HHS Public Access

Author manuscript

Biostat Epidemiol. Author manuscript; available in PMC 2020 April 06.

Published in final edited form as:

Biostat Epidemiol. 2020 ; 4(1): 6–14. doi:10.1080/24709360.2019.1572344.

Clinical data quality: a data life cycle perspective

Chunhua Weng

Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract

Clinical data is the staple of modern learning health systems. It promises to accelerate biomedical discovery and improves the efficiency of clinical and translational research but is also fraught with significant data quality issues. This paper aims to provide a life cycle perspective of clinical data quality issues along with recommendations for establishing appropriate expectations for research based on real-world clinical data and best practices for reusing clinical data as a secondary data source.

Keywords

Clinical data; data quality; learning health system

Clinical data – definition

The wide adoption of electronic health records (EHR), patient registries, clinical trial management systems, patient self-tracking devices, and other digital health technologies have made available large amounts of clinical data. The National Academy of Medicine defines clinical data broadly as data collected from any of these sources [1]: (1) data generated from clinical care processes; (2) data from clinical trials; and (3) patient registry data [1]. Example of available clinical data used for generating new insights include but are not limited to single-site electronic health records, administrative data, Health Maintenance Organization (HMO) networks data, Veterans Administration data, county birth and death statistics, industry-sponsored disease registries, Framingham Heart Study, and the National Heart, Lung, and Blood Institute (NHLBI)-funded studies available for public use (https://biolincc.nhlbi.nih.gov/studies/). Representing the resource most central to healthcare progress [2], clinical data contain rich information about healthcare processes, disease manifestations, diagnosis and treatment pathways, and treatment outcomes, and hence hold

CONTACT Chunhua Weng, chunhua@columbia.edu, Department of Biomedical Informatics, Columbia University, New York, NY, USA.

Notes on contributor

Dr Chunhua Weng is a tenured Associate Professor of Biomedical Informatics at Columbia University and an elected fellow of the American College of Medical Informatics (ACMI). She has also been co-leading the Biomedical Informatics Resource for the Columbia CTSA (The Irving Institute for Clinical and Translational Science) since 2011. She is also the chair of the Clinical Research Informatics Working Group for AMIA. Dr Weng holds a Ph.D. in Biomedical and Health Informatics from the University of Washington at Seattle. As an active researcher in the field of Clinical Research Informatics since 2000, Dr Weng has published on text knowledge engineering for optimizing clinical research eligibility criteria, electronic screening methods for clinical trial recruitment, EHR data quality assessment and data analytics, and deep EHR phenotyping.

Disclosure statement

No potential conflict of interest was reported by the author.

great promise to enable early clinical research feasibility assessment and to facilitate clinical trial recruitment, comparative effectiveness research, phenotype-driven rare or genetic disease diagnosis, and large-scale observational studies for robust evidence generation and validation.

In order to harness the value of this important data resource, many regional, national, or international clinical data research networks are burgeoning to enable large-scale clinical data sharing and federation. Notable examples include the EHR4CR project (www.ehr4cr.eu/) in Europe, the PCORnet Clinical Data Research Network (http://www.pcornet.org) in the United States, the Observational Health Data Sciences and Informatics (OHDSI) global data network (www.ohdsi.org), and the latest national initiative in the United States, which is to collect clinical data from multiple sources for the All of Us program (https://allofus.nih.gov). Many biomedical and clinical researchers are trying to leverage these clinical data collected from representative patient populations to accelerate efficient comparative effectiveness research and for robust evidence generation and evaluation using real-world data.

Clinical data – quality issues

Meanwhile, many studies have identified significant issues in clinical data quality [3]. Weiskopf et al. surveyed the literature and identified five substantively different dimensions of data quality [4]: (1) Completeness: Is truth about a patient present in the data? (2) Correctness: Is an element that is present in the data true? (3) Concordance: Is there an agreement between elements in the data, or between the data and another data source? (4) Plausibility: Does an element in the data makes sense in light of other knowledge about what that element is measuring? (5) Currency: Is an element in the dataa relevant representation of the patient state at a given point in time? Kahn et al. extended this taxonomy and developed a comprehensive data quality assessment framework by harmonizing the data quality terminology like this and many other data quality assessment frameworks [5]. In this framework, using a collaborative approach to reach consensus, the clinical data quality research community harmonized the major categories of data quality definitions into the following three: conformance, completeness, and plausibility, organized by verification and validation contexts, respectively. Verification focuses on how data values match expectations with respect to metadata constraints, system assumptions, and local knowledge, while validation focuses on the alignment of data values with respect to relevant external benchmarks. Each category has specific subcategories. For example, there exist value conformance (*e.g. allowable values for sex include only female or male*), relation conformance (*e.g. patient record number is linked to other tables as required*), and computational conformance (*e.g. recorded BMI is consistent with derived BMI from other information*). Plausibility can be further categorized into uniqueness plausibility (*e.g. patients from a single institution do not have multiple medical record numbers*), temporal plausibility (*e.g. death date should be later than birthdate*), and atemporal plausibility (*e.g. counts of unique patients by diagnoses are as expected*).

The aforementioned data quality frameworks have largely been developed based on the issues identified from structured clinical data. Since EHR text is not amenable for

computational analyses, data quality issues are harder to systematically investigate in text. It is well-known that EHR text is overflowing with copy-n-paste text [6]. Cohen et al. reported text redundancy in clinical notes and discussed its impact on text mining performance [7]. Tange et al. identified the granularity issues in EHR narratives and discussed their effect on speed and completeness of information retrieval [8]. Nuanced EHR narratives also contain non-patient specific, ambiguous, uncertain, or outdated information, which can affect the EHR information integrity [9].

Such data quality issues can significantly limit the suitability of clinical data for different types of clinical research studies. Shang et al. proposeda conceptual framework for assessing the data suitability of observational studies [10]. The key considerations regarding the data suitability for observational studies include explicitness of policy and data governance, relevance, availability of descriptive metadata and provenance documentation, usability, and quality.

Quality checking throughout the clinical data life cycle

Clinical data quality is subject to multiple threats occurring at different phases in the clinical data life cycle, which includes (1) data generation, (2) data transformation, (3) data reuse, and (4) post-reuse data quality reporting and data correction (Figure 1). Initially collected during clinical encounters to document patient care, the quality of clinical data hinges on the documentation behaviors of multidisciplinary care providers, which may have heterogeneous sampling biases or measurement biases [11], and the fitness of the heterogeneous data encoding standards imposed by different users or different institutions. Rusanov performed statistical analyses on the clinical data warehouse in New York Presbyterian Hospital and confirmed that sick patients tend to have more data than healthy patients so that clinical researchers who select patients with adequate clinical data can be implicitly biased towards sick patients in their recruitment strategy and introduce unintended biases into the clinical research designs [12,13]. Moreover, clinical data encoded using The International Classification of Diseases (ICD) versions 9 or 10 typically have less granularity than clinical data encoded using the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT). Furthermore, the existing clinical terminologies have limited concept coverage. Winnenberg et al. assessed the coverage of phenotypes in standard terminologies and found that phenotype concept coverages range from 54% in the Unified Medical Language System (UMLS) to 9% in ICD-10. Dhombres et al. found that only 30% of Human Phenotype Ontology concepts can be completely mapped to the SNOMED-CT standard [14]. Such poor coverage in the data encoding terminology can potentially lead to various intended documentation patterns, such as omitting important information due to the lack of appropriate codes or creating local codes that cannot be reused by others or outside the original institution, further imposing implicit information loss at the documentation phase to varying degrees and exacerbating the lack of interoperability among clinical datasets.

Later, clinical data from various sources get aggregated to form integrated clinical data repositories or networks. The Extraction-Transform-Load (ETL) process (https://en.wikipedia.org/wiki/Extract,_transform,_load) inevitably introduces further information

loss by transforming the original data into a selected standard-based format. Such information loss can be due to the lack of granularity in the concept representation or the lack of coding standards for the rich semantics in the target clinical data model. The quality of the resulting clinical data is influenced by the accuracy of the ETL processes, which can be error prone. Such ETL processes often span a long-time period after the original clinical data is collected and hence often the contextual information is already unavailable at the ETL process. They also involve a large number of users and computer systems in the process, sometimes from different institutions. The complexity and opaqueness of the ETL process and the lack of data provenance throughout this complex 'black-box' process can both further deteriorate the clinical data quality and introduce untraceable problems. Although there is no published evidence regarding the information loss rate at each ETL step towards the establishment of the large clinical data networks, it is well-known that only a subset of clinical data gets included in large clinical data networks, where the population representativeness of the patient cohort and data completeness needs careful investigation before researchers decide to use such clinical data for research studies.

After the ETL processes, clinical data will be made available to researchers for scientific discoveries. It is often not until this point is data quality issues starting to be noticed by different stakeholders, who have little clue what has happened to the data before this point. Data quality essentially is 'fitness-for-use,' but the uses, including reuses, do not come until late in the clinical data life cycle. Reuse of clinical data at this phase is out of the raw context, which unfortunately is not the ideal [15]. Sim et al. pointed out that

> sharing patient-level data from human studies would help investigators make more and better discoveries more quickly and with less duplication. However, this happy circumstance will occur only if investigators interpret the raw data properly within the context of the study that generated that data.

Detection of data quality issues at this phase requires substantial and sometimes difficult reverse engineering, out of context. Clinical data often have non-random missingness. Weiskopf analyzed four different types of completeness in clinical data: i.e. breadth completeness, density completeness, documentation completeness, and predictive completeness [16]. Only 0.6% of clinical data in an examined data warehouse met four completeness criteria, a finding not likely unique to a single institution. Possible sampling biases and measurement biases can be elusive to researchers.

Various data quality checks have been developed to detect clinical data quality issues at the reuse phase. Typical methods include uses of gold standards, data element agreement, element presence, data source agreement, distribution comparison, validity check, and log review [4]. Most of these checks are performed by technical professionals who understand databases and performs data management or curation on their jobs. Rarely are stakeholders such as researchers, administrators, and patients involved in the process, which represents a significant missed opportunity. Part of the challenges is the lack of effective mechanisms to facilitate the interdisciplinary communication and collaboration between technical professional and data stakeholders. Subsequently, often technical professionals leverage their empirical knowledge of the data pitfalls and design targeted data quality checks to detect and report corresponding errors using a knowledge-based approach. Such methods may not

systematically detect all possible data quality errors but only target those that are deemed most urgent or important problems for a specific organization or a specific task to invest in resources for detection and correction. The priorities selected by technical professionals may also not match the expectations of clinical data stakeholders, who in contrast, are receivers of the clinical data but currently do not have channels to provide feedback for clinical data quality nor contribute to the selection of focus areas for quality checking. Data quality checks across existing clinical data research networks are shown to exhibit great discrepancies [17]. For example, the OHDSI data network has 172 published quality data checks whereas the Kaiser Permanente clinical data research network has over 3400 data quality checks [17]. The data quality checks for OHDSI tend to be generic (e.g. 'counts of patients over time are consistent') while the data quality checks for Kaiser tend to be specific to individual data elements such as 'birth date' and specific data attributes such as formatting requirements in terms of the number of digits. At present, clinical data quality checking is prioritized primarily towards those data elements that can indicate the completeness of dataset, such as patient counts, diagnosis, or visits distributions over time, and surgery counts and frequencies. Fewer data quality checks address content validity, which is much harder to investigate but yet is critical for research using these data.

The last phase of the clinical data life cycle, the post-reuse phase, presents great individual and organizational barriers to data quality reporting [18]. Data quality issues or the true 'fitness-for-use' identified by individual researchers may occasionally be reported in the publications of the reuses but rarely communicated back to the data generators or data curators. There is no standard data quality reporting guideline, which is another knowledge gap for promoting best practices for data quality assurance and improvement.

In sum, data quality issues can be introduced by different people at various phases throughout the life cycle of clinical data for myriad reasons, including but not limited to the deficits in the documentation systems and clinical terminologies, the inevitable information loss during data transformation processes (e.g. ETL), and the lack of data provenance and contextual knowledge for reuse. At present, most of the time the key stakeholders of clinical data such as researchers are not part of data curation processes so that the general data preparation process is not adequately constructive forward-looking in addressing the data needs of data stakeholders, causing an inevitable mismatch of expectations and lack of 'fitness-for-use.' Currently, communication channels do not exist for an effective closed feedback loop from data consumers and stakeholders to data generators and data curators to critique or discuss the fitness of use for clinical data. There is also no guideline for standardizing data quality reporting and data quality results sharing, preventing the actionability of data quality knowledge among key stakeholders.

Recommendations

Clinical data quality is not a simple problem, and if the reuse of clinical data for data-driven discoveries is to become an accepted approach to medical research, the scientific community needs to be engaged early on for defining clinical data quality needs and for collaborating with technical professionals in developing validated, systematic methods of clinical data quality assessment and standards-based data quality reporting throughout its life cycle.

Therefore, solutions to clinical data quality issues call for a constructive, forward-looking mechanism with closed feedback loops and an interdisciplinary team science approach. Recommendations for improving awareness of clinical data quality issues and facilitating standards-based clinical data quality reporting using systematic methods are provided in Table 1.

The interdisciplinary conversation between biomedical researchers and technical professionals involved in data curation will be challenging because many biomedical researchers are not trained in database languages. Moreover, shared data quality conceptual frameworks and standard terminologies for characterizing clinical data quality issues should be adopted in discussions and reporting of clinical data quality. A taxonomy of data quality would enable a structured discourse and contextualize assessment methodologies. Systematic approaches should be developed to measure data quality and to develop and share best practices for the assessment of clinical data quality in the context of reuse for clinical research. Additionally, if we intend to embrace the concept of 'fitness-for-use,' it is important to consider the intended research use of clinical data when determining if clinical data are of sufficient quality. It will be important to develop a comprehensive understanding of the interrelationships of research tasks and data characteristics as they relate to data quality. For example, the completeness of a set of data elements required by one research protocol may differ from the completeness required for a different protocol. Many factors, including clinical focus, required resolution of clinical information, and desired effect size, can affect the suitability of a dataset for a specific research task. It is important that the clinical researchers begin to move away from ad hoc approaches to data quality assessment. Validated methods that can be adapted for different research questions are the ideal goal. Furthermore, ideally clinical data quality assessment is encouraged to adopt widely used clinical data standards such as the OHDSI OMOP Common Data Model (https://www.ohdsi.org/data-standardization/the-common-data-model/) so that algorithms for data quality assessment can be easily shared with other databases based on the same data standard and can be reused without re-development. The OHDSI community has contributed an open-source tool ACHILES (https://www.ohdsi.org/analytic-tools/achilles-for-data-characterization/) for sharable clinical data characterization. Ideally, the clinical data quality research community should collaborate in the continued development of such shared resources.

Conclusions

Regardless of quality concerns, clinical data has increasingly and will continue to play a crucial role in the development of learning health systems and in accelerating real-world evidence generation. The funding agency, the health policy leadership, the scientific community, the technical professionals involved in data curation, and the data consumers are strongly encouraged to collaborate and develop interdisciplinary team science approaches to systematically detect and report clinical data quality issues and disseminate results using clinical data standards-based methods transparently and non-ambiguously, throughout the clinical data life cycle.

Acknowledgments

Thanks to Michael Kashner, Steven Henley, and Richard Golden for providing valuable feedback to this work.

Funding

This project was funded by a special grant to Dr Michael Kashner from Veteran Affairs. Dr Weng is also supported by National Institutes Health grants R01LM009886 and OT3-TR002027-01S1.

References

[1]. IOM. Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care. Clinical data as the basic staple of health learning: creating and protecting a public good: workshop summary Washington (DC): National Academies Press (US); 2010 1, Clinical Data as the Basic Staple of the Learning Health System; 2010.

[2]. Detmer DE. Building the national health information infrastructure for personal health, health care services, public health, and research. BMC Meds Inform Decis Mak [2003 1 6];3:1.

[3]. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care [2013 8];51(8) Suppl 3:S30–S57. doi:10.1097/MLR.0b013e31829b1dbd. [PubMed: 23774517]

[4]. 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 1 1];20(1):144–151. doi:10.1136/amiajnl-2011-000681. [PubMed: 22733976]

[5]. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244. doi:10.13063/2327-9214.1244. [PubMed: 27713905]

[6]. Wang MD, Khanna R, Najafi N. Characterizing the source of text in electronic health record progress notes. JAMA Intern Med [2017 8 1];177(8):1212–1213. doi:10.1001/jamainternmed.2017.1548. [PubMed: 28558106]

[7]. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC Bioinform [2013 1 16];14:10. doi:10.1186/1471-2105-14-10.

[8]. Tange HJ, Schouten HC, Kester AD, et al. The granularity of medical narratives and its effect on the speed and completeness of information retrieval. J Am Med Inform Assoc [1998 Nov-Dec];5(6):571–582. [PubMed: 9824804]

[9]. Bowman S Impact of electronic health record systems on information integrity: quality and safety implications. Perspect Health Inf Manag 2013;10:1c.

[10]. Shang N, Weng C, Hripcsak G. A conceptual framework for evaluating data suitability for observational studies. J Am Med Inform Assoc [2017 9 8]. doi:10.1093/jamia/ocx095.

[11]. Pivovarov R, Albers DJ, Sepulveda JL, et al. Identifying and mitigating biases in EHR laboratory tests. J Biomed Inform [2014 10];51:24–34. doi:10.1016/j.jbi.2014.03.016. [PubMed: 24727481]

[12]. Rusanov A, Weiskopf NG, Wang S, et al. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak [2014 6 11];14:51. doi:10.1186/1472-6947-14-51. [PubMed: 24916006]

[13]. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annu Symp Proc 2013;2013:1472–1477. [PubMed: 24551421]

[14]. Dhombres F, Bodenreider O. Interoperability between phenotypes in research and healthcare terminologies–investigating partial mappings between HPO and SNOMED CT. J Biomed Semantics 2016;7:3. doi:10.1186/s13326-016-0047-3. [PubMed: 26865946]

[15]. Sim I, Chute CG, Lehmann H, et al. Keeping raw data in context. Science [2009 2 6];323(5915):713. doi:10.1126/science.323.5915.713a.

[16]. Weiskopf NG, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform [2013 10];46(5):830–836. doi:10.1016/j.jbi.2013.06.010. [PubMed: 23820016]

[17]. Callahan TJ, Bauck AE, Bertoch D, et al. A comparison of data quality assessment checks in six data sharing networks. EGEMS (Wash DC) 2017.

[18]. Callahan TJ, Barnard JG, Helmkamp LJ, et al. Reporting data quality assessment results: identifying individual and organizational barriers and solutions. EGEMS (Wash DC) 2017.

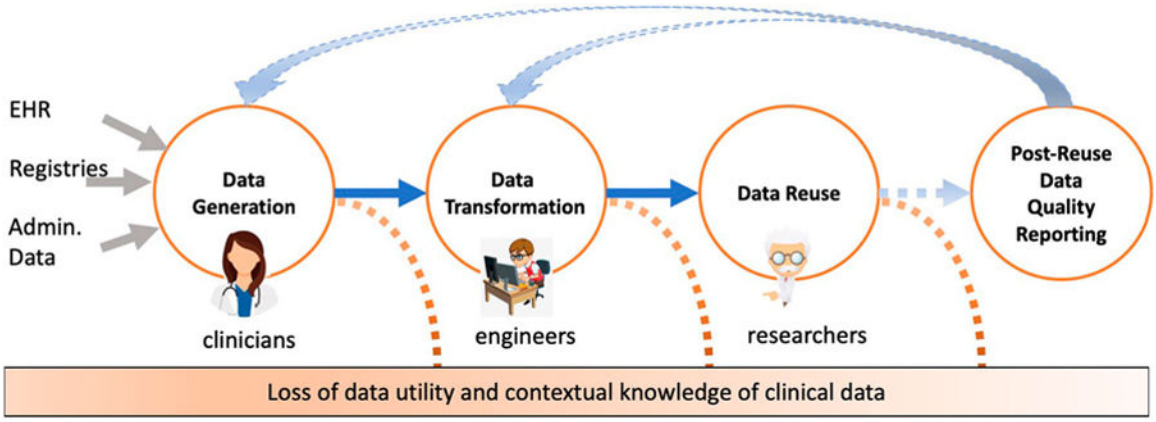

Figure 1.

The life cycle of secondary use of clinical data for research. It spans data generation, data transformation, data reuse, and optional data quality reporting and feedback provision to the data generation and data transformation processes. Data quality problems can be introduced at each phase in this life cycle. Knowledge of data quality (represented by the orange dotted lines) gets lost at each step along the way. Different stakeholders were only included at specific phases without much collaboration: e.g. clinicians are usually in the data generation step, software engineers or technicians are involved in data transformation processes without knowledge of intended uses of the data, and researchers are only involved at the data reuse phase without knowledge of data provenance. Heterogeneous data reporting occurs occasionally at the last step, with limited feedback loops established between the last step and the first two steps in the data life cycle.

Table 1.

Recommendations for clinical data quality assessment, reporting, and improvement.

Challenges or knowledge gaps	Stakeholders	Recommendations
Loss of information and contextual knowledge about clinical data during the clinical data life cycle	Informatics researchers and developers	Improve data provenance and make ETL processes transparent; Processing details for clinical data must be made transparent; traceability of all operations performed on clinical data from origination to analysis be published
Lack of constructive, forward-looking data quality requirements collection and communication framework	Clinical researchers	Clinical researchers be engaged early on when a new clinical data warehouse or clinical data network is established so that their clinical data needs can be fully considered at the design of the new clinical data resource
Misuse of clinical data of inadequate data quality that may lead to unreliable insights	Clinical researchers; Health policy leaders; Funding agencies	Clinical researchers be requested to report 'fitness-for-use' of the clinical data for the intended uses of these clinical data in the publications or grant applications
Lack of knowledge bases and best practices for data quality reporting	Informatics researchers	Harmonize the clinical data measures and metrics, compare clinical data quality assessment methods, and promote best practices for community development and collaboration support
Lack of standardized clinical data quality reporting	Clinical researchers; Health policy leaders; Informatics researchers	Clinical researchers and health policy leaders work together to harmonize and standardize clinical data reporting guidelines for adoption by researchers and enforcement by funding agencies and publishers
Lack of interoperable clinical data quality reporting tools	Informatics researchers and developers; Health policy leaders	Common data model (CDM)-based clinical data quality characterization tools such as ACHILES be developed by informatics researchers and developers and adopted by clinical researchers for sharing data quality reports
Lack of directions or actions based on data quality reports	Informatics researchers; Clinical researchers	Make clinical data quality reports actionable for various stakeholders for creating a culture of using quality clinical data for evidence generation, setting up acceptable thresholds or criteria for clinical data quality, making clinical data quality reports intuitive to comprehend or visualize, or rating clinical data resources by their quality
Lack of feedback loop and incentives for improving clinical data quality	Clinicians; Informatics researchers; Health Policy leaders	Reported clinical data quality problems be shared with data generators or clinicians, who can be provided opportunities and methods to improve or correct the data quality error, to improve their awareness of the data quality issues and their impact; Measure and make explicit the value of good clinical data to an organization; Measure and make explicit the impact of bad clinical data on an organization