Data Quality Assessment

Recommendations

Steering Committee Meeting, August 19-20, 2014
Bethesda, MD

Meredith Nahm Zozus, PhD
Assoc. Director, Clinical Research Informatics
Duke Translational Medicine Institute
Initiated by

1. Inventory of data sources and data quality assessment plans proposed in first round UH2 applications,

2. UH3 review criteria requiring data validation
Guiding Principles

• Need to demonstrate that data on which conclusions are based are capable of supporting them

• Should not assume use of a common data model for individual research projects

• Recommendations should be practical and reasonably achievable
Iterative Development

- Started with literature review
- Four versions reviewed by the PDSDQ core
- Last version/s reviewed by leadership of the EHR and Biostatistics Cores
- Presented to a PCORI data quality working group
- Reviewed by HCS Collaboratory for publication release
- Released in July 2014
Located in Knowledge Repository
Recommendations

1 - Key data quality dimensions to be measured

2 - Description of formal of assessments

3 - Reporting data quality assessment with research results
Recommendation 1

Accuracy, completeness, and consistency be formally assessed for data elements used in subject identification, outcome measures, and important covariates.

Why? *These are most impactful on the ability of data to support research conclusions.*
Recommendation 2

Specifics for measuring accuracy, completeness and consistency

**Completeness assessment recommendation:**
- four-part completeness assessment.
- Same column and data value completeness measures can be employed for monitoring completeness during the study.
- The completeness assessment applies to both prospectively collected and secondary use data.
- Additional requirements suggested by the Good Clinical Data Management Practices (GCDMP) document, such as on-screen prompts for missing data where appropriate, apply to data collected prospectively for a study.
An accuracy assessment requires selecting a source for comparison, making the comparison, and then quantifying the results. In Figure 1, sources for comparison are listed in descending order of their proximity to truth. If there are multiple options, those sources for comparison toward the top of the list in Figure 1 are preferred because the sources for comparison are closer to the truth. Thus, sources for comparison toward the top provide quantitative assessments of accuracy, whereas sources for comparison in the middle provide partial measures of accuracy and, depending on the data source used for the comparison, may enable identification of errors or may only indicate discrepancies. Sources for comparison toward the bottom identify only data discrepancies, i.e., items that may or may not represent an actual error. For example, if it has been shown that a percentage of missing values is

<table>
<thead>
<tr>
<th>Comparison to a source of “truth”</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison to an independent measurement</td>
<td>Partial accuracy</td>
</tr>
<tr>
<td>Comparison to independently managed data</td>
<td>Discrepancy detection</td>
</tr>
<tr>
<td>Comparison to an upstream data source</td>
<td></td>
</tr>
<tr>
<td>Comparison to a known standard</td>
<td></td>
</tr>
<tr>
<td>Comparison to valid values</td>
<td></td>
</tr>
<tr>
<td>Comparison to validated indicators</td>
<td></td>
</tr>
<tr>
<td>Comparison to aggregate statistics</td>
<td>Gestalt</td>
</tr>
</tbody>
</table>

The strength of the accuracy assessment depends not only on the proximity to truth of the source for comparison, but also on the importance and number of data elements for which accuracy can be assessed. Accuracy assessments are often performed on subsets of data elements or subsets of the population, rather than across the whole dataset.

Recommendation 2 cont.

The highest practical accuracy assessment in the hierarchy should be used.
Recommendation 2 cont.

**Consistency assessment recommendation:** Identification of:

a) areas where differences in clinical documentation, data collection, or data handling may exist between individuals, units, facilities, sites, or assessors, or over time and

b) measures to assess consistency and monitor it throughout the project.

- A systematic approach to identifying candidate consistency assessments should be used. Such an approach will likely be based on review of available data sources, accompanied by an approach for systematically identifying and evaluating the likelihood and impact of possible inconsistencies.
- This recommendation applies to both prospectively collected data and secondary use data.
**Recommendation 2 cont.**

**Impact assessment recommendation:** Use of completeness, accuracy, and consistency assessment results by the project statistician to test sensitivity of the analyses to anticipated or identified data quality problems, including a plan for reassessing based on results of data quality monitoring throughout the project.
Recommendation 3

Data quality assessments should be reported with research results.
Acknowledgements

The work presented here was funded by the Health Care Systems Research Collaboratory Coordinating Center grant number 1U54AT007748-01 through the National Center for Complementary & Alternative Medicine, a center of the National Institutes of Health.