OPEN ACCESS

Research Article

Data quality monitoring in clinical and observational epidemiologic studies: the role of metadata and process information

Management von Datenqualität in klinischen und beobachtenden epidemiologischen Studien: Die Rolle von Metadaten und Prozessinformationen

Adrian Richter[1]
Janka Schössow[1]
André Werner[1]
Birgit Schauer[1]
Dörte Radke[1]
Jörg Henke[1]
Stephan Struckmann[1]
Carsten Oliver Schmidt[1]

1 Institute for Community Medicine, University Medicine Greifswald, Germany

Abstract

High data quality is fundamental for valid inferences in health research. Metadata, i.e. "data that describe other data", are essential to implement data quality assessments but more guidance on which metadata to use is needed. Similarly, the selection and use of variables describing the measurement process should be exemplified to improve the design and conduct of observational health studies. This work provides a conceptual framework and overview of metadata and process information for systematic data quality reports based on implementations within the population-based cohort Study of Health in Pomerania (SHIP). In previous years, a prerequisite for automated data quality checks has been established by the augmentation of the data dictionary; the added information of up to 20 different characteristics for each variable is used for data quality assessments and triggers diverse data quality checks. Conceptually we distinguish static metadata, variable metadata, and process variables. Examples for static metadata are the expected probability distribution, plausibility limits, and the data type. Variable metadata may be reference limits of a laboratory marker. Information inherent to these metadata allows for the detection of data quality flaws by comparing observed with expected data characteristics. In contrast, process variables, such as the observer or device ID, also allow for the identification of sources of data quality issues. This is the case even if characteristics defined in metadata were not violated. Metadata and process variables can be used alone or in combination to implement a versatile and efficient data quality assessment. A comprehensive setup of metadata and process variables is an extensive task, particularly in studies involving large data collections. Nonetheless, the gain in transparency and efficacy of data curation and quality reporting after this setup is considerable.

Keywords: data quality, metadata, process variables, data monitoring, health research, cohort studies

Zusammenfassung

Eine hohe Datenqualität ist eine wesentliche Voraussetzung für valide Entscheidungen in der Gesundheitsforschung. Metadaten bzw. „Daten über andere Daten" sind für die Implementierung eines Datenqualitätsmonitorings essentiell. Klare Empfehlungen und Benennungen von Metadaten für spezifische Aspekte von Datenqualität werden in relevanter Literatur jedoch nicht gegeben. Gleichfalls ist nicht klar, welche Informationen über den datengenerierenden Prozess gesammelt werden sollten, um Studiendesign und -durchführung zu verbessern. In dieser Arbeit wird unter konzeptioneller Perspektive ein Überblick zu Metadaten und Prozessinformationen gegeben, welche in der Kohortenstudie

Study of Health in Pomerania (SHIP) verwendet werden. Zurückliegend wurde in SHIP das allgemein gebräuchliche *Data Dictionary* um Informationen erweitert, welche für Datenqualitätsbewertungen verwendet werden und diese auch steuern können; bis zu 20 unterschiedliche Charakteristika von Variablen können spezifiziert werden. Konzeptionell werden hierfür statische von variablen Metadaten sowie Prozessvariablen unterschieden. Zum Beispiel sind die Verteilungsform, Plausibilitäts- und Zulässigkeitsgrenzen sowie der Dateneingabetyp statische Metadaten. Variierende Referenzgrenzen von z.B. Laborparametern werden als variable Metadaten betrachtet. Diese Information erlaubt die Identifizierung von Beeinträchtigungen der Datenqualität durch einen Vergleich von beobachteten und erwarteten Charakteristika der Daten. Prozessvariablen wie die ID des Untersuchers oder des Messgeräts erlauben hingegen die Identifikation von möglichen Quellen für Fehler, selbst wenn keine Metadaten verletzt wurden. Metadaten und Prozessvariablen können jeweils allein oder in Kombination verwendet werden, um vielseitige und effiziente Qualitätsbewertungen umzusetzen. Die Erstellung notwendiger Metadaten und die Definition von Prozessvariablen bedeuten einen erheblichen Aufwand, insbesondere für größere Studien. Der Zugewinn an Transparenz und Effektivität bei der Qualitätsberichterstellung ist jedoch erheblich.

Schlüsselwörter: Datenqualität, Metadaten, Prozessvariablen, Datenmonitoring, Gesundheitsforschung, Kohortenstudien

Introduction

Metadata is considered as "data that describe other data" [1]. It plays a key role for the assessment of data quality in different scientific disciplines. Definitions and use of metadata are manifold [2], [3], [4], [5], [6]. In health research, metadata may cover conceptual aspects such as descriptions of the sampling scheme of a study, or it can relate to specific characteristics of single measurement variables [7] such as the variable name, plausibility limits, or the data type. Most software for either electronic data capture or data quality assessments such as RedCap [8], Square^2 [9] or OPAL [10] make systematic use of metadata. A German guideline on data quality in medical research also presumes an existing metadata concept [11]. However, these and other works [12], [13], [14], [15] do not provide clear guidance on the extent, structure and use of metadata for systematic data quality assessments.

More attention also needs to be given to assessments of the data generation process. Methods from statistical process control and industrial statistics suggest considering factors that might affect the data generating process. Respective factors are called *process variables* and are systematically controlled in designs of experiments [16]. Similarly to manufacturers and engineers, principal investigator (PIs) and scientists of observational studies with primary data collections have control over the data generating process. This characteristic differentiates primary from secondary data collections and enables for interventions during ongoing studies.

Accordingly, adequate assessments of data quality in health studies should make use of metadata and carefully monitor the process under which measurements are obtained. A simple use case illustrates this necessity. In the population-based Study of Health in Pomerania (SHIP) [17] participants are examined by different examiners in a dedicated center including the drawing of blood samples to determine for example c-reactive protein (CRP). A missing CRP value may have, among others, the following reasons for missingness: the actual value was below the detection limit of a device, a participant refused to provide a blood sample, or the examination was aborted. Related process information are the examiner, the time of the day, the transporting time elapsed between drawing of the blood sample and the final storage in the biorepository. Recording and investigating the frequencies of reasons for missing values in combination with associated process variables may point at possible targets of intervention, e.g. a training of examiners or the re-calibration of devices.

This work provides an overview of metadata and process variables along with conceptual considerations to support the implementation of systematic and automated data quality assessments based on our experience in the SHIP study.

Methods

The methodological background for this work originates from two decades of experience with data management and data monitoring in the Study of Health in Pomerania (SHIP) [18]. The SHIP study comprises two cohorts (SHIP and SHIP-TREND) with in total 8,728 participants. To date, four SHIP and two SHIP-TREND waves have been completed. More than 40,000 variables originate from computer-assisted personal interviews, self-reported question-

naires, biomaterials (blood, urine, faeces, saliva), imaging data (e.g. ultrasound and MRI), and a wide range of clinical examinations, including dental, dermatological, and cardiovascular measurements. Each electronic case report form (eCRF) in SHIP is based on metadata and collects process information. In addition, OMICS data complement the data collection, as well as subsequent secondary assessments, e.g. readings of magnetic resonance images.

Quality management within SHIP rests upon the storage of study data and metadata in a central data repository. For this purpose a PostgreSQL database backend is used [19]. Web applications support the creation of data dictionary elements (Shipdesigner), and electronic data capture (Shippie). The former is used for metadata setup which is used by the latter to control for errors in the data entry process. Subsequently, routines in SAS and in dedicated data quality assessment environments (SQuaRe, Square²) make use of metadata and process variables to conduct data quality checks [9], [19], [20].

This work summarizes metadata and process variables utilized in these applications and structures them (i) according to the type of input data and (ii) the data quality dimensions completeness and correctness. These data quality dimensions are sometimes referred to as intrinsic data quality [21], [22], i.e. data quality which can be evaluated without the use of contextual information such as a specific research question.

Results

Data structure and terminology

Our approach considers the relations of *study data* and *metadata*. Study data comprise identifiers for observational units, the clinical measurements, variables describing the process under which the data were collected and in some cases varying metadata. Each column of the study data contains varying data values or missings.

Pre-defined static characteristics apply for each column in the study data, such as a label, a value list or the data type, form the basis of static metadata on the variable level and may be stored in various forms. One option as implemented and used in SHIP is depicted in Figure 1, where the characteristics related to columns of the study data are stored in a separate table of static metadata. In SHIP each column of the study data is identifiable over a "key" which is defined in the static metadata.

Metadata for data quality assessments

Metadata used for data quality assessments describe desired or expected properties of the data [23] as well as additional information, for example, semantic annotation of variables based on uniform codes [24]. Each column of the study data has assigned static properties which are valid for the life cycle of the respective health research study [8]. Ideally, relevant metadata are defined before the data collections starts. Typical static metadata are the variable name and the data type (Figure 1, top right panel). Further examples are shown in Table 1. Such descriptive characteristics are usually denoted in the data dictionary (DD) of most studies.

In some cases applicable metadata may also vary across observations. For example, the detection limit of a new device has changed or reference limits of laboratory markers vary in a long-term study. In this case, a metadata variable needs to be included in the study data to assign varying reference limits to the target measurement (Figure 1, left panel). The link from the measurement variable (CRP) to the respective metadata variable (RefLimits_v101) is defined via an own column in the static metadata. The top right panel of Figure 1 mentions *key_ref_limits* which specifies the key of the variable containing the time-varying reference limits for CRP. Similar columns are denoted as key-columns which point to the associated metadata variable. Such structural information is required to implement automated procedures of data quality assessments.

Metadata may comprise information related to different aspects of data quality:

- data completeness (e.g. reasons for missing data, such as conditionally missing data in the category "birth complications" for males)
- data correctness (e.g. value lists, detection limits, admissible values, plausibility limits)
- the selection of statistical approaches to data quality checks (e.g. data types, distributional class)

In addition to the investigation of data characteristics a coherent and readable reporting is important. Particularly in studies involving thousands of study data variables the presentation quality is challenging but essential to impede misinterpretation [25]. Therefore, further static metadata such as labels and units of measurement can be defined to ensure a readable and standardized output. For example, the assignment of fixed colors for examiners or devices is recommended across graphical outputs (not shown in Figure 1).

Process variables for data quality assessments

Measurements in health research are vulnerable to various sources of distortion. Environmental conditions as well as examiners and devices may change over time. Process variables are needed to capture such information [16] along with the measurements in study data. Insofar process variables can be considered measurements themselves and may relate to:

- study conduct (e.g. observer, device ID, location)
- environmental conditions (e.g. examination times, processing times, room temperature, humidity).

Examples of process variables utilized in SHIP are provided in Table 2. Each eCRF in SHIP prompts the recording of such information, for example, regarding ultra-

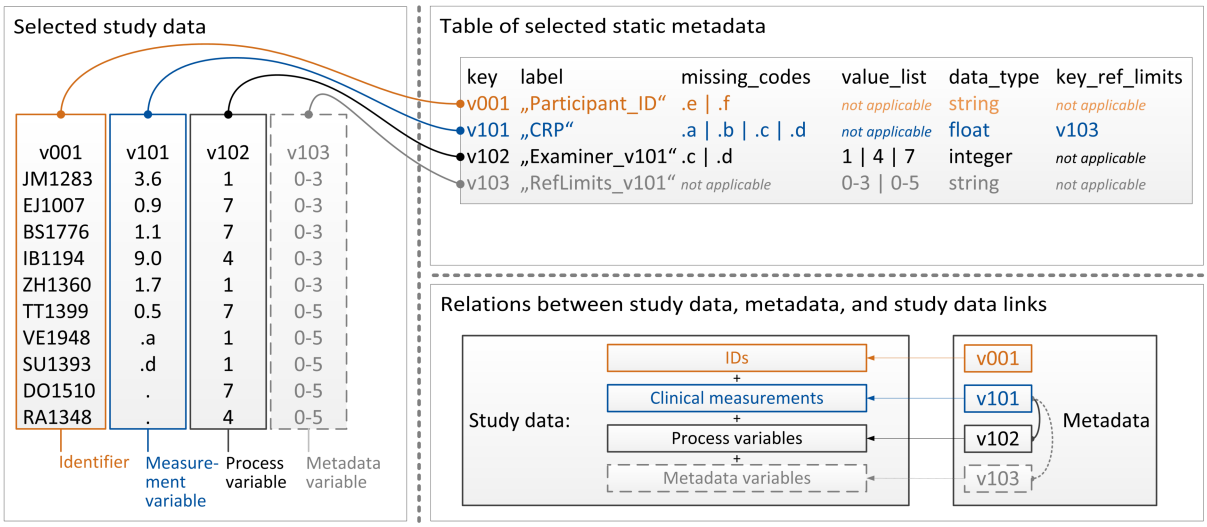

Figure 1: Left panel: study data usually comprise identifier, measurements (e.g. c-reactive protein (CRP), and process variables (e.g. examiner ID). In some cases metadata variables have to be added, if applicable metadata vary across observations. Top right panel: selection of static metadata with 1:1 relation to columns of study data. Bottom right panel: relations between study data, metadata, and links* between study data.

* Relations between two or more related study data variables should be defined in the metadata attributes, e.g. CRP-laboratory results and the examiner who drew the blood sample.

Table 1: Examples of metadata used for data quality monitoring

Name	Description	Application examples for data quality aspects
Metadata related to completeness		
Missing codes	List of specific reasons for missing measurements	Investigate types of nonresponse
Jump codes	Conditionally missing measurements (e.g. number of births in men)	Required to compute the correct denominator for nonresponse
Metadata related to data correctness		
Value list	List of specific categories/values	Inadmissible measurements
Precision	Number of decimals	End-digit preferences or ineligible rounding
Plausibility limits	Upper and/or lower soft limits	Implausible measurements
Admissibility limits	Upper and/or lower hard limits	Inadmissible measurements
Measurement/ detection limits	Limits of measurements inherent to devices	Inadmissible measurements
Distribution	Expected probability distribution (e.g. gaussian)	Unexpected probability distribution
Metadata related to selection of quality statistics		
Data type	e.g. date, date-time, float, integer, string	Used to select statistics, inadmissible measurements
Manual data transfer	Manual transfer of measurements, e.g. blood pressure measurement	e.g. selection of end-digit preference assessments

sound examination of the thyroid (https://medical-data-models.org/30755) [26] almost 25% of recorded variables comprise process information. The definition and identification of relevant process variables rests on appropriate background knowledge about factors that may influence the measurements. Their implementation in the measurement process may impose considerable additional efforts for the study conduct: additional measurement devices might be required with all related data quality management logistics.

There is a crucial difference regarding the definition and application of process variables and metadata: while metadata can be defined even after the data collection has been finished this is very difficult or impossible for process variables. Process variables should be identified and implemented prior to the start of a data collection to avoid missing and unrecoverable process information. Process information themselves are measurements of the conditions under which study data were generated.

Table 2: Examples of process variables used for data quality assessments

Name	Description	Application examples for data quality aspects
Visit	Number of visits in the examination center within a single examination wave	Compliance with procedural rules
Observer	Identifier of the examiner, reader, etc.	Observer influences
Device	Identifier of the device	Device influences
Device settings	e.g. display resolution	Device setting influences
Location	e.g. study site, building, or room	Location or center effects
Commit date	Date and time stamps related to a visit or examination, including start/end time, pre-analytic processing time	Compliance with procedural rules (e.g. blood taking only in morning hours); time trends (e.g. observer effects)
Repeated measures index	Repeated measurements nested within a single visit	Implausible measurements, retest reliability
Participant feedback	e.g. complaints, difficulties	Compliance with procedural rules (e.g. participant complains)
Environmental conditions	e.g. temperature, humidity, noise	Compliance with procedural rules (e.g. room temperature is 20–24°C)

Use of metadata for data quality assessments

Data quality assessments use metadata to investigate the compliance of observed data with expected properties [20], [23]. For example, if a categorical variable has per design four distinct values (Figure 1, "value_list") and the data show five, at least one invalid data value has been observed. Such data quality checks, also referred to as edit-, range- or cross checks [27], predominantly focus on the evaluation of entries in single data fields, i.e. each data field is checked against the desired properties of the data as coded in metadata. This means, for example regarding CRP in Figure 1, missing codes can be tabulated to infer on reasons for unavailable measurements (completeness), those being smaller than zero (inadmissibility) and those being greater than five (plausibility) are counted or flagged as potential correctness issues.

In the SHIP workflow, initial data quality checks for missing values and correctness predominantly rely on static metadata. This comprises checks during data entry in the Shippie electronic case reporting forms (eCRFs). After data capture, automated data quality controls are conducted based on SAS routines and batch jobs [19]. These checks are routinely conducted every night for the entire ongoing data collection. Feedback on issues is obtained through an MS Access data entry mask for each flagged data quality issue to ensure a timely response by the responsible quality manager. Only deviations from static metadata will be encountered at this stage, although some measurements might be inaccurate without violating predefined properties.

Use of process variables for data quality assessments

Data quality may be impaired although, according to metadata, formal discrepancies between observed data and expected properties are absent. Common examples are observer or device effects which can be impossible to detect in the overall distribution of a measurement. Process variables allow for the detection of data quality issues and their possible sources. The main focus of data quality assessments using process variables is their association with distributional characteristics of measurements. For example, very high room temperatures may explain lower performances in a spiroergometry. Seasonal changes in outcome variables that were identified using the examination date might be explainable this way.

Process variables are also required to assess the compliance with procedural rules, e.g. the analysis of sufficient resting time before a blood pressure measurement starts. Therefore, process variables may be automatically stored by recording start and end time of examinations. Other process variables can be used to check for appropriate ambient conditions under which measurements took place. For example, does the size of an arm cuff used for blood pressure measurement correspond with participants' arm circumference. These examples illustrate the different use of process variables compared to metadata for data quality assessments.

In the SHIP workflow, data quality issues measured by process variables are the main target of web applications (Square^2) dedicated to data quality assessment [9], [20]. Related reports are generated semi-automatically in defined intervals or on demand, using pdf as output format. Encountered issues are the basis for systematic feedbacks to the SHIP examination team and may trigger trainings. Contrary to the use of metadata for checks of single data fields, the use of process variables faces an important limitation. They require a sufficient number of cases to reliably detect data quality issues such as observer differences or time trends.

Combined use of metadata and process variables for data quality assessments

The combination of metadata and process variables enables for versatile data quality assessments. Univariate analyses may reveal disproportional numbers of missings or implausible measurements. In combination with process variables such as examiner, device, or preprocessing characteristics, a potential error source may be identified. For example, the measurement variable CRP in Figure 1 has two data values representing missing codes; one code has the denotation "polluted sample material". If such a missing code occurs frequently, the source should be investigated. For example, the pre-processing of probes in the laboratory or the handling of probes during/after blood drawing may cause a contamination of samples.

Discussion

This work provides an overview of metadata and process variables to monitor and improve data quality of observational studies. Accompanying conceptual considerations differentiate the features of static and variable metadata as well as process variables to support their handling in data quality assessments. The use of metadata in health research studies is crucial to follow guidelines [11] and to use metadata driven quality control with web applications such as RedCap, Square[2] or OPAL [8], [9], [10]. Similarly, process variables which were introduced from industrial statistics [16] are essential in health research studies with primary data collections, since varying conditions of the measurement process might distort the quality of the data.

The appropriate consideration of metadata and process variables may appear straightforward but in complex studies the setup is likely to be challenging and time consuming. Assigning unambiguous and understandable labels for thousands of variables requires consistent checks of the DD. In this context the use of unambiguous semantic annotation is particularly beneficial. For example, by unambiguous UMLS codes [24] which have been assigned to SHIP variables in a cooperation with the Portal for Medical Data Models (MDM) [26] to improve harmonized comparisons across studies. Some decisions for and definitions of static metadata are only possible with the intended outcome of the data quality reports in mind. Furthermore, it might be a matter of debate which plausibility or admissibility limits to define in a given study. However, starting a study with imperfect limits and comparing the data with those is more valuable than defining no limits at all, i.e. implementing no checks on measurement limits.

Metadata themselves can be a gateway of data quality flaws. For example, the static metadata *value list* and *missing codes* should be separate sets of values for each study variable, otherwise script-based routines might fail or the output of quality reports gets odd. Therefore, the coherent definition of data characteristics as metadata may consume considerable efforts if conducted for several thousand variables. However, augmenting the data dictionary with only some information required for data quality assessments (e.g. limits) already enables for essential data quality checks, particularly in larger and long-lasting studies.

The introduction and use of process variables for data quality assessments is of utmost importance. They may guide to means of interventions in the data generating process. A systematic understanding and selection of relevant process variables may require a review of existing literature because each examination needs to be considered independently. The collection of process variables can be elaborated and may add to the costs of a study. For example, monitoring ambient conditions in each examination room requires additional equipment and data base extensions, along with extensions of the data base and potentially the eCRFs. However, omitting the use of process variables eliminates one major advantage of primary data collections for data quality management: to trace back sources of errors and to regain control over a data generating process by close monitoring of this process.

The overview provided in this work has some limitations. Presented static and variable metadata are not comprehensive regarding other types of data (e.g. OMICS) [28]. We also omitted the reflection of longitudinal aspects of data quality assessments. In fact, some could be easily implemented into the presented concept, e.g. another static metadata may link different measurement variables for correctness checks such as "is age at follow-up higher than at baseline". However, longitudinal issues have special requirements regarding the format of the data (wide vs. long) and may require more complex statistical techniques for their assessment. Another limitation rests with the use of semantic annotation for study data variables. Such unambiguous codes facilitate correct interpretation of data but are currently not used for data quality assessments. Worthy of note is also the restricted use case for metadata presented in this work. Metadata are of importance beyond this application, for example, for the selection of data bases with similar populations and study focus.

High data quality means inherently that data should be fit for use. Completeness and correctness do not entirely account for this demand. Additional contextual information is necessary to evaluate the achieved data quality with respect to the intended use. For example, are all variables of interest available in a study with a sufficient sample size to analyze the effects of inflammation markers on back pain? Contextual information varies strongly with the research questions and is difficult to implement into a standardized metadata concept. Varying contextual information may also lead to different conclusions regarding the obtained data quality for the same data collection [22]. For example, sampling errors may impair the representativeness of data. This is of importance if we are interested in the prevalence of population

based risk factors. However, it may be less important if we are interested only in associations of risk factors.
Many aspects regarding the utility of metadata and process information provided in this overview are likely to be well known. However, comprehensive overviews are lacking and, in practice, their use seems inconsistent. This has contributed to critics regarding the transparency and reproducibility of research findings [29]. The provided overview may assist in the setup of data dictionaries for new studies or the augmentation of data dictionaries for existing studies. In particular smaller studies and those under development may profit from this overview in terms of transparency and several options for data quality management. Adding data quality related metadata to the DD provides an overview of applicable data quality checks. The largest gains in efficiency regarding the generation of data quality reports will be noticeable by larger, long-lasting studies requiring repeated data quality reporting.
Although metadata and process variables should be defined prior to the data collection, many pitfalls and concept flaws may only become obvious during data collection and after the system has gone productive. Therefore, improving the quality of primary health data may require adaptations of the metadata concept or the selection and measurement of process variables throughout the study.

Notes

Competing interests

The authors declare that they have no competing interests.

Funding

The development of work underlying this paper was supported by the Ministry for Education, Science and Culture of the State of Mecklenburg-Vorpommern, the European Social Fund (Grant UG 11 035A), and by the German Research Foundation (DFG, SCHM 2744/3-1).

References

1. Nadkarni PM. What Is Metadata? In: Metadata-driven software systems in biomedicine: designing systems that can adapt to changing knowledge. London, New York: Springer; 2011. (Health informatics). p. 1-16. DOI: 10.1007/978-0-85729-510-1_1
2. Schuurman N, Leszczynski A. Ontology-based metadata. Trans GIS. 2006;10(5):709-26. DOI: 10.1111/j.1467-9671.2006.01024.x
3. Vardigan M, Heus P, Thomas W. Data documentation initiative: Toward a standard for the social sciences. Int J Digit Curation. 2008;3(1):107-13. DOI: 10.2218/ijdc.v3i1.45
4. Vardaki M, Papageorgiou H, Pentaris F. A statistical metadata model for clinical trials' data management. Comput Methods Programs Biomed. 2009 Aug;95(2):129-45. DOI: 10.1016/j.cmpb.2009.02.004
5. Hughes B. Metadata Quality Evaluation: Experience from the Open Language Archives Community. In: Chen Z, Chen H, Miao Q, Fu Y, Fox E, Lim E, editors. Digital Libraries: International Collaboration and Cross-Fertilization. International Conference on Asian Digital Libraries. Berlin, Heidelberg: Springer; 2004. p. 320-9. DOI: 10.1007/978-3-540-30544-6_34
6. Huebner M, Le Cessie S, Schmidt CO, Vach W. A contemporary conceptual framework for initial data analysis. Obs Stud. 2018;4:71-192.
7. Finnie TJ, South A, Bento A, Sherrard-Smith E, Jombart T. EpiJSON: A unified data-format for epidemiology. Epidemics. 2016 Jun;15:20-6. DOI: 10.1016/j.epidem.2015.12.002
8. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377-81. DOI: 10.1016/j.jbi.2008.08.010
9. Schmidt CO, Krabbe C, Schössow J, Albers M, Radke D, Henke J. Square – A Web Application for Data Monitoring in Epidemiological and Clinical Studies. Stud Health Technol Inform. 2017;235:549-53.
10. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, Minion J, Boyd AW, Newby CJ, Nuotio ML, Wilson R, Butters O, Murtagh B, Demir I, Doiron D, Giepmans L, Wallace SE, Budin-Ljøsne I, Oliver Schmidt C, Boffetta P, Boniol M, Bota M, Carter KW, deKlerk N, Dibben C, Francis RW, Hiekkalinna T, Hveem K, Kvaløy K, Millar S, Perry IJ, Peters A, Phillips CM, Popham F, Raab G, Reischl E, Sheehan N, Waldenberger M, Perola M, van den Heuvel E, Macleod J, Knoppers BM, Stolk RP, Fortier I, Harris JR, Woffenbuttel BH, Murtagh MJ, Ferretti V, Burton PR. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014 Dec;43(6):1929-44. DOI: 10.1093/ije/dyu188
11. Nonnemacher M, Nasseh D, Stausberg J. Datenqualität in der medizinischen Forschung. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014. (TMF – Technologie- und Methodenplattform).
12. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. Int J Environ Res Public Health. 2014 May;11(5):5170-207. DOI: 10.3390/ijerph110505170
13. Edwards PN, Mayernik MS, Batcheller AL, Bowker GC, Borgman CL. Science friction: data, metadata, and collaboration. Soc Stud Sci. 2011 Oct;41(5):667-90. DOI: 10.1177/0306312711413314
14. Karr AF, Sanil AP, Banks DL. Data quality: A statistical perspective. Stat Methodol. 2006;3(2):137-73. DOI: 10.1016/j.stamet.2005.08.005
15. Nadkarni PM. Metadata-driven software systems in biomedicine: designing systems that can adapt to changing knowledge. London: Springer; 2011. (Health Informatics). DOI: 10.1007/978-0-85729-510-1
16. Montgomery DC. Design and analysis of experiments. New Jersey: Wiley; 2017.

17. Völzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, Friedrich N, Aumann N, Lau K, Piontek M, Born G, Havemann C, Ittermann T, Schipf S, Haring R, Baumeister SE, Wallaschofski H, Nauck M, Frick S, Arnold A, Jünger M, Mayerle J, Kraft M, Lerch MM, Dörr M, Reffelmann T, Empen K, Felix SB, Obst A, Koch B, Gläser S, Ewert R, Fietze I, Penzel T, Dören M, Rathmann W, Haerting J, Hannemann M, Röpcke J, Schminke U, Jürgens C, Tost F, Rettig R, Kors JA, Ungerer S, Hegenscheid K, Kühn JP, Kühn J, Hosten N, Puls R, Henke J, Gloger O, Teumer A, Homuth G, Völker U, Schwahn C, Holtfreter B, Polzer I, Kohlmann T, Grabe HJ, Rosskopf D, Kroemer HK, Kocher T, Biffar R, John U, Hoffmann W. Cohort profile: the study of health in Pomerania. Int J Epidemiol. 2011 Apr;40(2):294-307. DOI: 10.1093/ije/dyp394

18. Richter A, Schauer B, Henselin K, Junge M, Struckmann S, Sierocinsky E, Henke J, Schmidt CO. Which data and data structures are required to implement automated data quality monitoring in observational studies? Experiences from a population based cohort study. In: Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie, editor. 63. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS). Osnabrück, 02.-06.09.2018. Düsseldorf: German Medical Science GMS Publishing House; 2018. DocAbstr. 252. DOI: 10.3205/18gmds016

19. Werner A, Maiwald S, Henselin K, Westphal S, Henke J, Alte D, et al. Modular automatisierte Datenbereinigung in einer großen Bevölkerungsstudie [Modular automated data cleaning in a large population-based cohort]. In: Chenot JF, Minkenberg R, editors. Proceedings der 20. Konferenz der SAS®-Anwender in Forschung und Entwicklung (KSFE). Aachen: Shaker; 2016. p. 279-84.

20. Schmidt CO, Albers M, Henke J, Schipf S, Baumeister SE, Werner A, et al, editors. Quality monitoring in a complex epidemiologic study: Some lessons to be learned. DGEpi Jahrestagung; 2013 Sep 24-27; Leipzig.

21. Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. J Manag Inf Syst. 1996;12(4):5-33. DOI: 10.1080/07421222.1996.11518099

22. Watts S, Shankaranarayanan G, Even A. Data quality assessment in context: A cognitive perspective. Decis Support Syst. 2009;48(1):202-11. DOI: 10.1016/j.dss.2009.07.012

23. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013 Aug;51(8 Suppl 3):S22-9. DOI: 10.1097/MLR.0b013e31829b1e2c

24. Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository. BMC Med Res Methodol. 2016 06;16:65. DOI: 10.1186/s12874-016-0164-9

25. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data Sci J. 2015;14:2. DOI: 10.5334/dsj-2015-002

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

27. Lu Z, Su J. Clinical data management: Current status, challenges, and future directions from industry perspectives. Open Access J Clin Trials. 2010;2:93-105. DOI: 10.2147/OAJCT.S8172

28. Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods. 2014 Nov;11(11):1138-40. DOI: 10.1038/nmeth.3115

29. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011 Aug;10(9):712. DOI: 10.1038/nrd3439-c1

Corresponding author:
Dr. Adrian Richter
Institut für Community Medicine, Walther-Rathenau-Straße 48, 17475 Greifswald, Phone: +49 3834 867710
adrian.richter@uni-greifswald.de

Please cite as
Richter A, Schössow J, Werner A, Schauer B, Radke D, Henke J, Struckmann S, Schmidt CO. Data quality monitoring in clinical and observational epidemiologic studies: the role of metadata and process information. GMS Med Inform Biom Epidemiol. 2019;15(1):Doc08. DOI: 10.3205/mibe000202, URN: urn:nbn:de:0183-mibe0002027

This article is freely available from
https://www.egms.de/en/journals/mibe/2019-15/mibe000202.shtml

Published: *2019-11-08*