Indicators of data quality: review and requirements from the perspective of networked medical research

Indikatoren zur Datenqualität: Stand und Anforderungen aus Sicht der vernetzten medizinischen Forschung

Abstract

Data quality is of highest importance for quantitative medical research. A common set of indicators for data quality is needed to cope with the future challenges in data management for biomedical informatics. A guideline for adaptive data management was developed in 2006, which offers indicators for data quality organized in three categories: integrity, organization, and trueness. The guideline was revised in 2014 bottom-up by extending its content with standards from a cancer registry, a cohort, and a data repository in Germany. In parallel, a systematic literature review identified indicators of data quality published in the literature since 2005 using Medline as literature database. The guideline differentiates in its second version 51 indicators (integrity: 30, organization: 15, trueness: 6). The literature review identified 34 indicators in 31 articles. A lack of indicators in the literature addressing the organizational aspects of data sets became visible comparing both sets. Furthermore, indicators useful for data sets used in health care practice, such as timeliness, were missing in the guideline's set. The comparison is a first step towards a common set of indicators. Beyond a consented denomination of the indicators, this set should offer an operational definition that supports a reliable application from different parties to different data sets. Furthermore, a systematic organization of the indicators would foster an appropriate selection of the individual indicators according to specific use cases.

Keywords: medical research, data quality, healthcare, guidelines, analytics, informatics

Zusammenfassung

Datenqualität ist für die quantitative medizinische Forschung von höchster Bedeutung. Ein einheitliches Set von Indikatoren zur Datenqualität wird benötigt, um die zukünftigen Herausforderungen an das Datenmanagement in der biomedizinischen Informatik zu bewältigen. Dazu wurde eine Leitlinie zum adaptiven Datenmanagement im Jahre 2006 erarbeitet, die Indikatoren zur Datenqualität über drei Ebenen organisiert: die Ebenen Integrität, Organisation und Richtigkeit. Inhaltlich wurde die Leitlinie im Jahre 2014 Bottom-up durch die Einbindung von Standards eines Krebsregisters, einer Kohorte und eines Data Repository aus Deutschland erweitert. Parallel wurden über ein systematisches Literaturreview publizierte Indikatoren der Datenqualität mit Medline als Literaturdatenbank recherchiert. Die Leitlinie weist in ihrer zweiten Version 51 Indikatoren aus (Integrität: 30, Organisation: 15, Richtigkeit: 6). Das Literaturreview identifizierte 34 Indikatoren in 31 Publikationen. Im Vergleich beider Quellen war das Fehlen von Indikatoren zu organisatorischen Aspekten in der Literatur auffällig. Der Leitlinie fehlten hingegen Indikatoren mit Bedeutung für die Krankenversorgung wie Rechtzeitigkeit. Der vorgenommene Vergleich stellt einen weiteren Schritt zur Festlegung einem einheitlichen Sets von Indikatoren zur

Jürgen Stausberg[1]
Ulrike Bauer[2]
Daniel Nasseh[3]
Ron Pritzkuleit[4]
Carsten O. Schmidt[5]
Thomas Schrader[6]
Michael Nonnemacher[1]

1 Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), Faculty of Medicine, University Duisburg-Essen, Germany

2 Competence Network for Congenital Heart Defects, German Centre for Cardiovascular Research, Berlin, Germany

3 Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE), Ludwig-Maximilians-Universität München, Germany

4 Institute for Cancer Epidemiology e.V., University Lübeck, Germany

5 Study of Health in Pomerania, Institute for Community Medicine, Greifswald, Germany

6 Resource Center OpEN.SC, Brandenburg University of Applied Sciences, Brandenburg, Germany

Datenqualität in der medizinischen Forschung dar. Neben einheitlichen Bezeichnungen sollte ein solches Set umsetzbare Definitionen beinhalten, die eine zuverlässige Anwendung auf unterschiedlichen Datenbeständen durch unterschiedliche Forschergruppen sicherstellt. Zusätzlich würde eine systematische Organisation der Indikatoren eine angemessene Auswahl von Indikatoren für unterschiedliche Anwendungsszenarien unterstützen.

Schlüsselwörter: medizinische Forschung, Datenqualität, Gesundheitswesen, Leitlinie, Analyse, Informatik

Introduction

Data are the treasure of quantitative research. Strenuous efforts are undertaken to obtain high-quality data [1]. Metadata are defined, data acquisition is standardized, data collection is supported by plausibility checks, data quality is reported to study sites, recorded data are compared to originals ones, to mention only some of the available methods to achieve high data quality. However, those methods are only beneficial if their success is controlled. Moreover, those methods could be tailored to the level of the assessed data quality [2].

Data are more and more used beyond their original context, for example data from the electronic patient record in clinical trials [3]. Then, an assessment of the data quality is needed to decide whether the data are appropriate to answer a specific research question or not [4]. The use of indicators or key performance measures is an established methodology in health care to assess quality [5]. Results that are closer to a predefined goal or closer to an optimum indicate better quality. Meanwhile, the use of quality indicators becomes accepted also for the assessment of data quality. For example, cancer registries have a long tradition in calculating measures such as case completeness, data completeness and validity [6].

There is a strong emphasis on synthesizing a conceptual framework covering terminological and ontological aspects in data quality research. Wang and Strong distinguished four dimensions of data quality through a systematic approach; intrinsic data quality, contextual data quality, representational data quality, and accessibility data quality [7]. Fifteen indicators were assigned to one dimension each and briefly described by a single sentence. However, this work was not really elaborated in view of health care. Botsis et al. reduced data quality to the aspects of incompleteness (i.e. missing information), inconsistency (i.e. information mismatch), and inaccuracy [8]. Weiskopf and Weng came up with completeness, correctness, concordance, plausibility, and currency as dimensions [9]. They defined currency as "a relevant representation of the patient state at a given point in time". Recency and timeliness were listed as related terms. Furthermore, Weiskopf and Weng extended the perspective of data quality dimensions with seven data quality assessment methods like data source agreement. Kahn et al. shortened the top-level dimensions of Weiskopf and Weng to the data quality categories conformance, completeness, and plausibility related to the data quality assessment contexts verification and validation [10].

More than ten years ago, a group within the TMF – Technology, Methods, and Infrastructure for Networked Medical Research, an umbrella organization for networked medical research in Germany, developed a guideline for an adaptive management of data quality [2], [11]. This work started with the aim to support the practice of data management, in opposite to the conceptual approaches introduced before. The methodology applied was influenced by quality research, in particular health care quality research. According to Donabedian,

1. quality can be described on the levels of structures, processes, and outcomes [12],
2. quality is measured using indicators [5], and
3. continuous quality improvement is driven by the quality circle of Deming [13].

Central to the guideline is, first, the measurement of data quality using a set of indicators, and second adapting source data verification and feedback to the level of data quality that becomes evident by the indicator results. The first version of the guideline, published in 2006, included 24 indicators organized in three categories: plausibility (10 indicators), organization (7), and trueness (7) [11]. Plausibility referred to Donabedian's level of structures, organization to the level of processes, and trueness to the level of outcomes. The indicators were identified based on a systematic literature review. Due to the focus of the TMF members, the guideline particularly addressed the needs of cohorts and registries.

In the revision of the guideline a different approach was applied [14]. On the one hand, the list of indicators was evaluated and extended bottom-up making use of real-world examples of cohorts, data repositories, and registries. On the other hand, the systematic literature review concerning data quality was updated in parallel. Objective of the current study was to compare both results to find gaps that could be closed in future work and to identify consensus that could help to establish a consistent and unambiguous matrix of indicators of data quality.

Material and methods

Guideline revision

A bottom-up approach was applied in order to evaluate and update the list of quality indicators [14]. Projects were identified that are suited as proxies for cohorts, registries, and data repositories representing the main types of quantitative research of the TMF members. The measures for data quality used by those projects were collected and mapped onto the list of indicators defined in the first version of the guideline. Measures missing in the guideline were added based on a consensual decision by the study participants.
The proxies were as follows.

- One statutory epidemiological cancer registry participated as proxy for registries, being an active member of the Association of Population-based Cancer Registries in Germany (GEKID).
- The Study of Health in Pomerania (SHIP) participated as a proxy for cohort studies. SHIP is a large population-based epidemiological study in the region of Western Pomerania, Germany.
- The Open European Nephrology Science Center (OpEN.SC) represented data repositories. Data repositories collect data from a wide range of studies without a predefined research question. The data are then provided to third parties.

Systematic literature review

The literature review followed a standardized approach according to the recommendations on Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [15]. Medline was used as literature database. Citations from 2005 to March 2013 in English and German were included. The queries covered 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. The selection of the relevant literature was conducted in two steps and controlled by an overlapping evaluation between three raters. Decisions in case of questionable citations were made in a consensus. The indicators from the systematic literature review were mapped to the guideline's list of indicators. Denominations and definitions were used both for the mapping.

Results

Quality indicators proposed by the guideline

The second version of the guideline was expanded to 51 quality indicators organized in three categories: integrity (former denomination plausibility, 30 indicators), organization (15) and trueness (6) [16]. For the first time, an indicator addressing the quality of metadata was included. This indicator belongs to the category integrity. In the second version, the structured description of each indicator was substituted by information about the appropriate context. Three possibilities were differentiated for that context:

1. an indicator can be calculated for an individual record,
2. an indicator can be calculated for an individual observational unit,
3. an indicator can be calculated for a complete data set.

Context 3 is the traditional one in the application of indicators. Table 1 shows the list of all 51 indicators. Each quality indicator is defined in a structured format dating back to recommendations of the Joint Commission on Accreditation of Healthcare Organisations (JCAHO) [5] using the following attributes: name, description, definition of terms, identifier, type of indicator (structure, process or outcome), literature references, context (see above), alternative definitions, comments, numerator, denominator, subcategories, method of calculation, interpretation of results, predictors and confounders (cf. Attachment 1 for an example).

Quality indicators identified in the literature review

The systematic literature review yielded 39 articles concerned with either indicators of data quality, feedback about data quality, or source data verification [7], [9], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53]. Thirty-one of the 39 articles included information about 34 different quality indicators. Table 2 shows the list of the quality indicators along with a reference to the indicators listed in Table 1. Ten indicators from the guideline could not be attached to any of the indicators mentioned in the literature (20% from 51 indicators). Four of those ten indicators had been introduced by representatives of the cancer registry, two by representatives of SHIP, and one by representatives of the data repository from OpEN.SC. Thirteen indicators mentioned in the literature were not addressed in the guideline (38% from 34 indicators): accessibility, appropriate amount of data, availability, believability, contextualization, granularity, inaccuracy, policy relevance, predictive value, relevancy, responsiveness of data items, spatial stability, and timeliness. Combining both sets, 64 indicators were available.

Discussion

The second version of the TMF guideline for the management of data quality in cohorts and registries offers 51 indicators for the assessment of data quality. A sys-

Table 1: List of indicators proposed by the guideline

Category/indicator (hierarchically organized)	ID
Category integrity	
Agreement with previous values	TMF-1001
Concordance	TMF-1002
Consistency • Endless survivor • Certain contradiction/error • Possible contradiction/warning	TMF-1003 TMF-1035 TMF-1004 TMF-1005
Distribution of values • Last digit preferences • Distribution of parameters recorded by the investigator • Distribution of parameters recorded by the device • Distribution of findings recorded by a medical reader • Distribution of parameters between study sites • Medical tests on weekends	TMF-1006 TMF-1007 TMF-1009 TMF-1010 TMF-1011 TMF-1052 TMF-1008
Missing entries • Missing modules • Missing values in data elements • Missing values in mandatory data elements • Missing values in optional data elements • Data elements with value unknown etc.	 TMF-1012 TMF-1013 TMF-1014 TMF-1015 TMF-1016
Data elements with existing entries for all observational units	TMF-1017
Outliers (continuous data elements)	TMF-1018
Values that exceed measurement limits	TMF-1019
Values from standards	TMF-1020
Illegal values • Illegal values of qualitative data elements • Illegal values of qualitative data elements used for the coding of missings • Illegal values used for the coding of missing modules • Illegal values of qualitative data elements used for the coding of results exceeding measurement limits	 TMF-1021 TMF-1022 TMF-1023 TMF-1024
Data elements with unspecific values	TMF-1025
Observational unit with unknown primary tumor	TMF-1026
Evidence of known correlations	TMF-1027
Coverage of metadata from investigations	TMF-1050
Category organization	
Currency	TMF-1028
Duplicates	TMF-1029
Recruitment rate	TMF-1030
DCO rate (Death Certificate Only)	TMF-1051
Refusal rates • Refusal rate of investigations • Refusal rate of modules • Refusal rate of single data elements	 TMF-1031 TMF-1032 TMF-1033
Drop-out rate	TMF-1034
Synonyms	TMF-1036
Homonyms	TMF-1037
Notifications per observational unit	TMF-1038
Sole notifications from pathologists	TMF-1039
Rejected notifications	TMF-1040
Data sources per observational unit	TMF-1041
Observational units with follow-up	TMF-1042
Category trueness	
Accuracy	TMF-1043
Agreement with source data referring to data elements	TMF-1044
Agreement with source data referring to observational units	TMF-1045
Completeness	TMF-1046
Compliance with operating procedures	TMF-1047
Representativeness	TMF-1048

Table 2: List of indicators identified in the literature review

Indicator	Frequency	References	ID TMF
Accessibility	1	[21]	
Accuracy	10	[18], [20], [23], [25], [35], [38], [41], [44], [48], [53]	1043
Inaccuracy	1	[7]	
Agreement	1	[20]	1044, 1045
Appropriate amount of data	1	[30]	
Availability	1	[33]	
Believability	2	[30], [41]	
Comparability	3	[19], [25], [32]	1027, 1050
Completeness	4	[25], [33], [43], [52]	1046
Incompleteness	1	[7]	1026
Comprehensiveness	21	[9], [18], [21], [22], [23], [25], [27], [28], [31], [32], [33], [35], [37], [38], [40], [41], [42], [44], [45], [46], [48]	1012, 1013, 1014
Concordance	3	[9], [36], [42]	1002
Consistency	5	[21], [37], [40], [41], [45]	1003
Inconsistency	1	[7]	1004, 1019, 1035
Contextualization	1	[40]	
Correctness	14	[9], [22], [27], [29], [30], [32], [33], [37], [40], [41], [42], [45], [46], [51]	1021, 1022, 1023, 1024
Currency	6	[19], [21], [25], [31], [32], [43]	1028
Definition	1	[21]	1016
Generalizability	1	[29]	1017, 1048
Granularity	1	[21]	
Objectivity	1	30	1020
Plausibility	2	[9], [42]	1005, 1006, 1007, 1008, 1018
Policy relevance	1	[29]	
Precision	1	[21]	1025
Predictive value	1	[40]	
Prevention of duplicates	2	[25], [40]	1029, 1036
Rate of enrolment	2	[31], [52]	1030
Relevancy	2	[21], [23]	
Reliability	4	[28], [29], [40], [53]	1001, 1009, 1010, 1011, 1052
Responsiveness of data items	1	[25]	
Spatial stability	1	[40]	
Timeliness	6	[9], [21], [30], [33], [40], [42]	
Usefulness of data items	1	[25]	1015
Validity	4	[19], [25], [28], [43]	1039, 1051

tematic literature review revealed 34 quality indicators used in 31 articles. Three indicators were mentioned ten times or more: accuracy, comprehensiveness and correctness. The mapping between both sets of indicators failed for ten of the TMF indicators (1031, 1032, 1033, 1034, 1037, 1038, 1040, 1041, 1042, and 1047). Nine out of those ten indicators are defined in the category organization. There seems to be a lack of understanding in the literature concerning the importance of measures related to the organization of cohorts and registries. Furthermore, seven of the missing indicators in the category organization were applied in data management of real data sets. Indicator TMF-1047 "Compliance with operating procedures" is not only missing in the reviewed literature, but also in the introduced conceptual frameworks [7], [9], [10]. This neglects the requirements of empirical research to be compliant with predefined procedures, e.g. the timeline of follow-ups defined in the study protocol. One might assume that the elaboration of data quality focused in the past on the data itself neglecting the importance of process-related issues to some extent.

Thirteen out of the 34 indicators from the literature remained without a corresponding TMF indicator. Twelve were less frequently mentioned in the literature with only one or two citations. Timeliness was mentioned six times. Timeliness is an important issue particularly in diagnosis and treatment as well as for reimbursement. Timeliness is of minor importance for research purposes. However, to offer a comprehensive set of quality indicators, timeliness should be added. The other twelve are to some extend overlapping with other indicators proposed in the literature. For example, there is an unclear relationship between accuracy, believability, correctness, and validity. Proposals had been made in the literature to clarify the definitions [54]. However, the differentiation is still unclear. Contextualization, policy relevance, responsiveness of data items, and spatial stability extended the list of TMF indicators as well as the health care related conceptual frameworks mentioned before. These indicators could be assigned to the dimension contextual data quality defined by Wang and Strong. They already stated that contextual data quality "was not explicitly recognized in the data quality literature" [8]. Possibly, this conflicts with the paradigm of empirical research and health care, to define the tasks first and collect the required data second. Then, the usefulness of the data is guaranteed by the predefined usage. However, in view of an increasing use of already existing data, contextual data quality might receive a greater importance in the future [55], [56], [57].

Some shortcomings have to be mentioned concerning the list of quality indicators in Table 1. The granularity of the indicator denominations and definitions varies, having broad measures as concordance on the one hand and particular measures as the rate of Death Certificate Only cases (DCO rate) on the other hand. The hierarchical organization is an attempt to address those differences. However, this solution is still suboptimal. Terms like "data element" and "value" are not always precise enough to represent the content of the indicator by its denomination. Therefore, the structured description offered in the long version of the guideline is essential to understand the meaning of an indicator.

Conclusions

The list of indicators for data quality derived in the presented project (cf. Table 1) covers many of the concepts used in the literature. It combines different perspectives, all relevant for data quality,

1. the perspective of data management responsible for data collection and data control,
2. the perspective of data users, being unable to influence the process of data acquisition, and
3. the perspective of process owners, defining the host projects and studies.

Therefore, that list could be the starting point for a harmonization of indicators of data quality urgently needed noticing the variety and sometimes incompatibility of the measures mentioned in the literature. It will be a next step to offer a synthesis of both lists presented here along with precise definitions. This could be a valuable mission for standardization organizations that deal with data in health care and health care research. This research should take into account proposals for a formal definition of indicators [58]. Formal definitions would enable an automatic application of indicators to data sets, for example offering a syntax for statistical software based on standards for metadata [59].

Notes

Funding

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/) and contract STA 454/17-1 by the Deutsche Forschungsgemeinschaft (DFG).

Competing interests

The authors declare that they have no competing interests.

Attachments

Available from
http://www.egms.de/en/journals/mibe/2019-15/mibe000199.shtml

1. mibe000199_Attachment1.pdf (73 KB)
   Definition of indicator completeness

References

1. 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 Nov-Dec;9(6):600-11. DOI: 10.1197/jamia.M1087
2. Nonnemacher M, Weiland D, Neuhäuser M, Stausberg J. Adaptive management of data quality in cohort studies and registers: proposal for a guideline. Acta Inform Med. 2007;15:225-30.
3. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, Dugas M, Dupont D, Schmidt A, Singleton P, De Moor G, Kalra D. Electronic health records: new opportunities for clinical research. J Intern Med. 2013 Dec;274(6):547-60. DOI: 10.1111/joim.12119
4. Malin JL, Keating NL. The cost-quality trade-off: need for data quality standards for studies that impact clinical practice and health policy. J Clin Oncol. 2005 Jul;23(21):4581-4. DOI: 10.1200/JCO.2005.01.912
5. Joint Commission on Accreditation of Healthcare Organisations (JCAHO). Primer on indicator development and application. Measuring Quality in Health Care. Oakbrook Terrace: JCAHO; 1990.
6. Brooke EM. The current and future use of registries in health information systems. Geneva: World Health Organization; 1974.
7. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit Transl Bioinform. 2010 Mar 1;2010:1-5.
8. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst. 1996;12:5-33. DOI: 10.1080/07421222.1996.11518099
9. 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 Jan;20(1):144-51. DOI: 10.1136/amiajnl-2011-000681

10. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016 Sep 11;4(1):1244. DOI: 10.13063/2327-9214.1244

11. Nonnemacher M, Weiland D, Stausberg J. Datenqualität in der medizinischen Forschung. Leitlinie zum adaptiven Management von Datenqualität in Kohortenstudien und Registern. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2007.

12. Donabedian A. Evaluating the quality of medical care. Milbank Mem Fund Q. 1966 Jul;44(3):166-206.

13. Deming WE. Out of the crisis. Cambridge: Cambridge University Press; 1982.

14. 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.

15. Ziegler A, König IR. Leitlinien fur Forschungsberichte: Deutschsprachige Übersetzungen von CONSORT 2010, PRISMA und STARD. [Guidelines for research reports: German translation of CONSORT 2010, PRISMA and STARD]. Dtsch Med Wochenschr. 2011 Feb;136(8):357-8. DOI: 10.1055/s-0031-1272535

16. Nonnemacher M, Nasseh D, Stausberg J. Datenqualität in der medizinischen Forschung. 2., aktual. u. erw. Aufl. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014.

17. Baigent C, Harrell FE, Buyse M, Emberson JR, Altman DG. Ensuring trial validity by data quality assurance and diversification of monitoring methods. Clin Trials. 2008;5(1):49-55. DOI: 10.1177/1740774507087554

18. Berner ES, Kasiraman RK, Yu F, Ray MN, Houston TK. Data quality in the outpatient setting: impact on clinical decision support systems. AMIA Annu Symp Proc. 2005:41-5.

19. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer. 2009 Mar;45(5):747-55. DOI: 10.1016/j.ejca.2008.11.032

20. Brender JD, Suarez L, Langlois PH. Validity of parental work information on the birth certificate. BMC Public Health. 2008 Mar;8:95. DOI: 10.1186/1471-2458-8-95

21. Bronnert J, Clark JS, Cassidy BS, Fenton S, Hyde L, Kallem C, Watzlaf V. Data quality management model (updated). J AHIMA. 2012 Jul;83(7):62-71.

22. Brouwer HJ, Bindels PJ, Weert HC. Data quality improvement in general practice. Fam Pract. 2006 Oct;23(5):529-36. DOI: 10.1093/fampra/cml040

23. Chiba Y, Oguttu MA, Nakayama T. Quantitative and qualitative verification of data quality in the childbirth registers of two rural district hospitals in Western Kenya. Midwifery. 2012 Jun;28(3):329-39. DOI: 10.1016/j.midw.2011.05.005

24. Choquet R, Qouiyd S, Ouagne D, Pasche E, Daniel C, Boussaïd O, Jaulent MC. The Information Quality Triangle: a methodology to assess clinical information quality. Stud Health Technol Inform. 2010;160(Pt 1):699-703.

25. Couchoud C, Lassalle M, Cornet R, Jager KJ. Renal replacement therapy registries – time for a structured data quality evaluation programme. Nephrol Dial Transplant. 2013 Sep;28(9):2215-20. DOI: 10.1093/ndt/gft004

26. De S. Hybrid approaches to clinical trial monitoring: Practical alternatives to 100% source data verification. Perspect Clin Res. 2011 Jul;2(3):100-4. DOI: 10.4103/2229-3485.83226

27. Duda SN, Shepherd BE, Gadd CS, Masys DR, McGowan CC. Measuring the quality of observational study data in an international HIV research network. PLoS ONE. 2012;7(4):e33908. DOI: 10.1371/journal.pone.0033908

28. Dyck MJ, Culp K, Cacchione PZ. Data quality strategies in cohort studies: lessons from a study on delirium in nursing home elders. Appl Nurs Res. 2007 Feb;20(1):39-43. DOI: 10.1016/j.apnr.2006.01.004

29. França E, de Abreu DX, Rao C, Lopez AD. Evaluation of cause-of-death statistics for Brazil, 2002-2004. Int J Epidemiol. 2008 Aug;37(4):891-901. DOI: 10.1093/ije/dyn121

30. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012 Jul;50 Suppl:S21-9. DOI: 10.1097/MLR.0b013e318257dd67

31. Krzych LJ, Lees B, Nugara F, Banya W, Bochenek A, Cook J, Taggart D, Flather MD. Assessment of data quality in an international multi-centre randomised trial of coronary artery surgery. Trials. 2011 Sep;12:212. DOI: 10.1186/1745-6215-12-212

32. Larsen IK, Småstuen M, Johannesen TB, Langmark F, Parkin DM, Bray F, Møller B. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. Eur J Cancer. 2009 May;45(7):1218-31. DOI: 10.1016/j.ejca.2008.10.037

33. Loane M, Dolk H, Garne E, Greenlees R; EUROCAT Working Group. Paper 3: EUROCAT data quality indicators for population-based registries of congenital anomalies. Birth Defects Res Part A Clin Mol Teratol. 2011 Mar;91 Suppl 1:S23-30. DOI: 10.1002/bdra.20779

34. Macefield RC, Beswick AD, Blazeby JM, Lane JA. A systematic review of on-site monitoring methods for health-care randomised controlled trials. Clin Trials. 2013 Feb;10(1):104-24. DOI: 10.1177/1740774512467405

35. Maruszewski B, Lacour-Gayet F, Monro JL, Keogh BE, Tobota Z, Kansy A. An attempt at data verification in the EACTS Congenital Database. Eur J Cardiothorac Surg. 2005 Sep;28(3):400-4; discussion 405-6. DOI: 10.1016/j.ejcts.2005.03.051

36. McKenzie K, Walker S, Besenyei A, Aitken LM, Allison B. Assessing the concordance of trauma registry data and hospital records. Health Inf Manag. 2005;34(1):3-7. DOI: 10.1177/183335830503400103

37. Messenger JC, Ho KK, Young CH, Slattery LE, Draoui JC, Curtis JP, Dehmer GJ, Grover FL, Mirro MJ, Reynolds MR, Rokos IC, Spertus JA, Wang TY, Winston SA, Rumsfeld JS, Masoudi FA; NCDR Science and Quality Oversight Committee Data Quality Workgroup. The National Cardiovascular Data Registry (NCDR) Data Quality Brief: the NCDR Data Quality Program in 2012. J Am Coll Cardiol. 2012 Oct;60(16):1484-8. DOI: 10.1016/j.jacc.2012.07.020

38. Mphatswe W, Mate KS, Bennett B, Ngidi H, Reddy J, Barker PM, Rollins N. Improving public health information: a data quality intervention in KwaZulu-Natal, South Africa. Bull World Health Organ. 2012 Mar;90(3):176-82. DOI: 10.2471/BLT.11.092759

39. Nahm ML, Pieper CF, Cunningham MM. Quantifying data quality for clinical trials using electronic data capture. PLoS ONE. 2008 Aug;3(8):e3049. DOI: 10.1371/journal.pone.0003049

40. Sáez C, Martínez-Miranda J, Robles M, García-Gómez JM. Organizing data quality assessment of shifting biomedical data. Stud Health Technol Inform. 2012;180:721-5.

41. Salati M, Brunelli A, Dahan M, Rocco G, Van Raemdonck DE, Varela G; European Society of Thoracic Surgeons Database Committee. Task-independent metrics to assess the data quality of medical registries using the European Society of Thoracic Surgeons (ESTS) Database. Eur J Cardiothorac Surg. 2011 Jul;40(1):91-8. DOI: 10.1016/j.ejcts.2010.11.004

42. Shabestari O, Roudsari A. Challenges in data quality assurance for electronic health records. Stud Health Technol Inform. 2013;183:37-41.

43. Sigurdardottir LG, Jonasson JG, Stefansdottir S, Jonsdottir A, Olafsdottir GH, Olafsdottir EJ, Tryggvadottir L. Data quality at the Icelandic Cancer Registry: comparability, validity, timeliness and completeness. Acta Oncol. 2012 Sep;51(7):880-9. DOI: 10.3109/0284186X.2012.698751

44. Stevens W, Stevens G, Kolbe J, Cox B. Comparison of New Zealand Cancer Registry data with an independent lung cancer audit. N Z Med J. 2008 Jun 20;121(1276):29-41.

45. Taggart J, Liaw ST, Dennis S, Yu H, Rahimi A, Jalaludin B, Harris M. The University of NSW electronic practice based research network: disease registers, data quality and utility. Stud Health Technol Inform. 2012;178:219-27.

46. Thoburn KK, German RR, Lewis M, Nichols PJ, Ahmed F, Jackson-Thompson J. Case completeness and data accuracy in the Centers for Disease Control and Prevention's National Program of Cancer Registries. Cancer. 2007 Apr;109(8):1607-16. DOI: 10.1002/cncr.22566

47. Tolonen H, Dobson A, Kulathinal S; WHO MONICA Project. Assessing the quality of risk factor survey data: lessons from the WHO MONICA Project. Eur J Cardiovasc Prev Rehabil. 2006 Feb;13(1):104-14. DOI: 10.1097/00149831-200602000-00017

48. Tuble SC. Perfusion Downunder Collaboration Database – data quality assurance: towards a high quality clinical database. J Extra Corpor Technol. 2011 Mar;43(1):P44-51.

49. Tudur Smith C, Stocken DD, Dunn J, Cox T, Ghaneh P, Cunningham D, Neoptolemos JP. The value of source data verification in a cancer clinical trial. PLoS ONE. 2012;7(12):e51623. DOI: 10.1371/journal.pone.0051623

50. Venet D, Doffagne E, Burzykowski T, Beckers F, Tellier Y, Genevois-Marlin E, Becker U, Bee V, Wilson V, Legrand C, Buyse M. A statistical approach to central monitoring of data quality in clinical trials. Clin Trials. 2012 Dec;9(6):705-13. DOI: 10.1177/1740774512447898

51. Verhulst K, Artiles-Carloni L, Beck M, Clarke JT, Neto JC, Cox GF, Fernhoff PM, Guffon N, Kong Y, Martins AM, Tylki-Szymanska A, Whitley CB, Wijburg FA, Wraith EJ, Koepper CM. Source document verification in the Mucopolysaccharidosis Type I Registry. Pharmacoepidemiol Drug Saf. 2012 Jul;21(7):749-752. DOI: 10.1002/pds.2200

52. Wu Y, Takkenberg JJ, Grunkemeier GL. Measuring follow-up completeness. Ann Thorac Surg. 2008 Apr;85(4):1155-7. DOI: 10.1016/j.athoracsur.2007.12.012

53. Xian Y, Fonarow GC, Reeves MJ, Webb LE, Blevins J, Demyanenko VS, Zhao X, Olson DM, Hernandez AF, Peterson ED, Schwamm LH, Smith EE. Data quality in the American Heart Association Get With The Guidelines-Stroke (GWTG-Stroke): results from a national data validation audit. Am Heart J. 2012 Mar;163(3):392-8, 398.e1. DOI: 10.1016/j.ahj.2011.12.012

54. Brennan PF, Stead WW. Assessing data quality: from concordance, through correctness and completeness, to valid manipulatable representations. J Am Med Inform Assoc. 2000 Jan-Feb;7(1):106-7. DOI: 10.1136/jamia.2000.0070106

55. Dugas M, Lange M, Müller-Tidow C, Kirchhof P, Prokosch HU. Routine data from hospital information systems can support patient recruitment for clinical studies. Clin Trials. 2010 Apr;7(2):183-9. DOI: 10.1177/1740774510363013

56. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol. 1996 Apr;25(2):435-42. DOI: 10.1093/ije/25.2.435

57. Winter A, Funkat G, Haeber A, Mauz-Koerholz C, Pommerening K, Smers S, Stausberg J. Integrated information systems for translational medicine. Methods Inf Med. 2007;46(5):601-7. DOI: 10.1160/ME9063

58. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Application of An Ontology for Characterizing Data Quality For a Secondary Use of EHR Data. Appl Clin Inform. 2016;7(1):69-88. DOI: 10.4338/ACI-2015-08-RA-0107

59. Ngouongo SM, Löbe M, Stausberg J. The ISO/IEC 11179 norm for metadata registries: does it cover healthcare standards in empirical research? J Biomed Inform. 2013 Apr;46(2):318-27. DOI: 10.1016/j.jbi.2012.11.008

Corresponding author:

Prof. Dr. med. Jürgen Stausberg
Institute for Medical Informatics, Biometry and Epidemiology, Faculty of Medicine, University Duisburg-Essen, Hufelandstrasse 55, 45122 Essen, Germany
stausberg@ekmed.de

Please cite as

Stausberg J, Bauer U, Nasseh D, Pritzkuleit R, Schmidt CO, Schrader T, Nonnemacher M. Indicators of data quality: review and requirements from the perspective of networked medical research. GMS Med Inform Biom Epidemiol. 2019;15(1):Doc05.
DOI: 10.3205/mibe000199, URN: urn:nbn:de:0183-mibe0001998

This article is freely available from

http://www.egms.de/en/journals/mibe/2019-15/mibe000199.shtml

Published: *2019-07-09*