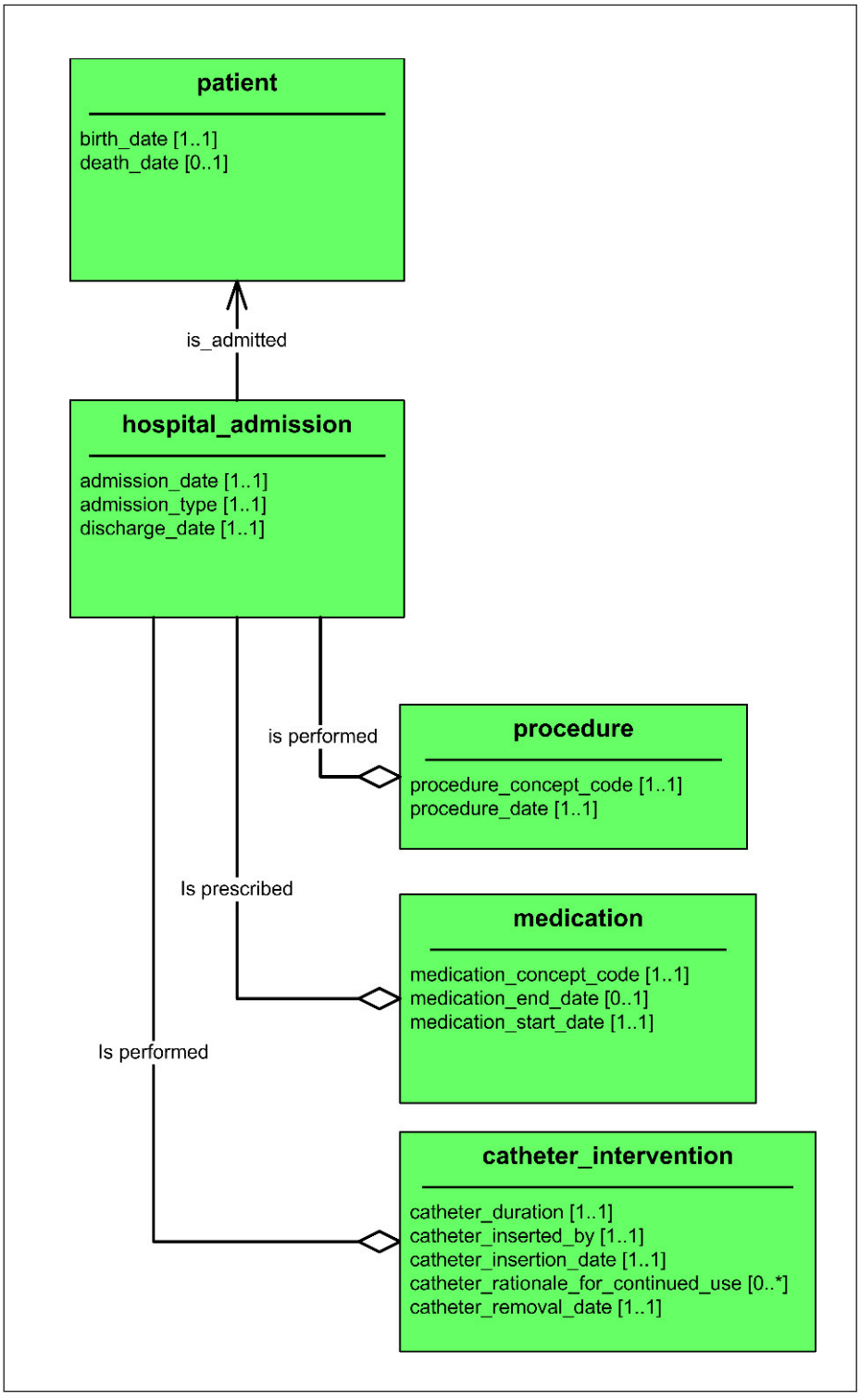

Fig. 3 Domain Ontology

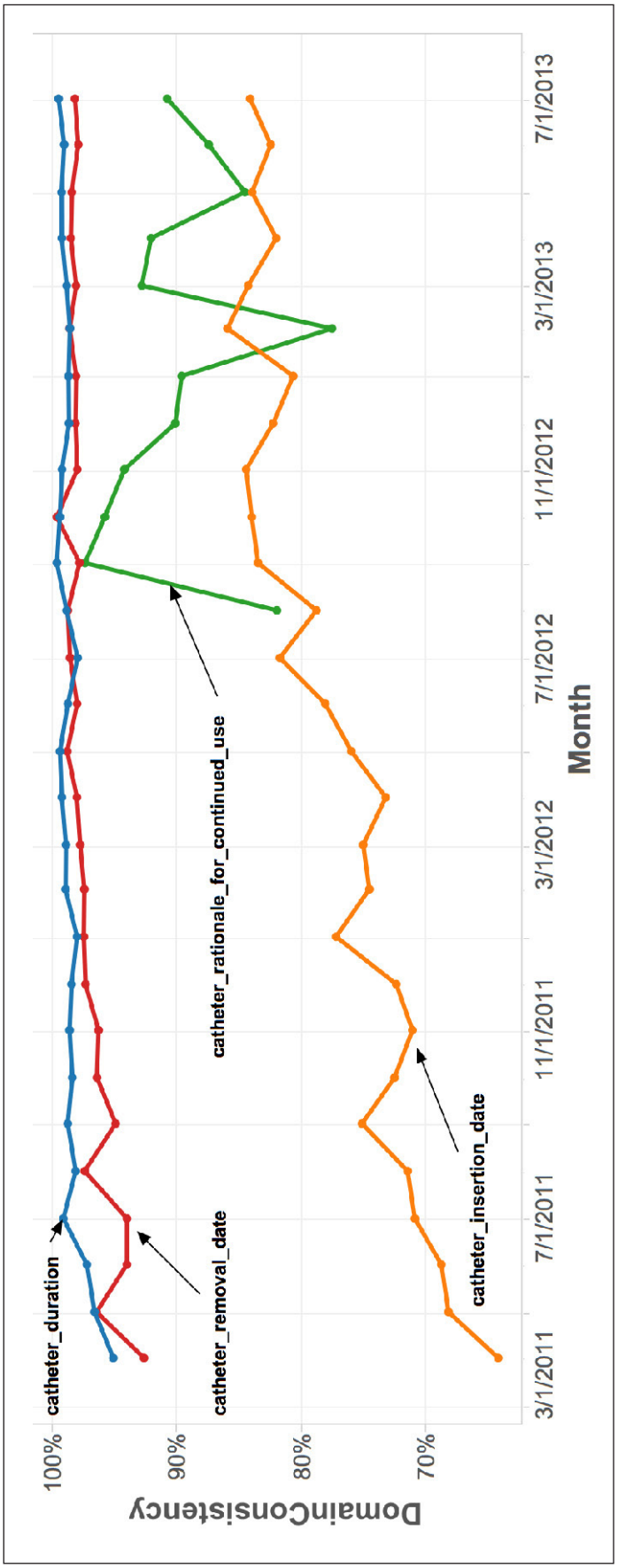

Fig. 4 DomainConsistency for Selected DomainConcepts, by Month

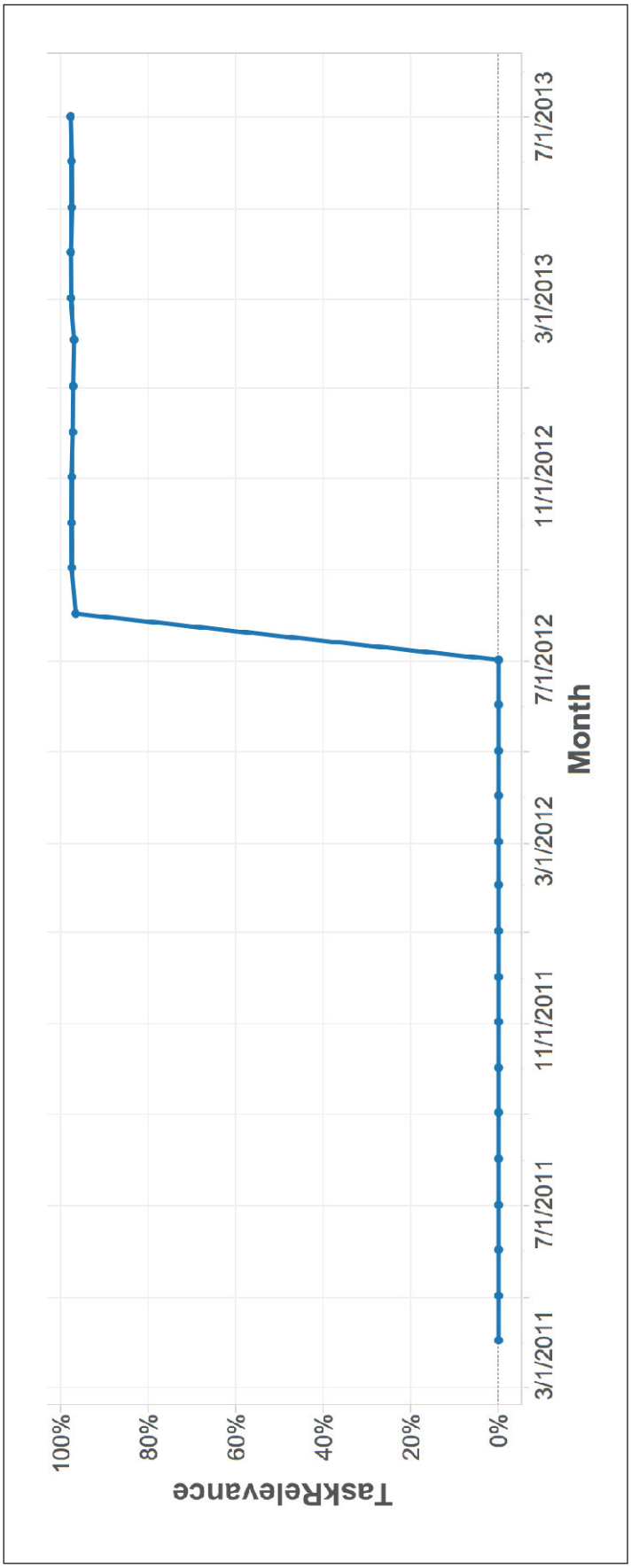

Fig. 5 TaskRelevance by Month

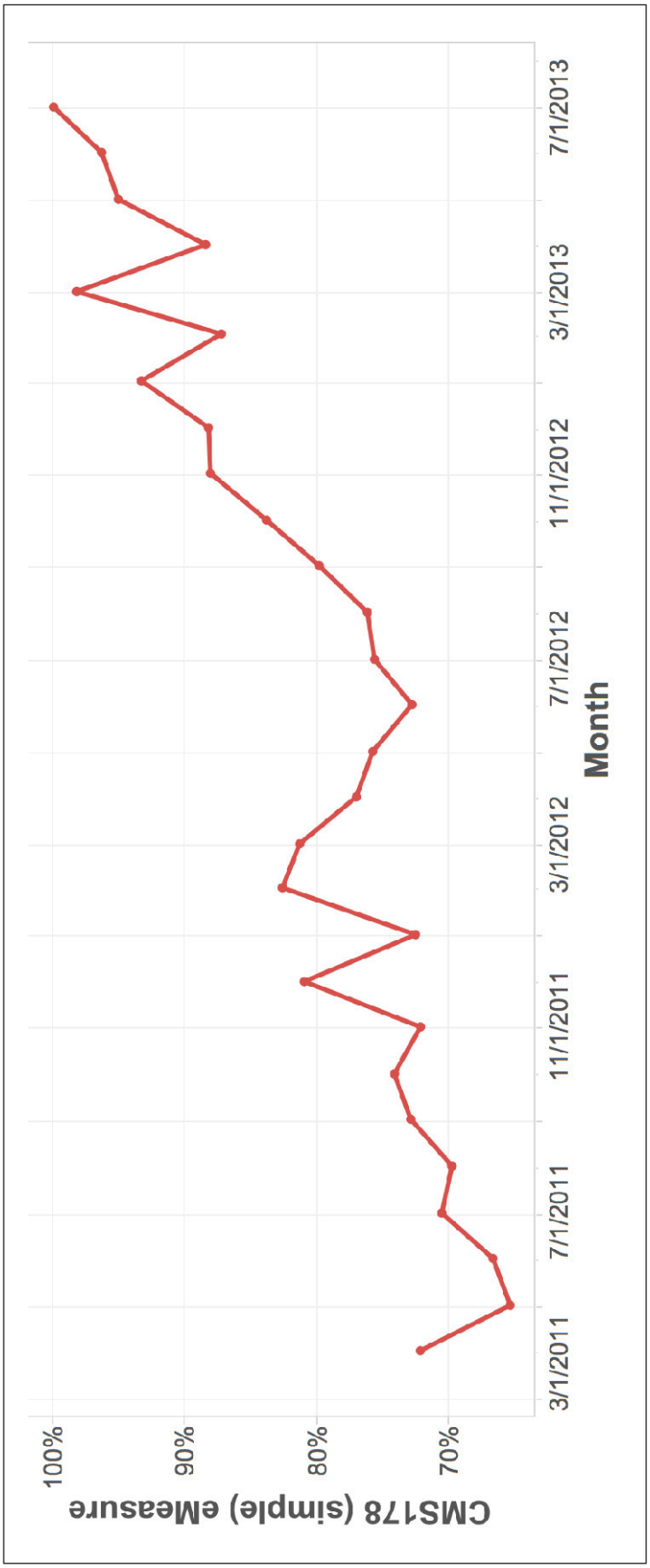

Fig. 6 CMS178 (simple) eMeasure by Month

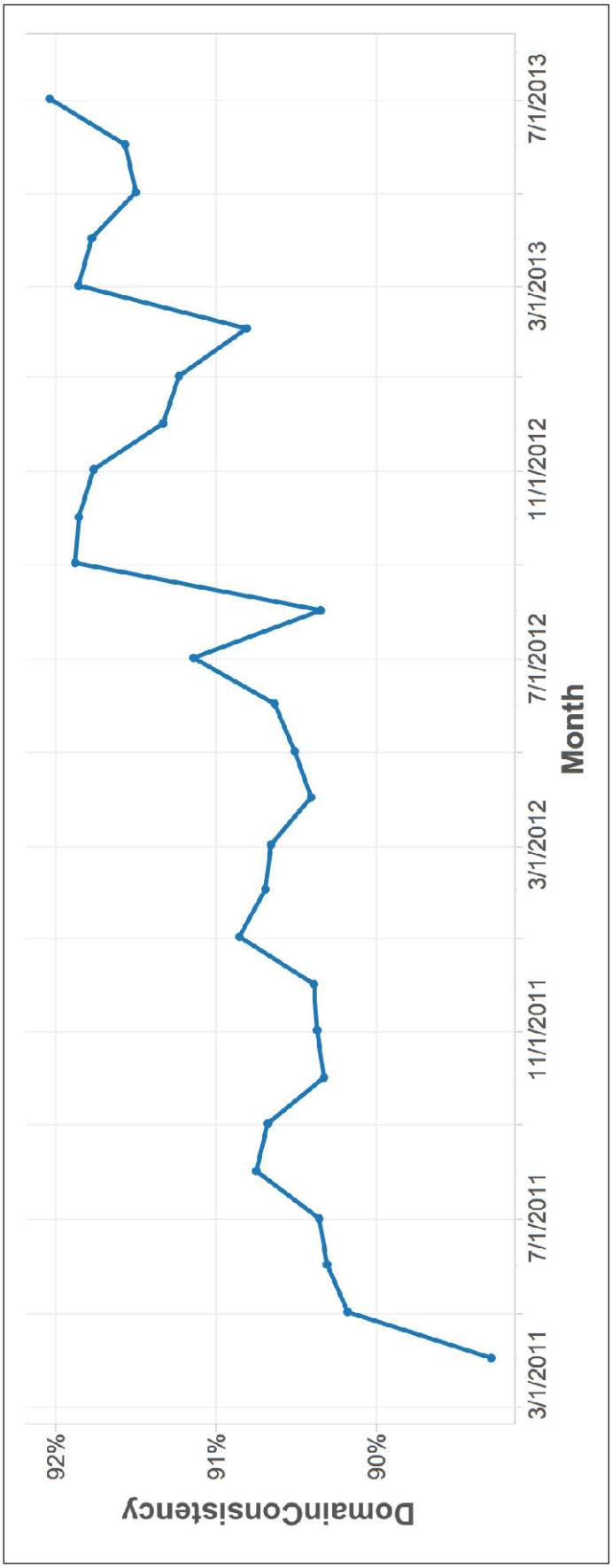

Fig. 7 DomainConsistency for the Dataset

Table 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.

Table 2 Data Quality Ontology – Measure Detail with Constraints

Measure	Definition	Constraint
ConsistencyMeasure		
Representation-Consistency	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()
Domain-Consistency	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
Domain-Constraints	All of the constraints defined for the DomainConcept are satisfied.	for each constraint in value.DomainConcept.constraints: constraint is True
Coding-Consistency	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		
Representation-Complete	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)

Table 2 Continued

Measure	Definition	Constraint
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)

Table 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 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%

References

1. Kayyali B, Knott D, van Kuiken S. The big-data revolution in US health care : Accelerating value and innovation. McKinsey & Company 2013; 1–6.
2. Blumenthal D, Tavenner M. The "Meaningful Use" Regulation for Electronic Health Records. N. Engl. J. Med. Massachusetts Medical Society; 2010; 363: 501–504.
3. King J, Patel V, Furukawa MF. Physician Adoption of Electronic Health Record Technology to Meet Meaningful Use Objectives: 2009–2012. ONC Data Brief 7. 2012.
4. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Detmer DE. 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): 1–9.
5. Holve E, Segal C, Hamilton Lopez M. Opportunities and Challenges for Comparative Effectiveness Research (CER) With Electronic Clinical Data. Med Care 2012; 50: S11–S18.
6. Mirnezami R, Nicholson J, Darzi A. Preparing for Precision Medicine. N Engl J Med 2012; 366: 489–491.
7. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ, Saltz JH. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013; 51: S30–S37.
8. 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.
9. 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.
10. Liaw S, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, Jalaludin B, Yeo AET, Talaei-Khoei A. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inform 2013; 82: 10–24.
11. 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.
12. 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: 21–29.
13. Canadian Institute of Health. The CIHI Data Quality Framework. 2009.
14. Observational Medical Outcomes Partnership (OMOP) [Internet]. [cited 2015 Jul 15]. Available from: http://omop.org/
15. 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.
16. Platt R, Carnahan R, Brown J, Chrischilles E, Curtis L, Hennessy S, Nelson JC, Racoosin JA, Robb M, Schneeweiss S, Toh S, Weiner MG. The U.S. Food and Drug Administration's Mini-Sentinel Program. Pharmacoepidemiol Drug Saf 2012; 21: 1–8.
17. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, Schilling LM, Weiskopf NG, Williams AE, Zozus MN. Transparent Reporting of Data Quality in Distributed Data Networks. eGEMs (Generating Evid. Methods to Improv. patient outcomes) 2015; 3; 1–12.
18. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. A Data Quality Ontology for the Secondary Use of EHR Data. AMIA 2015 Annu Symp Proc American Medical Informatics Association; 2015; 1937–1946.
19. Studer R, Benjamins R, Fensel D. Knowledge engineering: Principles and methods. Data Knowl Eng 1998; 25: 161–198.
20. Staab S, Studer R. Handbook on Ontologies. Springer; 2010.
21. Centers for Medicare & Medicaid Services. Proposed Clinical Quality Measures for 2014 CMS EHR Incentive Programs for Eligible Hospitals & CAHs [Internet]. 2014 [cited 2015 Aug 1]. p. 1–20. Available from: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Eligible_Hospital_Information.html
22. CMS Clinical Quality eMeasure Logic and Implementation Guidance v1.3 [Internet]. 2014 [cited 2015 Aug 1]. Available from: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/2014_eCQM_Measure_Logic_Guidancev13_April2013.pdf
23. 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 [Internet]. 2014 [cited 2015 Aug 1]. Available from: https://ecqi.healthit.gov/system/files/ecqm/2014/EH/measures/CMS178v5_1.html#toc
24. 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: 1544–1551.

25. Pipino LL, Lee YW, Wang RY. Data quality assessment. Commun ACM 2002; 45: 211.
26. Heinrich B, Kaiser M, Klier M. How to Measure Data Quality? A Metric Based Approach [Internet]. 2007 [cited 2015 Aug 1]. p. 1–15. Available from: http://epub.uni-regensburg.de/23633/1/heinrich.pdf