Data Quality Overview

OHDSI Data Quality Tutorial

September 16, 2019

What is Data Quality?

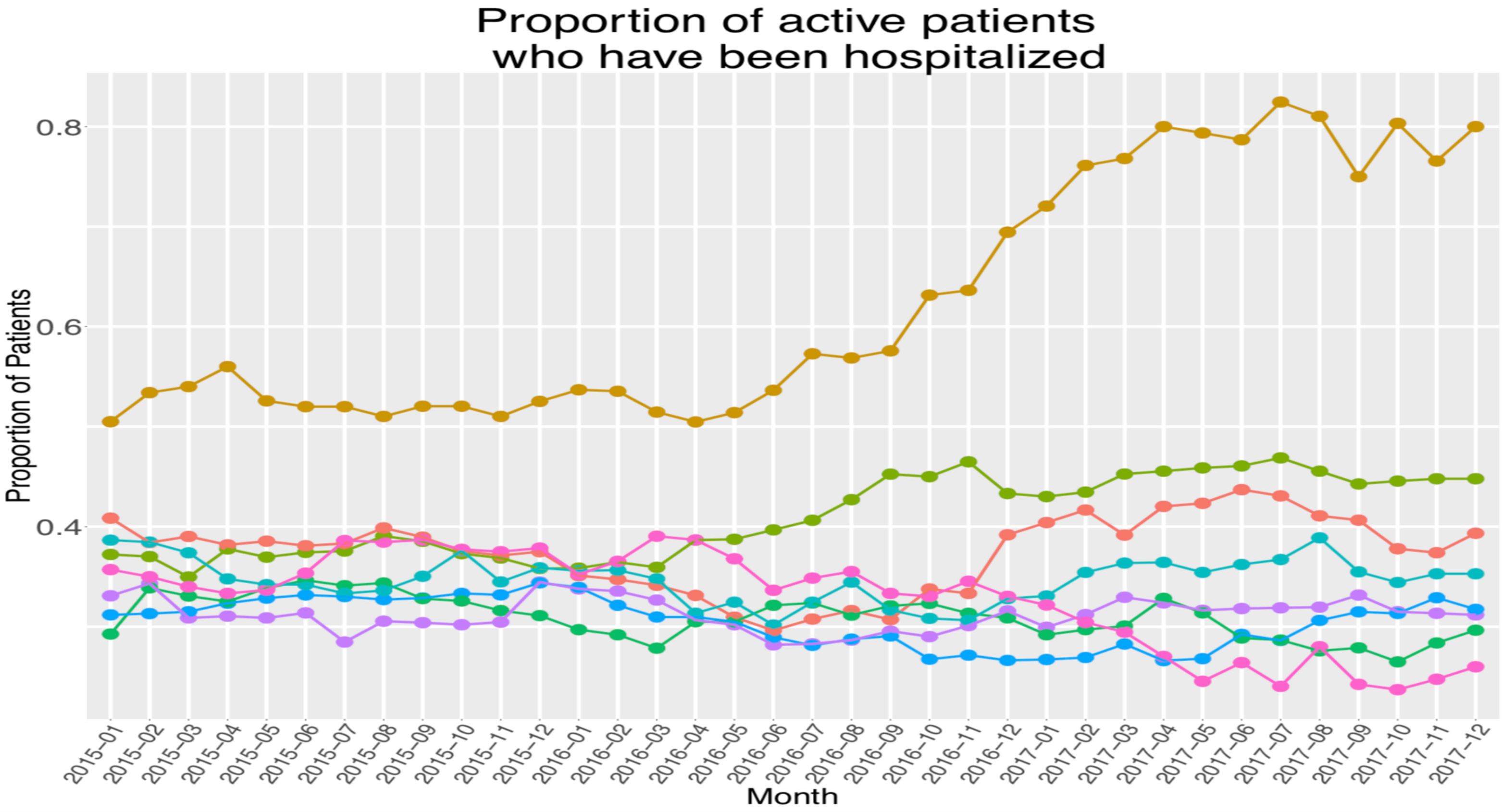
- A thing you think about briefly and roll up to *data cleaning* procedures mentioned as an afterthought in the Methods section of your JAMA paper ✗

- A complex discipline that has no formal consensus definition ✓
  - *”fit for use”*
  - *”... the ability to achieve desirable objectives using legitimate means. Quality data represents what was intended or defined by their official source, are objective, unbiased and comply with known standards.” (WHO)*

PEDSnet
A Pediatric Learning Health System

Poor Data Quality ➔ Poor Study!

- Poor data quality can lead to <u>spurious</u> cohort selection, <u>misclassification</u> of major variables, and <u>misleading</u> reporting of results
- *Does one site really have high hospitalizations or is there an underlying data quality problem?*



Data Quality: What's in a Name?

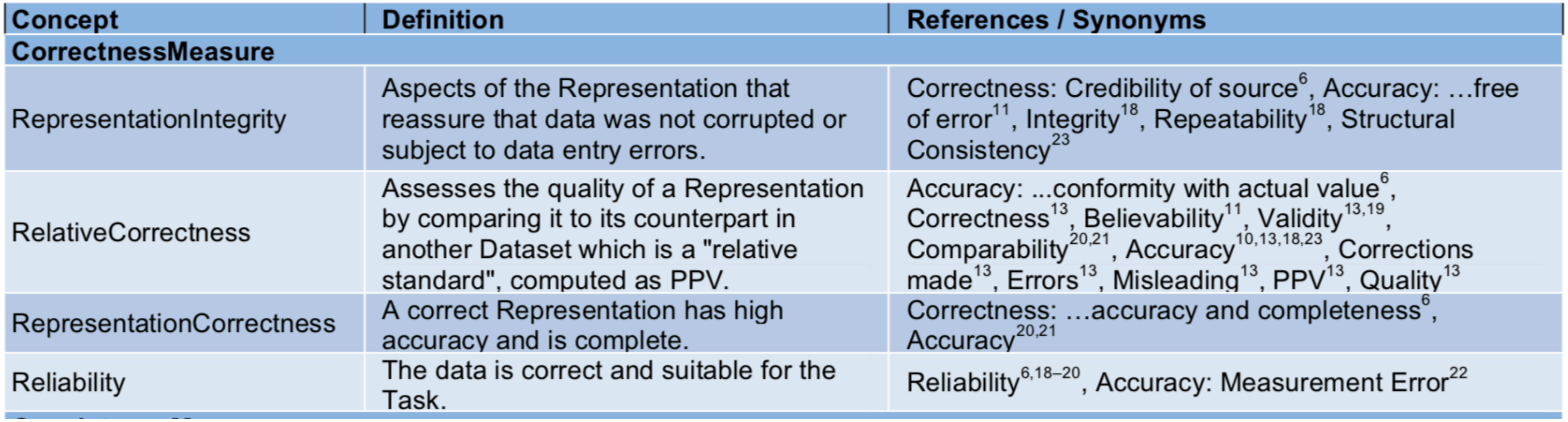
- Data quality and the problem of terminology
- How to assess ***fit for use?***

DQ 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

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.

Approaches to Data Quality

- Standardizing terminologies
  - e.g. *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]

Johnson, S. G., Speedie, S., Simon, G., Kumar, V. & Westra, B. L. A Data Quality Ontology for the Secondary Use of EHR Data. *AMIA Annu Symp Proc* 2015, 1937-1946.

- Network Benchmarking *(e.g., PCORnet, PEDSnet, etc)*
- Formal Statistical Tools and Methods
  - *tools:* clustering, distributions, correlations
  - *methods:* imputation, annotation, stratifying, handling confounding

Data Quality Landmark Papers: Kahn et al 2016

- Harmonized terminology
- Comprises:
  - Data Quality Categories / Subcategories
    - ***Conformance***:
      - value, relational, computational conformance
    - ***Completeness***
    - ***Plausibility***:
      - uniqueness, atemporal, temporal plausibility
  - Data Quality Contexts
    - ***Verification***
    - ***Validation***

Data Quality Landmark Papers: Kahn et al 2016

Verification	Validation
Conformance: *Do data values adhere to specified standards and formats?*	
Value Conformance	
Data values conform to internal formatting constraints and value sets? *e.g., Sex is only ASCII char and values M, F, or U.*	Data values conform to representational constraints based on external standards *e.g., Values for primary language conform to ISO standards.*
Relational Conformance	
Data values conform to relational constraints and constraints to data model or versioning. *e.g., Patient MRN links to other tables as required.*	Data values conform to relational constraints based on external standards. *e.g., Data values conform to all not-NULL requirements in a common multi-institutional data exchange format.*
Computational Conformance	
Computed values conform to computational or programming specifications *e.g., Database- and hard- calculated BMI values are identical*	Computed results yield values that match validation values provided by external source *e.g., Computed BMI percentiles yield values similar to those provided by CDC*

PEDSnet
A Pediatric Learning Health System

Data Quality Landmark Papers: Kahn et al 2016

Verification	Validation
Completeness: *Are Data Values Present?*	
The absence of data values at a single moment (or over time) agrees with local expectations *e.g., Encounter ID variable has missing values*	The absence of data values at a single moment (or over time) agrees with trusted reference standards or external knowledge *e.g., A drop in ICD-9CM codes matches implementation of ICD-10CM*

Data Quality Landmark Papers: Kahn et al 2016

Verification	Validation
Plausibility: *Are data values believable?*	
Uniqueness Plausibility	
Data values that identify a single object or are not duplicated *e.g., Patients do not have multiple MRNs*	Data values that identify a single object in external source are not duplicated *e.g., A site CMS facility ID does not refer to multiple institutions*
Atemporal Plausibility	
Data values and distributions agree with internal measurement or local knowledge; independent measurements of the same fact are in agreement; repeated measurements of the same fact show expected variability *e.g., height values are positive; oral and axillary temperatures are similar*	Data values and distributions agree with an external source; *e.g., HbA1c values are the same at a site as national reference; readmission rates by age groups for Medicare patients agree with CMS values;*
Temporal Plausibility	
Observed or derived values conform to expected temporal properties; measures are expected based on internal knowledge *e.g., Admission date occurs before discharge date; ED visits spike during flu season*	Observed or derived values conform to values across *external* comparators; conform to *external* knowledge. *e.g., Immunization sequence match CDC recommendations; Medications per patient-day matches claims data*

Examples of DQ Checks from Kahn et al (2016)

Atemporal Plausibility	• 48% of labs outside of normal range
Temporal Plausibility	• Unexpected change in number of records from month to month
Completeness	• 62% of *route_concept_id* is missing
Value Conformance	• ICD9 codes in *condition_concept_id*
Relational Conformance	• *visit_date* and *visit_datetime* inconsistency in

Data Quality Landmark Papers: Weiskopf et al 2017

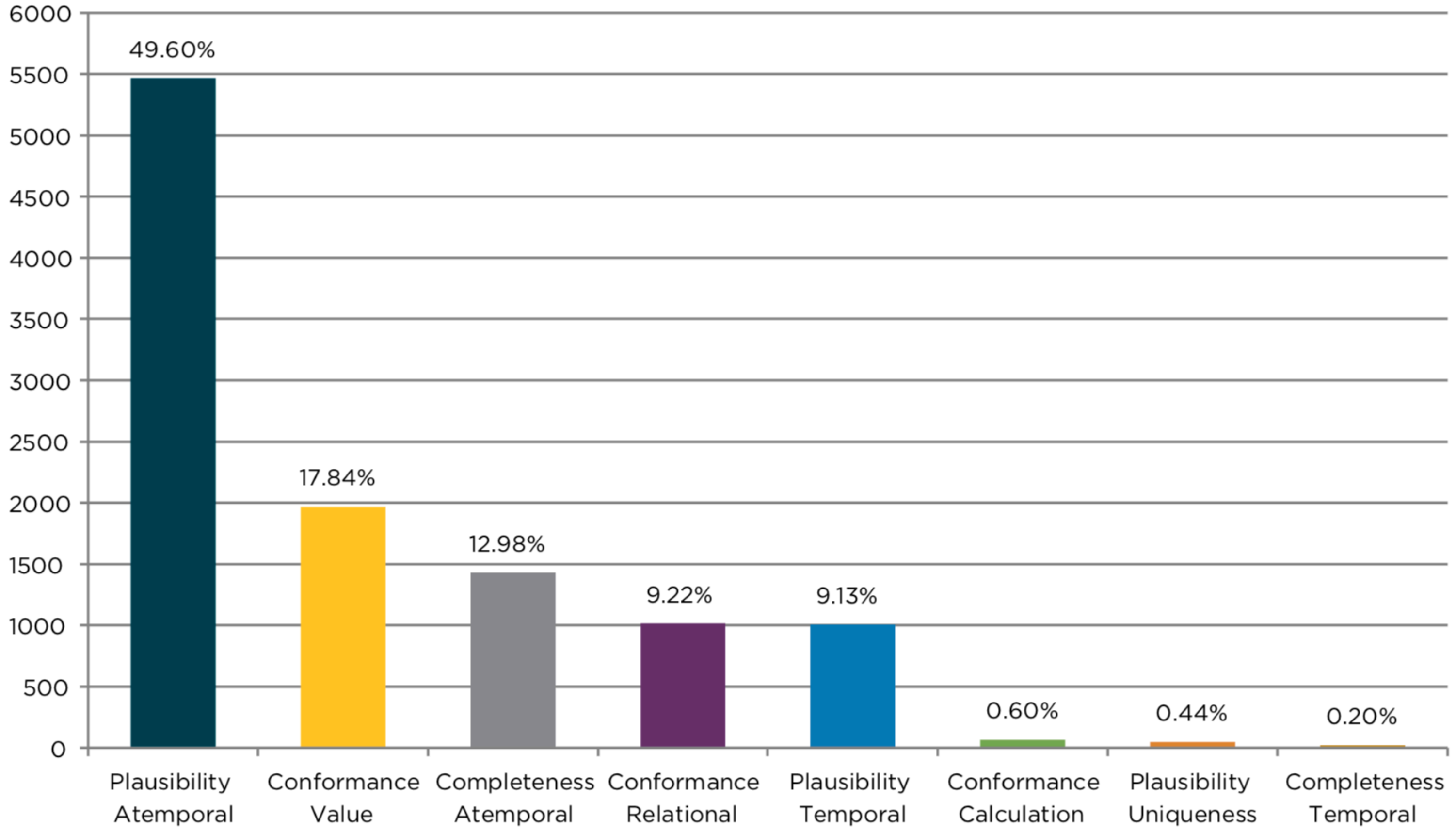
	Complete	Correct	Current
Patients	There are sufficient data points for each patient.	The distribution of values is plausible across patients.	All data were recorded during the timeframe of interest.
Variables	There are sufficient data points for each variable.	There is concordance between variables.	Variables were recorded in the desired order.
Time	There are sufficient data points for each time.	The progression of data over time is plausible.	Data were recorded with the desired regularity over time.

Data Quality Landmark Papers: Callahan et al 2017

- Compared DQA of 6 networks using the Kahn framework
  - **CESR:** Kaiser Permanente's Center for Effectiveness and Safety Research
  - **PHIS:** Pediatric Health Information System
  - **OHDSI:** Observational Health Data Sciences and Informatics
  - **MURDOCK:** Duke University School of Medicine's Measurement to Understand the Reclassification of Cabarrus/Kannapolis
  - **SENTINEL**
  - **PEDSnet**

Data Quality Landmark Papers: Callahan et al 2017

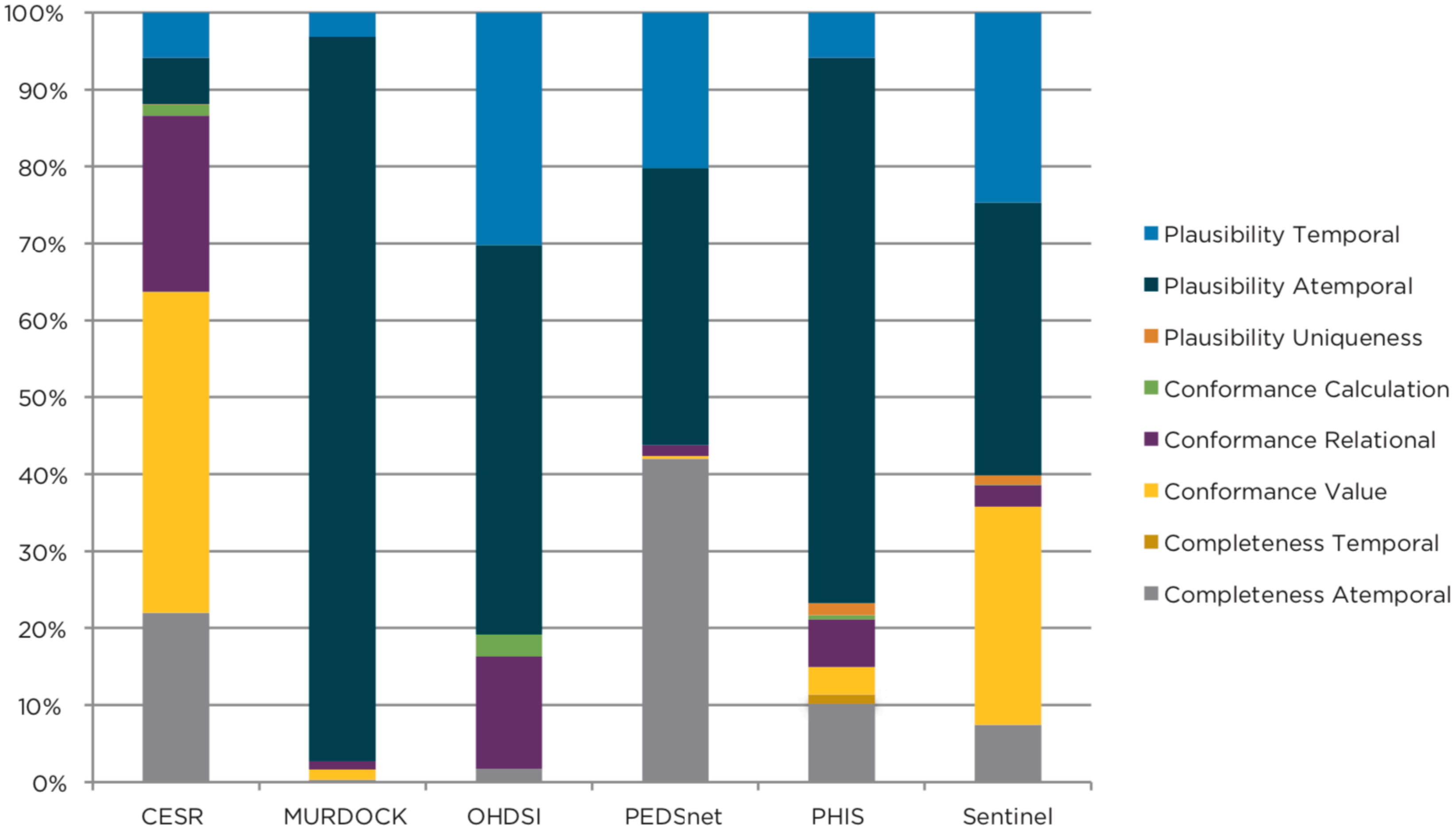
Figure 1. Harmonized DQA Terminology Mapped DQ Check Coverage



PEDSnet
A Pediatric Learning Health System

Data Quality Landmark Papers: Callahan et al 2017

Figure 2. Harmonized DQA Terminology Coverage of Mapped DQ Checks by Organization



PEDSnet
A Pediatric Learning Health System

Data Quality Landmark Papers: Callahan et al 2017

- Nearly 100% of the checks were in the *verification* context, not the *validation*
- Over 11,000 checks included
- Organizations vary in their approach to DQA, both in methodology as well as maturity