Methods for quality assessment in secondary use of EHR data

Nicole Weiskopf, PhD, Chunhua Weng, PhD, FACMI
Department of Biomedical Informatics
Columbia University

For Medical Informatics Initiative Germany
May 3rd, 2018, 8:30am-9:00am NYC Time
Outline

1. How shall we define EHR data quality issues?

2. How shall we measure EHR data quality, such as completeness and bias?

3. How shall we convert DQ assessment methods into actionable knowledge?

4. How can make DQ assessment more systematic?
EHR data reuse has many advantages

• Decreased cost and increased efficiency
  • Recruitment
  • Retention
  • Data collection
• Large volume of data
• Representative patient population
  • Patient-centered outcomes
EHR data are subject to quality problems

“What with the advent of the information era in medicine, we are pouring out a torrent of medical record misinformation. Medical records, which have long been faulty, contain more distorted, deleted, and misleading information than ever before.”

EHR data are subject to quality problems

“...the quality and comprehensiveness of the clinical data were not up to research standards or the analytical methods used to overcome these limitations were inadequate....”
What is data quality?

“Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features.”

Objective #1: to identify the most common dimensions of EHR data quality in the literature and map them these dimensions to methods of data quality assessment.
Literature review of EHR data quality assessment methods

Data collection:
- Literature review
- Inclusion criteria:
  - Original research using DQA methods
  - Data derived from EHR
  - Peer-reviewed
- Search: DQ terms & EHR terms
- Reviewed 230 articles
- Performed ancestor search
- Final pool: 95 relevant articles

## Dimensions of Data Quality Derived from Literature

<table>
<thead>
<tr>
<th>completeness</th>
<th>correctness</th>
<th>concordance</th>
<th>plausibility</th>
<th>currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>accessibility</td>
<td>accuracy</td>
<td>agreement</td>
<td>accuracy</td>
<td>recency</td>
</tr>
<tr>
<td>accuracy</td>
<td>corrections made</td>
<td>consistency</td>
<td>believability</td>
<td>timeliness</td>
</tr>
<tr>
<td>availability</td>
<td>misleading</td>
<td>reliability</td>
<td>trustworthiness</td>
<td></td>
</tr>
<tr>
<td>missingness</td>
<td>PPV</td>
<td>variation</td>
<td>validity</td>
<td></td>
</tr>
<tr>
<td>presence</td>
<td>quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quality</td>
<td></td>
<td></td>
<td>validity</td>
<td></td>
</tr>
<tr>
<td>rate of recording</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensitivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Dimensions of Data Quality Derived from Literature

<table>
<thead>
<tr>
<th>completeness</th>
<th>correctness</th>
<th>concordance</th>
<th>plausibility</th>
<th>currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>accessibility</td>
<td>accuracy</td>
<td>agreement</td>
<td>accuracy</td>
<td>recency</td>
</tr>
<tr>
<td>accuracy</td>
<td>corrections made</td>
<td>consistency</td>
<td>believability</td>
<td>timeliness</td>
</tr>
<tr>
<td>availability</td>
<td>misleading</td>
<td>reliability</td>
<td>trustworthiness</td>
<td></td>
</tr>
<tr>
<td>missingness</td>
<td>PPV</td>
<td>variation</td>
<td>validity</td>
<td></td>
</tr>
<tr>
<td>presence</td>
<td>quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quality</td>
<td>validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rate of recording</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensitivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Is a truth about a patient present in the EHR?*
### Dimensions of Data Quality Derived from Literature

<table>
<thead>
<tr>
<th>Completeness</th>
<th><strong>Correctness</strong></th>
<th>Concordance</th>
<th>Plausibility</th>
<th>Currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>Accuracy</td>
<td>Agreement</td>
<td>Accuracy</td>
<td>Recency</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Corrections made</td>
<td>Consistency</td>
<td>Believability</td>
<td>Timeliness</td>
</tr>
<tr>
<td>Availability</td>
<td>Misleading</td>
<td>Reliability</td>
<td>Trustworthiness</td>
<td></td>
</tr>
<tr>
<td>Missingness</td>
<td>PPV</td>
<td>Variation</td>
<td>Validity</td>
<td></td>
</tr>
<tr>
<td>Presence</td>
<td>Quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality</td>
<td>Validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate of recording</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Is an element that is present in the EHR true?
## Dimensions of Data Quality Derived from Literature

<table>
<thead>
<tr>
<th>completeness</th>
<th>correctness</th>
<th>concordance</th>
<th>plausibility</th>
<th>currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>accessibility</td>
<td>accuracy</td>
<td>agreement</td>
<td>accuracy</td>
<td>recency</td>
</tr>
<tr>
<td>accuracy</td>
<td>corrections made</td>
<td>consistency</td>
<td>believability</td>
<td>timeliness</td>
</tr>
<tr>
<td>availability</td>
<td>misleading</td>
<td>reliability</td>
<td>trustworthiness</td>
<td></td>
</tr>
<tr>
<td>missingness</td>
<td>PPV</td>
<td>variation</td>
<td>validity</td>
<td></td>
</tr>
<tr>
<td>presence</td>
<td>quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quality</td>
<td>validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rate of recording</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensitivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Is there agreement between elements in the EHR, or between the EHR and another data source?*
Dimensions of Data Quality
Derived from Literature

<table>
<thead>
<tr>
<th>completeness</th>
<th>correctness</th>
<th>concordance</th>
<th>plausibility</th>
<th>currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>accessibility</td>
<td>accuracy</td>
<td>agreement</td>
<td>accuracy</td>
<td>recency</td>
</tr>
<tr>
<td>accuracy</td>
<td>corrections made</td>
<td>consistency</td>
<td>believability</td>
<td>timeliness</td>
</tr>
<tr>
<td>availability</td>
<td>misleading</td>
<td>reliability</td>
<td>trustworthiness</td>
<td></td>
</tr>
<tr>
<td>missingness</td>
<td>PPV</td>
<td>variation</td>
<td>validity</td>
<td></td>
</tr>
<tr>
<td>presence</td>
<td>quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quality</td>
<td>validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rate of recording</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensitivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Does an element in the EHR makes sense in light of other knowledge about what that element is measuring?*
## Dimensions of Data Quality

**Derived from Literature**

<table>
<thead>
<tr>
<th>completeness</th>
<th>correctness</th>
<th>concordance</th>
<th>plausibility</th>
<th>currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>accessibility</td>
<td>accuracy</td>
<td>agreement</td>
<td>accuracy</td>
<td>recency</td>
</tr>
<tr>
<td>accuracy</td>
<td>corrections made</td>
<td>consistency</td>
<td>believability</td>
<td>timeliness</td>
</tr>
<tr>
<td>availability</td>
<td>misleading</td>
<td>reliability</td>
<td>trustworthiness</td>
<td></td>
</tr>
<tr>
<td>missingness</td>
<td>PPV</td>
<td>variation</td>
<td>validity</td>
<td></td>
</tr>
<tr>
<td>presence</td>
<td>quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quality</td>
<td>validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rate of recording</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensitivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Is an element in the EHR a relevant representation of the patient state at a given point in time?*
<table>
<thead>
<tr>
<th></th>
<th>Completeness</th>
<th>Correctness</th>
<th>Concordance</th>
<th>Plausibility</th>
<th>Currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold Standard</td>
<td>24</td>
<td>34</td>
<td></td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>Data Element Agreement</td>
<td>7</td>
<td>16</td>
<td>7</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>Element Presence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Data Source Agreement</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Distribution Comparison</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>Validity Checks</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Log Review</td>
<td>1</td>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td><strong>61</strong></td>
<td><strong>56</strong></td>
<td><strong>16</strong></td>
<td><strong>7</strong></td>
<td><strong>95</strong></td>
</tr>
</tbody>
</table>
Gold Standard

• A dataset drawn from another source or multiple sources, with or without information from the EHR, is used as a gold standard.
• Used for: correctness and completeness
• 39% of articles used a gold standard
Gold Standard


• Information from physicians (Lewis 2004, Lo Re 2009, van Staa 2000)


• Trained standard patients (Berner 2005, Peabody 2004)

• Alternate data sources (Madsen 1998, Staes 2006)

• Automatically recorded data (Vawdrey 2007)
Data Element Agreement

• Two or more elements within an EHR are compared to see if they report the same or compatible information.
• Used to assess completeness, correctness, concordance, and plausibility
• 26% of articles used agreement between data elements
Data Element Agreement


  - Data Quality Probes (Brown & Warmington 2002 & 2003)

Element Presence

• A determination is made as to whether or not desired or expected data elements are present.
• Used to assess completeness
• 24% of articles used element presence
Data Source Agreement

• Elements from the EHR are compared to data from other sources to determine if they are in agreement
  • Billing data (Roos 1989)
  • Shared data warehouse (Noel 2010)
  • Order system (Scobie 1995)
  • Survey data (Conroy 2005)

• Used to assess concordance

• 13% of articles used agreement between data sources
Distribution Comparison

- Distributions or summary statistics of aggregated data from the EHR are compared to the expected distributions for the clinical concepts of interest.

- Used to assess completeness, correctness, concordance, and plausibility

- 12% of articles used comparisons of data distributions
Validity Checks

• Data in the EHR are assessed using various techniques that determine if values “make sense.”
  • Changes in sequential data (Haerian 2009, Noel 2010)
  • End-digit preference (Alsanjari 2011, de Lusignan 2004)
  • Range checks (Basden 1980, Noel 2010, Staes 2006)
  • Zero value checks (Benson 2001)
  • Diagnoses appropriateness (Pearson 1996)

• Used to assess plausibility and correctness

• 8% of articles used validity checks
Log Review

• Information on data entry practices (e.g. dates, times, edits) is examined.

• Used to assess
  • Correctness (Benson 2001)

• 5% of articles used log review
Discussion: Terminology and Dimensions

• Inconsistent terminology
  • Overlap

• Agreement with previous frameworks of DQ
  • Wang & Strong (1996): match 4 of 15 dimensions
  • IOM (1997): match 2 of 3 relevant dimensions

• Fundamental vs. “proxy” dimensions
  • Concordance and plausibility
  • May be useful for when it’s not possible to directly assess fundamental dimensions
Discussion: DQ Assessment Methods

• Reliance on gold standards
  • GS often not available; de-identified databases
  • Not truly “gold”

• Most assessments relied upon ad hoc methods
  • Limited discussion of how to generalize methods

• Incorporation of “fitness for use”
  • Study and data needs should be made explicit prior to assessing DQ
Objective 2

Defining and measuring completeness of electronic health records for secondary use

Nicole G Weiskopf, George M Hripcsak, Alex Rusanov, Sushmita Swaminathan, Chunhua Weng
Completeness of EHR data is variable

• Hogan and Wagner (1997); 20 articles
  • Completeness: 1.1%-100%

• Thiru et al. (2003); 52 articles:
  • Sensitivity: 0.26 – 1.00

• Chan et al. (2010); 35 articles:
  • Completeness of BP: 0.1% – 51%

Why is there so much inconsistency?
Hypothesis:

EHR data completeness is task-dependent
A potential data point may be observed or unobserved...
...and recorded or unrecorded
Patient Record Representation

Data types

Time

Unobserved data point

* Observed data point

Desired data points

Observed data point

Unobserved data point

Data types

Time

Desired data points
A record is complete when it contains all observations
A record is complete when it contains all desired or expected types of data.
A record is complete when it contains a specified frequency of data points over time.
A record is complete when it contains sufficient information to predict a clinical phenomenon of interest.
Data

• NewYork-Presbyterian Hospital
  • Milstein Hospital, Allen Hospital, Morgan Stanley Children’s Hospital
• 300,000 unique patients per year
• 3.9 million unique patient records in electronic clinical data warehouse
• Population
  • 56% female
  • average age of 51 years
  • 32% Hispanic, 10% Asian, 19% Black, and 39% White
Completeness has increased with improved adoption of HIT

Data: narrative notes
Many visits are not accompanied by common data types

Breadth of Data by Data Type

- visits
- diagnoses
- labs
- meds

3.9 million patients

number of days

1 10 100 1000
By any definition, approximately half or more of all patient records in our EHR are incomplete
Overall, **55.7%** of patients in the CDW have at least one point of clinical data, and **26.9%** meet the criteria for at least one definition of completeness. Patients with **documentation complete records**—meaning they had at least one visit with an associated note—accounted for **18.5%** of all patients. In terms of **density**, only **11.8%** have a complete record when completeness is defined as at least 15 laboratory results or medication orders adjusted for temporal variance. When completeness is defined as a **breadth of five data types** of interest (date of birth, sex, medication order, laboratory test, and diagnosis), **11.4%** of patients have complete records. Finally, the presence or absence of a gap of 180 days or more could be correctly **predicted** for **8.4%** of patients. Only **0.6%** of patient records could be considered complete according to the implementations of all four definitions.
Hypothesis: sick patients are likely to have more complete records
Methods

• Population
  • 5,000 patients who have received anesthesia services

• Data
  • American Society of Anesthesiology (ASA) Physical Status Classification (1 to 6 scale; limited to 1 to 4)
  • Density of laboratory results (no time correction)
  • Density of medication orders (no time correction)
Distribution of ASA scores changes with threshold for completeness
Distribution of ASA scores changes with threshold for completeness

Threshold: 1 data point

Proportion of Complete Records

<table>
<thead>
<tr>
<th>ASA</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20%</td>
</tr>
<tr>
<td>2</td>
<td>80%</td>
</tr>
<tr>
<td>3</td>
<td>40%</td>
</tr>
<tr>
<td>4</td>
<td>0%</td>
</tr>
</tbody>
</table>
Distribution of ASA scores changes with threshold for completeness

Threshold: 3 data points
Distribution of ASA scores changes with threshold for completeness
Distribution of ASA scores changes with threshold for completeness

Threshold: 10 data points
Distribution of ASA scores changes with threshold for completeness

Threshold: 15 data points

Proportion of Complete Records

ASA

0% 20% 40% 60% 80%

1 2 3 4
Counts of available data points differ across ASA scores

**Medication Orders**

<table>
<thead>
<tr>
<th>ASA Score</th>
<th>Average Number of Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

**Laboratory Results**

<table>
<thead>
<tr>
<th>ASA Score</th>
<th>Average Number of Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
</tr>
</tbody>
</table>
Take Home:

• Be mindful of the potential limitations of a dataset prior to committing to its use
• Be explicit in your choice of completeness definition before assessing quality and suitability of a dataset
• Be transparent about data quality findings when reporting results
Take Home:

• Be mindful of the potential limitations of a dataset prior to committing to its use
• Be explicit in your choice of completeness definition before assessing quality and suitability of a dataset
• Be transparent about data quality findings when reporting results
Take Home:

• Be mindful of the potential limitations of a dataset prior to committing to its use
• Be explicit in your choice of completeness definition before assessing quality and suitability of a dataset
• Be transparent about data quality findings when reporting results
Take Home:

• Be mindful of the potential limitations of a dataset prior to committing to its use
• Be explicit in your choice of completeness definition before assessing quality and suitability of a dataset
• Be transparent about data quality findings when reporting results
EHR Data Quality Assessment Framework Based on Clinical Researcher Needs
Dimensions of EHR Data

“Completeness: breadth, depth, and scope of information contained in the data.”

## Data Quality Assessment Framework

<table>
<thead>
<tr>
<th></th>
<th>A: complete</th>
<th>B: correct</th>
<th>C: timely</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1: patients</strong></td>
<td>1A: There are sufficient data points for each patient.</td>
<td>1B: The distribution of values is plausible across patients for each type of data.</td>
<td>1C: All patients had data recorded during the timeframe of interest.</td>
</tr>
<tr>
<td><strong>2: variables</strong></td>
<td>2A: There are sufficient data points for each variable.</td>
<td>2B: There is concordance between different variables for each patient.</td>
<td>2C: Variables were recorded in desired order for each patient.</td>
</tr>
<tr>
<td><strong>3: time</strong></td>
<td>3A: There are sufficient data points for each time.</td>
<td>3B: The progression of data over time is plausible for each patient.</td>
<td>3C: Data points were recorded with the desired frequency over time for each patient.</td>
</tr>
</tbody>
</table>
Using DQA Framework

Study Design

Questions to Assess Data Requirements

Context

Decision

Assessment Methods

Action

<table>
<thead>
<tr>
<th></th>
<th>A: complete</th>
<th>B: correct</th>
<th>C: timely</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>There are sufficient data points for each patient.</td>
<td>The distribution of values is plausible across patients for each type of data.</td>
<td>All patients had data recorded during the timeframes of interest.</td>
</tr>
<tr>
<td>2A</td>
<td>There are sufficient data points for each variable.</td>
<td>There is concordance between different variables for each patient.</td>
<td>Variables were recorded in desired order for each patient.</td>
</tr>
<tr>
<td>3A</td>
<td>There are sufficient data points for each time.</td>
<td>The progression of data over time is plausible for each patient.</td>
<td>Data points were recorded with the desired frequency over time for each patient.</td>
</tr>
</tbody>
</table>
Extension of the Harmonized Data Quality Framework: Early Thoughts

Jimmy Rogers, MS; Chunhua Weng, PhD
Department of Biomedical Informatics, Columbia University
Current Practice: IT Centered DQ

“Do I apply all DQ rules? Or are only some of the DQ rules applicable?”

“Without domain knowledge and clinical or research tasks, how would I know what rules are relevant to measure the fitness of use of clinical data?”
Potential Stakeholder DQ Needs

• “I want to apply the same set of DQ rules that Michael applied to his database to my CDW and compare the quality of our CDWs”

• “I want to get a set of DQ rules for measuring the completeness of problem lists” (or replace problem lists with ICD-9 diagnosis codes)

• “find me all the DQ rules for diabetes”

• “find me all the DQ rules for breast cancer”

• “For this project, relevant DQ rules include completeness and correctness of timestamps; can I review and select from all related rules?”

• “DQ rules for elevated troponin may vary among different contexts; In clinical setting, elevated troponin has a threshold of 0.01 ng/ml, while in this research study, abnormal troponin should be > 0.04 ng/ml”

• “find me all the DQ rules identifying data errors, questionable data, potentially questionable data, respectively”
The Harmonized Framework for Data Quality
(Kahn MG et al, 2016)

- **Conformance**
  - Value → comply with pre-specified formatting constraints
  - Relational → comply with primary/foreign key relationships
  - Calculation → computational accuracy and feasibility

- **Completeness**
  - Atemporal → value present at a single point
  - Temporal → value present at multiple points across time

- **Plausibility**
  - Uniqueness → presence of duplicate values
  - Atemporal → values and distributions of values are feasible at a single point
  - Temporal → values and distributions of values are feasible across time

Table 2. DQ Check Coverage by DQA Category by Organization

<table>
<thead>
<tr>
<th>DQ HARMONIZATION TERMINOLOGY CATEGORIES</th>
<th>ORGANIZATIONS</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>TOTAL N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conformance</td>
<td>Value</td>
<td>CESR (41.76)</td>
<td>MURDOCK (1.34)</td>
<td>OHDSI (0.00)</td>
<td>PEDSnet (0.34)</td>
<td>PHIS (3.57)</td>
<td>SENTINEL (28.31)</td>
</tr>
<tr>
<td>Relational</td>
<td></td>
<td>786 (22.89)</td>
<td>36 (11.2)</td>
<td>25 (14.53)</td>
<td>13 (1.49)</td>
<td>114 (6.21)</td>
<td>42 (2.82)</td>
</tr>
<tr>
<td>Calculation</td>
<td></td>
<td>50 (1.46)</td>
<td>0 (0.00)</td>
<td>5 (2.91)</td>
<td>0 (0.00)</td>
<td>10 (0.54)</td>
<td>1 (0.07)</td>
</tr>
<tr>
<td>Completeness</td>
<td>Atemporal</td>
<td>754 (21.96)</td>
<td>9 (0.28)</td>
<td>3 (1.74)</td>
<td>367.5 (42.00)</td>
<td>186.5 (10.16)</td>
<td>111 (7.46)</td>
</tr>
<tr>
<td></td>
<td>Temporal</td>
<td>0 (0.00)</td>
<td>0 (0.00)</td>
<td>0 (0.00)</td>
<td>0 (0.00)</td>
<td>22 (1.20)</td>
<td>0 (0.00)</td>
</tr>
<tr>
<td>Plausibility</td>
<td>Uniqueness</td>
<td>1 (0.03)</td>
<td>0 (0.00)</td>
<td>0 (0.00)</td>
<td>0 (0.00)</td>
<td>29 (1.58)</td>
<td>18 (1.21)</td>
</tr>
<tr>
<td></td>
<td>Atemporal</td>
<td>207 (6.03)</td>
<td>3,031 (94.13)</td>
<td>87 (50.58)</td>
<td>315 (36.00)</td>
<td>1,300 (70.84)</td>
<td>527 (35.44)</td>
</tr>
<tr>
<td></td>
<td>Temporal</td>
<td>202 (5.88)</td>
<td>101 (3.14)</td>
<td>52 (30.23)</td>
<td>176.5 (20.17)</td>
<td>108 (5.89)</td>
<td>367 (24.68)</td>
</tr>
</tbody>
</table>

Provided DQ Checks: 3,434, 3,220, 172, 875, 1,835, 1,487, 1,487, 11,023
Towards A System for Indexing, Querying, and Retrieval of DQ Rules

• The harmonized DQ framework (Kahn et al.): establishes a good foundation with high-level concepts and can be extended

• The DQ ontology (Johnson et al.): focus on concept definition rather than support of real-world DQ application and utilization

• A Data Quality Assessment Guideline for EHR data reuse (Weiskopf et al.): focus on decision support and rule selection for various tasks, not on rule management

All expert-driven, not data driven
Can we enrich DQ rules with

- adoption status
- Reference standard
- creator
- disease domain
- source
- Measurement thresholds
- lab name
- Error type
- Data source
- Error cause

...Can add more metadata tags
Our Goal

• A knowledge base of human reviewable and machine executable DQ rules for
  • DQ Knowledge management and sharing
  • Stakeholders to perform knowledge-based DQ

• Supports the following tasks around DQ rules
  • **Indexing**: how can we organize rules in an understandable way?
  • **Retrieval**: which DQ rules suits my task?
  • **Auditing**: how are rules related? Is there redundancy? When will be a new rule be necessary?
  • **Machine learning**: generate rules systematically and automatically?

• Be
  • Stakeholder friendly
  • Fine grained
  • extensible
References


3. Weiskopf NG, Bakken S, Hripcsak G, *Weng C, A Data Quality Assessment Guideline for Electronic Health Record Data Reuse*, eGEMS, accepted


6. Huser, Vojtech; DeFalco, Frank J.; Schuemie, Martijn; Ryan, Patrick B.; Shang, Ning; Velez, Mark; Park, Rae Woong; Boyce, Richard D.; Duke, Jon; Khare, Ritu; Utidjian, Levon; and Bailey, Charles (2016) *Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets*, eGEMs: Vol. 4: Iss. 1, Article 24. DOI: [http://dx.doi.org/10.13063/2327-9214.1239](http://dx.doi.org/10.13063/2327-9214.1239)


