NIH Collaboratory *Rethinking Clinical Trials®*

Health Care Systems Research Collaboratory

Assessing Data Quality of Clinical Data for PCTs

Background

The credibility and reproducibility of pragmatic clinical research depends on the investigator's demonstration that the data are of sufficient quality to support the research conclusions. This document highlights recommendations for assessing the quality of data generated from routine patient care for use in PCTs. The full version of this white paper, along with a full list of references, and other guidelines are available on *Rethinking Clinical Trials®: A Living Textbook of Pragmatic Clinical Trials*,.

Dimensions of Data Quality Assessment

Accuracy, completeness, and consistency closely affect the capacity of data to support research conclusions (Table 1).

Table 1. Data Quality Dimensions Determining Fitness for Use of Research Data

Dimension	Conceptual definition	Operational examples
Completeness	Presence of the necessary data	Presence of necessary data elements, percent of missing values for a data element, percent of records with sufficient data to calculate a required variable (e.g., an outcome)
Accuracy	Closeness of agreement between a data value and the true value*	Percent of data values found to be in error based on a gold standard, percent of physically implausible values, percent of data values that do not conform to range expectations
Consistency	Relevant uniformity in data across clinical investigation sites, facilities, departments, units within a facility, providers, or other assessors	Comparable proportions of relevant diagnoses across sites, comparable proportions of documented order fulfillment (e.g., returned procedure report for ordered diagnostic tests)

**Consistent with the International Organization for Standardization (ISO) 8000 Part 2 definition of accuracy, replaced "property value" in the ISO 8000 definition with "data value" for consistency with the language used in clinical research*

Data Quality Assessment Recommendations for PCTs

1 - Key data quality dimensions

We recommend that accuracy, completeness, and consistency be formally assessed for data elements used in subject identification, outcome measures, and important covariates

2 - Description of formal of assessments for completeness, accuracy, consistency, and impact

See full paper for details and options. See below for different approaches to assess accuracy.

3 – Reporting data quality assessment with research results

Results of data quality assessments should be reported with research results. Data quality assessments are the only way to demonstrate that data quality is sufficient to support the research conclusions, and as such should be accessible to consumers of research.

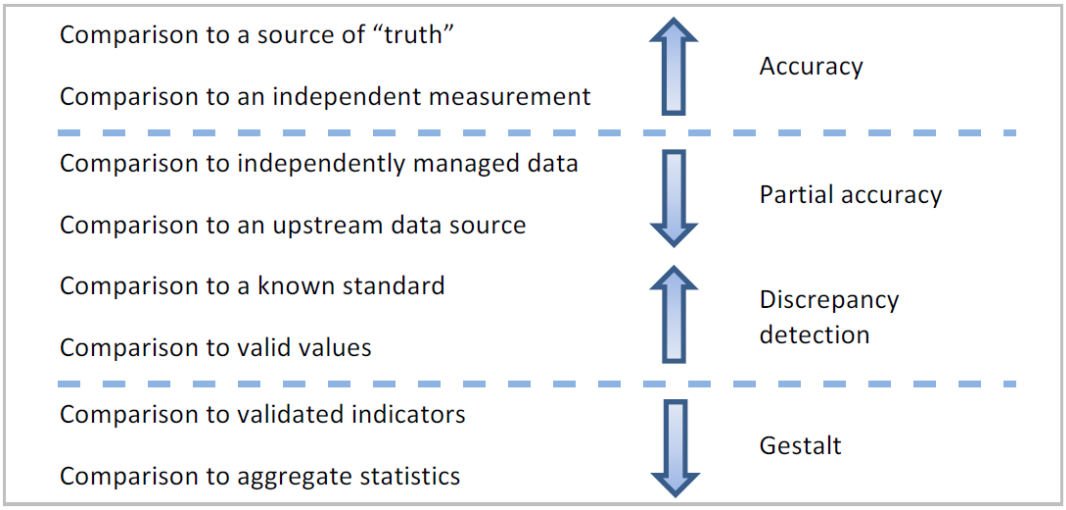
Use of workflow and data flow diagrams to inform data quality assessment

We encourage the creation and use of data flow and workflow diagrams to aid in identifying accuracy and in conducting consistency assessments. If that is not practical, the following questions could be reviewed with personnel at each research site.

1. Talk through each of the data elements used for cohort identification. Explain how and where each is documented in the clinic or unit (i.e., what information system, what screen, at what point in the clinical process, and by whom)?
2. When you send us the data or connect data to a federated system, what data store will you create/use? Describe all data transformations.
3. For each data element used in the cohort identification, are there difference in data capture or documentation practices across clinics or for different subsets of your population?
4. For each data element used in cohort identification, are there any subsets of data that may be documented differently, such as data from specialist or hospital reports external to your group versus data from your practice, or internal vs. external clinical laboratories?

Data Accuracy Assessment Approaches - Comparison Hierarchy

Comparison of data to sources listed above the top line provides full assessment of data accuracy; sources listed below the top line provide only partial assessments of accuracy. Sources above the bottom line can be used to detect actual data discrepancies, whereas sources below the bottom line can only indicate that discrepancies may exist. Items at the top of the list identify actual errors, whereas items in the middle only identify discrepancies that may or may not in fact be an error. Items toward the bottom merely indicate that discrepancies may exist.



This work was supported by a cooperative agreement (U54 AT007748) from the NIH Common Fund for the NIH Health Care Systems Research Collaboratory. The views presented here are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.