
Jürgen STAUSBERGa,1 , Daniel NASSEHb, Michael NONNEMACHERc

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.

Indicators 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 MKAPPS.C.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.](image-url)

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 Zealand, two from Asia and one each from South Africa and Brasilia. Three articles 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


Appendix: Selected literature