Quantifying the Effect of Data Quality on the Validity of an eMeasure

Steven G. Johnson[1] Stuart Speedie[1] Gyorgy Simon[1] Vipin Kumar[2] Bonnie L. Westra[1,3]

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

Address for correspondence Steven G. Johnson, PhD, University of Minnesota, Institute for Health Informatics, 330 Diehl Hall, 505 Essex Street SE., Minneapolis, MN 55455, United States (e-mail: joh06288@umn.edu).

Appl Clin Inform 2017;8:1012–1021.

Abstract

Objective The objective of this study was to demonstrate the utility of a healthcare data quality framework by using it to measure the impact of synthetic data quality issues on the validity of an eMeasure (CMS178—urinary catheter removal after surgery).

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

Results Catheter duration had the most impact on the CMS178 eMeasure with every 1% reduction in data quality causing a 1.21% increase in the number of missing events. For birth date and admission type, every 1% reduction in data quality resulted in a 1% increase in missing events.

Conclusion This research demonstrated that the impact of data quality issues can be quantified using a generalized process and that the CMS178 eMeasure, as currently defined, may not measure how well an organization is meeting the intended best practice goal. Secondary use of EHR data is warranted only if the data are of sufficient quality. The assessment approach described in this study demonstrates 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 validity and assess whether the values that are reported should be trusted.

Keywords
- data quality
- electronic health record
- data quality assessment
- ontology
- quality

Background and Significance

The U.S. healthcare system continues to invest in information technology to improve health outcomes.[1] 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.[2] Now that health information is in electronic form, it is more available for research.[3] 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.[4]

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.[5] eMeasures are just now becoming computable within EHR systems.[6,7] There are 297 active eMeasures listed

received
March 10, 2017
accepted after revision
August 28, 2017

DOI https://doi.org/10.4338/ACI-2017-03-RA-0042.
ISSN 1869-0327.

in the U.S. Department of Health and Human Services Measures Inventory[8] and many of these (93) are required to be reported by providers to meet the requirements of meaningful use.[9–11]

Computing a valid eMeasure value depends on how well the data are recorded in the EHR;[12] however, EHR vendors have not always ensured that data are captured at a quality sufficient to compute the eMeasure.[13] Data may be adequate to document care delivery but may be insufficient to support the valid computation of an eMeasure.[14] Data may be missing, incorrect, out of range, or inappropriate for secondary uses of a data field. The results from a Center for Medicare and Medicaid Services (CMS) pilot study showed that eMeasures matched manually abstracted measures less than half the time, primarily due to missing data.[15] Prior to 2014, encounters with missing data used in the eMeasure calculation were considered to fail. In 2014, CMS changed its approach and no longer considers missing data as failing the eMeasure.[16] When data are missing, the patient's record cannot be used in the calculation of the eMeasure and some ability to quantify the best practice that the eMeasure was intended to assess is lost. The validity of a measurement is the degree to which it measures what it purports to measure.[17] Measures are deemed valid by comparison to measures computed from a "gold standard" dataset.[18] But in practice, the best comparison that is usually available is a relative gold standard,[19] which is the approach used in this research.

Secondary uses of EHR data could be trusted more if the impact of the underlying data quality was assessed.[20] The Electronic Data Methods (EDM) Data Quality Collaborative recommends that researchers report on the quality of their data.[21] Frameworks for assessing healthcare data quality exist, but they are limited to specific projects.[22–24] We previously developed a generalized healthcare data quality framework (HDQF) that consists of a comprehensive data quality ontology and associated data quality assessment method that can be used to quantify data quality.[25]

Objective

The objective of this study was to demonstrate the utility of the HDQF by using it to measure the impact of synthetic data quality issues on the validity of an eMeasure (CMS178).

Methods

Data Source

Data were obtained from a clinical data repository (CDR) at the University of Minnesota. Institutional review board approval was received to extract a 72,127-encounter de-identified random sample of patients admitted between March 2011 and July 2013 to be used as the data source for this study.

Framework and Analysis

The HDQF is an ontology that contains data quality concepts, definitions, and relationships and an assessment method that produces quantities to characterize data quality along several dimensions. The HDQF has several benefits. The 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.[26] Concepts in the ontology are linked to two other ontologies: a Task ontology that describes the concepts, relationships in the data, and calculations necessary to carry out a particular use of the data and a Domain ontology that describes the semantics of the data by specifying constraints (rules) and relationships between concepts that the data should satisfy to accurately represent a clinical area.

The HDQF assessment method is used to calculate the proportion of the constraints that are satisfied for each type of data in a dataset (called *DomainConcepts* in the ontology).[27] The denominator is the number of data values for each DomainConcept in a population and the numerator is the number of data values which have all constraints satisfied. The research described in this article looks at two important aspects of data quality defined in the HDQF: RepresentationComplete and DomainConstraints. RepresentationComplete measures the degree to which data in a dataset is not missing (i.e., admission_date should not be blank). DomainConstraints assesses how well the data conforms to the Domain ontology (i.e., admission_date should be less than discharge_date).

The steps for applying the HDQF to this dataset were:

1. Define the Domain and Task ontologies.
2. Measure data quality.
3. Degrade the data.
4. Model and assess the impact.

Define the Domain and Task Ontologies

A Domain ontology was defined for this study and is shown in ►Table 1.

The constraints defined in the ontology are the same as those used in previous research.[27] They were intentionally kept simple to illustrate how to apply the HDQF. An example of a constraint is that the admission_date must be earlier than the discharge_date.

The Domain ontology (►Table 1) and a simplified CMS178 eMeasure ($CMS178_{simple}$, calculation shown in ►Fig. 1) were used for this research 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."[28] Patients who 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.[29] CMS178 calculates the proportion of patient encounters that satisfy this best practice.

The denominator includes all hospital patients (aged 18 and older) who 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, intravenous positive inotropic,

Table 1 Domain ontology with constraints

DomainConcept	Type	DomainConstraint
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 < 1,000 d
Admission_type	Code	
Discharge_date	Date	Admission_date ≤ discharge_date
Procedure		
Procedure_concept_code	Code	
Procedure_date	Date	Procedure_date ≥ admission_date
Medication		
Medication_concept_code	Code	
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 d Catheter_duration < 1,000 d
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

and vasopressor agents or paralytic agents. The numerator is the number of denominator surgical patients whose urinary catheter was removed within 48 hours of surgery. The steps to compute the CMS178 numerator and denominator for the baseline data are shown in ►Fig. 1. When computing CMS178, missing data need to be handled in a consistent manner. Prior to 2014, the definition of CMS178 required that an encounter would fail the eMeasure if there were any missing data. That requirement was removed from CMS178 definitions starting in 2014. The data quality inclusion policy is left up to the implementer of the CMS178 calculation. There are two approaches to data quality issues: (1) exclude encounters with missing data and (2) impute some reasonable value for the missing data (i.e., variable mean, a default value, etc.). For this research, the former approach was used. The eMeasure was computed as:

$$CMS178_{simple} = \frac{Numerator}{Denominator} = \frac{2725}{3541} = 0.77$$

Measure Data Quality

For this research, two aspects of data quality, RepresentationComplete and DomainConstraints, were studied. RepresentationComplete quantifies the extent of missing data. It is the proportion of encounters that have non null data values for a DomainConcept divided by the total number of encounters. DomainConstraints quantify the degree to which the data satisfy 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 value is the proportion of encounters in a population where that is true.

In this study, the HDQF was extended by describing a method for quantifying the degree that data quality issues for each DomainConcept impact a Task. The impact is quantified by deliberately injecting synthetic data quality issues into the underlying EHR data in a systematic way and observing how those changes affect the Task.

In this study, validity is a relative measure and was operationalized by comparison to a relative gold standard. The baseline, unmodified sample EHR data were used as the relative gold standard. The variable, missing_events, was computed to quantify the validity of the CMS178 eMeasure after the data are modified. This variable represents the number of patients who satisfied the CMS178 inclusion criteria and had a catheter removed within 48 hours in the baseline data but, after the data were degraded, were

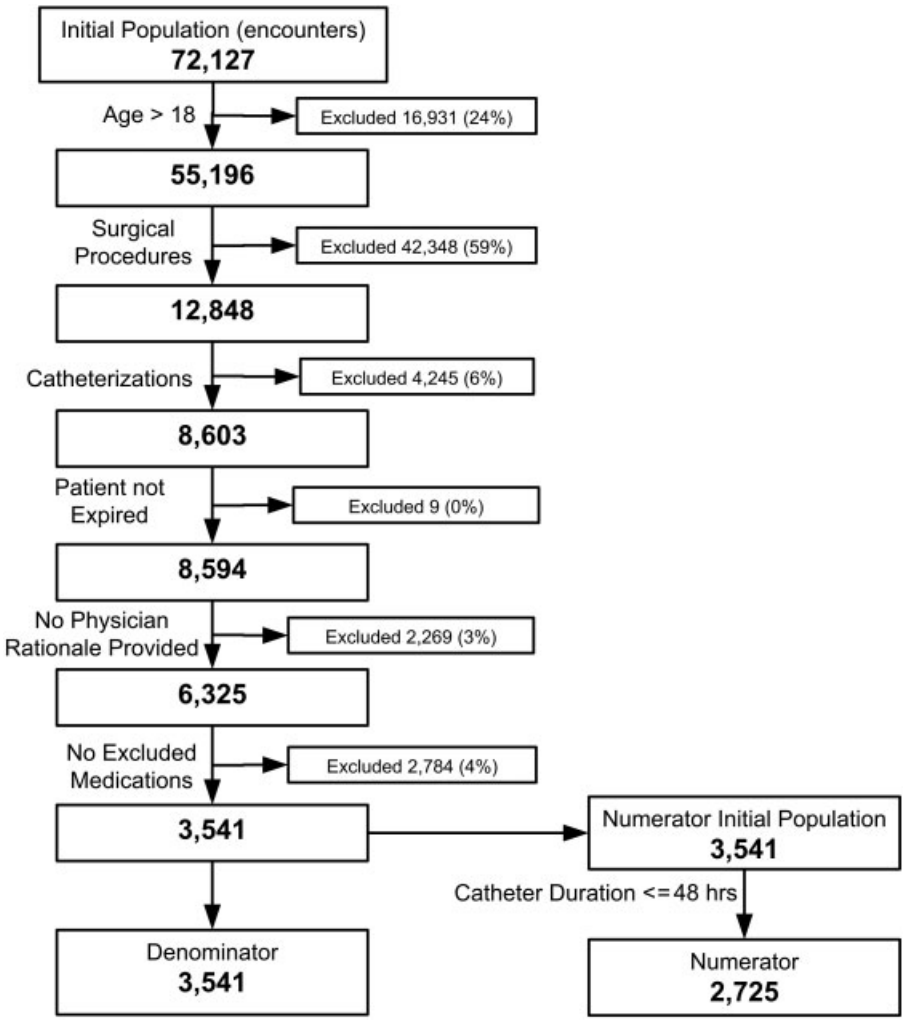

Fig. 1 Computation of CMS178 numerator and denominator for baseline data (undegraded).

subsequently not counted as satisfying the CMS178 numerator criteria. These are events of interest that were missing due to the induced data quality issues. The missing_events variable quantifies the impact that data quality issues have on the eMeasure.

Degrade the Data

Each type of data quality issue requires a specific method for degrading the data. For RepresentationComplete, missing data are simulated by removing data values. For DomainConstraints, the data are degraded by changing data values to no longer satisfy the Domain ontology rules. For example, for discharge_date, the underlying data were changed to occur before the admission_date by a random number of days. While "actual" data quality issues may not occur in this manner, these synthetic data quality issues are used to illustrate how the HDQF can be applied to assess the impact of data quality issues on a Task. Since the data quality inclusion policy for this research was to remove encounters that contained data quality issues, domain constraint violations have the same effect on the Task as missing data.

The full degradation process consists of iteratively applying the degradation method (i.e., removing data or violating constraints) to the data for each of the DomainConcepts listed in ►Table 1. The Task was performed (in this case, computing CMS178 and missing_events) and RepresentationComplete and DomainConstraint were recomputed on the degraded data. The RepresentationComplete and DomainConstraints statistics for every DomainConcept, the CMS178 eMeasure, and missing_events were recorded in an analysis database (see ►Fig. 2).

Two approaches to degrading data were examined: (1) independent and (2) correlated. Each process was performed to yield 1,200 observations with which 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 DomainConcept 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 DomainConstraints 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

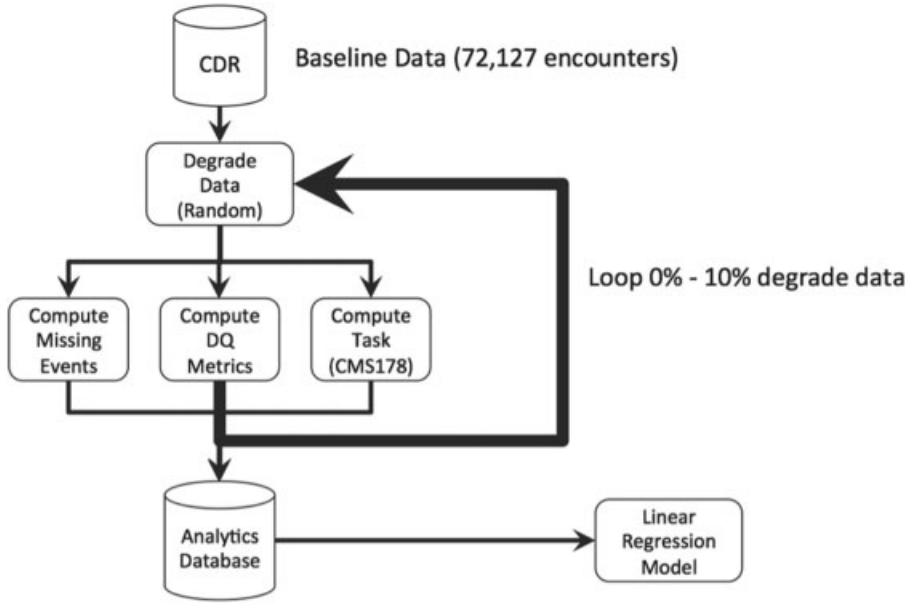

Fig. 2 Degradation process.

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 did not 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 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. This approach 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.

Model and Assess the Impact

A variable, missing_events_percent, is the number of missing events divided by the numerator of the base (undegraded) data. A linear regression model was fit to missing_events_percent as the dependent variable, with quantities for RepresentationComplete and DomainConstraints for each DomainConcept as the predictor variables. A linear regression model was computed using stepwise backward elimination. 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 by the same percentage. If data in an EHR are 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 the number of missed events would be reduced.

Results

RepresentationComplete and DomainConstraint issues were evaluated using the independent and correlated degradation methods. Results from a pairwise Pearson correlation and chi-square association between all 15 DomainConcepts showed three clusters of highly correlated variables:

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.

The resulting linear regression models for RepresentationComplete are shown in ►Table 2.

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. To illustrate, ►Table 3 shows the baseline number of encounters for the numerator and denominator and what those values are when 10% of the data are degraded for two example variables, catheter_duration and birth_date.

Table 2 Linear regression models for missing_events_percent based on predictor variables of representation completeness of Domain variables generated by applying independent and correlated degradation

Predictor	Coefficient	SE coefficient	*t*-Value	*p*-Value
Independent degradation				
Birth_date	−0.9949	0.0069	−144.7	< 0.0001
Admission_type	−0.9941	0.0069	−144.6	< 0.0001
Medication_start_date	−0.3927	0.0072	−54.3	< 0.0001
Catheter_duration	−1.2136	0.0083	−145.9	< 0.0001
Catheter_rationale_for_continued_use	−0.1226	0.0070	−17.6	< 0.0001
Correlated degradation				
Birth_date	−0.9955	0.0066	−151.2	< 0.0001
Admission_type	−0.9972	0.0066	−151.5	< 0.0001
Medication_ start _date	−0.4155	0.0073	−56.8	< 0.0001
Catheter_duration	−1.1863	0.0080	−149.0	< 0.0001
Catheter_rationale_for_continued_use	−0.1178	0.0067	−17.7	< 0.0001

Table 3 Impact of 10% degradation versus baseline

	Baseline	10% degrade (RepresentationComplete)	
		Birth_date	Catheter_duration
Numerator	2,725	2,447	2,450
Denominator	3,541	3,185	3,200
Missing events	0	278	275
CMS178	0.770	0.768	0.766

The resulting linear regression models for DomainConstraints are shown in ►Table 4.

The CMS178 eMeasure was also computed, as data quality was being degraded to show how it changed as the number of missing_events increased. A graph of CMS178 compared with RepresentationComplete for the dataset as data quality is degraded for a DomainConcept (in this case, catheter_duration) is shown in ►Fig. 3.

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

Discussion

The objective of this study was to demonstrate the utility of the HDQF by using it to measure the impact of synthetic data quality issues on the validity of an eMeasure (CMS178). The results of this study support two primary findings: (1) the impact of data quality issues for different variables 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.

Table 4 Linear regression models for missing_events_percent based on predictor variables of DomainConstraints for Domain variables generated by applying independent and correlated degradation

Predictor	Coefficient	SE coefficient	*t*-Value	*p*-Value
Independent degradation				
Birth_date	−1.0099	0.0067	−150.3	< 0.0001
Admission_type	−1.0020	0.0067	−149.2	< 0.0001
Medication_start_date	−0.4018	0.0071	−56.9	< 0.0001
Catheter_duration	−1.1970	0.0081	−147.5	< 0.0001
Catheter_rationale_for_continued_use	−0.1219	0.0068	−17.9	< 0.0001
Correlated degradation				
Birth_date	−0.9839	0.0061	−161.5	< 0.0001
Admission_type	−1.0000	0.0061	−164.1	< 0.0001
Medication_start_date	−0.3858	0.0064	−60.2	< 0.0001
Catheter_duration	−1.1685	0.0074	−158.7	< 0.0001
Catheter_rationale_for_continued_use	−0.1136	0.0062	−18.4	< 0.0001

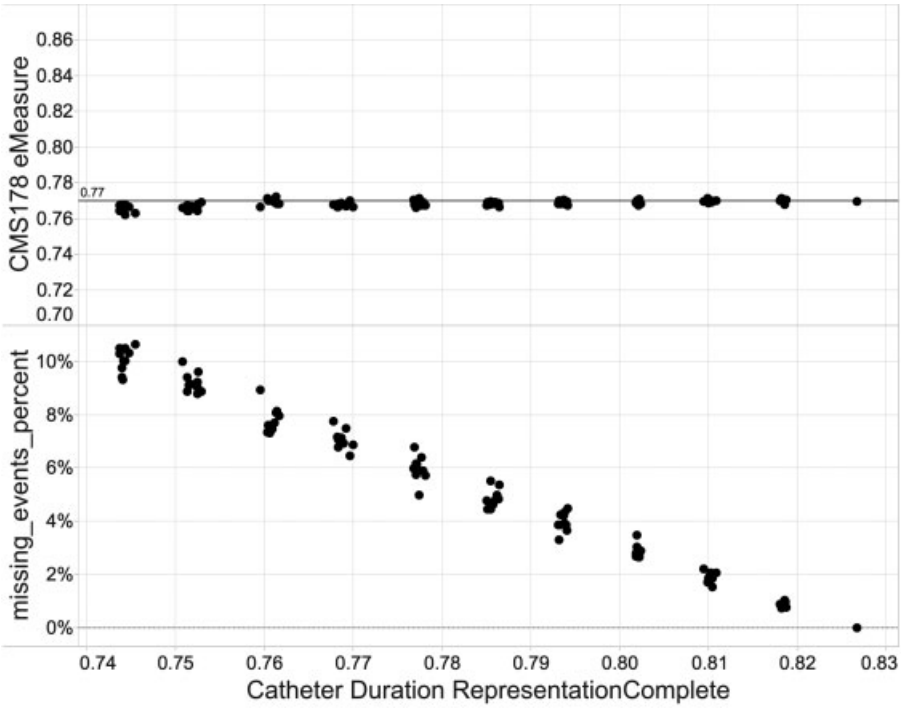

Fig. 3 CMS178 eMeasure and missing_events_percent versus catheter duration RepresentationComplete data quality.

In support of the first finding, the impact of the data quality of each DomainConcept on the eMeasure is reflected in the results of the linear regression models. The data quality issues related to missing data and DomainConstraint violations that were examined had an impact on the number of encounters that were included in both the numerator and denominator of the eMeasure and this is captured in missing_events. The coefficients in the models can be interpreted to quantify the magnitude of the impact on missing_events for a 1 unit improvement in data quality. For example, for the independently degraded RepresentationComplete in ►Table 2, for every 1% reduction in RepresentationComplete data quality for admission_type, there were 0.9941% (essentially 1.0%) more events missed. But a 1% reduction in data quality for catheter_rationale_for_continued_use results only in 0.12% of cases being missed.

►Table 3 helps illustrate how different DomainConcepts impact missing_events. The table shows denominator and numerator values for two examples of DomainConcepts: birth_date and catheter_duration. It shows that there were 278 missing_events when the birth_date field was degraded by 10%. CMS178 excludes patients who are younger than 18 years; so, when the birth_date is removed, the encounter is no longer included in either the denominator or the numerator because the age cannot be computed. 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 that there is a one-for-one impact on the number of missing events. 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, a 10% degradation of the data causes 275 missing events. The data quality inclusion policy for this research is that if data are missing, then the encounter should be removed from the CMS178 calculation and so would be removed from both the numerator and denominator. In this study, 17% of the encounters already had a missing catheter_duration (83% had values). So 1.17% of the catheter_duration data must be degraded to remove 1% more missing events. The coefficient of −1.21 in the linear model approximates this. For this simple eMeasure, these relationships could potentially be discovered algebraically without using a regression model, but for more complex Tasks this is not likely to be the case.

In the independent model, catheter_duration has the most impact on missing events. For every 1% decrease in RepresentationCompleteness of these variables (i.e., more missing data), there is approximately a 1.21% increase in the number of missing_events. The variables admission_type and birth_date were also very impactful variables. For every 1% decrease in data quality, 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 cannot be calculated because birth_date is missing, the encounter is removed from both the numerator and denominator. The eMeasure proportion stays roughly the same, but missing_events increase. The same is true for admission_type as a nonsurgical case is removed from both the numerator and denominator and the CMS178 eMeasure stays the same. As catheter_duration has the largest impact on the number of missed events, any data quality initiatives should focus on improving its data quality first.

Degrading each DomainConcept independently compared with degrading in a correlated manner produced the same set of variables that were most impactful. In the independent and correlated models, 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. 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. 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 DomainConstraints yielded the same set of variables that were impactful as RepresentationComplete. This is due to the data quality inclusion policy of removing data that violate constraints, so it has the same impact as missing data. As with RepresentationComplete, catheter_duration has the largest impact on the number of missed events (−1.197%) and a 1% improvement in data quality for admission_type and birth_date results in a 1% reduction in missing events.

The second finding is that the CMS178 eMeasure may not adequately measure catheter removal within 48 hours of surgery. As seen in ►Fig. 3, even though missing_events increase as the underlying data are degraded, CMS178 itself does not appreciably change. This is due to the fact that the eMeasure is a proportion. As the data are 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. The way the eMeasure is currently defined may not give CMS an accurate quantification of how well an organization is removing catheters within 48 hours of surgery due to the possibility of excluding cases that either meet or do not meet the inclusion criteria due to missing data or data violations. CMS may want to revisit their pre-2014 approach of reporting missing data.

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 validity. If data quality measures are too low in a particular area, it may be advisable not to report the eMeasure or at least indicate the level of data quality (using RepresentationComplete and DomainConstraint metrics) when the eMeasure is reported.

Limitations and Future Work

There are some limitations to this research. This research did not attempt to quantify every type of data quality issue and looked only at two types of problems: RepresentationComplete and DomainConstraints. There are other types of data quality issues that should be explored. This research also did not attempt to assess which data quality issues were actually occurring in the EHR data; it only defines data degradation methods to illustrate how the HDQF can be applied. 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 DomainConstraint to be violated. But an error, such as a typo, could occur that only affects the day of the month, which would not necessarily violate the DomainConstraint. Other approaches are needed to quantify those types of errors. The impact of an issue is also dependent on the specific Domain model that is defined as well as the amount of data that is degraded. This study modified up to 10% of the data, but further research is needed to determine the typical proportion of data errors in a CDR.

Another limitation is that RepresentationComplete should be expanded to encompass different types of missing data. The definition of completeness is contextual and dependent on how data will be used.[30] RepresentationComplete should differentiate between data that are missing completely at random, missing at random, and missing not at random.

This study used the CMS178 eMeasure as an example Task to study in detail the process for assessing the impact data quality has on the validity of Task results. 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.

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 may be the best, most robust approach. But degrading each DomainConcept independently has the fastest execution time. Further research is needed with additional Tasks to understand when each degradation technique can be applied. 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.

Conclusion

Access to a significant amount of structured electronic health data allows researchers to identify evidence-based best practices that improve patient outcomes. The secondary use of data is warranted only if the data are 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. This research described application of a HDQF to quantify the impact of RepresentationComplete and DomainConstraint data quality issues on the validity of an eMeasure and the assessment approach can be generalized for other data analysis Tasks. The research also raises some questions about how the CMS178 eMeasure is currently defined. It may not adequately assess how well an organization is removing catheters within 48 hours of surgery. 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 and assess whether the values that are being reported should be trusted.

Clinical Relevance Statement

Secondary use of EHR data is warranted only if the data are of sufficient quality to support the secondary use. The research described in this article quantified the impact of RepresentationComplete and DomainConstraint data quality issues on the validity of an eMeasure and the assessment approach can be generalized for other data analysis Tasks.

Multiple Choice Question

The process of degrading data to build a model of the impact of data quality issues on an eMeasure uses which of the following as variables for the linear regression model:

A. Data quality metrics for DomainConcepts as the dependent variables and missing events percent as an independent variable.
B. The eMeasure as the dependent variable and missing events as the independent variable.
C. The data quality metrics for DomainConcepts as the independent variables and missing events percent as the dependent variable.
D. You can't quantify the impact of the data quality issues.

Correct answer: The correct answer is C. The linear model is predicting the number of missing events based on the data quality metrics (i.e., RepresentationComplete and DomainConstraints) of each of the DomainConcepts. As those metrics vary, the coefficients in the linear regression model quantify the effect that those data quality issues have in causing the missing events.

Protection of Human and Animal Subjects

De-identified EHR data were 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).

Funding

The clinical data repository used in 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).

Conflict of Interest

None.

References

1 Blumenthal D. Launching HITECH. N Engl J Med 2010;362(05):382–385
2 Zerhouni EA. Translational and clinical science–time for a new vision. N Engl J Med 2005;353(15):1621–1623
3 Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017;26(01):1–15
4 Zozus MN, Hammond WE, Green BB, et al. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research [Internet]. NIH Collab. [cited March 1, 2016]. 2014:1–26. Available at: https://www.nihcollaboratory.org/Products/Assessing-data-quality_V1 0.pdf
5 Conway PH, Mostashari F, Clancy C. The future of quality measurement for improvement and accountability. JAMA 2013;309(21):2215–2216
6 Torda P, Tinoco A. Achieving the promise of electronic health record-enabled quality measurement: a measure developer's perspective. EGEMS (Wash DC) 2013;1(02):1031
7 Amster A, Jentzsch J, Pasupuleti H, Subramanian KG. Completeness, accuracy, and computability of National Quality Forum-specified eMeasures. J Am Med Inform Assoc 2015;22(02):409–416
8 Agency for Healthcare Research and Quality. Measures Inventory [Internet]. 2015 [cited November 3, 2015]. Available at: http://www.qualitymeasures.ahrq.gov/hhs/matrix.aspx
9 Agency for Healthcare Research and Quality. Clinical Quality Measures [Internet]. 2015 [cited November 3, 2015]. Available at: 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
10 Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010;363:501–504
11 Centers for Medicare & Medicaid Services (CMS). EHR Incentive Programs: 2015 through 2017 Overview [Internet]. 2015 [cited November 8, 2015]. Available at: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/Stage3Overview2015_2017.pdf
12 Persell SD, Wright JM, Thompson JA, Kmetik KS, Baker DW. Assessing the validity of national quality measures for coronary artery disease using an electronic health record. Arch Intern Med 2006;166(20):2272–2277
13 HealthCatalyst. The Unintended Consequences of Electronic Clinical Quality Measures [Internet]. 2015 [cited November 8, 2015]. Available at: https://www.healthcatalyst.com/electronic-clinical-quality-measures-impact-data-quality
14 Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010;67(05):503–527
15 Centers for Medicare & Medicaid Services (CMS). Hospital Inpatient Quality Reporting (IQR) eCQM Validation Pilot Summary. 2016
16 Centers for Medicare & Medicaid Services (CMS). Clinical Quality Measures for CMS's 2014 EHR Incentive Program for Eligible Hospitals: Release Notes, April 1, 2013. 2014
17 Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. Psychol Rev 2004;111(04):1061–1071
18 Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc 1997;4(05):342–355
19 Kahn MG, Eliason BB, Bathurst J. Quantifying clinical data quality using relative gold standards. AMIA Annu Symp Proc 2010;2010:356–360
20 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
21 Kahn MG, Brown JS, Chun AT, et al. Transparent reporting of data quality in distributed data networks. EGEMS (Wash DC) 2015;3(01):1052
22 Observational Medical Outcomes Partnership (OMOP) [Internet]. [cited July 15, 2015]. Available at: http://omop.org/
23 Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(01):54–60
24 Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf 2012;21(Suppl 1):1–8

25 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 2015:1937–1946
26 Staab S, Studer R. Handbook on Ontologies. Springer; 2010
27 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(01):69–88
28 CMS Clinical Quality eMeasure Logic and Implementation Guidance v1.3 [Internet]. 2014 [cited August 1, 2015]. Available at: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/2014_eCQM_Measure_Logic_Guidancev13_April2013.pdf
29 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
30 Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013;46(05):830–836