A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research

Michael G. Kahn, MD, PhD,† Marsha A. Raebel, PharmD,‡§ Jason M. Glanz, PhD, MS,‡‖ Karen Riedlinger, MPH, MT (ASCP),¶ and John F. Steiner, MD, MPH‡*

Introduction: Answers to clinical and public health research questions increasingly require aggregated data from multiple sites. Data from electronic health records and other clinical sources are useful for such studies, but require stringent quality assessment. Data quality assessment is particularly important in multisite studies to distinguish true variations in care from data quality problems.

Methods: We propose a "fit-for-use" conceptual model for data quality assessment and a process model for planning and conducting single-site and multisite data quality assessments. These approaches are illustrated using examples from prior multisite studies.

Approach: Critical components of multisite data quality assessment include: thoughtful prioritization of variables and data quality dimensions for assessment; development and use of standardized approaches to data quality assessment that can improve data utility over time; iterative cycles of assessment within and between sites; targeting assessment toward data domains known to be vulnerable to quality problems; and detailed documentation of the rationale and outcomes of data quality assessments to inform data users. The assessment process requires constant communication between site-level data providers, data coordinating centers, and principal investigators.

Discussion: A conceptually based and systematically executed approach to data quality assessment is essential to achieve the potential of the electronic revolution in health care. High-quality data allow "learning health care organizations" to analyze and act on their own information, to compare their outcomes to peers, and to address critical scientific questions from the population perspective.

Key Words: data quality, data quality assessment, single-site studies, multisite studies

(*Med Care* 2012;50: S21–S29)

Answers to important clinical research questions increasingly require data that are aggregated across multiple sites. Some research studies require large sample sizes to detect small but important treatment benefits or harms.[1] Comparative effectiveness research (CER) studies attempt to associate naturally occurring clinical practice variations with differences in clinical outcomes. In both cases, findings arising from a multisite study are more likely to be generalizable than findings from a single site. In multisite research studies, site-level differences in disease incidence, predictive variables, and health outcomes can either represent true "small area" variation in practice patterns and outcomes,[2,3] or variability in data collection methods across sites. Distinguishing between true and artifactual variation that arises from problems with data quality is an essential first step in a multisite comparative effectiveness study.

Assessing data variability across sites is particularly important in studies that use data from the electronic health records (EHRs). EHR data are gathered during routine practice by individuals with a wide range of backgrounds and with different levels of commitment to data quality. As a result, EHR data are rarely subjected to the stringent data quality assessment that is routinely applied to prospective observational or interventional research.[4] If data quality issues with EHR data across sites can be identified and addressed[5–7] using a standardized approach, this rich data source can increase the efficiency and reduce the costs of observational CER studies and pragmatic clinical trials.

Data quality assessments (DQAs) for multisite studies are typically performed in 2 stages. In stage 1, source datasets are created at each site and evaluated using a "fit-for-use" perspective. A comprehensive approach to the initial stage of data quality assessment in an EHR-based, multisite study requires both within-site and across-site data quality assessment using consensus standards, as multisite assessment often identifies data quality issues not evident from single-site

From the *Department of Pediatrics; †Anschutz Medical Center, Colorado Clinical and Translational Sciences Institute, University of Colorado, Aurora; ‡Institute for Health Research, Kaiser Permanente Colorado, Denver; §School of Pharmacy, University of Colorado; ‖Department of Epidemiology, Colorado School of Public Health, Aurora, CO; and ¶Northwest Kaiser Center for Health Research, Portland, OR.

Supported by a contract from AcademyHealth. Additional support was provided by AHRQ 1R01HS019912-01 (Scalable Partnering Network for CER: Across Lifespan, Conditions, and Settings), AHRQ 1R01HS019908 (Scalable Architecture for Federated Translational Inquiries Network), and NIH/NCRR Colorado CTSI Grant Number UL1 RR025780 (Colorado Clinical and Translational Sciences Institute).

The authors declare no conflict of interest.

Reprints: Michael G. Kahn, MD, PhD, Children's Hospital Colorado, 13123 East 16th Avenue, B400, Aurora, CO 80045. E-mail: michael.kahn@ucdenver.edu.

ISSN: 0025-7079/12/5007-0S21

assessment alone. These assessments typically include comparisons of simple cross-tabulations and descriptive statistics such as means, medians, and histograms across sites. Once agreed-upon standards are met, smaller analytic datasets designed to test specific hypotheses are created. Data quality assessment stage 2 is then conducted on these analytic datasets, typically focusing on the independent and dependent variables directly related to the research question. When necessary, medical record review is used in stage 2 to validate outcomes, exposures, and/or covariates.

Individual researchers tend to focus only on "data cleaning" necessary to test a hypothesis of interest within an already-assembled dataset—in other words, on stage 2. Although some investigators, data analysts, and data coordinating centers develop procedures for assessing data quality during stage 1, a comprehensive conceptual framework for this initial stage is lacking. In this paper, we present a pragmatic conceptual model for stage 1 characterization of data quality using a "fit-for-use" perspective.[8,9] We then describe an approach for single-site and multisite data quality assessment in large observational studies that employ EHR data. Examples from actual multisite studies illustrate data quality issues and highlight approaches to detecting and resolving these problems.

CONCEPTUAL MODEL FOR CLINICAL DATA QUALITY ASSESSMENT

Many frameworks have been proposed in the information sciences literature to conceptualize the multiple dimensions of data quality.[10–18] All approaches acknowledge that data quality is a multidimensional process that can require trade-offs between dimensions because of organizational and data-use priorities and resource constraints. None of these approaches have been directly applied to clinical or health services research studies, however.

Recent publications in information sciences have focused on achieving data quality standards that makes the data "fit-for-use by data consumers."[19] Unfortunately, competing descriptive models have resulted in the lack of a clear unified framework for measuring data quality or procedures to improve data quality.[12] In a comprehensive review, Wand and Wang listed 26 data quality dimensions. Five of these dimensions (accuracy, reliability, timeliness, relevance, and completeness) appear in most published data quality frameworks.[20] Wang and Strong proposed a comprehensive "fit-for-use" data quality assessment model that encompassed both a data-element and a data-system perspective.[19] In its original form, the model categorized 118 data quality features into 4 high-level categories: intrinsic features, contextual features, representational features, and accessibility. These 4 categories were used to define 15 data quality dimensions. Given the absence of conceptual frameworks for data quality in clinical or health services research, we have modified and simplified this model in Table 1, and have also provided clinical examples.

Clinical data quality assessment must always be considered within the design and implementation of the particular study. No standardized definitions apply across all data contexts. Yet, several uniform elements should always be considered when evaluating fitness for use. These include (a) the intended data application; (b) the quality characteristics of highest importance within the application; (c) the user's expectations of useful information; and (d) the resources available. The weighting of each element varies from situation to situation.

A PROCESS FOR SINGLE-SITE AND MULTISITE DATA QUALITY ASSESSMENT

Multisite research projects are comprised of data elements collected, extracted, examined, and formatted at individual sites. Combining site-specific data into a multisite dataset is rarely a straightforward process. Within each research study, specific data quality assessment routines are selected based on local knowledge of potential data problems, previous data quality concerns, or the requirements of a central data coordinating center. The data quality assessment process then proceeds through multiple iterations of within-site and cross-site assessment, as illustrated in Figure 1. Data problems identified within a single site may result in data reextraction and quality reassessment. Problems with data accuracy and with the programming used to extract and transform the data manifest at this stage. When data are aggregated across sites, additional data quality assessments may identify new anomalies, necessitating additional data quality assessment cycles at the original sites. After correction or explanation of data anomalies at each site, the data quality assessment cycle continues until datasets exceed a preestablished quality threshold (Fig. 1).

Table 2 illustrates the results of applying the stage 1 data quality assessment process shown in Figure 1 to an EHR-based study that originally included 4 sites. In the example, the exposure was a drug dispensing within a specified timeframe and the outcome required an ICD9-CM coded diagnosis. One site had far fewer patients with the drug exposure than would be expected given its population size. Data quality assessment revealed that the site was not capturing claims for all dispensed prescriptions. When exclusion criteria were applied, this site also lost 65% of the initial cohort, in comparison with 19%–27% at the other 3 sites. Finally, no patients at this fourth site appeared to have experienced the study outcome, whereas the other three sites had between 75 and 157 outcome events. Evaluation of this discrepancy revealed that diagnostic claims were incomplete at the outlier site. The site ultimately was dropped from the study.

APPROACHES FOR DATA QUALITY ASSESSMENT

Single-site Data Quality Assessment

The conceptual model in Table 1 ensures that relevant data quality dimensions are considered, but does not specify approaches that researchers can use to determine data quality. Specific assessment methods must be selected and executed to determine how well a given dataset meets quality expectations. Unlike the complex statistical models used to test specific study hypotheses, data quality assessment typically relies on simple distributions, cross-tabulations, and

TABLE 1. A "Fit-For-Use" Data Quality Model Adapted For Clinical Research

From Wang and Strong			CER Adaptation	
Category	**Domains**	**Technical Definition**	**Clinical Data Redefinition**	**Examples**
Intrinsic: data quality features that are inherent to data alone	Accuracy	The extent to which data are correct, reliable, and free of error	Data values represent the true state of a patient within the limitations of the measurement methods	Mishandled specimen (eg, potassium levels in hemolyzed serum); test performance against a gold standard
	Objectivity	The extent to which data are unbiased and impartial	The methods used to obtain data values are well described and represent best practices. Component values represent the total clinical measurement	CLIA-approved laboratory tests Use of standardized psychometric instruments to assess patient status
	Believability	The extent to which data are regarded as true, real, and credible	Independent measurements make clinical sense	Sex agreement with sex-specific features (pregnancy, prostate cancer). Collections of related measures are physiologically consistent (AST, ALT, bilirubin, PTT in liver failure) Clinical trials Standard Operating Procedures
Conceptual: data quality features that are relevant in the context of the task for which the data are to be used	Timeliness	The extent to which the age of the data is appropriate for the task at hand	Serial measurements over time sufficient to detect clinical state	Blood pressure measurements for diagnosing hypertension linked in time to clinic visits
	Appropriate amount of data	The extent to which the quantity or volume of available data is appropriate	Data are present or absent as expected	Degree and distribution of missingness

Other categories of data quality (representation and accessibility) defined by Wang and Strong are not included here but address important additional system-level considerations for data users.

Modified from Wang and Strong.[19]

graphical visualizations to aid in inspecting data. In Table 3 we present a comprehensive set of data quality rules and quality assessment methods, adapted from Maydanchik.[18] In this approach, 5 categories of rules—attribute domain constraints, relational integrity, historical integrity, state-dependent rules, and attribute dependency rules—are operationalized through multiple assessment methods.

Attribute domain constraints focus on individual variables, looking for anomalies in data values, distributions, units, and missingness. For example, a date of birth in the year 1390 would be identified as an invalid value (perhaps a digit transposition for the year 1930). A comparison of simple descriptive statistics for prescription claims in Table 2 identified a missing data problem at 1 site.

Relational integrity rules look for inconsistencies across multiple-related variables, seeking to detect data quality issues in comparing elements from 1 data table to related elements in another data table. These rules are often called "double-checks" and "triple-checks." As an example, a count of 8 prescriptions in a summary table should correspond to 8 filled prescriptions in the original pharmacy table.

Historical data rules assess temporal relationships. The large number of assessment methods in this category emphasizes the complex temporal in most datasets. Historical rules examine sequences, gaps, patterns and dependencies across multiple data values, and variables. For example, an individual who died in 2008 cannot have a hospitalization in 2009.

State-dependent objects rules extend the analysis of temporal data to include logical consistency, where the sequence of temporal events conforms to knowledge about the expected or allowed evolution of a process or set of states over time. For example, a series of prenatal visits would be expected to culminate in an outcome (such as a birth), and then be followed by a postpartum checkup. Similarly, assessing the impact of changes in coding schemes or allowed values over time would be included in this class of data quality assessments.

Attribute dependency rules are the most complex because they combine real-world knowledge about how objects and processes are measured and represented in a dataset. These rules examine conditional dependencies and expected correlations across subsets of data and aggregates. For example, a postpartum checkup would be unlikely to occur 18 months after the delivery date.

Rote application of all data quality assessment methods in Table 3 to a dataset would result in thousands of actual data quality measures. The resources in a research study are never sufficient to assess all data elements against all data quality dimensions. Thus, prioritization is critical. For a given study, the set of critical data quality assessment methods will vary on the basis of the features of the variables that drive the key scientific questions. However, in general, the DQA methods listed under the "attribute domain constraints" and "relational integrity rules" in Table 3 can be applied broadly across all data elements in a dataset or database.

As an example, the Observational Source Characteristics Analysis Report (OSCAR) tool developed by the Observational Medical Outcomes Partnership (OMOP) program

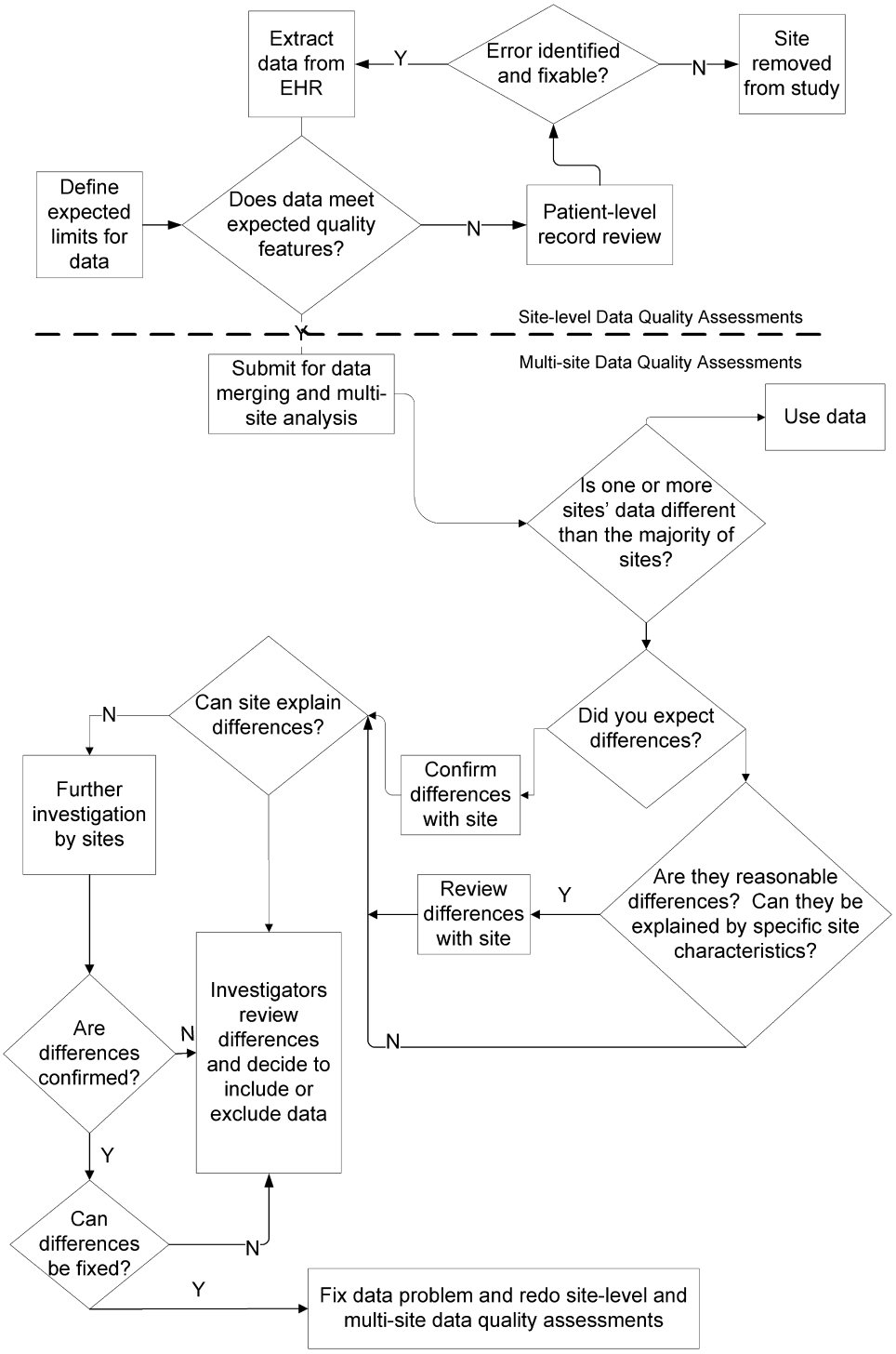

FIGURE 1. Detailed process model for multisite data quality assessment. Multiple data quality assessment cycles occur at both the single-site and multisite level. Each decision point (diamond) requires an examination of ≥ 1 data quality measures against predefined acceptability criteria. Many actions require continued engagement with sites to distinguish expected from unexpected site-specific data variability. HER indicates electronic health record; N, no; Y, yes.

generates summary statistics for all categorical and continuous variables, resulting in thousands of assessments without any prioritization.[21] The related Generalized Review of OSCAR Uniform Checking (GROUCH) tool extracts only those OSCAR measures that fail to meet prespecified data quality specifications.[22] These thresholds can be set for all categorical or continuous variables or for individual variables in the dataset. The OSCAR tool is comprehensive but

TABLE 2. Example of Data Variability in Exposures and Outcomes That was not Detected With Single-site Data Quality Assessment but was Detected During Multisite Data Quality Assessment

	No. Patients at Each of 4 Sites with Exposure of Interest			
	A	**B**	**C**	**D**
Exposure (Drug)				
Initial cohort	8730	6984	11708	3136
Final cohort	6820	5116	9446	1102
% of initial cohort retained	78	73	81	35
	No. Patients at Each of 4 Sites with the Outcome of Interest (%)			
	A	**B**	**C**	**D**
Outcome (Diagnosis)				
Emergency department visit with relevant diagnosis code	68 (43)	17 (23)	57 (44)	0
Hospitalization with relevant diagnosis code	89 (57)	58 (77)	74 (56)	0

Interpretation of data variability analysis

Sites B and D had similarly sized membership populations; sites A and C had approximately similarly sized membership populations; all sites had similar age distributions for potential study cohort.

Site D had a far lower number of patients with drug exposure in initial cohort (despite having approximately the same size membership as site B. Although this could be appropriate, upon quality assessment, it was determined that the site had incomplete capture of pharmacy dispensing claims for a portion of the study period.

After applying exclusion criteria, sites A, B, and C had similar proportions of patients retained in the final cohort; site D retained a far lower proportion of patients that was determined to be due to incomplete pharmacy claims capture.

Site D—no outcomes because of incomplete capture of diagnostic codes (could have been detected with single-site assessment but was not).

nonselective; the GROUCH tool provides prioritization and specificity.

MULTISITE DATA QUALITY ASSESSMENT

Local data entry processes, data quality validation checks, data storage models, and data extraction routines affect the data structure at a single site. The term *syntactic variability* expresses data variability caused by differences in the representation of data elements. For example, weights may be recorded and stored in different locations within an EHR, and in different formats or units. Failure to extract data from all locations and to transform into a common format would result in incomplete data. Syntactic issues tend to be detectable and resolvable using single-site data.

When data are combined from multiple sites, data that are syntactically identical (same format, same units) can show important differences if data elements that supposedly represent the same concept actually represent different concepts at each site. The term *semantic variability* expresses data variability caused by differences in the meaning of data elements. Differences in data collection, abstraction and extraction methods, or measurement protocols can result in semantic variability. For example, failure to distinguish between fasting and random blood glucose, finger-stick or venipuncture sampling, or serum or plasma measurements would result in glucose values that do not represent the same concept. Semantic variability is difficult to detect using single-site data alone because data semantics tend to be consistent within an institution. Only when data are combined from multiple sites can such semantic differences be detected.

The assessment methods in Table 3 were developed to assess single-site data quality. Few standardized approaches have been proposed to extend such single-site data quality assessment methods to multisite data. Two standardized approaches in clinical and health services research are: (1) The OMOP GROUCH data quality analysis tool described previously; and (2) The HMO Research Network (HMORN) Virtual Data Warehouse (VDW). The OMOP GROUCH tool implements 35 data quality rules. Eleven rules explicitly compare OSCAR-generated data quality measures from 1 site against all other sites.[22] The HMORN VDW research resource maintains a standardized quality assessment battery of SAS programs that enables all HMORN sites to evaluate their VDW data tables in comparison with a standardized set of "pass/fail" metrics common across all HMORN sites and to evaluate within-site attribute domain constraints, relational integrity, and historical integrity.[23]

In multisite data comparisons, calculation of simple descriptive statistics such as expected event rates, frequency distributions, and time trends allows detection of typical semantic anomalies such as: wide variation in counts or event rates, differences in distributions (eg, histograms) and temporal trends, including sudden deviation from previous trends, and the degree of missing data (as exemplified in Table 2). Such differences may be quite pronounced. In Figure 2, we provide a comparison of fasting serum glucose tests completed per 1000 patients at 7 different sites over a 6-year period. The data in this figure starkly demonstrate 4 different quality anomalies.

An observed site-level deviation in stage 1 does not necessarily indicate a data quality issue. True differences in populations, measurement processes, clinical workflows, or treatment strategies can result in significant differences across data sources. Such naturally occurring variation in clinical practice has been a source of important research for many years.[3,24,25] The data quality assessment process in Figure 1 will detect these differences, but cannot always explain them. As highlighted in Figures 1 and 2, observed anomalies need to be interpreted in collaboration with the local site to determine if the differences can be explained by factors other than data

TABLE 3. Data Quality Assessment Methods and Rules

Data Quality Rule	Definition	Assessment Methods	Data Quality Assessment Methods Examples
Attribute domain constraints	Rules that validate individual attribute values based on restrictions on allowed values	Attribute profiling	Basic aggregate statistics (counts, means, medians, minimum, and maximum values
			Examination of highest and lowest values
			Value distributions (histograms)
		Optionality	No. null (missing) values
			Use of default values to denote missingness (9999, 1/1/1900)
		Format	DDMMYYYY vs. MMDDYYYY vs. YYYYMMDD
			SSN: NNN-NN-NNNN
		Valid values	Sex = "M" or "F" or "U" only
			Route of administration = "PO", "IM", "IV", "other" only
			Blood glucose cannot be a negative value
			Age cannot be more than 120 y
		Precision	Units of measure
			Rounding rules
			Date/time precision
Relational integrity rules	Rules that ensure accurate relationships between entities (tables), instances (records), and attributes (fields) across multiple tables	Identity	Different unique IDs (keys) refer to distinct things (person, place, concept or event)
			The same unique ID (eg, SSN) refers to same entity (eg, person)
		Reference	A reference in 1 table to data in another table points to a row that exists in the second table.
		Cardinality	Relationship cardinality profiling—the count of the actual number of occurrences for each relationship in the database
			References to data in a table refer to no more than the allowed number of occurrences
		Inheritance	Entities are grouped into types and subtypes correctly (eg, all patients, parents, spouses, and employees are also persons)
Historical data rules	Rules involving time-varying data	Currency	The effective date for the earliest record meets a preestablished date
			The date for the most recent record meets a preestablished date
		Retention	The overall duration or number of records per case meets a preestablished threshold (time or number of records)
		Granularity	Measures across time all have the same units or duration (eg, months, years)
		Continuity	Gaps or overlaps between records do not exceed prespecified thresholds.
		Timeline patterns	Timestamps fall into expected intervals (eg, weekly, once a month)
			Intervals between successive timestamps do not exceed a minimum or maximum duration
		Value patterns	Successive values follow the expected direction of change (increase or decrease)
			Size of change in successive values per unit time is reasonable
		Event dependencies	The frequency of events per unit of time does not exceed a prespecified threshold
			The frequency of events per unit of time meets context-sensitive threshold (eg, elderly diabetics see their physicians more frequently than healthy teenagers)
		Event conditions	One or more events (causes) exist for each observed effect
			Certain events always occur together (coincidental events)
		Event attributes	All attributes relevant to an event description are present (eg, a surgical event requires a patient and a surgeon).
State-dependent objects rules	Rules that ensure that changes in the lifecycle of an object follow expected transitions	State-transition profiling	Examination of valid transition states over time
			A terminal state cannot be followed by another state (eg, an expired patient cannot be readmitted)
			A valid action or event is associated with a corresponding state transition (eg, cardiac arrest precedes patient state transition to expired)
		State domain	An object's state can only be a valid value (eg either admitted or discharged from hospital)
		Action domain	The set of actions that can be applied to an object can only be a valid value (eg, a test can be performed or canceled, but not both)
		Terminator domain	States in which an object can start or stop its lifecycle can only be a valid value
		State-actions	Each action is consistent with the change in state that it engenders (eg, myocardial infarction cannot lead to subsequent "no cardiac disease" state)

(*Continued*)

TABLE 3. Data Quality Assessment Methods and Rules (*continued*)

Data Quality Rule	Definition	Assessment Methods	Data Quality Assessment Methods Examples
		Continuity	Sequence or timing of start of each state record must follow the end of the previous state record
		Duration	The minimum or maximum length of time an object can stay in a specific state (eg, admission for myocardial infarction cannot be less than 1 d).
			Zero-length rule: end date of a state must be later than the start date
Attribute dependency rules	Rules for describing real-world objects	Redundant attributes	Same attributes in different data sources should have identical values
			Historical measures of a nontime varying attribute should have identical values
		Derived attributes	An aggregated value must equal the total of the atomic level values
			An aggregated value must follow appropriate rules when component data elements are missing
		Partially dependent attributes	The allowed values of 1 attribute are limited by the assigned value of another attribute (eg, sex = female eliminates prostate cancer as a valid diagnosis)
		Conditional optionality	The allowed values of 1 attribute determines if another attribute must be (cannot be) present (eg, a discharge disposition = "to home" implies expiration date = null)
			Mutually exclusivity—the presence of any value in 1 attribute precludes (requires) a value in another attribute
		Correlated attributes	Values in 1 attribute changes the likelihood of values in another attribute (eg, sex = male and age = 65 and smoker = yes increases the likelihood of discharge diagnosis = myocardial infarction)

IM indicates intramuscularly; IV, intravenously; PO, orally; SSN, social security number.
Modified from Maydanchik.[18]

quality. Site data owners have intimate knowledge about local workflows, data collection conventions, and changes in technologies (new systems, new updates, expanded implementations) that usually are not known to multisite collaborators.

In dynamic datasets that are regularly updated, information gleaned from previous data quality assessments within and between sites can identify areas that need extra evaluation during subsequent data quality assessments. For example, the HMORN VDW conducts across-site quality assessment yearly, whereas within-site assessment of previously identified data quality concerns occurs during the intervening months. By guiding resources and attention to known data quality concerns, data efficiency and consistency from each primary site improve over time.

Detailed documentation of the rationale for conducting data quality assessment and the outcomes of those assessments is essential, because every data quality assessment plan is a compromise between limited time and resources and the desire for the highest possible data quality. Investigators must be aware that, in many cases, data quality decisions are made by database managers and programmers who provide support to multiple studies that rely on the same data sources. Investigators may not be aware of crucial data quality decisions unless they are recorded and accessible.

IMPLICATIONS FOR CLINICAL RESEARCH DATA NETWORKS

A more standardized and comprehensive approach to stage 1 data quality assessment is likely to improve the validity of multisite, EHR-based comparative effectiveness research. When data variability across multiple sites reflects true differences in clinical practice, researchers gain the opportunity to measure medication comparative safety and effectiveness, procedures and tests utilization in an observational setting. Once variability is identified and data quality problems are eliminated, traditional stage 2 data quality approaches such as manual chart review (ideally using the EHR, but at times requiring manual record review) can be used to assure the validity of critical exposure and outcome variables. Numerous studies using manual record review as the "gold standard" have demonstrated that automated clinical data from managed-care organizations can inaccurately measure disease incidence, with positive predictive values as low as 20%.[26–30] These inaccuracies can be independent or dependent of treatment status, thus leading to differential or nondifferential misclassification bias that can result in both false-positive and false-negative study findings.[31] Although manual medical record review can provide a "gold standard" for validating automated clinical data, it is expensive and time consuming, and thus cannot be used routinely in stage 1 data quality assessments. Developing automated stage 1 methods to identify and resolve data anomalies across multiple sites is therefore essential to both the efficiency and the validity of CER studies that rely on cross-site comparisons.

In recent years, decentralized data models for multisite research (distributed research networks, or DRNs) have been developed that bypass the need to transfer data containing protected health information to a central data warehouse.[32–40] In a DRN, a centralized data coordinating center

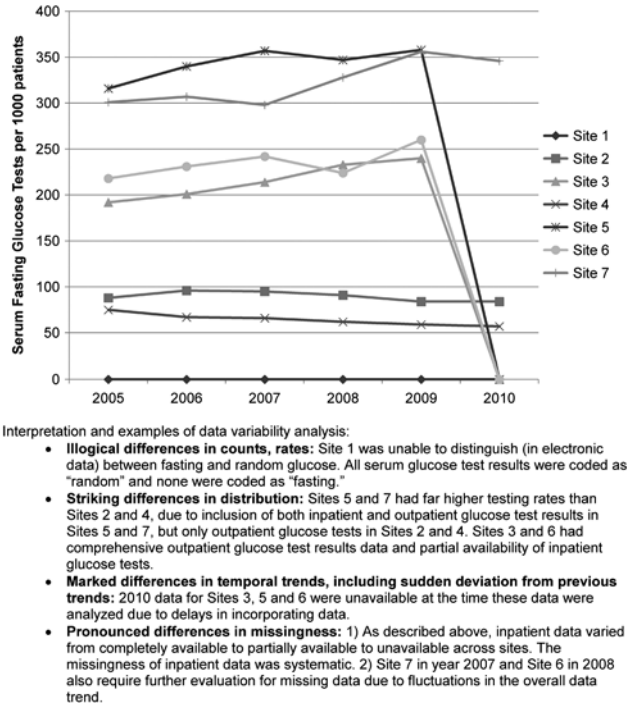

FIGURE 2. Variability over time and across sites in number of serum fasting glucose tests performed per 1000 patients.

may still be responsible for data quality assessment. But as the data are not pooled, the data coordinating center must devote more attention to planning and conducting quality assessment. In a DRN, the data coordinating center distributes programming code to all participating sites. A programmer then reviews and runs the code on local data, and sends the results back to the data coordinating center for evaluation. Such a process can potentially be less efficient than a traditional centralized model initially. However, the preparatory work required in DRN assessment may lead to progressive improvements in data quality over time, as programmers have to anticipate data issues and write their code to reduce inefficiencies.

In 2010, the President's Council of Advisors on Science and Technology (PCAST) report recommended the use of mandatory "metadata" tags attached to every data element.[41] These tags provide additional information about the data element, such as where the data was created (data provenance) and privacy permissions and restrictions. These metadata tags could be expanded to include the data quality measures as listed in Table 3, either as a single summary data quality statistic or as individual values for specific data quality measures that were applied to the data before release. An example of data quality metadata tags for continuous variables would be tags that include the attribute's mean, median, SD, interquartile range, and percentage missing across the original dataset. The simple distribution measures created by the OMOP OSCAR tool could be a useful initial set of data quality metadata tags consistent with the PCAST recommendation.

Data quality metadata tags are analogous to other metadata documentation that is required to support effective data sharing across institutions. Because data sharing is a National Institutes of Health priority,[42] appropriate resources should be included in the study budget directed explicitly to metadata documentation, including documenting data quality assessment results. In the future, new informatics tools could be developed that could perform data quality and automatically attach the appropriate DQA metadata tags and assessment results directly to the dataset.

DISCUSSION

Few publications have described the use of data quality models in health care data.[43–46] One case study in a large national intensive care unit registry examined the causes of error that influenced data accuracy and completeness at local sites and at a central coordinating center.[44] This paper categorized data errors into 3 categories (setup and organization, data collection, and quality improvement), described error-promoting processes at data collection sites and at the central coordinating center, and proposed a comprehensive framework for improving data quality. This paper did not address most elements of the conceptual model in Table 2, and did not describe in detail the approaches used to identify and correct those data quality issues.

Because the literature is sparse, and because systematic approaches to stage 1 data quality assessment have not been proposed for both single-site and multisite projects, a conceptual framework is useful for addressing data variability in a logical and comprehensive manner. This framework can be translated into a rigorously applied strategy for data quality assessment, which defines data quality dimensions and assessment methods to assess the variability in a multisite dataset. To execute this strategy in a multisite study, a strong collaborative relationship between members of the data coordinating center and the individual sites is essential because identifying and correcting data errors is of necessity iterative. If well-designed, this strategy assures that data quality issues discovered at individual sites can be used to improve data quality for subsequent studies.

If the variability in populations, treatment exposures, and clinical outcomes is due to true "small area variations" in clinical practice or health plan benefit design, CER studies can provide critical information about medical treatment effectiveness and safety or innovations in care delivery across broad populations. However, apparent data variability because of unrecognized data quality problems has the potential to invalidate study findings. Careful differentiation between real and spurious data variability at both the single-site and multisite level is essential if the promise of CER is to be achieved.

REFERENCES

1. Hanna KE. Think research: Using electronic medical records to bridge patient care and research. 2005. Available at: http://fastercures.org/objects/pdfs/comments/emr_whitepaper.pdf. Accessed April 22, 2012.
2. Wennberg DE, Wennberg JE. Addressing variations: is there hope for the future? *Health Aff. (Millwood)*. 2003;suppl web exclusives):W3-614–W3-617. doi: 10.1377/hlthaff.w3.614.

3. Wennberg JE. Practice variations and health care reform: connecting the dots. *Health Aff (Millwood)*. 2004;suppl variation):VAR140–VAR144. doi: 10.1377/hlthaff.var.140.
4. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Ann Intern Med*. 2009;151:359–360.
5. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. *J Am Med Inform Assoc*. 1997;4:342–355.
6. Aronsky D, Haug PJ. Assessing the quality of clinical data in a computer-based record for calculating the pneumonia severity index. *J Am Med Inform Assoc*. 2000;7:55–65.
7. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev*. 2010;67:503–527.
8. Gryna FM, Chua RCH, De Feo JA, et al. *Juran's Quality Planning and Analysis: For Enterprise Quality*. 5th ed. Boston: McGraw-Hill; 2007.
9. Juran JM, Gryna FM. *Quality Planning and Analysis: From Product Development Through Use*. 3rd ed. New York: McGraw-Hill; 1993.
10. Scannapieco M, Catarci T. Data quality under the computer science perspective. *Ital J Arch Comput*. 2002;1–12.
11. Batini C, Cappiello C, Francalanci C, et al. Methodologies for data quality assessment and improvement. *ACM Comput Surv*. 2009;41:1–52.
12. Batini C, Scannapieca M. *Data Quality: Concepts, Methodologies and Techniques*. Berlin; New York: Springer; 2006.
13. Wang R, Storey V, Firth C. A framework for analysis of data quality research. *IEEE Tran on Knowl Data Eng*. 1995;7:623–640.
14. Tayi G, Ballou D. Examining data quality. *Comm ACM*. 1998;41:54–57.
15. Abate M, Diegert K, Allen H. A hierarchical approach to improving data quality. *Data Quality Journal*. 1998;4. Available at: http://www.dataquality.com/998abate.htm. Accessed April 22, 2012.
16. Pipino L, Lee Y, Wang R. Data quality assessment. *Comm ACM*. 2002;45:211–218.
17. Redman TC. *Data Quality: The Field Guide*. Boston: Digital Press; 2001.
18. Maydanchik A. *Data Quality Assessment*. Bradley Beach, NJ: Technics Publications; 2007.
19. Wang R, Strong D. Beyond accuracy: what data quality means to data consumers. *J Manag Inf Syst*. 1996;12:5–34.
20. Wand Y, Wang R. Anchoring data quality dimensions in ontological foundations. *Comm ACM*. 1996;39:86–95.
21. Observational Medical Outcomes Partnership. OSCAR—Observational Source Characteristics Analysis Report (OSCAR) Design Specification and Feasibility Assessment. 2011. Available at: http://omop.fnih.org/OSCAR. Accessed July 31, 2011.
22. Observational Medical Outcomes Partnership. Generalized Review of OSCAR Unified Checking. 2011. Available at: http://omop.fnih.org/GROUCH. Accessed July 31, 2011.
23. Bauck A, Bachman D, Riedlinger K, et al. Developing a consistent structure for VDW QA checks. 2011. Available at: http://www.hmoresearchnetwork.org/archives/2011/concurrent/A1-02-Bauck.pdf. Accessed September 3, 2011.
24. Wennberg JE. Future directions for small area variations. *Med Care*. 1993;31:YS75–YS80.
25. Wennberg JE. Unwarranted variations in healthcare delivery: implications for academic medical centres. *BMJ*. 2002;325:961–964.
26. France EK, Glanz J, Xu S, et al. Risk of immune thrombocytopenic purpura after measles-mumps-rubella immunization in children. *Pediatrics*. 2008;121:e687–e692.
27. Hambidge SJ, Glanz JM, France EK, et al. Safety of trivalent inactivated influenza vaccine in children 6 to 23 months old. *JAMA*. 2006;296:1990–1997.
28. Mullooly J, Drew L, DeStefano F, et al. Quality assessments of HMO diagnosis databases used to monitor childhood vaccine safety. *Methods Inf Med*. 2004;43:163–170.
29. Mullooly JP, Donahue JG, DeStefano F, et al. Predictive value of ICD-9-CM codes used in vaccine safety research. *Methods Inf Med*. 2008;47:328–335.
30. Glanz JM, Newcomer SR, Hambidge SJ, et al. Safety of trivalent inactivated influenza vaccine in children aged 24 to 59 months in the vaccine safety datalink. *Arch Pediatr Adolesc Med*. 2011;165:749–755.
31. Greenland S, Robins JM. Confounding and misclassification. *Am J Epidemiol*. 1985;122:495–506.
32. Brown J, Holmes JH, Maro J, et al. Design specifications for network prototype and cooperative to conduct population-based studies and safety surveillance. 2009; Effective Health Care Research Report No 13. Available at: http://www.effectivehealthcare.ahrq.gov/ehc/products/54/150/2009_0728DEcIDE_DesignSpecNetCoopPopSafety.pdf. Accessed April 22, 2012.
33. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. *Ann Intern Med*. 2009;151:341–344.
34. Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. *Pharmacoepidemiol Drug Saf*. 2001;10:373–377.
35. Wagner EH, Greene SM, Hart G, et al. Building a research consortium of large health systems: the Cancer Research Network. *J Natl Cancer Inst Monogr*. 2005;35:3–11.
36. McMurry AJ, Gilbert CA, Reis BY, et al. A self-scaling, distributed information architecture for public health, research, and clinical care. *J Am Med Inform Assoc*. 2007;14:527–533.
37. Hornbrook MC, Hart G, Ellis JL, et al. Building a virtual cancer research organization. *J Natl Cancer Inst Monogr*. 2005;35:12–25.
38. Chen RT, Glasser JW, Rhodes PH, et al. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. The Vaccine Safety Datalink Team. *Pediatrics*. 1997;99:765–773.
39. Lazarus R, Yih K, Platt R. Distributed data processing for public health surveillance. *BMC Public Health*. 2006;6:235–246.
40. Moore KM, Duddy A, Braun MM, et al. Potential population-based electronic data sources for rapid pandemic influenza vaccine adverse event detection: a survey of health plans. *Pharmacoepidemiol Drug Saf*. 2008;17:1137–1141.
41. President's Council of Advisors on Science and Technology. Report to the President Realizing the full potential of health information technology to improve healthcare for America: The path forward. 2010. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf. Accessed April 22, 2012.
42. Office of Extramural Research. NIH data sharing policy. 2007. Available at: http://grants.nih.gov/grants/policy/data_sharing/. Accessed February 25, 2012.
43. Davidson B, Lee Y, Wang R. Developing data production maps: meeting patient discharge data submission requirements. *Int J Healthcare Technol Manag*. 2004;6:223–240.
44. 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.
45. Riain C, Helfert M. An evaluation of data quality related problems patterns in healthcare information systems. Paper presented at: Lisbon, Portugal: IADIS virtual multi conference on computer science and information systems; 2005.
46. Mettler T, Rohner P, Baacke L. Improving data quality of health information systems—a holistic design-oriented approach. European Conference on Information Systems; 2008; National University of Ireland Galway.

