NIH Collaboratory *Rethinking Clinical Trials®*

Health Care Systems Research Collaboratory

Assessing Data Quality for Healthcare Systems Data Used in Clinical Research

Table of Contents

Objective

Quality assessment of healthcare data used in clinical research is a developing area of inquiry. The methods used to assess healthcare data quality in practice are varied, and evidence-based or consensus "best practices" have yet to emerge.[1] Further, healthcare data have long been criticized for a plethora of quality problems. To establish credibility, studies that use healthcare data are increasingly expected to demonstrate that the quality of the data is adequate to support research conclusions.

Pragmatic clinical trials (PCTs) in healthcare settings rely upon data generated during routine patient care to support the identification of individual research subjects or cohorts as well as outcomes. Knowing whether data are accurate depends on some comparison, e.g., comparison to a source of "truth" or to an independent source of data. Estimating an error or discrepancy rate, of course, requires a representative sample for the comparison. Assessing variability in the error or discrepancy rates between multiple clinical research sites likewise requires a sufficient sample from each site. In cases where the data used for the comparison are available electronically, the cost of data quality assessment is largely based on time required for programming and statistical analysis. However, when labor-intensive methods such as manual review of patient charts are used, the cost is considerably higher. The cost of rigorous data quality assessment may in some cases present a barrier to conducting PCTs. For this reason, we seek to highlight the need for more cost-effective methods for assessing data quality.

PRAGMATIC CLINICAL TRIAL (PCT): We use the definition articulated by the Clinical and Translational Science Awards pragmatic clinical trials infrastructure (PCTi) workshop: "A prospective comparison of a community, clinical, or system-level intervention and a relevant comparator in participants who are similar to those affected by the condition(s) under study and in settings that are similar to those in which the condition is typically treated."[2]

Because of the potential cost implications and the fear of taking the "pragmatic" out of PCTs, we find it difficult to make these recommendations. However, the principles underlying recommendations for applying data quality assessment to research that uses healthcare data are irrefutable. The credibility and reproducibility of research depends on the investigator's demonstration that the data on which conclusions are based are of sufficient quality to support them. Thus, the objective of this document is to provide guidance, based on the best available evidence and practice, for assessing data quality in PCTs conducted through the National Institutes of Health (NIH) Health Care Systems Research Collaboratory.

Prepared by: Meredith N. Zozus, PhD; W. Ed Hammond, PhD; Beverly B. Green, MD, MPH; Michael G. Kahn, MD, PhD; Rachel L. Richesson, PhD, MPH; Shelley A. Rusincovitch; Gregory E. Simon, MD, MPH; Michelle M. Smerek. Reviewed by: NIH Version: 1.0, last updated July 28, 2014
Collaboratory Phenotypes, Data Standards, and Data Quality Core

The NIH Health Care Systems Research Collaboratory

The NIH Health Care Systems Research Collaboratory (http://www.nihcollaboratory.org) or "Collaboratory" is intended to improve the way clinical trials are conducted by creating new approaches, infrastructure, and methods for collaborative research. To develop and demonstrate these methods, the Collaboratory also supports the design and rapid execution of high-impact PCT Demonstration Projects that 1) address questions of major public health importance and 2) engage healthcare delivery systems in research partnership. Organizationally, the Collaboratory comprises a series of these Demonstration Projects funded for 1 planning year, with competitive renewal to allow transition into actual trial conduct, and a Coordinating Center to provide support for these efforts. Within the Coordinating Center, seven Working Groups/Cores serve to identify, develop, and promote solutions for issues central to conducting PCTs: 1) electronic health record use in research; 2) phenotypes, data standards, and data quality; 3) patient-reported outcomes; 4) healthcare system interactions; 5) regulatory and ethical issues; 6) biostatistics and study design; and 7) stakeholder engagement. The Cores have the bidirectional objectives of promoting the exchange of information on methods and approaches among Demonstration Projects and the Coordinating Center, as well as synthesizing and disseminating best practices derived from Demonstration Project experiences to the larger research community. Supported by the NIH Common Fund, the Collaboratory's ultimate goal is to ensure that healthcare providers and patients can make decisions based on the best available clinical evidence.

The Collaboratory provides an opportunity to observe data quality assessment plans and practices for PCTs conducted in healthcare settings. The Collaboratory's Phenotypes, Data Standards, and Data Quality (PDSDQ) Core[3] includes representatives from the Collaboratory Coordinating Center and Demonstration Projects, researchers with related interests, and NIH staff. In keeping with the bidirectional goals of the PDSDQ Core, an action research paradigm was used in which the Core interacted with Demonstration Projects, observed data quality assessment plans and practices, participated where invited, and synthesized experience to generalize information for others embarking on similar research. We report here the observations and iteratively developed (and still-evolving) data quality assessment methodology from the initial planning grant year for the Collaboratory's first seven Demonstration Projects. These results have been vetted by the PDSDQ Core and other Collaboratory participants and represent the experience of this group at the time of development; however, they do not represent official NIH opinions or positions.

Data Quality Assessment Background

Depending on the scientific question posed by a given study, PCTs may rely on data generated during routine care or on data collected prospectively for the study. Therefore, data quality assurance and assessment for such studies necessarily includes methods for both situations: 1) collection of data specifically for a study, where the researcher is able to influence or control the original data collection and 2) use of data generated in routine care, where the researcher has little or no control over the data collection. For the former, significant guidance is available via the Good Clinical Data Management Practices (GCDMP) document,[4] and we do not further discuss quality assurance and assessment methods for these types of prospective research data. Instead, this guidance will focus on the use or re-use of data generated from routine patient care, based on the following:

1. Existing literature on data quality assessment for healthcare data that are re-used for research
2. Experience during the first year of the Collaboratory

In this document, we rely on a multidimensional definition of data quality. The dimensions of accuracy and completeness are the most commonly assessed in health-related research.[5] A recent review identified five dimensions that have been assessed in electronic health record (EHR) data used for research; they include completeness, correctness, concordance, plausibility, and currency.[6] Accuracy, completeness, and consistency (Table 1) most closely affect the capacity of data to support research conclusions and are therefore the focus of our discussion here. A brief review of the literature on defining data quality is provided in Appendix I, and specific dimensions used here are defined below. Unfortunately, definitions of data quality dimensions are highly variable in the literature. The sections below outline conceptual definitions of these dimensions followed by operational examples.

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,[7] replaced "property value" in the ISO 8000 definition with "data value" for consistency with the language used in clinical research.*

Based on the literature relevant to data quality assessment in the secondary use of EHR data and our experience thus far with the Collaboratory (described in Appendices II and III), we offer a set of data quality assessment recommendations for Collaboratory projects. First, we summarize important dimensions, common or reported approaches to characterizing them, and characteristics of an ideal operationalization. Next, we streamline specific recommendations for researchers using data generated from routine care.

Data Quality Assessment Dimensions

Completeness

Conceptually, completeness is the presence of necessary data. The operationalization of completeness presented below was adapted from recent theoretical work by Weiskopf et al.,[8] in which a comprehensive assessment of completeness covers four mutually exclusive areas:

1. **Data element completeness:** The presence of all necessary variables in a candidate dataset; i.e., "Are the right 'columns' present?" Data element completeness is assessed by examining metadata, such as a data dictionary or list of data elements contained in a dataset and their accompanying definitions, and comparing this information against the variables required in the analytic or statistical plan. With adequate data documentation, data element completeness can be assessed without examining any data values.
2. **"Column" data value completeness:** The percentage of data values present for each data element. Note, however, that often (as in normalized structures) more than one data element may be stored in a database column. The word *column* is used to help the reader visualize the concept and because normalized data structures are often flattened to a 1-column-per-data-element format to generate and report data quality–related statistics. Column data value completeness is assessed by structuring the dataset in a "1-column-per-data-element" format and calculating the percentage of non-missing data for each column, with non-missing defined as "not null and not otherwise coded to a null flavor." Null flavors (e.g., not applicable, not done) are defined in the International Organization for Standardization (ISO) 21090[9] and Health Level Seven International (HL7)[10] data type definition standards.
3. **Ascertainment completeness:** The percentage of eligible cases present; i.e., "Do you have the right 'rows' in the dataset?" Ascertainment usually cannot be verified with absolute certainty. Assessment options are typically comparison based and include but are not limited to: 1) chart review in a representative sample and 2) comparison to one or more independent data sources covering the same population or a subset of that population. Ascertainment completeness is affected by data quality problems, by phenotype definition and execution, and by factors that bias membership of a dataset. Other issues commonly evaluated in an ascertainment

assessment include the presence and extent of duplicate records and records for patients that do not exist (for example: an error in the medical record number creates a new case; a patient gives a name other than his or her own), or duplicate events such as a single procedure being documented more than once. Ascertainment completeness and phenotype validation significantly overlap in goals and can be accomplished together.

4. **"Row" data value completeness:** The percentage of cases/patients with sufficient data values present for a given data use. Row data value presence is assessed using study-specific algorithms programmed to calculate the percentage of cases with all data or with study-relevant combinations of missing and non-missing data (e.g., in the case of body mass index [BMI], the percent missing of "either weight OR height" might be calculated, because missing either data point renders the case unusable for calculating BMI).

DATA ACCURACY: The closeness of agreement between a data value and the true value.
—adapted from ISO 8000[6]

A comprehensive completeness assessment consists of all four components. In terms of effort, column completeness is accomplished through a review of data elements available in a data source, and column data value completeness and row data value completeness are straightforward computational activities. Ascertainment completeness, however, can be a resource-intensive task (e.g., chart review on a representative sample; electronic comparison among several data sources). Additional guidance and discussion regarding data completeness in the setting of EHR data extracted for pragmatic clinical research is available here.[11]

Completeness, although necessary to establish fitness for use in clinical research, is not sufficient to evaluate the competence of a dataset to support research conclusions. Assessment of accuracy and consistency are also necessary.

Accuracy

In keeping with ISO 8000 standards,[7] we define data accuracy as the property exhibited by a data value when it reflects the true state of the world at the stated or implied point of assessment. It follows that an inaccurate or errant datum does not reflect the true state of the world at the stated or implied point of assessment.[12] Data errors are instances of inaccuracy.

Detection of data errors is accomplished through comparison; for example, comparison of a dataset to some other source of information (Figure 1). The comparison may be between the data value and a "source of truth," a known standard, a set of valid values, a redundant measurement, independently collected data for the same concept, an upstream data source, a validated indicator of possible errors, or aggregate statistics. As the source for comparison moves farther from a "source of truth," we move from identification of data errors to indications that a datum may be in error.

We use the term *error* to denote any deviation from accuracy regardless of the cause. For example, a programming problem in data transformation that renders an originally accurate value incorrect is considered to have caused a data error. Because data are subject to multiple processing steps, some count the number of errors (consider, for instance, a data value that has sustained two problems that each would have individually caused an error for a total of two errors). From an outcomes perspective, it is the number of fields in error that matters rather than the number of errors; thus, in data quality assessment, the number of data values in error is counted rather than the number of errors. Different agreement statistics may be applied depending on whether the source of comparison is considered a source of truth or gold standard versus an independent source of information.

Operationally, an instance of inaccuracy or data error is any discrepancy identified through such a comparison that cannot be explained by documentation.[12,13] The caveat "not explained by documentation" is important because efforts to identify data discrepancies (i.e., potential errors) can be undertaken on data at different stages of processing. Such processing sometimes includes transformations on the data such as imputations that purposefully change the value. In these cases, a data consumer should expect the changes to be supported by documentation and be traceable through all of the data processing steps.

REPRESENTATIONAL INADEQUACY: The degree to which a data element differs from the desired concept.

Accuracy has been described in terms of two basic concepts: 1) representational adequacy/inadequacy, defined as the extent to which an operationalization is consistent with/differs from the desired concept (validity), including but not limited to imprecision or semantic variability, hampering interpretation of data and 2) information loss and degradation, including but not limited to reliability, change over time, and error.[14]

Representational inadequacy is the degree to which a data element differs from the desired concept. For example, a researcher seeking obese patients for a study uses BMI to define the obesity phenotype, knowing that a small percentage of bulky but lean bodybuilders may be included. Representational inadequacy is best addressed at the point in research design when data elements and sources are selected.

Representational inadequacy can be affected by local work and data flows of data elements used in a study, e.g., differences in local coding practices causing differences in datasets across institutions. Thus, harmonization of data elements across sites is emphasized in NIH review criteria for Collaboratory PCTs (Appendix I). Documenting work and data flows for each data element, from the point of origin to the analysis dataset (traceability), has long been required in regulated research,[4] reported as a best practice in the information quality literature, and implemented in healthcare settings.[15] Comparisons of data definitions,

INACCURACY/DATA ERROR: Any discrepancy that cannot be explained by documentation.

workflows, and data flows across research sites are as important in assessing representational inadequacy of healthcare data as is the use of validated questionnaires in assessing subjective concepts. Some differences in workflow will not affect representation, while others may; the only way to know is to understand the workflow at each site and evaluate the effect, if any, of representation. Such documentation for data collected in healthcare settings may not be as precise as that for clinical trial data collection processes. For example, it can be difficult to assess the data capture process of patient-reported outcomes (PROs) in healthcare settings due to differences in individual departments, clinics, and hospitals within an individual healthcare organization. The workflow can also vary over time as refinements are made.

Results of such assessments for representational inadequacy are often qualitative and used either as formative assessments in research design or to describe limitations in reported results. Comparisons of aggregate or distributional statistics (e.g., marginal) as performed by the Observational Medical Outcomes Project (OMOP),[16] have also been used to identify representational variations in datasets caused by differences in local practice among the institutions providing data.[16] Using both process-oriented and data-based approaches in concert to confirm representational adequacy of data elements is recommended. A process-oriented approach may be used at the time of site selection; once consistency is confirmed, a data-based approach may be used to monitor consistency during the study.

Information loss or degradation is the loss of information content over time and can arise from errors or purposeful decisions in data collection and processing (for example: data reduction such as interval data collected as ordinal data; separation of data values from contextual data elements; or data values that lose accuracy or relevance over time). Information loss and degradation may be prevented or mitigated by decisions made during research design. Because such errors and omissions are sensitive to many organizational factors (e.g., local clinical documentation practices, mapping decisions made for warehoused data), they should be assessed for any data source. Thus, workflow and data flow documentation also help to assess sources of information loss and degredation.[15]

Assessing data accuracy, primarily with regard to information loss and degradation, involves comparisons, either of individual values (as is commonly done in clinical trials[14] and registries[5]) or of aggregate or distributional statistics.[14,16-18]

1. **Individual value comparisons:** At the individual-value level, the comparison could be to the truth (if known), to an independent measurement, to a validated indicator, or to valid (expected, physically plausible, or logically consistent) values.[14,17,19] In practice, the options for comparison (Figure 1) represent a continuum from truth to measurements of lesser proximity to the truth, such as an accepted gold standard or valid values. Thus, accuracy assessment usually provides a disagreement rate, and much less often, an actual error rate. Further, in some prospective settings,[4] the identification of data discrepancies is done for the purpose of resolving them; in other settings, where data correction is not possible, data discrepancies are

identified for the purpose of reporting a data discrepancy rate or to inform statistical analysis.

2. **Aggregate and distributional comparisons:** Aggregate and distributional comparisons (such as frequency counts or measures of central tendency and dispersion) can be used as a surrogate accuracy assessments. For example, differences in aggregate or distributional measures between a research dataset and an independent data source with a similar population may indicate possible data discrepancies, while similar measures would increase confidence in the research data. Differences in central tendency and dispersion measures in age or a socioeconomic status measure may indicate significant differences in the populations in two data sets. Aggregate and distributional comparisons can be also be performed within a dataset,[16-18] between multiple sites in a multicenter study,[17,18] or between subgroups as measures of consistency.

In the absence of a source of truth, comprehensive accuracy assessment of multisite studies includes use of individual value, aggregate, and distributional measures.[17] To emphasize the importance of these within and between dataset comparisons, a third dimension, consistency (described below), was added. The difference between the two dimensions here lies not in the measures, but in the purpose of the comparisons and in the choice of data on which to run them.

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 inversely correlated with data accuracy, then percent missing may be an indicator of lower accuracy.

The hierarchy of sources for comparison shown in Figure 1 provides a list of possible comparisons ranging (from bottom to top) from those that are achievable in every situation but provide less information about true data accuracy, to the ideal but rarely achievable case that provides an actual data error rate. This hierarchy simplifies the selection of sources for comparison: where more than one source for comparison exists, the highest practical comparison in the list should be used.

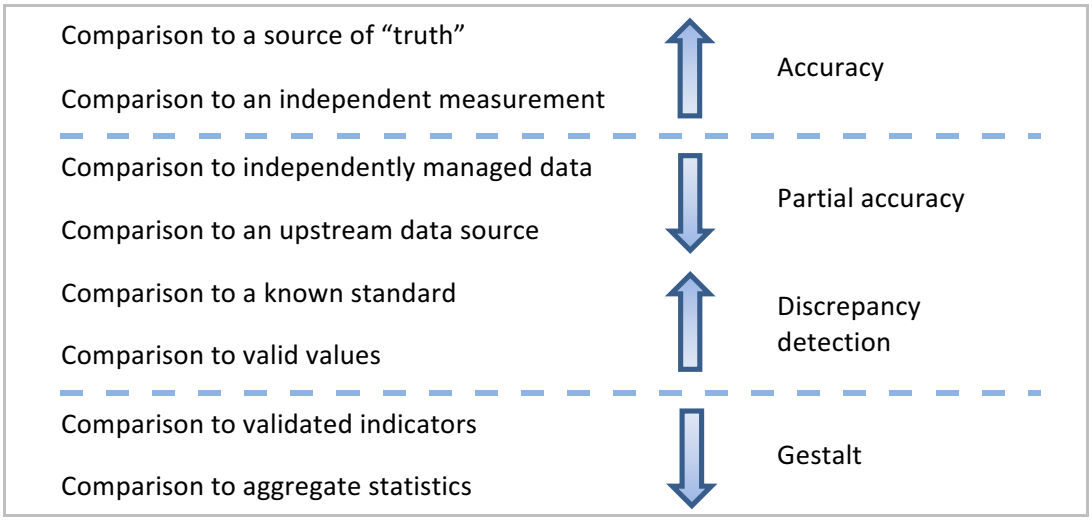

Figure 1. Data Accuracy Assessment Comparison Hierarchy. *Comparison of data to sources listed above the top line provides full assessment of data accuracy; sources listed below the top line provide only partial assessments of accuracy. Sources above the bottom line can be used to detect actual data 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.*

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. Common subsets assessed include data elements used in subject or cohort identification, data elements used to derive clinical outcomes, and patients for whom an independent source of data (such as registry or Medicare claims data) is readily available for comparison. Accuracy assessments should be done for cohort identification data elements, outcome data elements, and covariates. Accuracy assessments for a given study may use different sources for comparison.

Comparisons for data accuracy assessments will likely differ based on the underlying nature of the phenomena about which the data values were collected. Examples of different phenomena include anatomic or pathologic phenomena, physiologic or functional phenomena, imaging or laboratory findings, patients' symptomatic experiences, and patients' behaviors or functioning. The data values collected about these phenomena may be the result of inherently different processes, including but not limited to measurement of a physical quantity, direct observation, clinical interpretation of available information, asking patients directly, or psychometric measurements. These are not complete lists, and we do not provide a deterministic map of phenomena and measurement processes to associated error sources. We simply note that different phenomena and measurement or collection processes are sometimes characteristically prone to different sources of error.

Such associations should be considered when data elements and comparisons for data quality assessment are chosen.

Consistency

Consistency is defined here as relevant uniformity in data across clinical investigation sites, facilities, departments, units within a facility, providers, or other assessors. Inconsistencies, therefore are instances of difference. In other frameworks,[18,20,21] the label *consistency* is used for several different things, such as uniformity of data over time or conformance of data values to other values in the dataset (e.g., gender correctness of gender-specific diagnoses and procedures, procedure dates before discharge date). Here, we view these valid value comparisons as surrogate indicators of accuracy (Figure 1).

There are many ways that data can be inconsistent; for example, clinical documentation policies or practices may vary over time within a facility, between facilities, or between individuals in a facility. Consider a study where the outcome measure is whether or not patient behavior regarding medication taking changes. If some sites document filled prescriptions from pharmacy data sources while others rely on patient reporting, the outcome measure would be inconsistent between the sites. Actions should be taken to improve similarity in documentation or to use other documentation that is common across all sites. Otherwise, such inconsistencies may introduce bias and affect the capacity of the data to support study conclusions. Thus, the consistency dimension comes into play particularly in multisite or multifacility studies and when such differences may exist in clinical documentation, data collection, or data handling within a study. Comparisons of multisite data over time to examine expected and unexpected changes in aggregate or distributional data can also be useful. For example, changes in EHR systems, such as new data being captured, data no longer being captured, or even implementation of a new system, are commonplace and affect data. Assessing consistency during a study (data quality monitoring) is the only way to ensure that such changes will be detected.

CONSISTENCY: Relevant uniformity in data across clinical investigation sites, facilities, departments, units within a facility, providers, or other assessors.

Targeted consistency assessments are important during the feasibility-assessment phase of study planning. For example, to ascertain whether data are sufficiently consistent across facilities to support a proposed study, consistency assessments may be operationalized by qualitative assessments such as review of clinical documentation policies and procedures, interviews with facilities covering clinical documentation procedures and practice, or direct observation of workflow. Initial consistency checks can also be established using aggregate or distributional statistics. Once data collection has started, consistency should be monitored over time or across individuals, units, or facilities by aggregate or distributional statistics.

OMOP[16] and Mini-Sentinel[18] both provide publically available consistency checks that are executable against the OMOP and Mini-Sentinel common data models, respectively.

Although PCTs are less likely to utilize a common data model, the OMOP and Mini-Sentinel programs provide excellent examples of checks that can be used to evaluate consistency across investigational sites, facilities, departments, clinical units, providers, or assessors in PCTs.

As with accuracy assessments, consistency assessments should be conducted for data elements used in subject or cohort identification, outcome data elements, and covariates.

Data Quality Assessment Recommendations for Collaboratory Projects

We have defined critical components of data quality assessment for research using data generated in healthcare settings that we consider to be necessary in demonstrating the capacity of data to support research conclusions. Our recommendations below for data quality assessment for Collaboratory research projects are based on these key components:

Recommendation 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.

Recommendation 2 - Description of formal of assessments

1. **Completeness assessment recommendation:** Use of a four-part completeness assessment. The same column and data value completeness measures can be employed for monitoring completeness throughout the project. The completeness assessment applies to both prospectively collected and secondary use data. Additional requirements suggested by the GCDMP, such as on-screen prompts for missing data where appropriate, apply to data collected prospectively for a study.
2. **Accuracy assessment recommendation:** Identification and conduct of project-specific accuracy assessments for subject/cohort identification data elements, outcome data elements, and covariates. The highest practical accuracy assessment in the hierarchy shown in Figure 1 should be used. The same measures may be applicable for monitoring data accuracy throughout the project. Additional requirements suggested by the GCDMP, such as on-screen prompts for inconsistent data where appropriate apply to prospectively collected data.
3. **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.

4. **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 – Reporting data quality assessment with research results

As recommended elsewhere, results of data quality assessments should be reported with research results.[1,22] Data quality assessments are the only way to demonstrate that data quality is sufficient to support the research conclusions. Thus, data quality assessment results must be accessible to consumers of research.

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

In our initial recommendations (Appendix III), we encouraged the creation and use of data flow and workflow diagrams to aid in identifying accuracy and in conducting consistency assessments; however, this strategy has both advantages and disadvantages. Among the advantages is that the diagrams are helpful in other aspects of operationalizing a research project and in managing institutional information architecture. Thus, they may already exist, and if not, they will likely be used for other purposes. Understanding workflow around clinical documentation of cohort identifiers, outcomes data, and covariates is necessary for assessing potential inconsistencies between sites.

Workflow knowledge is also required in cases where the clinical workflow will be modified for the research, e.g., collecting study-specific data within clinical processes or using routine clinical data to trigger research activities. In the Collaboratory STOP CRC Demonstration Project, documentation of a patient's colonoscopy "turns off" further fecal occult blood test screening interventions for a period of time. Logic decisions similar to these would be clearly documented in the workflow and data flow analysis. On our test project, the process of creating and reviewing the diagrams prompted discussion of potential data quality issues as well as strategies for prevention or mitigation of problems.

Alternatively, if workflow diagrams do not exist for a facility, creation of these diagrams solely for the purpose of such an analysis may not be feasible. Consider a study with 30 small participating investigational sites from different institutions. Creation of workflow and data flow diagrams *de novo* for a study would consume significant resources. In such cases where the effort associated with creating and reviewing such diagrams is not practical, we offer the following set of questions that could be reviewed with personnel at each facility. These questions were developed based on our experience with the testing of the initial recommendations.

1. Talk through each of the data elements used for cohort identification. Can you explain how and where each one is documented in the clinic/on the 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? Importantly, please describe all data transformation between the source system and the data store used for this research.
3. For each data element used in the cohort identification, do you know of any difference in data capture or clinical documentation practices across clinics at your site or for different subsets of your population?
4. For each data element used in cohort identification, do you know of 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 laboratory data from analyzers on site versus those that you receive from external clinical laboratories?

The four questions above should be applied to other important data elements such as outcome measures and covariates.

Concluding Remarks

Moving forward, attention to data quality will be critical and increasingly expected, as in the case of the data validation review criteria for the Collaboratory. Although generalized computational approaches have shown great promise in large national initiatives such as Mini-Sentinel and OMOP, they are currently dependent on the existence of a common data model. However, as healthcare institutions across the country embark upon data governance initiatives, and as standard data elements become a reality for healthcare and health-related research, more and better machine-readable metadata are becoming available. Ongoing research in this arena will work toward leveraging this information to increase automation of data quality assessment and create metadata-driven, next-generation approaches to computational data quality assessment.

Funding

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.

References

1. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care 2013;51(8 suppl 3):S22–S29. PMID: 23793049. doi: 10.1097/MLR.0b013e31829b1e2c.
2. Saltz J. Report on Pragmatic Clinical Trials Infrastructure Workshop. Available at: https://www.ctsacentral.org/sites/default/files/documents/IKFC%201%204%202013.pdf. Accessed July 28, 2014.

Prepared by: Meredith N. Zozus, PhD; W. Ed Hammond, PhD; Beverly B. Green, MD, MPH; Michael G. Kahn, MD, PhD; Rachel L. Richesson, PhD, MPH; Shelley A. Rusincovitch; Gregory E. Simon, MD, MPH; Michelle M. Smerek. Reviewed by: NIH Version: 1.0, last updated July 28, 2014
Collaboratory Phenotypes, Data Standards, and Data Quality Core

3. Richesson, RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc 2013;20:e226–e231. PMID: 23956018. doi: 10.1136/amiajnl-2013-001926.

4. Society for Clinical Data Management. Good Clinical Data Management Practices (GCDMP). Available at: http://www.scdm.org/sitecore/content/be-bruga/scdm/Publications/gcdmp.aspx. Accessed July 2, 2014.

5. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 2002;9:600–611. PMID: 12386111. doi: 10.1197/jamia.M1087.

6. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20:144–151. PMID: 22733976. doi: 10.1136/amiajnl-2011-000681.

7. International Organization for Standardization. ISO 8000-2:2012(E) Data Quality – Part 2: Vocabulary. 1st ed. June 15, 2012.

8. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013;46:830–836. PMID: 23820016. doi: 10.1016/j.jbi.2013.06.010.

9. International Organization for Standards. ISO 21090. Health Informatics – Harmonized Data Types for Information Interchange. 2011.

10. Health Level Seven International. HL7 Data Type Definition Standards. Available at: http://www.hl7.org/implement/standards/product_section.cfm?section=2&ref=nav. Accessed July 2, 2014.

11. NIH Health Care Systems Collaboratory Biostatistics and Study Design Core. Key issues in extracting usable data from electronic health records for pragmatic clinical trials. Version 1.0 (June 26, 2014). Available at: https://www.nihcollaboratory.org/Products/Extracting-EHR-data_V1.0.pdf. Accessed July 28, 2014.

12. Nahm M, Bonner J, Reed PL, Howard K. Determinants of accuracy in the context of clinical study data. International Conference on Information Quality (ICIQ), Paris, France, November 2012. Available at: http://mitiq.mit.edu/ICIQ/2012/. Accessed July 2, 2014.

13. Nahm ML, Pieper CF, Cunningham MM. Quantifying data quality for clinical trials using electronic data capture. PLoS ONE 2008;3:e3049. PMID: 18725958. doi: 10.1371/journal.pone.0003049.

14. Tcheng J, Nahm M, Fendt K. Data quality issues and the electronic health record. Drug Information Association Global Forum 2010;2:36–40.

15. Davidson B, Lee WY, Wang R. Developing data production maps: meeting patient discharge data submission requirements. International Journal of Healthcare Technology and Management 2004;6:223–240.

16. Observational Medical Outcomes Partnership. Generalized Review of OSCAR Unified Checking. 2011. Available at: http://omop.org/GROUCH. Accessed July 2, 2014.

17. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care 2012;50(suppl):S21–S29. PMID: 22692254. doi: 10.1097/MLR.0b013e318257dd67.

18. Mini-Sentinel Operations Center. Mini-Sentinel Common Data Model Data Quality Review and Characterization Process and Programs. Program Package version: 3.1.2. September 2013. Available at: http://mini-sentinel.org/data_activities/distributed_db_and_data/details.aspx?ID=131. Accessed July 2, 2014.

19. Brown PJ, Warmington V. Data quality probes - exploiting and improving the quality of electronic patient record data and patient care. Int J Med Inform 2002;68:91–98. PMID: 12467794. doi: 10.1016/S1386-5056(02)00068-0.

20. Weiskopf NG, Enabling the Reuse of Electronic Health Record Data through Data Quality Assessment and Transparency. Doctoral Dissertation, Columbia University, June 13, 2014.

21. Sebastian-Cole L. Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Waltham, MA: Morgan Kaufmann (Elsevier); 2013.

22. Kahn MG, Brown J, Chun A, et al. A consensus-based data quality reporting framework for observational healthcare data. Submitted to eGEMS Journal, December 2013. Draft version available at: http://repository.academyhealth.org/cgi/viewcontent.cgi?article=1001&context=dqc. Accessed July 2, 2014.

Appendix I

Defining data quality

The ISO 8000 series of standards focuses on data quality.[1] Quality is defined as the "…degree to which a set of inherent characteristics fulfills requirements."[2] Thus, data quality is the degree to which a set of inherent characteristics of the data fulfills requirements for the data.

Describing data quality in terms of characteristics inherent to data means that we subscribe to a multidimensional conceptualization of data quality.[3] Briefly, these inherent characteristics, also called dimensions of data quality, include concepts such as accuracy, relevance, accessibility, contemporaneity, timeliness, and completeness. The initial work establishing the multidimensional conceptualization of data quality identified over 200 dimensions in use across surveyed organizations from different industries.[4] For most data uses, only a handful of dimensions are deemed important enough to formally measure and assess. The dimensions measured in data quality assessment should be those necessary to indicate fitness of the data for a particular use. In summary, data quality is assessed by identifying important dimensions and measuring them.

Defining the quality of research data

The Collaboratory has embraced the definition of quality data from the 1999 Institute of Medicine Workshop Report titled, *Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making,*[5] in which fitness for use (i.e., quality data) in clinical research is defined as "**data that sufficiently support conclusions and interpretations equivalent to those derived from error-free data.**"[5] The job, then, of assessing the quality of research data begins with identifying those aspects of data that bear most heavily on the capacity of the data to support conclusions drawn from the research.

Immediately prior to the April 2013 Collaboratory Steering Committee meeting, the program office released the review criteria for Demonstration Projects applying for funds for trial conduct:

- **Criterion 1:** "Are data collection methods adequately validated?"
- **Criterion 2:** "Validated methods for the electronic health record information?"
- **Criterion 3:** "Demonstrated quality assurance and harmonization of data elements across healthcare systems/sites?"
- **Criterion 4:** "Plans adequate for data quality control during the UH3 (trial conduct) phase?"

In keeping with the Institute of Medicine definition of quality data, the goal of these requirements is to provide reasonable assurance that data used for Collaboratory Demonstration Projects are capable of supporting the research conclusions.

Prepared by: Meredith N. Zozus, PhD; W. Ed Hammond, PhD; Beverly B. Green, MD, MPH; Michael G. Kahn, MD, PhD; Rachel L. Richesson, PhD, MPH; Shelley A. Rusincovitch; Gregory E. Simon, MD, MPH; Michelle M. Smerek. Reviewed by: NIH Version: 1.0, last updated July 28, 2014
Collaboratory Phenotypes, Data Standards, and Data Quality Core

The requirements were not further defined at the time of release. To aid in operationalizing data quality assessment, the PDSDQ Core drafted definitions for each criterion. These draft definitions (provided below) reflect the consensus of the Core and do not necessarily represent the opinions or official positions of the NIH.

Briefly, Criterion 1 pertains to data prospectively collected for research only (i.e., in addition to data generated in routine care). Criterion 2 applies to data generated in routine care. Criterion 3 pertains to *a priori* activities to assure consistency in data collection and clinical documentation across clinical sites. Criterion 4 requires plans to assess and control data quality throughout trial conduct. The criteria can be decomposed into data quality activities and data sources to which they apply (Figure A1). The third axis of consideration is the data quality dimensions important for a given study.

Data Source ➔ / ↓ Data quality activity	Routine care	Data collected solely for the research
Validation of data collection methods		✔
Data quality assurance	✔	
Harmonization of data elements	✔	✔
Data quality control	✔	✔

Figure A1. Graphic Representation of Review Criterion

Historically, in clinical trials conducted for regulatory review for marketing authorization, identification of data discrepancies was followed by a communication back to the source of the data in an attempt to ascertain the correct value.[6] This process of identifying and resolving data discrepancies is a type of data cleaning. Correction of data discrepancies is best applied to prospective trials with prospectively collected data. As described above, some trials conducted in healthcare settings will collect "add-on data" (i.e., data necessary for the research that are not captured in routine care).

Our initial Demonstration Project data quality assessment inventory (data available upon request) confirmed that multiple Demonstration Projects are collecting prospective data. Five Demonstration Projects planned to collect PROs and one added screens in the local EHR to capture study-specific data. All projects also used routine care data and administrative data. Details of the Demonstration Project data quality assessment inventory are provided in Appendix II.

Data quality–related review criteria

The following four UH3 review criteria (February 12, 2013 UH3 Transition Criteria Draft) were provided by the National Center for Complementary and Alternative Medicine. The PDSDQ Core has defined the criteria as outlined below.

Criterion 1: Are data collection methods adequately validated?

Scope: This criterion applies to data collected prospectively for the project (i.e., collected outside of routine clinical documentation).

Purpose: The purpose of this criterion is to provide assurance that project-specific data collection tools, systems, and processes produce data that can support the intended analysis and ultimately the research conclusions.

> *Data collection methods:* The processes used to measure, observe, or otherwise obtain and document study assessments.
>
> *Adequate:* Evidence that the error rate has been characterized and will not likely impact the intended analysis and ultimately the conclusions.
>
> *Validated:* Shown to consistently represent and record the intended concept. For questionnaires and rating scales, this refers to evidence that the tool measures the intended concept in the intended population. With respect to measurement of physical quantities or observation of phenomena, this refers to the ability of the measurement or observation to consistently and accurately capture the actual state of the patient. With respect to data processing, this refers to evidence of fidelity in operations performed on the data.

Criterion 2: Validated methods for the electronic health record information?

Scope: This criterion applies to data collected during routine care (i.e., during or associated with a clinical encounter or assessment). It applies to patient-reported data collected in conjunction with routine care (e.g., intake forms, questionnaires, or rating scales used in routine care and collected through healthcare information systems such as patient portals or EHRs). NOTE: Questionnaires administered through stand-alone systems created for a research study are not included in this criterion.

Purpose: The purpose of this criterion is to provide assurance that health system data used for the project can support the intended analysis and ultimately the research conclusions.

> *See definition of validated above.*
>
> *EHR information:* For our purposes, this definition encompasses data from information systems used in patient care and self-monitoring; this includes such data obtained through organizational data warehouses.

Criterion 3: Demonstrated quality assurance and harmonization of data elements across healthcare systems/sites?

Scope: This criterion applies to data elements collected for the project, including both those collected through healthcare systems and those collected through add-on systems for the study.

Purpose: The purpose of this criterion is to provide assurance that the meaning and format of data are consistent across facilities and that the methods of measurement, observation, and collection uphold the intended consistency.

Quality assurance (within this criterion): All the planned and systematic activities implemented within the quality system that can be demonstrated to provide confidence that a product or service will fulfill requirements for quality. Here, quality assurance pertains to activities undertaken to 1) assess existence of and potential for inconsistent data across participating facilities and 2) technical, managerial, or procedural controls in place to maintain consistency throughout the UH3 phase. NOTE: The U.S. Food and Drug Administration has defined quality assurance as independent.

Harmonization of data elements across health systems/sites: Use of or mapping organizational data to common data elements.

Common data elements: Data elements with the same semantics and representation as defined by the ISO 11179 standard.[7]

Data element: As defined by the ISO 11179 standard, a data element is pairing of a concept and a set of valid values.[7]

Criterion 4: Are plans adequate for data quality control during the UH3 phase?

Scope: This criterion applies to data collected for the project, including both those collected through healthcare systems and those collected through add-on systems for the study.

Purpose: The purpose of this criterion is to provide assurance that data quality monitoring and control processes are in place to maintain the desired quality levels and consistency between data collection facilities/sites.

Quality control: The operational techniques and activities used to fulfill requirements for quality. Quality control activities are usually thought of as those activities performed as part of routine operations to measure, monitor, and take corrective action necessary to maintain the desired quality levels within acceptable variance (e.g., re-abstracting a sample of charts on a quarterly basis to measure inter-rater reliability and provide feedback to abstractors).

References

1. International Organization for Standardization. ISO 8000-2:2012(E) Data Quality – Part 2: Vocabulary. 1st ed. June 15, 2012.
2. International Organization for Standardization. ISO 9000:2005, definition 3.1.1, ISO 9000:2005(E) Quality management systems — Fundamentals and vocabulary. 3rd ed. September 15, 2005.
3. Lee YW, Pipino LL, Funk JD, Wang RY. Journey to Data Quality. Cambridge, MA: MIT Press; 2006.

4. Wang R, Strong D. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems 1996;12:4–34. Available at: http://www.jstor.org/stable/40398176. Accessed July 2, 2014.

5. Institute of Medicine Roundtable on Research and Development of Drugs, Biologics, and Medical Devices. Davis JR, Nolan VP, Woodcock J, Estabrook RW, eds. Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making. Workshop Report. Washington, DC: National Academies Press; 1999. Available at: http://books.nap.edu/openbook.php?record_id=9623. Accessed July 2, 2014.

6. Society for Clinical Data Management. Good Clinical Data Management Practices (GCDMP). Available at: http://www.scdm.org/sitecore/content/be-bruga/scdm/Publications/gcdmp.aspx. Accessed July 2, 2014.

7. International Organization for Standardization. ISO 11179. Information Technology – Metadata registries (MDR).

Appendix II: Data Quality Assessment Plan Inventory

The initial plan of the PDSDQ Core was to inventory planned data quality assessment practice from the UH2 applications, review the statements of data quality assessment plans with the Demonstration Projects, summarize the plans in the context of the existing literature, and support the projects as needed in following the existing plans or in formulating and undertaking new plans if desired.

The PDSDQ Core conducted a data quality assessment inventory to characterize data quality assessment plans across the initial seven UH2 funded Demonstration Projects. The data quality assessment inventory was conducted in March 2013 and reported at the April 29-30, 2013 Collaboratory Steering Committee meeting (data available upon request).

Given the Collaboratory's focus on PCTs in its Demonstration Projects, we expected variability in the data sources used as well as the extent to which any project relied upon any one data source. To characterize their use, data sources commonly used by the Demonstration Projects were classified into five categories (Table A1):

1. **External PRO:** PRO or other questionnaire data collected outside of an EHR, such as those using a separate personal health record system, interviews, or paper questionnaires.
2. **PRO in EHR:** PRO or other questionnaire data collected using an EHR system.
3. **Research-specific EHR screens:** Data collection fields, modules, or screens rendered for users as if they were a part of the EHR system.
4. **Clinical data from an institutional data warehouse:** Medications, reports from laboratory and diagnostic tests, clinical notes, and structured clinical data such as vital signs originating from a patient care–facing system that are accessed through an institutional clinical data warehouse rather than directly from the transactional system.
5. **Administrative data from an institutional data warehouse:** Coded diagnoses and procedures used for reimbursement.

There was also variability in the extent to which Demonstration Projects relied on existing versus prospectively collected data. Five of seven projects were identified as collecting research data in addition to routine care data, four projects are using PRO data (one of which includes patient and staff interviews), and one project is adding data collection screens to the EHR.

Table A1. Data Source Summary

Project	External PRO	PRO in EHR	Research-specific EHR screens	Data warehouse/ EHR clinical data	Data warehouse admin. data
Collaborative care for chronic pain in primary care*	Paper, interview	X		X	X
Nighttime dosing of anti-hypertensive medications: a pragmatic clinical trial	Personal health record or interview			X	X
Decreasing bioburden to reduce healthcare-associated infections and readmissions*,†			X	X	X
Strategies and opportunities to stop colon cancer in priority populations*	Patient interviews			X	X
A pragmatic trial of lumbar image reporting with epidemiology (LIRE)‡				X	X
Pragmatic trial of population-based programs to prevent suicide attempt		X		X	X
Pragmatic trials in maintenance hemodialysis		X		X	X

**Includes staff interview data. †Includes data from external laboratory. ‡Includes externally enhanced data.*

Variation in data quality assessment methods corresponds with variation in data sources. To characterize data quality assessment practices across Demonstration Projects, initial applications were reviewed and each project provided any updates describing their planned data quality assessment practices (Table A2).

Due to the dependence of data quality assessment on the type of data and the available sources for comparison, opportunistic data quality assessments that leverage available sources for comparison should be expected, rather than uniformity with respect to comparisons.

Table A2. Data Quality Assessment Activities

Project	Collection control	Completeness		Accuracy	
		Ascertainment	% Column complete	Individual	Aggregate
Collaborative care for chronic pain in primary care	Procedural; technical		Part of ETL into warehouse	Part of ETL into warehouse	
Nighttime dosing of anti-hypertensive medications: a pragmatic clinical trial	Procedural (abstraction forms)	100-case chart review; AC PPV/NPV ≥90%	1. Yes on n=1000 cases, AC <5% per DE; 2. Site-to-site variability, completeness	1. Comparison to patient self-report; 2. Comparison to NDI; 3. IRR abstraction threshold; 4. Endpoint review; 5. Out-of-range values	
Decreasing bioburden to reduce healthcare-associated infections and readmissions	Procedural; technical		Yes (monthly monitoring)	Health system validated	
Strategies and opportunities to stop colon cancer in priority populations		Independent data; % chart review		Call audit; % chart review	
A pragmatic trial of lumbar image reporting with epidemiology (LIRE)			Yes		Site-to-site variability
Pragmatic trial of population-based programs to prevent suicide attempt		% chart review			
Pragmatic trials in maintenance hemodialysis	Procedural		Yes	Valid values	

AC, ascertainment completeness; DE, data error; ETL, extract-transform-load; IRR, interrater reliability; NPV, negative predictive value; PPV, positive predictive value

In most cases, data quality assurance and control activities for data collected *de novo* were not described in detail.

This inventory was reported to the Collaboratory in the context of the existing literature.

Appendix III: Initial Data Quality Assessment Recommendations for Collaboratory Projects

At the April 2013 Collaboratory Steering Committee meeting, an approach for addressing the data validation requirements was presented (data available upon request). This approach included:

1. **Completeness assessment:** A four-dimensional completeness assessment that could be conducted by all Demonstration Projects.
2. **Accuracy assessment:** Identification and conduct of project-specific data quality (accuracy) assessments.
3. **Impact assessment:** Use of the completeness and accuracy assessment results by the project statistician to test sensitivity of the analyses to anticipated data quality problems.

A Total Data Quality Management approach[1,2] was applied to identify and prioritize project-specific data quality needs (step 2 above). The following project information was reviewed:

1. Data elements collected and used for the project's statistical analysis
2. Workflow diagrams for clinic processes that generate data used in the study
3. Data flow diagrams for data elements used in the study

Higher priority was to be given to cohort identification and outcome data elements. The initial data element list from each project application was reviewed and updated where needed; specification of the source system for each data element was added. The workflow and data flow diagrams concentrated on processes used to generate data utilized in the study without regard to whether these processes were part of routine clinic practice or specific to the study. The development and discussion of the diagrams were used to surface potential sources of inconsistency or data error.

The Core proposed one-on-one work with, or individual work by, each Demonstration Project team to determine the type of accuracy assessment attainable and the targeted data validation assessments valuable for each project.

Testing the recommendations with the STOP CRC project

At the Steering Committee meeting, one project, Strategies and Opportunities to Stop Colon Cancer in Priority Populations (STOP CRC), came forward to work through the proposed approach with the Core. A series of several calls held over 2 months were conducted as the project and Core worked through the above approach. The calls were attended by a co–principal investigator of the STOP CRC trial, the project informatician overseeing the study's multifacility EHR implementation, the first author of this report, and two informaticians from the Coordinating Center. As planned, the development and discussion of the diagrams were used to surface potential sources of inconsistency or data error.

The data quality assessment work was reported on monthly calls and was summarized both in writing and by a template for reporting a data quality assessment.

Summary of findings from testing with the STOP CRC project

A workflow diagram existed and was contributed by the STOP CRC project team, as was an updated data dictionary. A Coordinating Center informatician conducted interviews with the STOP CRC project co–principal investigator and informatician to understand the workflow and complete the data flow diagram. The majority of the time on the calls was spent 1) educating the Collaboratory Coordinating Center informaticians about the study and the local data policies, procedures, and systems; 2) discussing data sources and reviewing the workflow and data flow diagrams; 3) discussing possible data quality problems based on the co–principal investigator's experience with a similar project, as well as potential solutions; and 4) creating a plan for initial and ongoing assessments of data quality and completeness. The STOP CRC statistician attended the final call to discuss the data validation plan results, impact assessments, and plans for ongoing data quality assurance during the multisite trial.

Due to the fact that data quality assessment plans existed for each project and were deemed acceptable through the grant review, the systematic approach at designing a data quality assessment plan was offered on a voluntary basis. Because of the inclusion of the data validation criteria in the review criteria for the trial conduct funding decision, we anticipated that most if not all Demonstration Projects would have shown interest in the offered approach and support. Only one project engaged the Core for a systematic assessment. No other projects reported making use of the draft review criteria definitions or data quality assessment information, plans, or templates produced.

References

1. Davidson B, Lee WY, Wang R. Developing data production maps: meeting patient discharge data submission requirements. International Journal of Healthcare Technology and Management 2004;6:223–240.
2. Lee YW, Pipino LL, Funk JD, Wang RY. Journey to Data Quality. Cambridge, MA: MIT Press; 2006.