EDM Forum
EDM Forum Community

eGEMs (Generating Evidence & Methods to improve patient outcomes) Publish

3-23-2015

Transparent Reporting of Data Quality in Distributed Data Networks

Michael G. Kahn
University of Colorado, michael.kahn@ucdenver.edu

Jeffrey S. Brown
Harvard Pilgrim Health Care Institute / Harvard Medical School, jeff_brown@harvardpilgrim.org

Alein T. Chun
Cedars-Sinai Health System, alein.chun@cshs.org

Bruce N. Davidson
Hoag Memorial Hospital Presbyterian, brucendavidson@icloud.com

See next pages for additional authors

Recommended Citation

Kahn, Michael G.; Brown, Jeffrey S.; Chun, Alein T.; Davidson, Bruce N.; Meeker, Daniella; Ryan, Patrick B.; Schilling, Lisa M.; Weiskopf, Nicole G.; Williams, Andrew E.; and Zozus, Meredith Nahm (2015) "Transparent Reporting of Data Quality in Distributed Data Networks," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 3: Iss. 1, Article 7.
DOI: https://doi.org/10.13063/2327-9214.1052
Available at: http://repository.edm-forum.org/egems/vol3/iss1/7

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

Transparent Reporting of Data Quality in Distributed Data Networks

Abstract

Introduction: Poor data quality can be a serious threat to the validity and generalizability of clinical research findings. The growing availability of electronic administrative and clinical data is accompanied by a growing concern about the quality of these data for observational research and other analytic purposes. Currently, there are no widely accepted guidelines for reporting quality results that would enable investigators and consumers to independently determine if a data source is fit for use to support analytic inferences and reliable evidence generation.

Model and Methods: We developed a conceptual model that captures the flow of data from data originator across successive data stewards and finally to the data consumer. This “data lifecycle” model illustrates how data quality issues can result in data being returned back to previous data custodians. We highlight the potential risks of poor data quality on clinical practice and research results. Because of the need to ensure transparent reporting of a data quality issues, we created a unifying data-quality reporting framework and a complementary set of 20 data-quality reporting recommendations for studies that use observational clinical and administrative data for secondary data analysis. We obtained stakeholder input on the perceived value of each recommendation by soliciting public comments via two face-to-face meetings of informatics and comparative-effectiveness investigators, through multiple public webinars targeted to the health services research community, and with an open access online wiki.

Recommendations: Our recommendations propose reporting on both general and analysis-specific data quality features. The goals of these recommendations are to improve the reporting of data quality measures for studies that use observational clinical and administrative data, to ensure transparency and consistency in computing data quality measures, and to facilitate best practices and trust in the new clinical discoveries based on secondary use of observational data.

Acknowledgements

The EDM Forum Data Quality Collaborative was funded by the EDM Forum, which is supported by the Agency for Healthcare Research and Quality (AHRQ) through the American Recovery & Reinvestment Act of 2009, Grant U13 S19564-01. Dr. Kahn was also supported by AHRQ 1R01HS019908 (Scalable Architecture for Federated Translational Inquires Network), NIH/ NCATs Colorado CTSI grant number UL1 TR000154 (Colorado Clinical and Translational Sciences Institute) and PCORI Contract ME-1303-5581. The authors would also like to acknowledge Dr. Erin Holve and Dr. Marianne Hamilton Lopez from AcademyHealth for providing strategic guidance and assistance throughout the project.

Keywords

Informatics, data use and quality, research networks, observational data, comparative effectiveness

Disciplines

Health Information Technology | Health Services Research

Authors

Michael G Kahn, *University of Colorado*; Jeffrey S Brown, *Harvard Pilgrim Health Care Institute / Harvard Medical School*; Alein T Chun, *Cedars-Sinai Health System*; Bruce N Davidson, *Hoag Memorial Hospital Presbyterian*; Daniella Meeker, *University of Southern California*; Patrick B Ryan, *Observational Health Data Sciences and Informatics*; Lisa M Schilling, *University of Colorado, Denver*; Nicole G Weiskopf, *Oregon Health Sciences University*; Andrew E Williams, *Maine Medical Center Research Institute*; Meredith N Zozus, *Duke University*.

eGEMs

Transparent Reporting of Data Quality in Distributed Data Networks

Michael G. Kahn, MD, PhD;[i] Jeffrey S. Brown, PhD;[ii,iii] Alein T. Chun, PhD, IQCP;[iv] Bruce N. Davidson, PhD, MPH, MPI, IQCP;[v] Daniella Meeker, PhD;[vi] Patrick B. Ryan, PhD;[vii] Lisa M. Schilling, MD, MSPH;[i] Nicole G. Weiskopf, PhD;[viii] Andrew E. Williams, PhD;[ix] Meredith Nahm Zozus, PhD[x]

Abstract

Introduction: Poor data quality can be a serious threat to the validity and generalizability of clinical research findings. The growing availability of electronic administrative and clinical data is accompanied by a growing concern about the quality of these data for observational research and other analytic purposes. Currently, there are no widely accepted guidelines for reporting quality results that would enable investigators and consumers to independently determine if a data source is fit for use to support analytic inferences and reliable evidence generation.

Model and Methods: We developed a conceptual model that captures the flow of data from data originator across successive data stewards and finally to the data consumer. This "data lifecycle" model illustrates how data quality issues can result in data being returned back to previous data custodians. We highlight the potential risks of poor data quality on clinical practice and research results. Because of the need to ensure transparent reporting of a data quality issues, we created a unifying data-quality reporting framework and a complementary set of 20 data-quality reporting recommendations for studies that use observational clinical and administrative data for secondary data analysis. We obtained stakeholder input on the perceived value of each recommendation by soliciting public comments via two face-to-face meetings of informatics and comparative-effectiveness investigators, through multiple public webinars targeted to the health services research community, and with an open access online wiki.

Recommendations: Our recommendations propose reporting on both general and analysis-specific data quality features. The goals of these recommendations are to improve the reporting of data quality measures for studies that use observational clinical and administrative data, to ensure transparency and consistency in computing data quality measures, and to facilitate best practices and trust in the new clinical discoveries based on secondary use of observational data.

Introduction

Multi-institutional comparative effectiveness research (CER), pragmatic clinical trials, and patient centered outcomes research (PCOR) are new models for clinical investigation that have emerged, in part, due to the availability of clinically rich observational electronic data captured during routine care.[1–4] Electronic health records (EHRs) support the capture of detailed clinical, operational, and administrative data as part of routine clinical and business processes.[5–7] In addition, billing, claims, and drug fulfillment databases; specialized registries; and patient-reported and social media data expand the scope, richness, and completeness of available patient-level longitudinal health data. As EHR use becomes the norm,[8] the availability of observational clinical and administrative electronic data from a variety of practice- and patient settings provides an opportunity to transform clinical research and other analytics, allowing critical insights into the effectiveness of clinical- and system-level interventions on health outcomes, disease progression, and patient quality of life; medical product safety surveillance in real-world settings; clinical quality and patient safety interventions; and other secondary uses of data.[9–11] This "last mile" of clinical evidence generation and research translation focusing on patient-centered, clinically relevant outcomes with real-world patients in real-world settings should complement traditional randomized experiments and provide independent evidence in populations and settings where traditional clinical trials are not feasible.[12–14]

Substantial investments in national large-scale distributed research networks and other data sharing initiatives support the use of existing clinical data for multisite observational studies that have received substantial funding over the past five years, such as the Agency for Healthcare Research and Quality (AHRQ) distributed research programs, the FDA Mini-Sentinel network and, most recently, the Patient-Centered Outcomes Research Institute (PCORI)'s National Patient-Centered Clinical Research Network

[i]University of Colorado, [ii]Harvard Pilgrim Health Care Institute, [iii]Harvard Medical School, [iv]Cedars-Sinai Health System, [v]Hoag Memorial Hospital Presbyterian, [vi]University of Southern California, [vii]Observational Health Data Sciences and Informatics, [viii]Oregon Health Sciences University, [ix]Maine Medical Center Research Institute, [x]Duke University

(PCORnet).[1,15–21] In addition, the National Institutes of Health (NIH) has included data quality as a review criterion.[22] Propelled by the Meaningful Use provisions in the American Recovery and Reinvestment Act (ARRA) Health Information Technology for Economic and Clinical Health (HITECH) Act, there is increased interest in using patient-level data analytics to support a broad range of health care decisions for patients, providers, payers, and policymakers.[23,24] Increased patient engagement, in the form of expanded access to and management of their personal health care data, highlights patients' growing awareness of the value of their EHR data.[25–27] These synergistic activities have unified under the banner of collaborative Learning Health Systems that continuously generate new evidence and rapidly respond to that evidence to achieve optimal outcomes while minimizing inappropriate or ineffective resource utilization.[28–31] Learning Health Systems leverage clinical and administrative data from various disparate electronic-data sources for a broad range of uses including collaborative quality and process improvement programs, chronic disease management, public health surveillance, and more formal observational and CER and PCOR studies.[1,32,33]

Key to achieving any of the Learning Health System goals is access to high-quality clinical and administrative data so that meaningful conclusions can be derived from interventional and observational studies. Numerous publications illustrate the challenges of using observational data for CER, including recording biases, workflow differences, and issues with variations in data collection, such as invalid, inconsistent, and missing data.[34–38] These issues are further compounded when data need to be harmonized and combined from various intra-institutional data sources and then combined across multiple institutions. Appropriate data characterization must account for variation across institutions, data domains (e.g., demographics, observations, medications, laboratory results), and time.[34,36,39,40]

Given the potential limitations in observational data and its growing use to support new clinical insights, it is critical to develop a comprehensive data-quality reporting framework for assessing and describing data quality. Associated with this framework should be a comprehensive set of data-quality reporting recommendations that are applied to data received from any data source that are used to support new insights and new evidence that may have an impact on clinical care or health care policy. A common, comprehensive set of data quality assessments is necessary to promote trust by establishing reproducible processes and improving the transparency and integrity of the appropriate use of observational data in these settings.

Existing reporting recommendations include the Cochrane Collaboration for describing the quality of published evidence[41] and STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) recommendations for reporting results from observational clinical studies.[42] Many guidelines for reporting clinical research results are found on the Equator Network website.[43] None of the existing recommendations focus on reporting the results from data quality analyses. In an effort to address the acknowledged problems with observational data quality and the need for transparency to support transparency, reproducibility, and the development of common data quality reporting recommendations, the Data Quality Collaborative (DQC) was established with support from the Electronic Data Methods (EDM) Forum.

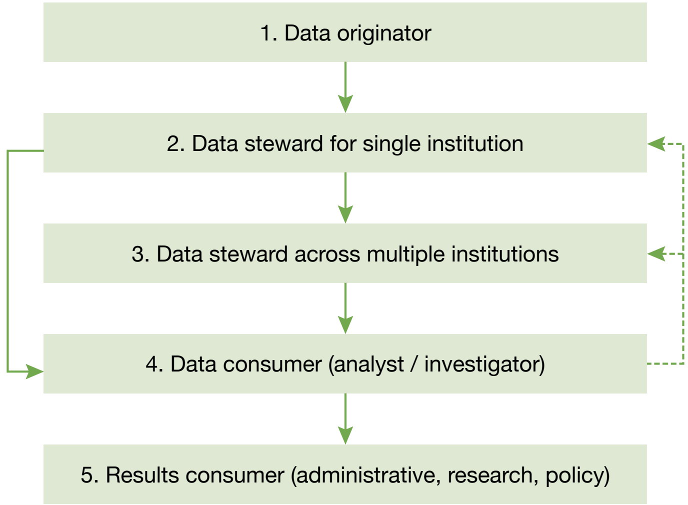
Defining Data Quality in Health Care

Data quality is a complex, multidimensional concept that defies a one-size-fits-all description.[44–48] Data quality is *context dependent*, which means the same data elements or data sources may be deemed high quality for one use and poor quality for a different use. The intended use determines which variables are relevant or have a high impact in a specific context. A key data-quality concept is "fitness for use"— a term originally used in industrial quality control but adapted by the Information Quality community to describe the context-dependent nature of data quality. Data are considered fit for use "if they are free of defects and possess desired features…for their intended uses in operations, decision making, and planning."[49] While it is unrealistic to expect observation data to ever be "free of defects," the emphasis on "desired features…for their intended uses" emphasizes that different settings ("intended uses") may have different criteria ("desired features"). For example, the presence of extreme values may be irrelevant in determining a median value for creating a rough estimate of the number of patients who might be eligible for a clinical study. Yet the same extreme values may have significant undue influence on the results of certain clustering algorithms or other analytic methods whose behaviors are sensitive to outliers.

Information Science professionals published formal models of data quality more than 20 years ago.[46,50–53] These publications focused on business-oriented definitions. Members of the DQC have recognized the need to develop similar data-quality assessment models for clinical data sets,[54–57] especially in the context of large multi-institutional distributed research networks.

Figure 1 highlights key data "actors" who have distinct roles in managing observational data and in evaluating the quality of data that are entrusted to their oversight or use. Data originators (Box 1 in Figure 1) refer to individuals who support the initial data capture or data recording systems, such as the EHR or patient survey tool, at an institution. Data stewards (Box 2 in Figure 1) refer to individuals who extract and manage data from originating data systems, such as members of a data warehouse or a report writing team. Data stewards for multiple institutions (Box 3 in Figure 1) are found in centralized data-coordinating centers and are responsible for integrating data across a network. While data stewards may have other data-oriented responsibilities, such as implementing data governance or enforcing honest broker- and regulatory compliance, we focus here only on their data management and data quality assessment activities. Data consumers receive data from previous data stewards, perform analyses, and create visualizations that attempt to highlight new discoveries or features for results consumers (Box 5 in Figure 1). We do not limit the uses and users of data in Box 4 to just clinical researchers, although our final recommendations define a category of analysis-specific recommendations that are organized in the language of hypothesis-driven clinical research (cohort, outcome, exposure, confounding variables). In Figure 1, we use the term "data stewards" to refer to all individuals who have some role in the creation, management, and use of observational clinical data (Boxes 1–4). Each data steward has unique roles and responsibilities regarding data oversight, data quality, data quality assessment, and data quality reporting. We refer to these roles and responsibilities as a "chain of data stewardship." Different aspects of data quality assessment may apply at each level in the chain of data stewardship. At the end of each arrow is a data consumer who receives data from the data steward in the previous box.

Figure 1. Chain of Data Stewardship with Key Data Stewards


1. Data originator
2. Data steward for single institution
3. Data steward across multiple institutions
4. Data consumer (analyst / investigator)
5. Results consumer (administrative, research, policy)

Notes: Dashed lines represent data quality issues referred back to previous data stewards.

The Value of Data Quality and Data Quality Assessment

Studies examining the cost of poor data quality have focused on non-health care businesses. Redman estimated that up to 5 percent of data within an organization are of poor quality.[51] A survey of 29 New Zealand organizations from government, banking and financial, utilities, and service organizations estimated that poor data quality has an average cost impact as high as 10 percent of an organization's annual revenues.[58] Respondents to a recent United Kingdom survey revealed estimates that 17 percent of all data are inaccurate, resulting in wasted budgets, loss of potential customers, reduced customer satisfaction, and inefficient use of staff.[59] Loshin categorized the negative impacts attributable to poor data quality into four areas: (1) direct and indirect financial losses; (2) reduced customer, employee, or supplier confidence and satisfaction; (3) reduced productivity due to increased workloads or reduced efficiency; and (4) increased risk and compliance.[60]

In health care delivery, data quality issues—particularly in the context of clinical documentation systems within EHR systems—have been demonstrated to have negative impacts on clinical decision support and patient safety[61–63] and in clinical research.[35,40,56,64–66] In epidemiology contexts, the misclassification of key data elements in administrative billing systems, such as the exposure or outcome of interest, can bias estimates of treatment effects if analyses are not properly calibrated for the sensitivity and specificity of the data capture process.[67] Yet there appears to be no literature that has studied the cost-benefit and business case for improved health care data quality. Data quality characterization is itself a data collection and analytic activity with costs inherent in the process of data quality assessment, monitoring, and governance. Thus, developing a detailed cost-benefit analysis of improved administrative and clinical data quality on clinical care, outcomes, and research remains an area of new research.

Even though research budgets for collecting randomized clinical trial (RCT)-quality data are very high due to the use of dedicated data collection tools and personnel, these data still have documented data quality deficiencies.[64,68,69] The secondary use of observational data, such as from EHRs and administrative claims, can dramatically reduce the cost of data collection and thus increase the efficiency of the research process, but incurs a cost of lower data quality. Known limitations of secondary use of observational data, such as only including patients when and if they access the health system, must be considered in study design, analysis, and interpretation. Assigning a value of improving observational data quality for research applications is challenging. Costs of incorrect research findings due to data quality issues are difficult to discover and quantify. Little information is available as data inspection is rarely replicated across studies, but replication or reanalysis studies do occasionally result in retractions.[70] Negative effects would be amplified if flawed conclusions were to alter policy decisions or clinical practice. Identifying the appropriate valuation and incidence parameters for costs and benefits of different approaches and rigor in data quality assessment might be assessed with a combination of empirical and expert-based assessments.

Methods

To promote the transparency of data-quality assessment reporting, the EDM Forum sponsored the creation of the DQC, which provided multiple convening activities (meetings and webinars) and an open-access, web-based information sharing environment, to bring together members across informatics, investigator, and methodology communities. The Collaborative focused on identifying a set of data-quality assessment-reporting recommendations that should be included as additional metadata to be associated with observational data. "Metadata" is "data about data." In this case, data quality measures are metadata that provide insights into the quality of the underlying database, data set, or data elements. Envisioned users of these data-quality reporting recommendations include data management staff responsible for releasing data for internal or external use, data analysts responsible for combining data across a Learning Health System or research network, clinical investigators using a data set for an analysis, and consumers—both scientific and lay pubic—of the inferences derived from analytic results.

The DQC conceptualized its work by defining key features that should be contained in a hypothetical "Table 1a" for data quality assessment reporting. “Table 1a” is intended to be analogous to the “Table 1” commonly found in publications on clinical studies. In those publications, the typical first table (“Table 1”) describes the key characteristics of a study population prior to presenting analytic findings, such as the distribution of age, sex, gender, race, socioeconomic class, and significant risk factors in the study population. In our context, rather than describing *clinical characteristics of a population*, our hypothetical first table (“Table 1a for data quality”) would report the key *data quality characteristics of a data set or data source* that might be used for multiple research and non-research purposes.

An initial draft set of recommendations was developed and revised by the DQC members (the authors) at weekly teleconference calls. The initial recommendations were derived by inspecting existing data-quality profiling methods used by DQC members, such as the Mini-Sentinel data characterization routines[71] and data quality rules embedded in the Observational Medical Outcomes Partnership (OMOP) data quality tools,[72,73] standard operating procedures for data quality assessment used in past or current projects or programs, published “best practices,” and data quality publications from both the clinical research and information sciences literatures.[37,39,51,56,74–77] This internal effort elicited approximately 50 initial potential recommendations. In December 2012 and June 2013, the EDM Forum’s DQC convened two face-to-face workshops, held eight months apart, that focused on reviewing the current draft for data-quality reporting recommendations. Workshop participants included the DQC and EDM Forum members and approximately 25 invited contributors who were identified through professional networks, publication authorship, and stakeholder recommendations to represent a broad range of data stakeholders—including data owners, data analysts, clinical investigators, statisticians, and policymakers (consumers). Approximately 50 percent of attendees attended both workshops. In addition, the EDM Forum disseminated a broad-based call for comments to the CER community via sponsored workgroups, CER-related listserv, electronic newsletters, and personal outreach. The EDM Forum provided online access to the evolving set of recommendations and invited comments to be posted. In 2012 and 2013, DQC members presented on two national webinars that were attended by over 100 participants, and presented panels at two national conferences describing multiple activities in data quality assessment, including draft data-quality reporting recommendations. All webinar and meeting participants were directed to the website for reviewing the draft recommendations and for posting comments or were encouraged to directly correspond with the lead author or EDM Forum staff. In June 2014, an updated version of the recommendations was again presented to relevant stakeholders at two EDM Forum-hosted workshops where additional input was solicited.

In response to multiple DQC meetings, public webinars and presentations, email outreach, and targeted solicitations, over 200 individuals were exposed to the data-quality reporting recommendations. In addition to in-meeting and webinar-based comments and discussion, approximately 20 responses were obtained either via the public-facing Web page or via direct email to a DQC member. In total, approximately 50 individual recommendations were obtained by the various stakeholder outreach efforts.

Delphi methods and other formal consensus methods were not employed to develop the initial or final recommendations. Recommendations that had strong consensus were added or removed from the evolving set. No formal voting process was used to determine the degree of consensus. Recommendations were continuously revised in response to stakeholder input and were reposted to the public website. Input was divided roughly evenly between requests for clarification and requests for simplification. No major additional categories were identified via public comments. Four versions of the recommendations were posted for review and comment. Using informal group consensus, recommendations that addressed similar issues were merged and recommendations that addressed issues beyond data quality—such as data access, security, and privacy concerns—were eliminated. The final recommendations reflect a compromise between an extensive list and the practical realities of implementation. For example, while it might be desirable to validate data elements against independent, external data sources, such as using United States census data as an external validation source for assessing demographic distributions in an observational data set, these data quality checks were considered out of scope for the data-quality reporting recommendations. The final set of recommendations was reduced to 20 data quality features for reporting.

We defined four contexts of data quality transparency that we used to organize the reporting recommendations: (1) data capture descriptions, (2) data processing descriptions, (3) data elements characterization, and (4) analysis-specific data elements characterization.

“Data capture descriptions” refers to information on the mechanisms by which data were observed or collected and recorded in the original electronic system, such as an EHR, and how data items are transformed and stored in the output data set or database. In the informatics literature, the originating data is often called the “source data” and the output data set is called the “target data.” Source data definitions specify the context in which data elements were measured and recorded, such as operational clinical systems versus specialized research data collection environments. Target data definitions describe the data environment used to transform and store the target data. In this phase, the data quality recommendations focus mostly on ensuring that the data collection context, assumptions, limitations, and other related contextual information are reported. Issues revealed at the data capture phase can render the rest of the data either useless or irrelevant for the intended data use.

“Data processing descriptions” refers to information on the transformational processes that have been applied to the source data, such as unit conversion; missing data imputation schemes; calculation of derived values; and mapping from original, possibly site-specific, codes to different code lists, during the creation of the target data. Complete transparency of all operations performed on data from initial data collection to analysis—such as the elimination of unrealistic extreme values or inferences of clinical state or status derived from complex data processing—is emphasized in this group of recommendations. Data processing transparency may not be entirely attainable because "raw" data may be subjected to internal or proprietary processes within vendor systems where only postprocessed data are available. However, descriptions of data processing operations performed during the creation of the target data within the analytic systems are necessary for reproducible research.

"Data elements characterizations" refers to information on observed data features of the target data, such as data distributions and missingness. These descriptions include both single-variable and multivariable data-characterization assessments. Most data quality programs focus on developing meaningful quantitative measures to characterize the observed features of one or more data variables. Data-quality reporting features in this grouping focus on reporting various computed statistical or distributional metrics.

"Analysis-specific data elements characterizations": Because fitness for use is context specific, it is difficult to define an exhaustive set of data characterization methods that are both specific and sufficient for all intended uses. The previous set of recommendations highlights the overall characteristics of the entire database. This set of recommendations specifies further data quality reporting on core data elements that are unique to a specific analysis or study that can have an impact on the validity of evidence generation and inference in this section. Variables used to determine cohort identification, exposures, outcomes, and potential confounders in a given analysis or study would be included here. While the specific variables involved in each analysis may vary, our recommended data characterizations should be reported with the same specificity and transparency as are all other core data-quality assessment measures.

Results

Table 1 in this paper presents 20 recommendations for reporting results on data quality. We have formatted the data-quality reporting recommendations to emulate the STROBE reporting recommendations. The existing STROBE recommendations include two items that are directly relevant to data quality reporting:

- STROBE Item 8: Data sources and measurements: For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.
- STROBE Item 12(c): Explain how missing data were addressed.

Our data-quality reporting recommendations extend the STROBE requirements around these two existing reporting categories with more detailed and structured reporting requirements to meet these recommendations.

Six recommendations apply to documentation of data capture systems. Three of the six recommendations describe features of the source data systems; the remaining three describe features of the target data system. The three source recommendations describe the context under which the original source data were collected: the data collection environment (clinical care versus research versus survey), the data collection technology (EHR, interviews, instruments), and the original intended purpose for data collection (clinical care, quality improvement, research). These source descriptions provide qualitative insights into the incentives of the data collectors in ensuring complete and accurate data capture that may conform to specific versus broad data definitions. For example, data collected in clinical care environments will tend to apply more practical operational definitions whereas data collected in research environments will tend to apply more specific formal definitions. In addition clinical care systems are known to have fewer real-time data-entry validation checks than are implemented in research data collection systems. The three target system recommendations provide insight into how source data are transformed (not included in target system, reformatted into a different structure such as from an integer into a text string, mapped to different codes such as from ICD9 into SNOMED-CT). Specific restrictions, such as only including in the target system patients seen at an institution since 2009 with at least one in-person contact, are also part of these recommendations. These reporting recommendations provide critical information about how the original data will be represented and how these new representations may have an impact on interpreting data in the target system.

Five recommendations focus on data processing and data provenance. Data provenance is concerned with ensuring that all of the processes applied to a data element from initial creation to final storage are made explicit. For example, a data element may be transformed, recoded, combined with other variables to create derived variables, removed as an outlier, replaced with a fixed value or flag, imputed when missing, and altered in other ways. Data provenance documents these manipulations so that data users fully understand all of the transformation steps that were applied to a data element that appears in a data set. Included in this set of recommendations are descriptions of how each transformation step was validated to ensure that the processing accurately performs the intended changes. Critical transformation-validation information includes confirming that the transformation process correctly handles missing, extreme, and incorrectly formatted values.

The characterization of data elements group lists only four reporting recommendations. Yet in actual implementations, this set of recommendations is likely to represent the component that consumes the largest amount of resources. Unlike previous recommendations that were mostly textual descriptions that describe key contextual features that can have an impact on data quality and data interpretation from source to target systems, the recommendations in this group are more computational and are focused on describing distributional characteristics of the target data set.

Table 1. Data Quality Assessment Documentation and Reporting Recommendations

	Item #	Recommendation
Data Capture		
1. Original data source		
Data origin	1	A description of the source of the original or raw data prior to any subsequent processing or transformation for secondary use. Examples would be “clinical practices via AllScripts EHR 2009,” “interviewer-administered survey,” or “claim for reimbursement.”
Data capture method	2	A description of the technology used to record the data values in electronic format. Examples would be “EHR screen entry via custom form,” “automated instrument upload,” and “interactive voice response (IVR).”
Original collection purpose	3	A description of the original context in which data were collected. Examples would be “clinical care and operations,” “reimbursement,” or “research”—and in which kinds of facilities data were collected—such as “ambulatory clinic,” “same-day surgery clinic,” and “clinical research center.”
2. Data steward information		
Data steward	4	A description of the type of organization responsible for obtaining and managing the target data set. Examples could be “PBRN,” “Registry,” Medical group practice,” and “State agency.”
Database model/data set structure	5	A description of how the data tables and variables are structured and linked in the target database or data set. Includes information on variable types (integer, date, string), min/max ranges if defined, and allowed values for enumerated categorical variable. Includes rules for mandatory/optional fields (variables), especially for fields used to link rows across tables.
Data dictionary/data set definitions	6	A description of data definitions used for data elements, including the URL to documentation if available on the Internet, that provides table- and field-level descriptions of data types and content for each element, and any required context for interpreting data within a patient or across the population. Whereas Recommendation #5 focuses on how the data are *structured* (data syntax), this requirement focuses on descriptions on what the data *mean* (data semantics) as described in the data definitions.
Data Processing/Data Provenance		
Data extraction specifications, including use of natural language processing to extract variables from text documents	7	Documentation on how the target data was obtained from the source data. Examples would be “direct data entry by medical personnel,” “indirect data entry by medical record chart abstraction guidelines,” and “natural language processing algorithms.” Should include the URL to the documentation of the data creation specifications if available on the Internet.
Mappings from original values to standardized values	8	Documentation on how original data values were transformed to conform to the target data model format. Documentation should list source values and describe the logic or mappings used to transform from the original source to the required target values.
Data management organization’s data transformation routines, including constructed variables	9	Documentation of any additional data alterations that were performed by the data management team in creating the final data set, such as replacing missing values by imputed values, removal of extreme values, and creation of additional computed values, such as BMI from raw height and weight observations. Should include the URL to documentation if available on the Internet. The documentation should allow an independent reader to trace a value in the target data set to the original source value(s) and should explain all operations performed on the data.
Data processing validation routines	10	Documentation of all data validation rules to which the data were subjected. Rules should identify both data elements and validation algorithms. Examples include comparisons of row counts between source and target data sets and an explanation for any differences in row count or documentation, and a listing of differences in the distribution of categorical data values across source-to-target mappings. Should include the URL to documentation if available on the Internet.
Audit trail	11	Documentation of all changes made to data values, user/system making the change and date/time of the change in the process of “cleaning” a data set prior to use. Reason for the change should be evident from data transformation routines or documented issues (e.g., correction of isolated error, replacement of missing values with standardized “missing value” flag).

Table 1. Data Quality Assessment Documentation and Reporting Recommendations (Cont'd)

	Item #	Recommendation
Data Elements Characterization		
Data format	12	For required data variables verify the format, proper storage, and that required elements are not missing. Examples include verifying that floating point values are not rounded to integer values, conversions across units of measures are correct, and that precision and rounding rules are as expected based on transformations.
Single element data descriptive statistics	13	For each variable, calculate the following descriptive statistics: • Available or not (#/% missing) • For continuous variables: min, max, mean, median, range, percentiles, etc. • For categorical variables—frequencies & proportions by category • If a specific distribution is anticipated, report on goodness-of-fit tests
Temporal constraints	14	Evaluate whether expected temporal constrains are violated or not. Examples include: • Start date and times occur before stop dates and times, • Distribution of intervals between successive measurements, • For time-series—changes in adjacent values and expected directionality in changes meet expectations, and • Conformance to state transition/sequencing rules.
Multiple variables cross validations/ consistency	15	Across two or more data variables that are known to be linked:[80] • Report violations of data model cardinality rules. A cardinality rule determines when zero, one, or more than one data rows in one table can be linked to one or more data rows in another table. • Report violations of data model primary/foreign key rules. A primary/foreign key requires that a row in one table (the foreign key) must point to a row in another table (the primary key). The primary key row must be present. • Report violations of cross-variables dependency rules. A cross-variables dependency states that one row can only exist if another row or value exists. For example, the state of pregnancy should exist only if the patient sex is female. • Report violations of co-occurrence rules. Systolic and diastolic blood pressures should always occur as a pair. • Report violations of co-measurement rules (two distinct measurements of the same observation). Age and date of birth should agree. • Report violations of mutual exclusivity rules. A patient should not be recorded as being dead and alive at the same time.
Analysis—Specific Data Quality Documentation (As Applied by Investigators or Analytic Team)		
Data cleansing/customization	16	Analytic- or study-specific additions to Item# 9
Data quality checks of key variables used for cohort identification	17	Analytic or study specific additions to Items #13–15 that focus on variables that identify cohorts, detect outcomes, define exposures, and participate as covariates. Where these variables may be affected by other related (perhaps causal) variables, these influential variables should also be included. The list of variables contained in these assessments will vary by intended analysis/clinical study. However variables assessed should be organized according to the following categories: cohort, outcome, exposure, confounding.
Data quality checks of key variables used for outcome categorization	18	
Data quality checks of key variables used to classify exposure	19	
Data quality checks of key confounding variables	20	

Notes: "Source data" refers to the original originating data. "Target data" refers to the data as received by the data user.

These recommendations are also the most technical and statistical, using methods such as goodness-of-fit testing for variables that have an expected distribution, state-transition checks for variables that are expected to exhibit a specific sequence of values over time (e.g., inpatient admission event should be followed only by a discharge or death event; a death event should not be followed by a clinic encounter), or primary and foreign key constraints such as the provider listed in a patient's record (a foreign key) must *always* be a provider already present in the provider table (the primary key).

The reporting requirements in this section are not as prescriptive as are the previous reporting requirements, listing *classes* or types of data-quality reporting recommendations rather than specific reporting elements. For example, Recommendation 15 states: "conformance to co-occurrence rules." A co-occurrence rule examines the presence of two or more variables that are expected to be recorded at the same time. Systolic and diastolic blood pressure measurements are a simple example of co-occurrence. This recommendation does not list specific co-occurrence rules to check. A small data set may have just a few variables that are expected to co-occur and other cross validation rules mentioned in Recommendation 15, whereas a very large data set may have hundreds of rules. In actual practice, large data sets tend to start with a small set of validation rules and expand the set of rules over time as more sophisticated data-quality checking capabilities are implemented. The recommendations are sequenced in order of increasing complexity—from single variables to multiple variables and from constant measures to temporal measures. The intent of the recommendations in this group is to ensure that those data-quality validation checks that are performed are reported and that these recommendations can be used to organize the reported results.

While the "data elements characterization" recommendations deal with classes or types of data-quality reporting recommendations rather than specific reporting elements, the fourth reporting recommendation group, "analysis-specific data elements characterization," is even less prescriptive. This grouping recognizes that it is not possible to anticipate all of the ways in which data are to be used in a complex analytic context, so highly specialized and very specific data-quality checks and reporting should be performed when the data are used in a very focused analytic context. The reporting recommendations in this section highlight that additional data-quality assessment checks should be targeted to key variables that are used to identify cohorts, detect outcomes, define exposures, and participate as covariates. Where these variables may be affected by other related (perhaps causal) variables, these influential variables should also be included. The data-quality reporting recommendations in this section focus on determining fitness for a clearly defined use, where the research study hypotheses or intended analytic model set the context of use and the study design and analytic methods frame the criteria for fitness. We further explore the relationship between fitness for use and these recommendations in Section 4.

Discussion

Current data-quality assessment approaches are based on individualized ad hoc methods and characterizations. Substantial "data cleaning" efforts occur behind the scenes where the details often are not disclosed with the final data sets. The recommendations listed in Table 1 in this paper are intended to replace ad hoc (if at all) data-quality reporting methods with an organized framework. We have developed a set of data-quality reporting recommendations based on community input from a convenience sample of clinical research practitioners engaged in a wide range of large-scale, multi-institutional data networks spanning prospective regulated clinical trials, pragmatic clinical trials, and observational outcome studies. The data-quality reporting recommendations provide a logical and comprehensive set of explicit disclosures that focus on increasing the transparency of these important data quality processes. Although the focus and framing of this work is on reporting data-quality results in the setting of clinical research and new knowledge discovery, the same principles can be applied to any use of observational clinical data.

"Fitness for use" is a key concept in determining if a data set can be (or should be) used in a specific task. When viewed from the perspective of a general data resource, such as an institutional research data warehouse, the fitness for use concept applies only weakly. In this context, broad global assessments of data completeness, consistency, and accuracy may apply. As the data set size increases, the number of *potential* criteria for fitness for use can grow, resulting in an exponential explosion of nonspecific data-quality checks that are based on an overall understanding of how the various data elements should "hang together" in a general clinical context that may not be relevant for a more focused intended use. While these data quality checks are important in an initial assessment of fitness for use for a specific analysis or study, these checks may not be sufficient to understand the fitness for a highly specialized use. An analysis or study may need to calculate and report on additional data-quality checks on key analytic variables as highlighted in the "analytic-specific data characterizations" recommendations. This work does not provide explicit guidance on how to determine if a data set is, in fact, fit for its intended use, because that determination is highly context specific. The intent of these recommendations is to ensure that the information on data quality is reported with sufficient specificity and completeness to enable data users to determine if a data set is sufficient for their needs and to enable consumers of the results from a study or analysis to independently determine if the data set was fit for how it was used.

While we describe the concept of a "Table 1a for Data Quality Reporting" to frame our work, we do not limit the proposed contents of our recommendations to a set of elements that would meet the physical constraints of typeset journal pages. Rather, we acknowledge the growing use of electronic annexes that allow detailed information to be associated with a publication. As a long-term objective, we seek to create a computer-readable data structure that allows detailed data-quality assessment metrics to be incorporated into analytic programs that use the associated data sets.

Our recommendations can be applied by all data stewards and owners along the chain of data stewardship (Figure 1), with all stewards documenting and adding their data quality assessment results as they pass data sets along the chain to the next data steward or user. If our recommendations are adopted, a data set will accumulate additional data-quality documentation over its lifetime, ensuring that the needed details are available to all subsequent downstream data consumers. With each additional data-quality assessment result added to a data set, confidence in that data set should increase, making data assets with extensive data-quality assessment measures more valuable over time.

Limitations

While the EDM Forum engages a broad community of CER and PCOR stakeholders, it represents only a small fraction of observational clinical-data stewards and data users. Our outreach efforts were based on a convenience sample of individuals who participated in either the EDM Forum-hosted workshops or webinars, or other contacts. No purposeful sampling method was used to ensure engagement from specific stakeholder communities. Thus we cannot be certain that we reached saturation[78] in identifying all relevant reporting recommendations.

At all stages of developing the data quality recommendations, participants were instructed not to focus on the level of effort or resources needed to generate the data quality measures nor to consider the size of the files containing the results. Yet in order for these recommendations to move into practice, these real and significant logistic and resource limitations need to be addressed. One possibility that we are exploring is to develop a toolbox of common data-quality assessment routines that could be used by data owners to create the data characterization elements enumerated in these recommendations. Using common assessment routines would not only reduce the computational burden but would also ensure that reported data-quality measures were computed in the same way and therefore are comparable. Issues with differences in data models, terminology, and representation make this possibility difficult to realize.

We have not designed a format for our proposed Table 1a for data quality reporting. Technical specifications, such as ISO and IEC 11179, have been created to structure metadata, which would include the information contained in our recommendations.[79] However, metadata specifications are written in formats that are not widely supported and are not easily understood by humans.

Current data-quality assessment methods precompute a wide range of data descriptive variables that attempt to anticipate the types of data quality issues that might be most relevant to preconceived intended data uses. Analysis-specific data-quality checks specialize these computations based on the unique requirements of an analysis. Our recommendations are constructed with this approach in mind. However, interactive data-visualization techniques are being developed that allow users to dynamically explore data distributions and relationships beyond static data-distribution measures and graphical plots, thus enabling more flexible data characterizations than is possible with preconfigured data-quality measurements. It is not clear how we would include data-quality reporting recommendations if data-quality assessment activities were performed within an interactive data-visualization tool rather than by the current methods of creating static metrics or graphical visualizations.

Future Work

Future directions include empirical evaluations of the level of effort and value associated with implementing the full spectrum of the recommended documentation in a series of case studies based on past research projects. Existing, large national data-sharing environments—such as Mini-Sentinel, HMO Research Network, and OMOP—have extensive data-model documentation and data quality programs or tools that are publically available. However these data-quality assessment tools only work on data sets that are structured according to the specific data structure. An effort is underway to evaluate the already-available documentation and the data quality characterizations and checks contained in existing tools against the recommendations in Table 1.

As described in Section 3, modeling the return on investment for data quality assessment and the institutional and scientific risks associated with the inappropriate use of poor quality data for different stakeholders in a formal cost-benefit analysis is a critical missing component. A cost-benefit analysis might also reveal the value to stakeholders in investing in a toolbox of common data-quality assessment tools and common best practices that could be used or customized by data owners and users. Using common assessment routines would ensure that reported data-quality measures were comparable, and would reduce the marginal costs of repeating data-quality assessments. The use of well-established best practices could accelerate the implementation of more sophisticated data-quality assessment programs.

Our focus has been mostly on data-quality reporting recommendations when using observational clinical data for generating new evidence as a means of improving transparency and trust in the conclusions derived from these activities. Table 1 can also be a starting point for establishing a minimal but useful set of data-quality reporting components to be included with any use of observational clinical data. However, while reporting the strengths and weaknesses of a data set may improve transparency and trust, the same activity can also have serious unintended negative consequences, such as revealing internal data quality problems or embarrassing data contributors who subsequently withdraw from a data-sharing network. We need a nonpunitive culture that embraces transparency as a means of improving data quality over time. The significance of these nontechnical barriers must be acknowledged and addressed in order for these data quality-reporting recommendations to become commonplace practice.

Acknowledgements

The EDM Forum Data Quality Collaborative was funded by the EDM Forum, which is supported by the Agency for Healthcare Research and Quality (AHRQ) through the American Recovery & Reinvestment Act of 2009, Grant U13 HS19564-01. Dr. Kahn was also supported by AHRQ 1R01HS019908 (Scalable Architecture for Federated Translational Inquires Network), NIH/NCATs Colorado CTSI grant number UL1 TR000154 (Colorado Clinical and Translational Sciences Institute) and PCORI Contract ME-1303-5581. The authors would also like to acknowledge Dr. Erin Holve and Dr. Marianne Hamilton Lopez from AcademyHealth for providing strategic guidance and assistance throughout the project.

References

1. Lopez MH, Holve E, Sarkar IN, Segal C. Building the informatics infrastructure for Comparative Effectiveness Research (CER): A review of the literature. Med Care. 2012 Jul;50 Suppl:S38–48.
2. Selby JV, Beal AC, Frank L. The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda. JAMA J Am Med Assoc. 2012;307(15):1583–4.
3. Clancy C, Collins FS. Patient-Centered Outcomes Research Institute: the intersection of science and health care. Sci Transl Med. 2010;2(37):37cm18–37cm18.
4. Washington AE, Lipstein SH. The Patient-Centered Outcomes Research Institute—promoting better information, decisions, and health. N Engl J Med [Internet]. 2011 [cited 2013 Aug 29];365(15). Available from: http://www.nejm.org/doi/full/10.1056/nejmp1109407
5. Institute of Medicine (U.S.). Committee on Improving the Patient Record, Dick RS, Steen EB. The computer-based patient record: An essential technology for health care, revised edition. Washington, D.C.: National Academy Press; 1997. xii, 190 p. p.
6. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. J Am Med Inf Assoc. 2007 Jan 1;14(1):1–9.
7. Weiner MG, Embi PJ. Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? Ann Intern Med. 2009;151(5):359–60.
8. Borycki EM, Newsham D, Bates DW. eHealth in North America. Yearb Med Inform. 2013;8(1):103–6.
9. Holve E, Segal C, Hamilton Lopez M. Opportunities and challenges for Comparative Effectiveness Research (CER) with electronic clinical data: A perspective from the EDM forum. Med Care. 2012 Jul;50 Suppl:S11–8.
10. Embi PJ, Payne PRO. Evidence Generating Medicine: Redefining the Research-Practice Relationship to Complete the Evidence Cycle. Med Care. 2013 Aug;51:S87–91.
11. Southworth MR, Reichman ME, Unger EF. Dabigatran and Postmarketing Reports of Bleeding. N Engl J Med. 2013;368(14):1272–4.
12. Concato J, Shah N, Horwitz R. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000 Jun 22;342(25):1887–92.
13. Bloomrosen M, Detmer D. Advancing the Framework: Use of Health Data - A Report of a Working Conference of the American Medical Informatics Association. J Amer Med Inf Assoc. 2008 Dec;15(6):715–22.
14. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Tourchia M. Developing a protocol of observational comparative effectiveness research: A user's guide [Internet]. Rockville, MD; 2013. Report No.: AHRQ Publication No. 12(13)-EHC099. Available from: www.effectivehealthcare.ahrq.gov/Methods-OCER.cfm
15. Chen RT, Glasser JW, Rhodes PH, Davis RL, Barlow WE, Thompson RS, 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 Jun;99:765–73.
16. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, et al. Building a virtual cancer research organization. J Natl Cancer Inst Monogr. 2005;12–25.
17. Pace WD, Cifuentes M, Valuck RJ, Staton EW, Brandt EC, West DR. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009 Sep 1;151(5):338–40.
18. Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, et al. Design of a national distributed health data network. Ann Intern Med. 2009 Sep 1;151(5):341–4.
19. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf. 2012 Jan;21 Suppl 1:1–8.
20. Holve E, Segal C, Lopez MH, Rein A, Johnson BH. The Electronic Data Methods (EDM) Forum for Comparative Effectiveness Research (CER). Med Care. 2012 Jul;50 Suppl:S7–10.
21. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, et al. The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration. EGEMs Gener Evid Methods Improve Patient Outcomes [Internet]. 2014 Mar 24 [cited 2014 Apr 16];2(1). Available from: http://repository.academyhealth.org/egems/vol2/iss1/2
22. RFA-RM-13-012: NIH Health Care Systems Research Collaboratory - Demonstration Projects for Pragmatic Clinical Trials Focusing on Multiple Chronic Conditions (UH2/UH3) [Internet]. [cited 2013 Dec 9]. Available from: http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-13-012.html
23. Blumenthal D, Tavenner M. The "Meaningful Use" Regulation for Electronic Health Records. N Engl J Med. 2010;363(6):501–4.
24. Blumenthal D. Launching HITECH. N Engl J Med. 2010 Feb 4;362:382–5.
25. Lopez MH, Holve E, Rein A, Winkler J. Involving Patients and Consumers in Research: New Opportunities for Meaningful Engagement in Research and Quality Improvement. 2012; Available from: http://repository.academyhealth.org/edm_briefs/2/
26. Bates DW, Bitton A. The future of health information technology in the patient-centered medical home. Health Aff (Millwood). 2010;29(4):614–21.
27. Ralston JD, Coleman K, Reid RJ, Handley MR, Larson EB. Patient experience should be part of meaningful-use criteria. Health Aff (Millwood). 2010;29(4):607–13.
28. Etheredge LM. A rapid-learning health system. Health Aff Proj Hope. 2007 Apr;26(2):w107–18.
29. Roundtable on Evidence-Based Medicine. The Learning Healthcare System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine). Olsen L, Aisner D, McGinnis JM, editors. 2007; Available from: http://www.nap.edu/openbook.php?record_id=11903

30. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, et al. Rapid-Learning System for Cancer Care. J Clin Oncol. 2010 Jun 28;28(27):4268–74.
31. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010 Nov 10;2:57cm29.
32. Grossman C, McGinnis JM. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. National Academies Press; 2011.
33. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System--a national resource for evidence development. N Engl J Med. 2011 Feb 10;364(6):498–9.
34. 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 JAMIA. 2000 Feb;7(1):55–65.
35. Arts DGT, De Keizer NF, Scheffer G-J. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc JAMIA. 2002 Dec;9(6):600–11.
36. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: Data quality issues and informatics opportunities. AMIA Summits Transl Sci Proc. 2010;2010:1–5.
37. Canadian Institute for Health Information. The CIHI data quality framework [Internet]. Ottawa, Ont.: CIHI; 2009. Available from: http://www.cihi.ca/CIHI-ext-portal/pdf/internet/DATA_QUALITY_FRAMEWORK_2009_EN
38. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research: Med Care. 2013 Aug;51:S30–7.
39. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013;51:S22–9.
40. Hartzema AG, Reich CG, Ryan PB, Stang PE, Madigan D, Welebob E, et al. Managing data quality for a drug safety surveillance system. Drug Saf Int J Med Toxicol Drug Exp. 2013 Oct;36 Suppl 1:49–58.
41. Higgins JP, Green S, Collaboration C. Cochrane handbook for systematic reviews of interventions [Internet]. Wiley Online Library; 2008 [cited 2013 Aug 29]. Available from: http://www.cochrane.org
42. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008 Apr;61(4):344–9.
43. Simera I, Altman DG, Moher D, Schulz KF, Hoey J. Guidelines for reporting health research: the EQUATOR network's survey of guideline authors. PLoS Med. 2008 Jun 24;5(6):e139.
44. Oliveira P, Rodrigues F, Henriques PR. A Formal Definition of Data Quality Problems. IQ [Internet]. 2005 [cited 2013 Jul 4]. Available from: http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202005/Papers/AFormalDefinitionofDQProblems.pdf
45. Kim W, Choi B-J, Hong E-K, Kim S-K, Lee D. A taxonomy of dirty data. Data Min Knowl Discov. 2003;7(1):81–99.
46. Wand Y, Wang R. Anchoring data quality dimensions in ontological foundations. Comm ACM. 1996;39:86–95.
47. Sanna F. Discovering Data Quality [Internet]. 2006. Available from: http://www.dmreview.com/portals/portalarticle.cfm?articleId=1056199&topicId=230005
48. Chapman AD. Principles of Data Quality V1.0. Report for the Global Biodiversity Information Facility [Internet]. GBIF Secretariat; 2005. Available from: http://www.gbif.org/orc/?doc_id=1229
49. Wang R, Strong D. Beyond accuracy: What data quality means to data consumers. J Manag Inf Syst. 1996;12:5–34.
50. Batini C. Data Quality: Concepts, Methodologies and Techniques. Berlin: Springer; 2006. 262 p.
51. Redman TC. Data Quality: The Field Guide. Boston: Digital Press; 2001. 241 p.
52. Redman TC. Data quality for the information age. Boston: Artech House; 1996. 303 p.
53. Winkler WE. Methods for evaluating and creating data quality. Inf Syst. 2004 Oct;29(7):531–50.
54. 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 Jul;50 Suppl:S21–9.
55. 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 Jan 1;20:144–51.
56. Nahm M. Data quality in clinical research. Clinical Research Informatics. London: Springer-Verlag; 2012. p. 175–201.
57. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, et al. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. Int J Med Inf. 2013 Jan;82(1):10–24.
58. Office of the National Coordinator. Report to Congress: Update on the adoption of health information technology and related efforts to facilitate the electronic use and exchange of health information [Internet]. Office of the National Coordinator for Health Information Technology; 2014 Oct [cited 2013 Dec 2] p. 54. Available from: http://www.healthit.gov/sites/default/files/rtc_adoption_and_exchange9302014.pdf
59. Joel Curry. The Data Advantage: How accuracy creates opportunity. A Experian QAS 2013 Research Report [Internet]. 2013 [cited 2013 Dec 2]. Available from: http://www.experian.co.uk/assets/marketing-services/white-papers/wp-qas-the-data-advantage.pdf
60. Loshin D. The Practitioner's Guide to Data Quality Improvement. Morgan Kaufmann; 2010. 432 p.
61. Kohn LT, Corrigan JM, Donaldson MS. To err is human: Building a safer health system [Internet]. Washington D.C.: National Academy Press; 2000.
62. Riain C, Helfert M. An evaluation of data quality related problems patterns in healthcare information systems. In: Isaias P, Nunes M, dos Reis A, editors. IADIS virtual multi conference on computer science and information systems. 2005.

63. Brennan PF, Stead WW. Assessing data quality: from concordance, through correctness and completeness, to valid manipulatable representations. J Am Med Inf Assoc. 2000 Jan;7:106–7.
64. Nahm ML, Pieper CF, Cunningham MM. Quantifying data quality for clinical trials using electronic data capture. PloS One. 2008;3(8):e3049.
65. Baigent C, Harrell FE, Buyse M, Emberson JR, Altman DG. Ensuring trial validity by data quality assurance and diversification of monitoring methods. Clin Trials Lond Engl. 2008;5(1):49–55.
66. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. J Biomed Discov Collab. 2011;6:48–52.
67. Greenland S, Lash TL. Bias Analysis. Modern epidemiology. Philadelphia: Wolters Kluwer / Lippincott Williams & Wilkins; 2008.
68. Davis JR, Institute of Medicine (U.S.). Roundtable on Research and Development of Drugs Biologics and Medical Devices. Assuring data quality and validity in clinical trials for regulatory decision making : workshop report [Internet]. Washington, DC: National Academy Press; 1999. xii, 76 p. p. Available from: http://books.nap.edu/catalog/9623.html
69. Gassman JJ, Owen WW, Kuntz TE, Martin JP, Amoroso WP. Data quality assurance, monitoring, and reporting. Control Clin Trials. 1995 Apr;16(2 Suppl):104S – 136S.
70. Grieneisen ML, Zhang M. A Comprehensive Survey of Retracted Articles from the Scholarly Literature. PLoS ONE. 2012 Oct 24;7(10):e44118.
71. Mini-Sentinel Coordinating Center. Mini-Sentinel Standard Operating Procedure: Data Quality Checking and Profiling. 20; Available from: http://www.mini-sentinel.org/work_products/About_Us/Mini-Sentinel_SOP_Data-Quality-Checking-and-Profiling.pdf
72. Observational Medical Outcomes Partnership. OSCAR - Observational Source Characteristics Analysis Report (OSCAR) Design Specification and Feasibility Assessment [Internet]. 2011 [cited 2013 Apr 1]. Available from: http://omop.fnih.org/OSCAR
73. Observational Medical Outcomes Partnership. Generalized Review of OSCAR Unified Checking [Internet]. 2011 [cited 2013 Apr 1]. Available from: http://omop.fnih.org/GROUCH
74. Maydanchik A. Data quality assessment. Bradley Beach, NJ: Technics Publications; 2007. xiv, 321 p. p.
75. Magnusson D, Bergman LR, European Network on Longitudinal Studies on Individual Development. Data quality in longitudinal research. Cambridge [England] ; New York: Cambridge University Press; 1990. xii, 285 p. p.
76. Singh S. Evaluation of data quality. [London, England], International Statistical Institute by Oxford University Press,; 1987. p. 618-643. p.
77. Sadiq S, editor. Handbook of Data Quality. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013.
78. Creswell JW. Qualitative, Quantitative, and Mixed Methods Approaches (Crewell, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches) 4th edition. Fourth Edition. Thousand Oaks, California: SAGE Publications, Inc; 2013. 273 p.
79. ISO. ISO/IEC 11179 - Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes. International Standard Organization; 2013.
80. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care. 2012 Jul;50 Suppl:S60–7.