German Medical Data Sciences 2022 - Future Medicine
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI220809

Assessment of the Consistency of Categorical Features Within the DZHK Biobanking Basic Set

Khalid YUSUF [a*], Kais TAHAR[a*], Ulrich SAX[a,b], Wolfgang HOFFMANN[c,d], and Dagmar KREFTING[a,b,c,1]
[a]Department of Medical Informatics, University Medical Center Göttingen, Göttingen,Germany
[b]*Campus Institute Data Science (CIDAS), Georg August-University, Göttingen, Germany*
[c]*DZHK (German Centre for Cardiovascular Research)*
[d]*Institute for Community Medicine, Department Epidemiology of Health Care and Community Health, University Medicine Greifswald, Germany*

Abstract
Data quality in health research encompasses a broad range of aspects and indicators. While some indicators are generic and can be calculated without domain knowledge, others require information about a specific data element. Even more complex are indicators addressing contradictions, that stem from implausible combinations of multiple data elements. In this paper, we investigate how contradictions within interdependent categorical data can be identified and if they give additional information about possible quality issues, their cause, and mitigation options. The 19 data elements that represent four biosample types including their pre-analytic states within the DZHK Biobanking basic set are exported to the CDISC Operational Data Model (ODM), transformed and loaded into a tranSMART instance. Through the implementation of a data quality assessment workflow as a SmartR plug-in, statistical information about the domain-specific consistency of interdependent values are retrieved, assessed, and visualized. Data quality indicators have been selected for the assessment according to common recommendations found in the literature. Different contradictions could be discovered in the dataset including mismatch of interdependent values in the pre-analytic states of blood and urine samples, as well as primary and aliquoted samples. The overall assessment rating shows that 99.61% of the interdependent values are free of contradictions. However, measures within the EDC design to avoid contradictions may result in overestimated missing rates in automatic, item-based quality assessment checks. Through consistency checks on interdependent categorical features, we demonstrated that consistency flaws can be found in the categorical data of biobanking metadata and that they can help to detect issues in the data entry process. Our approach underscores the importance of domain knowledge in the definition of the consistency rules but also knowledge about the EDC implementation of such consistency rules to consider the impact on item-based quality indicators.

Keywords. Data quality, Biological specimen bank, metadata

* The first two authors should be regarded as joint first authors.

1 University Medical Center Göttingen, Germany; E-mail: dagmar.krefting@med.uni-goettingen.de.

1. Introduction

Data quality in health research encompasses a broad range of aspects and indicators [1-3]. According to [1], the DIN (German Institute for Standardization) EN ISO 14050 2010, defines the term data quality as the "properties of data with regard to their suitability for fulfilling specified requirements". Thus, quality indicators are quantifiable measures to describe data quality for a specific purpose. Some indicators are generic and can be calculated without domain knowledge, e.g. the completeness of a mandatory data element within a data set. Others require information about a specific data element, for example, to define limits for acceptable values. Even more complex are indicators addressing contradictions, that stem from implausible combinations of multiple data elements. A typical example is the rule, that the diastolic blood pressure must not be higher than the systolic blood pressure when measured simultaneously.

A key asset of the DZHK is the DZHK Heart Bank that provides liquid biomaterial samples and image data with comprehensive clinical data for research projects from all studies directly funded[2]. A common data set encompassing clinical data and the metadata describing the available bio-samples and images has been defined that allows interested researchers to access this information throughout the Heart Bank. The definition of the data elements including the Biobanking basic set (BBS) can be found in the publicly available data catalogue[3]. Biomaterial can be considered as a specifically valuable asset due to its high processing and maintenance costs as well as their finite nature - other than digital data, that can be copied without loss. Therefore, accuracy of associated data is of high importance, in particular in scientific cohort studies and registries [1]. To encourage consistency in the collection of high-quality data associated with biomaterial, Kiehntopf and Böer recommend data reconciliation techniques to be introduced in the biobank's information system [4, p. 64]. Previous studies have shown instances of data anomalies in biomaterial banks [5,6]. Also, the pre-analytic states of blood samples have been observed as one of the leading sources of errors in clinical laboratories [7,8]. As the DZHK BBS encompasses information about the different states of the samples, we implemented a data quality workflow to assess potential contradictions regarding the biosample processing states.

2. Material and Methods

2.1. Data

The Translational Registry for Cardiomyopathies (TORCH) has been one of the first registries conducted within the DZHK. It encompasses comprehensive data of 2300 patients with non-ischemic cardiomyopathies in order to analyze the parthenogenesis and therapeutic interventions for cardiomyopathy patients [9]. The data is recorded using the Electronic Data Capture (EDC) System secuTrial, including the BBS for this registry. Since most variables in biomaterial data are categorical, our case study investigates the categorical data elements within the BBS that capture information about EDTA (ethylenediaminetetraacetic acid) plasma, citrate plasma, serum, and urine. For each of

2 https://dzhk.de/en/dzhk-heart-bank/daten-und-bioproben/ressource-mit-fluessigproben-und-bilddaten/

3 https://dzhk.de/en/research/data-and-sample-collections/data-catalogue/biobanking-basic-set/

these sample types, three data elements refer to them: The number of primary receptables filled, that may range from 0 to 2, where 2 is only allowed in EDTA and Citrate plasma if the biomaterial kit is from specific manufacturer. The content of the primary receptables is then distributed to aliquots of 300 ul. For each sample type, a desired number of aliquots is defined. The second and the third data elements refer to the number of the filled aliquots: One element represents the fact, that all desired aliquots are filled, while the other element represents the quantity of aliquots filled if the desired number of aliquots is not reached. We also investigate the data elements that capture the pre-analytic properties of blood and urine samples. For blood, the four possible states are *normal, lipeamic, icteric* and *haemolytic*. They are determined by the color of the blood plasma. For urine, three different states are addressed: *normal, cloudy* and *bloody*, and are also determined by visual inspection. A total of 19 categorical elements are thus considered.

2.2. Selection of Data Quality Indicators

Two factors have been considered in the selection of data quality indicators for the assessment of consistency as follows: (1) The quality indicators are described consistently in current literature and (2) they are suitable for interdependent data elements. Among the prominent data quality metrics that feature mostly in literature are those identified in the data quality assessment guidelines by the TMF (Technology and Method Platform for Networked Medical Research) team [1] and the harmonized framework by Kahn et al. [2]. Schmidt et al. [3] also formulated a harmonized framework at the intersection of [1] and [2]. These resources were considered in the determination of suitable indicators for the assessment of the consistency of interdependent categorical features. Two data quality indicators are considered within the taxonomy of Schmidt et al: (1) logical contradictions, and (2) empirical contradictions. The primary reporting metrics of the indicators are thus the number of data fields N and the rate of inconsistent data %. This is also in accordance with the recommendation by Nonnemacher et al. [1]. An alternative definition of these quality indicators has been proposed by Kahn et al.: Atemporal plausibility through (a) common and (b) domain knowledge. While the latter would better reflect the atemporal character of the assessment, the former better reflects the aspect of a multi-item based quality indicator. Our decision is based on the consideration that the present work focuses on categorical variables thus excluding time-related aspects, and exclusively addresses multi-item contradictions.

2.3. Data Quality Assessment implementation in tranSMART

The data has been exported in secutrial's built-in ODM format and transformed to be loaded into the tranSMART (v16.2) analysis platform through its standard *batch_importer* method. The data quality assessment workflow is developed as a SmartR plug-in. SmartR plug-ins allow for the application of arbitrary operations on data stored in tranSMART [10]. Through the logical rules implemented in the R-Script components of the plug-in, interdependent data elements are tested against predefined logic. An inference is drawn by combining results from the data analysis with the common knowledge from the study design and domain knowledge offered by the experts. The statistical information derived from the R-Script analysis is then written as JSON object and parsed to the built-in D3-js component for visualization.

3. Results

3.1. Consistency between primary receptables and sample aliquots

As explained in section 2.1, figure 1 describes the flow of interdependent fields within the three data elements of each sample type: The number of primary receptables, the information if the desired number of aliquots are available, and - if the latter is not the case, the number of aliquots filled. In the paper-based Biomaterial Collection Form Basis Set, the number of primary receptables can be set to 0, 1 and additionally for plasma 2. For the aliquot number, there is a tickbox that allows to mark that all aliquots area filled. The number of desired aliquots is given besides the tickbox. Besides the tickbox, a field is given, with a header "Quantity". In the electronic form, in addition to the possible categorical values described above, further options not known (unbekannt) and not ascertained (nicht erhoben) are implemented for all items. The form has been implemented in a way, that the number of aliquots (Aliquot count) is only visible if the checkbox is set to no. Missing values are indicated by NA. The combinations of contradictory values are given in Table 1.

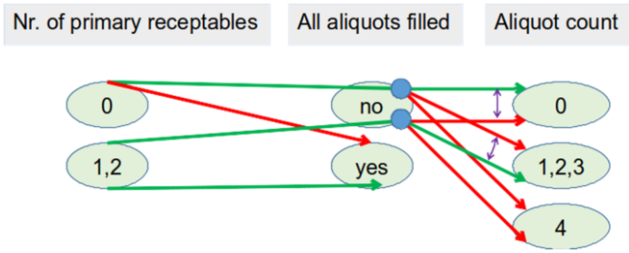

Figure 1*:* Flow of interdependent fields in the case report form (CRF). Impossible combinations due to implemented rules are not shown. Green arrows indicate plausible combinations, red arrows implausible combinations of item values. Violet arrow indicates combinations where plausibility can only be determined if the Nr. of primary receptables is considered, thus indicating a multidimensional consistency.

Table 1: Contradictory values in sample type quantities.

Rule	Nr. of Primary Receptables	All aliquots filled	Aliquot count
C1	0	yes	NA
C2	0	no	>0
C3	>0	yes	< desired aliquots
C4	>0	no	desired aliquots

Figure 2 shows the visualization of the data quality assessment for citrate plasma within tranSMART. For citrate, the maximum number of primary receptables is 2, and the number of desired aliquots is 4. For the visualization of the three dimensions a tile-based plot has been chosen. The color of the tile indicates the consistency, the number within the tile indicates the number of datasets where this value combination is found. Contradictions C1 and C2 are found in the left column of the tile matrix (*Primary receptables = 0*), C4 is found in the 2nd and 3rd column (*Primary receptables = 1,2*) within the 2nd row (*aliquot_nein*), although in very few cases. We would like to note that by EDC design, C3 is already inhibited, because the number of aliquots cannot be entered if the status that all aliquots are filled is set to *yes*. This implies that in the majority of

cases, the value for the Aliquot count is NA. Naive counting of NA values in individual data elements for completeness would result here in a high number of missing values. Another aspect is the handling of implausible combinations that are caused by missing values. In the last column (Primary receptables = NA) some entries for a full as well as an incomplete aliquot set can be observed. However, we have decided to count such implausibilities not as contradiction, as it would consider a missing value in the same way as an entered value.

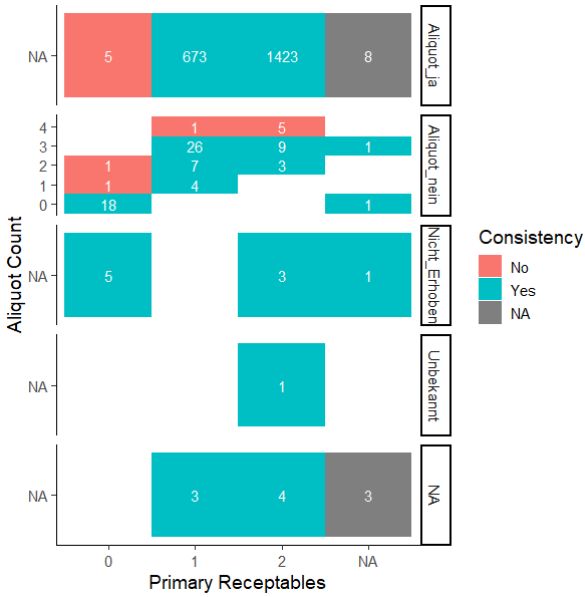

Figure 2: Visualization of the quality indicator consistency for citrate samples. Tiles represent the different states of the item value combinations; the color indicates the result of the consistency evaluation and the number gives the number found within the respective tile. (NA = missing).

3.2. Consistency of the pre-analytic sample states

The pre-analytic states of blood and urine both depend on the visual inspection of the respective liquid sample. The following combinations are inconsistent: All states are set to *no*, normal state and at least one other state are set to *yes*. According to domain expert, combinations of the states *lipemic, icteric and hemolytic* are consistent for blood samples and combinations of *cloudy* and *bloody* are consistent for urine samples.

The paper-based form has three tick-boxes for each property: *yes, no,* and *not ascertained*. Again, missing values are represented by NA in the electronic form. Furthermore, the states other than normal are only shown if normal is not set to yes, impeding the inconsistency that normal state and at least one other state are set to *yes.* Figure 3 shows the visualization for the blood sample properties. Here, combinations of four items are visualized. Only item values are shown that appear at least once in the respective combination. For example, the value *not ascertained (nicht erhoben)* is only found in combinations where all four properties are not ascertained and in one case, where hemolytic state is set to yes and all others to not ascertained. Due to the implemented consistency rule within the electronic form, all other states are set to NA if

state normal is set to yes. Again, we find the possible inconsistency where all states are set to *no* but again in very few cases. Again, for the vast majority of cases, the possibility to create inconsistent combinations of values has been prevented by EDC design: Values for diverging blood properties cannot be set if the property *normal* has been set to *yes*. As the paper-based form explicitly asks for individual assessment of each of the properties including the option to set *not ascertained* for each property, without knowledge of the EDC rules, the NA values for the diverging properties might be interpreted as missing in completeness assessments. Domain experts evaluated it as implausible that the blood-state is not ascertained at all, but we did not consider it here as contradictory, as *not ascertained* is a valid value.

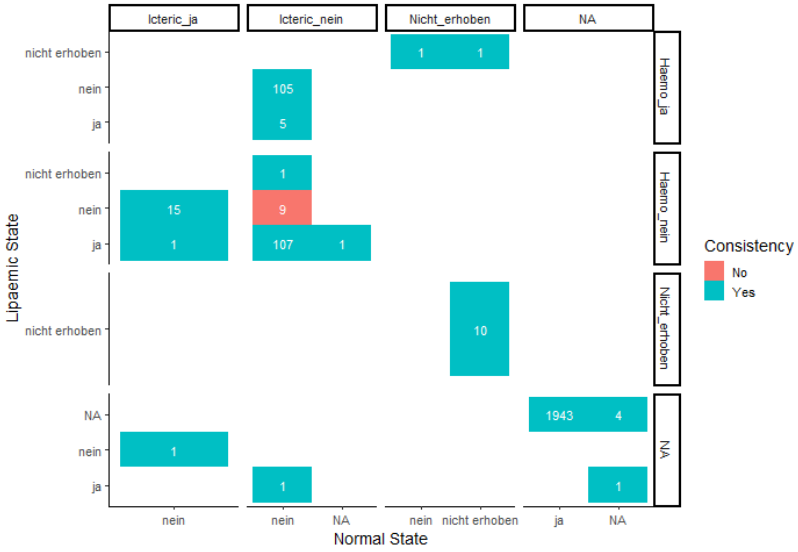

Figure 3: Visualization of the consistency of pre-analytic states of blood samples. Tiles represent the combination of the four different variables, indicated at the four axes. (haemo:hemolytic state, ja: yes, nein:no, nicht erhoben: not ascertained, NA: missing)

3.3. Overall Quality Assessment Report

The quality of the TORCH categorical dataset is assessed using the atemporal consistency metric for all mentioned variables. The overall assessment rating, given in Table 2, shows that 99.61% of the interdependent fields are free of contradictions. The consistency rates for different classes of interdependent categorical features are all in the same order of magnitude and show very few consistency flaws. We would like to mention that here the number of affected fields, i.e. the individual data item values are given rather than the number of inconsistent value combinations. The number of actual datasets that contain inconsistencies and may therefore not be used is obtained by the division of the number of variables involved (4 for blood-states, 3 for others).

Table 2: Data Quality Assessment Report. (AC = Atemporal Consistency)

ConsistencyCheck				
Feature	**Passed-Fields**	**Failed-Fields**	**Plausible-Rate (%)**	
EDTA	6600	18	99.73	
Serum	6612	6	99.91	Logical contradictions
Citrat	6579	39	99.41	
Urine	6567	51	99.23	
Urine-states	6603	15	99.77	Empirical contradictions
Blood-states	8788	36	99.59	
Summary	**41749**	**165**	**99.61**	

4. Discussion

The present work considers multi-directional consistency checks to detect inconsistencies within interdependent data elements in a biomaterial variable set. This approach helps to assess the consistency of interdependent values in three-way relationships (as seen in section 3.1) and four-way relationships (as demonstrated in section 3.2). With this approach we go beyond the item-based consistency checks for example executed in the works of Spengler et al. [11] and Blacketer et al. [12] and that are common in EDC design. Schmidt et al. [3] considered consistency checks involving two data values of the same measurement unit using their *con_contradictions* module in *DataquieR* package, but this is still not sufficient to check for multi-directional consistency issues across multiple interdependent elements as obtained in this study. While the quality indicators defined in this work fit well into the proposed taxonomy by Schmidt et al. [3], in future work, we would explore how the *con_contradictions* module can be adapted to fit specific consistency requirements like we have in our use case. The main contribution of our work is the assessment of multi-item contradictions in a dataset. In the investigated TORCH study, EDC design already prevents main multi-item contradictions by hiding parts of the form if a specific condition is set. However, in the current implementation *NAs* resulting from a conditional rule and *NAs* resulting from missing information can only be distinguished by knowledge of the specific rules. For automatic assessment of item-based quality indicators such as completeness it might be helpful to have a specific value that indicates that the item could not be entered in this case due to EDC rules, e.g. *not shown*. Also, unrestricted fields as observed in the entry of the count of aliquots will result in the entry of arbitrary values that can contradict other predefined dependent values in other elements.

Consistency checks on contradictions have shown to require domain and process knowledge but also knowledge about the EDC implementation. The implemented checks are specific to DZHK biomaterial data. Hence, more collaboration with domain experts is needed to specify convenient rules for other use cases. As we found very few remaining contradictions in the data set, a concrete benefit of the effort for the improvement of the quality of the TORCH data set might be disputable. But we would like to emphasize, that the strength of the DZHK Heart bank lies in the harmonization of processes and data sets. Therefore, the implemented quality indicators are applicable to all DZHK-funded studies.

5. Conclusion

We demonstrate methods and tools for interactive data quality reviews. These are of utmost importance for improving single data sets and especially integrated data sets. Through consistency assessment on interdependent categorical items, data quality flaws could be identified in the DZHK-BBS. Our results underscore the importance of domain knowledge in both the clinical aspects but also in the aspects of information system design for data quality assessment.

Declarations

Ethical approval and consent: Use of the TORCH dataset was approved by DZHK and TORCH.

Conflict of Interest: The authors declare, that there is no conflict of interest.

Author Contributions: This paper is based on the master thesis of KY (Master's Graduate), he conceived the study together with KT, US and DK (Supervisors). KY handled the software implementation and data analysis. KT performed the use case selection and requirement analysis as well as the software design and the supervision of practical implementation. WH contributed expertise in domain related questions. KY, KT and DK formulated the manuscript with contributions from all authors.

Acknowledgement: The work is supported by the German Centre for Cardiovascular research (DZHK). We equally recognize the support of Ms. Otte, the Chief Laboratory Technician at the UMG.

Availability of data and materials: The used data is available within the DZHK Heart bank[4] available for researchers through the common DZHK Use & Access process. The source-code of the plug-in implementation is publicly available in the Gitlab repository of the project[5].

References

[1] M. Nonnemacher, D. Nasseh, J. Stausberg, and U. Bauer, *Datenqualität in der medizinischen Forschung: Leitlinie zum adaptiven Management von Datenqualität in Kohortenstudien und Registern*, 2., Aktualisierte und erw. Aufl. Berlin: Med. Wiss. Verl.- Ges, 2014.

[2] M. G. Kahn *et al.*, "A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data," *EGEMs Gener. Evid. Methods Improve Patient Outcomes*, vol. 4, no. 1, p. 18, Sep. 2016, doi: 10.13063/2327-9214.1244.

[3] C. O. Schmidt *et al.*, "Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R," *BMC Med. Res. Methodol.*, vol. 21, no. 1, p. 63, Dec. 2021, doi: 10.1186/s12874-021-01252-7.

[4] M. Kiehntopf and K. W. Böer, *Biomaterialbanken: Checkliste zur Qualitätssicherung*. Berlin: Medizinisch-Wissenschaftliche Verl.-Ges, 2008.

4 https://dzhk.de/en/dzhk-heart-bank/submitting-applications/

5 Project Repository: https://gitlab.gwdg.de/medinfpub/data-quality-workflow

[5] M. R. La Frano *et al.*, "Impact of post-collection freezing delay on the reliability of serum metabolomics in samples reflecting the California mid-term pregnancy biobank," *Metabolomics*, vol. 14, no. 11, p. 151, Nov. 2018, doi: 10.1007/s11306-018-1450-9.

[6] L. M. Spekhorst *et al.*, "Cohort profile: design and first results of the Dutch IBD Biobank: a prospective, nationwide biobank of patients with inflammatory bowel disease," *BMJ Open*, vol. 7, no. 11, p. e016695, Nov. 2017, doi: 10.1136/bmjopen-2017-016695.

[7] C.-J. L. Farrell and A. C. Carter, "Serum indices: managing assay interference," *Ann. Clin. Biochem. Int. J. Lab. Med.*, vol. 53, no. 5, pp. 527–538, Sep. 2016, doi: 10.1177/0004563216643557.

[8] M. Cornes *et al.*, "European survey on preanalytical sample handling – Part 2: Practices of European laboratories on monitoring and processing haemolytic, icteric and lipemic samples. On behalf of the European Federation of Clinical Chemistry and Laboratory Medicine (EF," *Biochem. Medica*, vol. 29, no. 2, pp. 334–345, Jun. 2019, doi: 10.11613/BM.2019.020705.

[9] T. Schwaneberg *et al.*, "Data privacy management and data quality monitoring in the German Centre for Cardiovascular Research's multicentre TranslatiOnal Registry for CardiomyopatHies (DZHK-TORCH): TORCH data quality management and monitoring," *ESC Heart Fail.*, vol. 4, no. 4, pp. 440–447, Nov. 2017, doi: 10.1002/ehf2.12168.

[10] S. Herzinger *et al.*, "SmartR: an open-source platform for interactive visual analytics for translational research data," *Bioinformatics*, vol. 33, no. 14, pp. 2229–2231, Jul. 2017, doi: 10.1093/bioinformatics/btx137.

[11] H. Spengler, I. Gatz, F. Kohlmayer, K. A. Kuhn, and F. Prasser, "Improving Data Quality in Medical Research: A Monitoring Architecture for Clinical and Translational Data Warehouses," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Rochester, MN, USA, Jul. 2020, pp. 415–420. doi: 10.1109/CBMS49503.2020.00085.

[12] C. Blacketer, F. J. Defalco, P. B. Ryan, and P. R. Rijnbeek, "Increasing Trust in Real-World Evidence Through Evaluation of Observational Data Quality," Health Informatics, preprint, Mar. 2021. doi: 10.1101/2021.03.25.21254341.