Building Capacity for Health Informatics in the Future
F. Lau et al. (Eds.)

doi:10.3233/978-1-61499-742-9-418

Towards Data Value-Level Metadata for Clinical Studies

Meredith Nahm ZOZUS[a,1] and Joseph BONNER[b]
[a] *Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR;*
[b] *Ascension Health, National Data Governance Office, St. Louis, MO*

Abstract: While several standards for metadata describing clinical studies exist, comprehensive metadata to support traceability of data from clinical studies has not been articulated. We examine uses of metadata in clinical studies. We examine and enumerate seven sources of data value-level metadata in clinical studies inclusive of research designs across the spectrum of the National Institutes of Health definition of clinical research. The sources of metadata inform categorization in terms of metadata describing the origin of a data value, the definition of a data value, and operations to which the data value was subjected. The latter is further categorized into information about changes to a data value, movement of a data value, retrieval of a data value, and data quality checks, constraints or assessments to which the data value was subjected. The implications of tracking and managing data value-level metadata are explored.

Keywords. Clinical Research Informatics, reproducibility, data quality, metadata, traceability

1. Introduction

In research, when a replication or reproduction fails questions arise about how data were collected and managed. While the Food and Drug Administration (FDA) regulated therapeutic development industry is supported in this regard by regulation and guidance [1-7], in primarily academically oriented studies outside the therapeutic development industry, discussions are just starting about reasonable expectations for such documentation. Further, across all clinical studies, the number and variety of data sources have seen a ten-fold increase in the last decade [8]. These new data sources, many of which involve direct electronic measurement and capture, coupled with the rapid rise in secondary use of data for research push the boundaries on existing regulation and guidance.

We surmise that many have shied away from such dissuasion for academically oriented studies because of perceived burden of additional documentation. With rising expectations for research replication and reproducibility and static resources, avoidance is understandable. However, we offer another perspective, that when such documentation is created and maintained in a computable format, it can be leveraged to advantage by researchers and research institutions.

[1] Corresponding Author: Meredith Nahm Zozus, University of Arkansas for Medical Sciences, 4301 W. Markham St. #782, Little Rock, AR 72205-7199

2. Background

The International Standards Organization (ISO) in the Metadata Registries standard (ISO/IEC 11179) defines metadata as data that defines and describes other data [9]. However, the standard primarily describes methods for lifecycle management of metadata that defines data elements, -- lifecycle management of data element definition rather than lifecycle events of actual data values. Other treatments of metadata, for example, metadata in library science has historically been focused at the resource level - on bibliographic information for acquisition, storage, discovery, searching, access, viewing and downloading information resources managed by libraries [10]. Metadata for public records has a similar resource level focus. Similarly, disciplines such as business, environmental science, social media and information technology have discipline-specific definitions and categorizations of metadata reflective of the needs of the discipline and the functions that the metadata support [10]. In clinical research, the ClinicalTrials.gov registry contains metadata about clinical studies that supports discovery of ongoing studies by patients, results reporting and information retrieval. Others including the Clinical Data Interchange Standards Consortium (CDISC), the Biomedical Research Integrated Domain Group (BRIDG), the Ontology of Clinical Research (OCRe), and the Human Studies Database Project (HSDB) have also addressed study level metadata and there has been significant collaboration among the groups such that the resulting study level metadata is well harmonized. There is, however, no comprehensive description of or agreed upon metadata for the data upon which the clinical study conclusions are based. Here, we are concerned with this data value-level metadata and the ways in which it may support use and reuse of data from clinical studies.

3. Metadata in Clinical Studies

3.1. Uses of Metadata in Clinical Studies

As demands for interoperability and automation in clinical studies increase so does the need for good metadata. Metadata has been used to automate screen generation, for example, configuring data entry screens for web-based Electronic Data Capture (EDC) systems from a spreadsheet of descriptive metadata about data elements such as the prompt, data collection format and choice options for discrete data elements [11]. Similar descriptive metadata has also been used to automate exchange of data as in the case of Health Level Seven (HL7) messages [12]. Metadata detailing changes to data values has been used to support viewing the history of data values [13]. Metadata containing information about data discrepancies has been used to facilitate status reporting [14] and data quality assessment [15]. Others have used metadata to automate comparison of data files [16]. Additionally, algorithms for operations performed on data have provided traceability, and metadata about data transfer has been used to automate receipt and integration of incoming data files [17]. Further, these metadata can all be used to support automated data visualization of the source or origin of data, the provenance or path that data traverse from their origin to some specified state, the status of data processing, and data quality information.

While all of these represent advances in automation of information provision in the context of clinical studies, we emphasize that reaping benefit in terms of increased quality or time saved at scale requires software that leverages standards and standard metadata. Further, most metadata elements do not include descriptions of why data elements were generated, who or what made the observation that became a datum, how a dataset was assembled or retrieved, and who might have changed an element since its initial capture. Lacking metadata limits the type and extent of automation possible. Alternatively, full understanding ad associating the type and extent of automation possible with the necessary metadata enables direct assessment and balancing of the cost of additional metadata acquisition and management versus the benefit gained.

3.2. Sources of Metadata in Clinical Studies

Each of the aforementioned metadata uses reveals sources of metadata. Zozus *et al* suggested types of metadata required to support traceability in longitudinal studies based on Electronic Health Record (EHR) data [17]. We expand this list to cover a broader set of clinical studies such as retrospective studies based on electronic or manually abstracted data, prospective clinical trials, observational studies, and institutional data stores supporting such research (Table 1).

Table 1. Sources of Data Value-level Metadata

Audit trail	Date, time, and attribution of entries and changes as well as record of data values prior to changes.
Data element definition	Metadata registry managing the lifecycle of data element definition. The ISO/IEC 11179 standard is a good example of data element definition. Such definition is often extended through association with controlled terminology standards as well as ontologies containing formal logic-based semantics.
Data origin	Statement of the source of the data. The source is the occurrence of the phenomena which was observed, questioned or measured resulting in the data value. Contextual information such as location, date and time, observation method, and environmental conditions may also be important.
Data transfer*	Specifications for data received from or sent to an external system or organization as well as import and export logs.
Data transformation	Specifications for and logs of computer programs performing operations on data. It is assumed that changes to data based on these programs are captured in the audit trail.
Data quality assessment	Specifications for all rules or other logic to which data values were subject in attempts to identify data discrepancies as well as results of execution of the data quality assessment approach. It is assumed that changes to data based on these programs are captured in the audit trail.
Information retrieval	Specifications or source code for queries run to extract data and other information sufficient to recreate the extraction.

* Data origin and data transfer together are meant to cover the chain of custody through which the data value traveled from origin to the current system.

The sources listed in Table 1 comprehensively track the history of a data value and all operations performed on a data value. The intent is to capture typical transformations necessary and sufficient to recreate analysis data from raw data or vice versa, i.e., sufficient to support full traceability. An implication of traceability is that such metadata might be expected to be received with data. A second implication is that this metadata may be used to computationally compare data from two sources and facilitate understanding of why the data values may differ and at what point such divergence occurred.

3.3. Classification of Metadata for Clinical Studies

Uses and sources of metadata inform categorization of data value-level metadata for clinical studies. From Table 1 above, we might say that every data value has one and only one origin. Further, several categories from Table 1 describe operations performed on data values (changes/transformations, assessments, movement, and retrieval). As a result of these operations every data value has a history that delineates all past values and operations performed on the data value from its origin to its current state. Finally, every data value has a definition that unambiguously conveys its meaning.

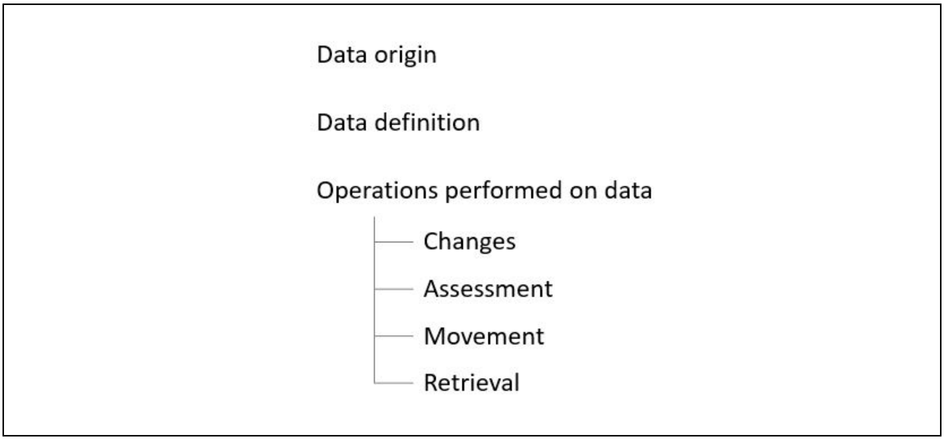

Figure 1. Categories of Data Value-level Metadata

These primitives of data value-level metadata support calculation of other second-generation data value-level metadata such as data value age, extent of changes to the data value, and the path through which the data have traveled. These second-generation metadata support new uses of metadata, for example viewing values in a health record that originated from the institution versus all data including values received from other institutions, or visually marking older or implausible data values.

4. Implications for Standard Data Value-level Metadata for Clinical Studies

Definition and categorization of data value-level metadata supports the metadata uses described above as well as those that are inevitable as the volume of data collected for biomedical research increases. Data value-level metadata is information is intrinsic to a data value, i.e., independent on any specific use to which the data value might be subjected, and remains inextricably linked to the data value. The existence of data value-level metadata necessarily means that data volume and storage requirements at least double – the minimal case being one piece of data value-level metadata per data value. Beyond structures for audit trails, management methods and storage approaches for data value-level metadata have yet to be articulated. Although the above is a starting point for discussion of classification and definition of data value-level metadata, it is far from a standard that can be leveraged in software used in biomedical research. Additionally, much of the data value-level metadata will be system generated. This will require new software functionality that beyond audit trail functionality does not exist today. Finally, data element level metadata requires human curation and so far in biomedical research with a few notable exceptions, we have not excelled in this area.

There are, however, considerable upsides. The first is that metadata curated at the data element level is reusable and applies to multiple data values for that data element. With additional software functionality, most data value-level metadata will be system generated. Once standardized software can leverage the metadata to provide significant value through signal detection and visualization, for example displaying heat maps showing data with a high frequency of changes, older data, or data with a high percentage of discrepancies. Other uses might include documentation of data use, recreation of information retrieval after the underlying data source changes, data transfer-to-data transfer comparison, and automated comparison of incoming data to specifications. While many of these have been demonstrated and shown to be of value, because we lack standard data value-level metadata, this level of automation is not widely expected or available today.

5. Discussion and Conclusion

Metadata stems from the Greek prefix “meta-“, denoting “about” or data that defines other data. The Greek prefix “para-“ denotes alterations or modifications. In survey research methodology word *paradata* defines data elements describing, for example, how long a particular survey lasted, the time of day the survey was delivered and degree of reluctance of an interviewee. In survey science disciplines *paradata* about a survey is considered a class of metadata [18-20]. We propose adding *paradata* to the biomedical research lexicon to refer to data about how a data-value was created, altered, or otherwise operated on. The Greek prefix “ortho-“ denotes something that is straight or upright. In general, data quality constraints and assessment refer to testing whether data values are in-line with expectations. We further propose the term *orthodata* referring to data quality constraints and assessment logic. Both *paradata* and *orthodata* should be considered as special classes of metadata. As the costs for discovering and developing new medical therapies rise, so rise the need for secondary use of electronic health data. Comprehensive metadata helps us choose and use electronic more confidently for research.

The amount of data that we collect and process for research is not decreasing. In fact, many consider the increasing volume, velocity, and variety, of data and data sources to be a major challenge facing biomedical research today. Once simple operations such as data checking, integration and cleaning that could be performed by humans and with minimal automation, require automation today. Automating operations on data requires metadata. Automating operations on data in a scalable manner requires standardized metadata as well as software that leverages the metadata. Thus, coming to a complete, correct, unambiguous and standardized categorization of data value-level metadata is a clear and present need.

Acknowledgements

This work was partially supported by grant R00-LM011128 from the National Institutes of Health (NIH) National Library of Medicine (NLM). The work and opinions presented here are those of the authors and do not necessarily represent the NIH of any federal agency.

References

[1] Title 42 CFR 93, *Public health service policies on research misconduct*. 2015.

[2] Title 21 CFR 58.130, *Conduct of a nonclinical laboratory study*. Sections (c) and (e) 2015.

[3] Title 21 CFR 58.35 *Quality assurance unit*, section (b) 2011.

[4] Title 21 CFR Part 11, *Electronic records; electronic signatures*, 1997.

[5] Title 21 CFR 58.3 *Good laboratory practices for nonclinical laboratory studies, definitions section* (k), 2015.

[6] United States Food and Drug Administration, *Guidance for Industry Electronic Source Data in Clinical Investigations*. September 2013.

[7] United States Food and Drug Administration, *Guidance for Industry Computerized Systems Used in Clinical Investigations*. May 2007.

[8] M. N. Zozus, A. Lazarov, L. Smith, T. Breen, S. Krikorian, P. Zbyszewski, K. Knoll, D. Jendrasek, D. Perrin, D. Zambas, T. Williams, C. Pieper, Analysis of Professional Competencies for the Clinical Research Data Management Profession: Implications for Training and Professional Certification. Submitted to *JAMIA*, August 2016.

[9] International Organization for Standardization (ISO) 11179-1 *Information technology – Metadata Registries (MDR) – Part 1 Framework*. ISO/IEC 11179-1:2004(E).

[10] J. Greenberg, J. Metadata and digital information. In Marcia J. Bates, Mary Niles Maack, eds. *Encyclopedia of Library and Information Science*, New York: Marcel Dekker, Inc. 2009.

[11] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, J. G. Conde. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* **42**(2) (2009), 377-81

[12] A. Hinchley. *Understanding Version 3 - A Primer on the HL7 Version 3 Healthcare Interoperability Standard - Normative Edition*, Munich, Alexander Mönch, 2007.

[13] W. I. Kuchinke, J. Aerts, S. C. Semler, C. Ohmann. CDISC standard-based electronic archiving of clinical trials. *Methods Inf Med 48*(5) (2009), 408-13. doi: 10.3414/ME9236. Epub 2009 Jul 20.

[14] J. J. Pan, M. Nahm, P. Wakim, C. Cushing, L. Poole, B. Tai, C. F. Pieper. A centralized informatics infrastructure for the National Institute on Drug Abuse Clinical Trials Network. *Clin Trials* **6**(1) (2009), 67-75.

[15] D. Yoon, E. K. Ahn, M. Y. Park, S. Y. Cho, P. Ryan, M. J. Schuemie, D. Shin, H. Park, R. W. Park. Conversion and data quality assessment of electronic health record data at a Korean tertiary teaching hospital to a common data model for distributed network research. *Healthc Inform Res* **22**(1) (2016), 54-8.

[16] E. A. Voss, R. Makadia, A. Matcho, Q. Ma, C. Knoll, M. Schuemie, F. J. DeFalco, A. Londhe, V. Zhu, P. B. Ryan. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc* **22**(3) (2015), 553-64.

[17] M. N. Zozus, R. L. Richesson, A. Walden, J. D. Tenenbaum, W. E. Hammond. Research reproducibility in longitudinal multi-center studies using data from electronic health records. *AMIA Jt Summits Transl Sci Proc*. 2016; 2016: 279–285. Published online 2016 Jul 20.

[18] F. Kreuter, M. Couper, L. Lyberg. The use of paradata to monitor and manage survey data collection. Section on Survey Research Methods – *JSM* 2010. (online at amstat.org).

[19] F. Scheuren. 2000. *Macro and Micro Paradata for Survey Assessment*. 1999 NSAF Collection of Papers.

[20] M. Couper. 1998. Measuring survey quality in a CASIC environment. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.