Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

Measuring Data Quality: A Review of the Literature between 2005 and 2013

Jürgen STAUSBERG[a,1] , Daniel NASSEH[b], Michael NONNEMACHER[c]
[a] *Essen, Germany*
[b] *IBE, Ludwig-Maximilians-Universität München, Germany*
[c] *IMIBE, Universitätsklinikum Essen, Germany*

Abstract. A literature review was done within a revision of a guideline concerned with data quality management in registries and cohort studies. The review focused on quality indicators, feedback, and source data verification. Thirty-nine relevant articles were selected in a stepwise selection process. The majority of the papers dealt with indicators. The papers presented concepts or data analyses. The leading indicators were related to case or data completeness, correctness, and accuracy. In the future, data pools as well as research reports from quantitative studies should be obligatory supplemented by information about their data quality, ideally picking up some indicators presented in this review.

Keywords. Cohort study, data quality, quality indicators, registry

Introduction

According to ISO 14005, data quality could be defined as the "characteristic of data that bears on their ability to satisfy stated requirements". Usually, in empirical research, the right data from the right population in an appropriate quality hold the answers to the research questions [1]. In health care, good data are the prerequisite for patient safety and successful outcomes [2]. Numerous procedures are available supporting high data quality in the phases of the planning, the implementation, and the operation of a medical record system or a study database [3]. Nevertheless, it becomes more and more important to estimate the quality of already available data, either to foster quality improvement methods like feedback and source data verification (SDV) or to decide about the usefulness of the data in regard to a particular research question.

Indictors for data quality are an aid to quantify the degree to which data satisfy stated requirements (see definition above). However, quality indicators are not an objective measure of quality. Quality indicators are primarily intended to support quality management; abnormalities of the results could or could not be caused by real errors. Therefore, indicators for data quality should be applied with care, knowing their strengths and weaknesses.

In 2006, the TMF – Technology, Methods, and Infrastructure for Networked Medical Research - an umbrella organization for networked medical research in Germany - published recommendations about data quality in medical research as a

[1] Corresponding Author: Jürgen Stausberg, Kordulastr. 13, 45131 Essen, Germany, e-mail: stausberg@ekmed.de.

guideline for the adaptive management of data quality in cohort studies and registries [4]. A review of relevant literature concerning data quality published until 2005 was part of these recommendations. In 2011, a revision process started, again funded by the TMF [5]. Amongst other things, it resulted in an update of the literature review about data quality covering the period from 2005 to 2013. The literature review focused on the core topics of the guideline, indicators for data quality, feedback about data quality, and SDV. In the following, we will present the respective results.

1. Methods

Medline was used as the literature database via http://www.pubmed.org/. We applied the same queries as in the first edition of the guideline using the following terms in several combinations: clinical trial, cohort, data accuracy, data collection, data quality, feedback, fraud, medical registry, quality assessment, quality control, registries, and source data verification. Furthermore, the queries included related citations for some outstanding papers and a few specifically chosen authors. The research was conducted on March the 6th in 2013. Citations before January the 1st of 2005 were excluded.

The selection of relevant literature was done in two steps beginning with an inspection and rating of the abstracts and other metadata offered by Medline. The three authors participated as raters (JS, DN, MN). Each citation within the initial set of results was rated as "relevant", "not relevant", or "unclear" according to the three topics indicators for data quality, feedback about data quality, or SDV. Preceding, a training phase was conducted. During that phase, discrepancies in the ratings of 100 randomly selected citations rated by all three raters were discussed and solved in a consensus. The training phase should establish a common sense about the criteria applied in the rating process. The remaining citations were split into three sets each handled by a different rater. However, each of the sets contained an additional overlap of 100 randomly selected citations in order to subsequently estimate the agreement between the raters. Unclear ratings were solved in a consensus between all three raters. Citations in other languages as English or German were excluded from the result list.

The full text articles of the citations successfully selected in the first step were obtained. The same rating categories as in the first step were applied in the second step. In case of relevance, the rater had to state the underlying criterion on which the decision was based, in particular indicator for data quality, feedback about data quality, or SDV. JS rated all remaining articles, DN and MN rated a half of the articles each. Unclear ratings were solved in a consensus between DN and JS or MN and JS.

Two of the authors (DN, MN) summarized the selected literature in a structured format including the following metadata:

- Criterion (multiple answers possible): <feedback about data quality|indicators for data quality|SDV>
- Origin of the authors (multiple answers possible)
- Type of project: <analysis of data quality|intervention|health technology assessment|presentation of a concept|systematic review>
- Site and type of intervention (for intervention studies only), data pool, quality indicators, conclusions of the authors, summary, commentary of the rater

Each step was supported by an application realized with Microsoft Access. The application offered a user-interface presenting the metadata of the citations on the one hand and allowing the input of the rater's decision on the other hand. Kappa was calculated as reliability measure with script MKAPPSSC.SPS with IBM Statistics 21. Kappa was interpreted as proposed by Landis and Koch [6].

2. Results

The literature search resulted in 2,067 citations. Excluding 151 duplicates, 1,916 citations remained for the subsequent literature selection (cf. figure 1). Throughout the first selection step, which was based on the inspection of abstracts, 110 citations were selected. Among that, three citations were excluded due to the use of other languages than English or German. Kappa showed a moderate agreement both with 0.42 in the training collection of 100 citations and 0.39 in the overlap of further 100 citations. Full text articles could be obtained from all 107 remaining citations. Forty articles were assessed as relevant in the second selection step. One article was excluded [5] because it presents preliminary results from the revision of the guideline. Thirty-nine articles remained for further description (cf. appendix). All remaining articles were in English. The detailed descriptions can be found in [7] in German.

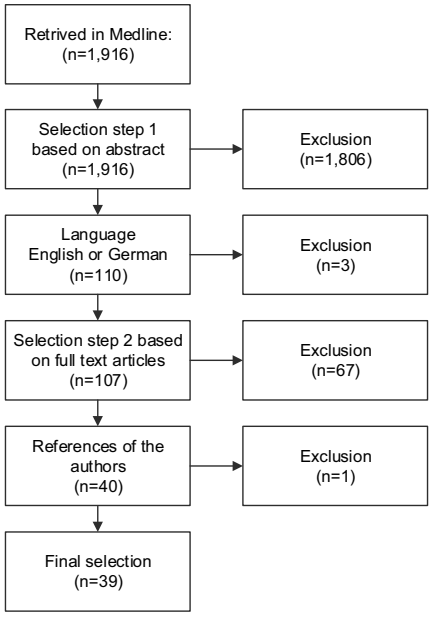

Figure 1. Flow chart of the literature selection process.

Seventeen of these articles directly addressed analysis of data quality of different datasets, 14 presented concepts regarding the handling of data quality, four summarized the subject in form of a review, and two performed a health technology assessment while two described an intervention to improve data quality. Fifteen articles stem from

Europe, 13 from North America, four from Australia or New Zeeland, two from Asia and one each from South Africa and Brasilia. Three article had a multinational authorship. Thirty-two articles were selected due to information about quality indicators, 14 due to information about SDV, and four due to information about feedback concerning data quality. We identified 34 different concepts used in the definition of quality indicators (number of articles in parentheses): accessibility (1), accuracy (10), agreement (1), appropriate amount of data (1), availability (1), believability (2), comparability (3), completeness (4), comprehensiveness (21), concordance (3), consistency (5), contextualization (1), correctness (14), currency (6), definition (1), generalizability (1), granularity (1), incompleteness (1), inconsistency (1), incorrectness (1), objectivity (1), plausibility (2), policy relevance (1), precision (1), predictive value (1), prevention of duplicates (2), rate of enrolment (2), relevancy (2), reliability (4), responsiveness of data items (1), spatial stability (1), timeliness (6), usefulness of data items (1), and validity (4).

3. Discussion

There is a huge amount of literature dealing with data quality in health care and medical research. Looking at our specific interests, we recognized an increase in publications comparing the recent results with a literature review covering Medline from the beginning up to 2005 [4]. The number of hits for “source data verification” increased from 6 (up to 2005) to 16 (2005 to 2013), the number of hits for the combination of “feedback” with “medical registry” or “cohort” increased from 28 to 132. However, there is only a small number of articles presenting interventional studies, i.e. studies that analyzed the contribution of a procedure like training, feedback, or SDV on data quality in a controlled setting. Most of the literature presented results from analyzing one or several data pools with regard to specific quality indicators. To their best, those analyses could be rated as observational studies. The geographical distribution of the articles was comparable with other fields in medical informatics [8].

The majority of the quality indicators mentioned in the articles could be organized according to the classical triad of case completeness (here completeness), data completeness (here comprehensiveness), and correctness (here correctness, to some extend accuracy). New ideas about indicators of data quality became visible as policy relevance, spatial stability, and usefulness of data items. A flat list is no longer appropriate for the organization of those manifold indicators. Consequently, common ontologies covering the objects in the field of data quality are needed.

Medical data are increasingly available beyond the time and the site of their initial recording. The amount of medical data is blowing up, in particular, but not limited to, the field of genomics. Electronically available medical data get an essential role in therapy planning and health care monitoring. Consequently, information about the quality of medical data should be regarded as an obligatory supplement of data pools. Our literature review offers several candidate measures for this supplement.

Acknowledgements

The project was funded under contract V020_05 by the TMF - Technology, Methods, and Infrastructure for Networked Medical Research (cf. /http://www.tmf-ev.de/).

References

[1] Malin JL, Keating NL. The cost-quality trade-off: need for data quality standards for studies that impact clinical practice and health policy. Journal of Clinical Oncology 2005; 23: 4581-4.
[2] Ammenwerth E, Aly A-F, Bürkle T, et al. Memorandum on the use of information technology to improve medication safety. Methods Inf Med. 2014; 53: 336-43.
[3] Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc. 2002; 9: 600-11.
[4] Nonnemacher M, Weiland D, Neuhäuser M, Stausberg J. Adaptive management of data quality in cohort studies and registers: proposal for a guideline. Acta Informatica Medica 2007; 15: 225-30.
[5] Stausberg J, Pritzkuleit R, Schmidt CO, Schrader T, Nonnemacher M. Indicators of data quality: revision of a guideline for networked medical research. Stud Health Technol Inform. 2012; 180: 711-5.
[6] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74.
[7] Nonnemacher M, Nasseh D, Stausberg J. Datenqualität in der medizinischen Forschung. 2., aktualisierte und erweiterte Auflage. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014.
[8] Moorman PW, Schuemie MJ, van der Lei J. An inventory of publications on electronic medical records revisited. Methods Inf Med. 2009; 48: 454-8.

Appendix: Selected literature

Baigent C 2008. Ensuring trial validity by data quality assurance and diversification of monitoring methods; Berner ES 2005. Data quality in the outpatient setting: impact on clinical decision support systems; Botsis T 2010. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities; Bray F 2009. Evaluation of data quality in the cancer registry: principles and methods. Part I; Brender JD 2008. Validity of parental work information on the birth certificate; Bronnert J 2012. Data quality management model; Brouwer HJ 2006. Data quality improvement in general practice; Chiba Y 2012. Quantitative and qualitative verification of data quality in the childbirth registers of two rural district hospitals in Western Kenya; Choquet R 2010. The Information Quality Triangle: a methodology to assess clinical information quality; Couchoud C 2013. Renal replacement therapy registries-time for a structured data quality evaluation programme; De S 2011. Hybrid approaches to clinical trial monitoring: Practical alternatives to 100% source data verification; Duda SN 2012. Measuring the quality of observational study data in an international HIV research network; Dyck MJ 2007. Data quality strategies in cohort studies: lessons from a study on delirium in nursing home elders; França E 2008. Evaluation of cause-of-death statistics for Brazil, 2002-2004; Kahn MG 2012. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research; Krzych LJ 2011. Assessment of data quality in an international multi-centre randomised trial of coronary artery surgery; Larsen IK 2009. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness; Loane M 2011. Paper 3: EUROCAT data quality indicators for population-based registries of congenital anomalies; Macefield RC 2013. A systematic review of on-site monitoring methods for health-care randomised controlled trials; Maruszewski B 2005. An attempt at data verification in the EACTS Congenital Database; McKenzie K 2005. Assessing the concordance of trauma registry data and hospital records; Messenger JC 2012. The NCDR Data Quality Brief; Mphatswe W 2012. Improving public health information: a data quality intervention in KwaZulu-Natal, South Africa; Nahm ML 2008. Quantifying data quality for clinical trials using electronic data capture; Sáez C 2012. Organizing data quality assessment of shifting biomedical data; Salati M 2011. Task-independent metrics to assess the data quality of medical registries using the ESTS Database; Sigurdardottir LG 2012. Data quality at the Icelandic Cancer Registry: comparability, validity, timeliness and completeness; Shabestari O 2013. Challenges in data quality assurance for electronic health records; Stevens W 2008. Comparison of New Zealand Cancer Registry data with an independent lung cancer audit; Taggart J 2012. The University of NSW electronic practice based research network: disease registers, data quality and utility; Thoburn KK 2007. Case completeness and data accuracy in the Centers for Disease Control and Prevention's National Program of Cancer Registries; Tolonen H 2006. Assessing the quality of risk factor survey data; Tuble SC 2011. Perfusion Down under Collaboration Database-data quality assurance; Tudur Smith C 2012. The value of source data verification in a cancer clinical trial; Venet D 2012. A statistical approach to central monitoring of data quality in clinical trials; Verhulst K 2012. Source document verification in the Mucopolysaccharidosis Type I Registry; Weiskopf NG 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research; Wu Y 2008. Measuring follow-up completeness; Xian Y 2012. Data quality in the American Heart Association Get With The Guidelines-Stroke.