MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220229

ODM-DQA-Reporter: A Generic Approach to Assess and Monitor Basic Data Quality of Medical Research Data in Operational Data Model (ODM) Format

Ayşenur Süer[a], Sarah Riepenhausen[a], Michael Storck[a], Leonard Greulich[a], Claudia Zeidler[b], Sonja Ständer[b], Martin Dugas[a]

[a] Institute of Medical Informatics, University of Münster, Münster, Germany
[b] Department of Dermatology and Center for Chronic Pruritus (KCP), University Hospital Münster, Münster, Germany

Abstract

A generic approach for assessment and continuous monitoring of data quality in ODM-based research data has been developed. The focus is on the two data quality indicators completeness and syntactic correctness. The main idea is to enable the generation of a data quality report without additional programming effort.

Keywords:

Data accuracy, Data management, Operational Data Model

Introduction

Assessment and continuous monitoring of data quality is generally a complex task and associated with high costs [1, 2], especially in multicenter studies [3]. Data quality is a prerequisite for data analysis and serves the purpose of insight and information, but also reliability and compliance with regulatory requirements [4]. Suboptimal data quality can yield inaccurate research results and entail high costs [3]. In this context, no explicit guidance exists on when a dataset is “fit for use” [2].

A variety of different data quality indicators can be used to measure data quality [1-2, 5]. A distinction is made between task-dependent (or subjective) and task-independent (or objective) dimensions [1]. Completeness and correctness/ accuracy are the most frequently assessed indicators [1]. In the context of our work, we currently focus on completeness and syntactic accuracy of medical research data in the Operational Data Model (ODM) format. The goal is to provide a generic approach to task- and domain-independent, basic assessment and support for continuous monitoring of gathered data across the data collection phase and to allow reporting without further programming effort.

Methods

The Operational Data Model (ODM) of the Clinical Data Interchange Standard Consortium (CDISC) is an FDA-compliant, XML-based standard for study definition and archiving [6] and enables structured capture of medical research data. This standard is accepted as an import and export format by electronic data capture systems such as x4T-EDC [7] and REDCap [8]. A large number of forms in ODM format is available in the Portal of Medical Data Models (MDM Portal) [9, 10].

In order to generate valid data quality reports for heterogeneous ODM-based research data of different sample sizes, a general analysis approach is necessary. In the first step, tools based on the OMOP Common Data Model like Achilles Heel [11] and PEDSnet Data-Quality-Analysis [12] were analyzed, as well as generic R-based methods [13] such as mosaicQA developed for epidemiological purposes [14] and MOQA [15]. The next step was to implement specific R scripts to assess the quality of ODM-based test data of different sample sizes (between N=25 and 10,000) generated using the in-house developed ODM Clinical Data Generator [16]. Data import was implemented in Java due to performance reasons. Grouping functions (with respect to subject keys, sites and metadata) were implemented to allow restriction to specific parts of the dataset (e.g. forms or subjects). The ODM standard does not provide a common definition for conditions that can be uniquely integrated into the metadata. Therefore, in the third step, a simple definition was developed in order to be able to consider conditional items, especially to determine completeness. Subsequently, the focus was on the generalization of the developed R scripts, mainly through parameterization. In the fifth step, the automated generation of a quality report in PDF format was realized. Finally, the data quality of three real-world datasets collected within the projects "Translational Pruritus Research" (N=639), "Registry for Primary Ciliary Dyskinesia" (N=1475) and "Pruritus Research Database" (N=10389) were assessed as proof of the generic approach.

Results

The developed generic approach allows an automatic generation of a data quality report based on the ODM standard without additional programming effort. Source code and exemplary data are available at https://imigitlab.uni-muenster.de/sueerays/odm-dqa-reporter. In addition to descriptive statistics on completeness at different levels and syntactic correctness, a metadata dictionary and graphical recruitment history can be integrated into the report. Grouping functions allow restriction to parts of the data set. Such groupings can be applied in multicenter studies or for validation of inclusion criteria. The following syntax for selectors has been defined for conditions:

SE-StudyEventOID[RepeatKey]/F-FormOID[RepeatKey]/IG-ItemGroupOID[RepeatKey]/I-ItemOID

Relative paths are allowed and 'RepeatKey' is optional. If the latter is not specified, the selector refers to the current examined repeat instance. In addition, the Boolean operators ==, !=, <, <=, >, >=, IN, NOTIN, AND and OR are supported. If more than one comparison value is specified, the values must be given comma-separated in the form [v1, v2, v3].

Furthermore, the real data sets could be analyzed and reports generated. This shows that the generic approach was successfully implemented.

Discussion

Currently, the focus of ODM-DQA reporter is on the data quality indicators completeness and syntactic correctness. The results are visualized using descriptive plots that are easy to understand (e.g., pie chart, bar chart, box plot). The integration of further indicators such as plausibility and consistency are in preparation. Further options would be the integration into ODM Data Analysis Tool [17] developed for automated generation of descriptive, univariate statistics or deployment as REDCap plugin.

Conclusions

ODM-DQA reporter is a generic, context-independent tool for automated assessment and review of data quality in medical research data provided in ODM format. Standard-compliant collection of research data simplifies the systematic data quality assessment. Both data managers and researchers can use this approach without additional programming effort.

Acknowledgements

The authors would like to thank their colleagues at the Institute of Medical Informatics (IMI), University of Münster, for their valuable and constructive criticism and support. In particular, thanks go to Philipp Neuhaus, Tobias Brix and Johannes Oehm, who helped in developing a definition for conditions in ODM.

References

[1] N.G. Weiskopf, and C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, *J Am Med Inform Assoc*. **20** (2013) 144–151. doi:10.1136/amiajnl-2011-000681.

[2] M.G. Kahn, J.S. Brown, A.T. Chun, B.N. Davidson, D. Meeker, P.B. Ryan, L.M. Schilling, N.G. Weiskopf, A.E. Williams, and M.N. Zozus, Transparent reporting of data quality in distributed data networks, *EGEMS (Wash DC)*. **3** (2015) 1052. doi:10.13063/2327-9214.1052.

[3] D. Juárez, E.E. Schmidt, S. Stahl-Toyota, F. Ückert, and M. Lablans, A Generic Method and Implementation to Evaluate and Improve Data Quality in Distributed Research Networks, *Methods Inf Med*. **58** (2019) 86–93. doi:10.1055/s-0039-1693685.

[4] D. Venet, E. Doffagne, T. Burzykowski, F. Beckers, Y. Tellier, E. Genevois-Marlin, U. Becker, V. Bee, V. Wilson, C. Legrand, and M. Buyse, A statistical approach to central monitoring of data quality in clinical trials, *Clin Trials*. **9** (2012) 705–713. doi:10.1177/1740774512447898.

[5] N.G. Weiskopf, S. Bakken, G. Hripcsak, and C. Weng, A Data Quality Assessment Guideline for Electronic Health Record Data Reuse, *EGEMS (Wash DC)*. **5** (2017) 14. doi:10.5334/egems.218.

[6] ODM-XML | CDISC, (n.d.). https://www.cdisc.org/standards/data-exchange/odm (accessed May 11, 2021).

[7] P. Bruland, C. Forster, B. Breil, S. Ständer, M. Dugas, and F. Fritz, Does single-source create an added value? Evaluating the impact of introducing x4T into the clinical routine on workflow modifications, data quality and cost-benefit, *Int J Med Inform*. **83** (2014) 915–928. doi:10.1016/j.ijmedinf.2014.08.007.

[8] P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and 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** (2009) 377–381. doi:10.1016/j.jbi.2008.08.010.

[9] M. Dugas, Portal für Medizinische Datenmodelle (MDM-Portal), (n.d). https://medical-data-models.org (accessed May 11, 2021).

[10] M. Dugas, P. Neuhaus, A. Meidt, J. Doods, M. Storck, P. Bruland, and J. Varghese, Portal of medical data models: information infrastructure for medical research and healthcare, *Database (Oxford)*. **2016** (2016). doi:10.1093/database/bav121.

[11] V. Huser, F.J. DeFalco, M. Schuemie, P.B. Ryan, N. Shang, M. Velez, R.W. Park, R.D. Boyce, J. Duke, R. Khare, L. Utidjian, and C. Bailey, Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets, *EGEMS (Wash DC)*. **4** (2016) 1239. doi:10.13063/2327-9214.1239.

[12] R. Khare, L.H. Utidjian, H. Razzaghi, V. Soucek, E. Burrows, D. Eckrich, R. Hoyt, H. Weinstein, M.W. Miller, D. Soler, J. Tucker, and L.C. Bailey, Design and Refinement of a Data Quality Assessment Workflow for a Large Pediatric Research Network, *EGEMS (Wash DC)*. **7** (2019) 36. doi:10.5334/egems.294.

[13] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, (2021). *https://www.R-project.org/* (accessed May 11, 2021).

[14] M. Bialke, H. Rau, T. Schwaneberg, R. Walk, T. Bahls, and W. Hoffmann, mosaicQA - A General Approach to Facilitate Basic Data Quality Assurance for Epidemiological Research, *Methods Inf Med*. **56** (2017) e67–e73. doi:10.3414/ME16-01-0123.

[15] M. Bialke, T. Schwaneberg, and R. Walk, MOQA: Basic Quality Data Assurance for Epidemiological Research, (2017). https://CRAN.R-project.org/package=MOQA (accessed May 11, 2021).

[16] T.J. Brix, L. Becker, T. Harbich, J. Oehm, M. Fechner, M. Dugas, and M. Storck, ODM Clinical Data Generator: Syntactically Correct Clinical Data Based on Metadata Definition, (In-press).

[17] T.J. Brix, P. Bruland, S. Sarfraz, J. Ernsting, P. Neuhaus, M. Storck, J. Doods, S. Ständer, and M. Dugas, ODM Data Analysis-A tool for the automatic validation, monitoring and generation of generic descriptive statistics of patient data, *PLoS One*. **13** (2018) e0199242. doi:10.1371/journal.pone.0199242.

Address for correspondence

Ayşenur Süer, Institute of Medical Informatics, University of Münster, Germany; aysenur.sueer@uni-muenster.de