German Medical Data Sciences 2023 — Science. Close to People.
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI230707

Impact of Clinical Study Implementation on Data Quality Assessments - Using Contradictions within Interdependent Health Data Items as a Pilot Indicator

Khalid O. YUSUF[a,1], Irina CHAPLINSKAYA-SOBOL[a], Anne SCHONEBERG[a], Sabine HANSS[a,c], Heike VALENTIN[d], Bettina LORENZ-DEPIEREUX[e], Stefan HANSCH[f], Karin FIEDLER[g], Margarete SCHERER[g], Shimita SIKDAR[g], Olga MILJUKOV[h], Jens-Peter REESE[h], Patricia WAGNER[i], Isabel BRÖHL[i], Ramsia GEISLER[g,i], Jörg J. VEHRESCHILD[g,i], Sabine BLASCHKE[j], Carla BELLINGHAUSEN[k], Milena MILOVANOVIC[l], and Dagmar KREFTING[a,b,c]

[a] *Department of Medical Informatics, University Medical Center Göttingen, Germany*
[b] *Campus Institute Data Science, Georg-August-University, Göttingen, Germany*
[c] *German Center for Cardiovascular Research, Partner Site Göttingen, Germany*
[d] *Trusted Third Party of the University Medicine Greifswald, Germany*
[e] *Institute of Epidemiology, Helmholtz Zentrum München, Munich, Germany*
[f] *Department for Infectious Diseases and Infection Control, University Hospital Regensburg, Germany*
[g] *Department II of Internal Medicine, Hematology/Oncology, Goethe University, Frankfurt, Frankfurt am Main, Germany*
[h] *University of Würzburg, Institute for Clinical Epidemiology and Biometry*
[i] *Department I for Internal Medicine, Faculty of Medicine and University Hospital of Cologne, University of Cologne, Cologne, Germany*
[j] *Emergency Department, University Medical Center Goettingen, Germany*
[k] *Goethe University Frankfurt, University Hospital Frankfurt, Medical Clinic I, Department of Respiratory Medicine / Allergology*
[l] *Malteser Krankenhaus St. Franziskus Hospital, Medical Clinic I, Flensburg, Germany*

Abstract: Introduction: Contradiction is a relevant data quality indicator to evaluate the plausibility of interdependent health data items. However, while contradiction assessment is achieved using domain-established contradictory dependencies, recent studies have shown the necessity for additional requirements to reach conclusive contradiction findings. For example, the oral or rectal methods used in measuring the body temperature will influence the thresholds of fever definition. The availability of this required information as explicit data items must be guaranteed during study design. In this work, we investigate the impact of activities related to study database implementation on contradiction assessment from two perspectives including: 1) additionally required metadata and 2) implementation of checks within electronic case report forms to prevent contradictory data entries. **Methods:** Relevant information (timestamps, measurement methods, units, and interdependency rules) required for contradiction

[1] Department of Medical Informatics, University Medical Center Göttingen, Germany; E-mail: olusolakhalid.yusuf@med.uni-goettingen.de.

checks are identified. Scores are assigned to these parameters and two different studies are evaluated based on the fulfillment of the requirements by two selected interdependent data item sets. **Results:** None of the studies have fulfilled all requirements. While timestamps and measurement units are found, missing information about measurement methods may impede conclusive contradiction assessment. Implemented checks are only found if data are directly entered. **Discussion:** Conclusive contradiction assessment typically requires metadata in the context of captured data items. Consideration during study design and implementation of data capture systems may support better data quality in studies and could be further adopted in primary health information systems to enhance clinical anamnestic documentation.

Keywords. Data quality, electronic data capture, metadata definition, contradictions

1. Introduction

In health research, a comprehensive and documented data quality assessment increases not only the reliability of the data but also the credibility of the research conclusions drawn from the analysis of such data [1–3]. Contradiction is a key data quality indicator that examines implausible value-combinations in a multi-item relationship [4,5]. Assessment is usually initialized with the identification of interdependent data items within a dataset, where contradictory dependencies are defined by established domain-specific rules [6]. However, the presence of a contradiction may rely on context-specific information, that can be considered as metadata to the data items. As an illustration, while body temperature thresholds are directly indicative of the presence or absence of fever, validating this comparison will rely on 1) the reference time to ensure both measurements are captured at the same timepoint and 2) temperature measurement methods, as they have different thresholds for fever. The configuration of an Electronic Data Capture (EDC) system plays a key role in this regard as the additional information required for this validation step must be defined during study design and populated with values, as it might be impossible to assess them retrospectively.

A widely used platform for electronic clinical study data capture in Germany is secuTrial® (interActive Systems GmbH, Berlin, Germany) [7]. While there is a central data management that implements relevant schemas for individual studies, the content of the electronic case report form (eCRF) and data monitoring are handled by the domain experts from the recruiting study centres. Challenges of unavailability of information required to answer important clinical questions have been reported in literature [1,8]. Recent reports on contradiction assessment on different data sets also showed limits in reaching conclusive contradiction findings due to missing items [9,10]. This paper therefore aims to evaluate the influence of the EDC system configuration on the assessment of contradictions within health datasets. A focus is on additional information (data or metadata items) required for contradiction assessment and the integration of interdependency rules within eCRFs. The investigation is targeted at strengthening metadata collection during study design and ensuring preemptive quality control during data entry.

2. Methods

2.1. Electronic Data Capture System and Items Definition

The EDC system secuTrial® is compliant with the requirements of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH E6 (R2) Good Clincial Practise) and the requirements of the Food and Drug Administration (FDA 21 CRF Part 11). As a rule, the implementation of a study database is realized according to an established quality management system based on Standard Operating Procedures (SOP), which also regulates the use of edit checks, i.e. checks that are performed during data entry. Databases are technically implemented based on the requirements of the study protocol, list(s) of study-specific items, as well as communication with the study coordination and SOPs. Items are organized in the aforementioned eCRFs and are then assigned to visits. The eCRFs include, but are not limited to, checks for completeness, plausibility like ranges, cross-checks for valid data entry, and for adherence to the visit schedule. All edit checks can be implemented as "soft" or "hard" checks. If the rules are hard, it is not possible to save the complete form. Soft checks can be ignored from a study team and the data can be saved, but in this case, a rule violation is registered in the validation protocol. Though hard rules save monitoring resources during the study and may guarantee higher data quality, it could also induce entry of plausible but invalid data. Soft rules are used for example for laboratory values, where outliers are to be expected. Hard and soft rules can be combined and applied according to different usage scenarios. Some checks are not supported by secuTrial®, for example checks between visit forms and repetition groups forms. A repetition group is a type of record in the form of a list that reoccur for each entry. A finalized eCRF undergoes a standardized testing procedure before transfer to the production environment.

2.2. Employed Data Collections

The National Pandemic Cohorts Network (NAPKON) and the German Centre for Cardiovascular Research (DZHK) use secuTrial® to capture and store clinical data of different studies related to the COVID-19 pandemic and cardiovascular diseases respectively [7,11]. From these studies we selected NAPKON's cross-sectoral Platform (SÜP) and DZHK's Home-Based Screening for Early Detection of Atrial Fibrillation in Primary Care Patients Aged 75 Years and Older (SCREEN-AF, in this manuscript referred as SAF). While the SÜP is directly captured, SAF has been documented in an external system and has been later imported to secuTrial®. We identified interdependent items from the publicly available dataset tables and database schema, i.e. (I1) diastolic blood pressure (DP), systolic blood pressure (SP), and indication of hypertension (HT); and (I2) diagnosis of Diabetes Mellitus (DM) and indication of insulin medication (INS). Regarding (I1) there are two interdependencies: DP is always lower at SP measured at the same time, so the time of measurement is required as additional information as well as the measurement unit. Furthermore, certain levels of SP and DP are indicators for HT. However, these thresholds depend on the applied regional rules (American Heart Association or European Society of Hypertension) and on the measurement context, i.e. in the clinic, at home, or ambulatory [12]. In (I2) insulin medication is only plausible in presence of diabetes mellitus. Onset of the diabetes mellitus and time of the insulin medication are relevant for this assessment.

The identified relevant metadata for contradiction assessment (rm_ca) can be grouped into timestamps (rm_ts), measurement methods (rm_mm) and measurement units (rm_mu). The rules used for the determination of hypertension are considered as MM here.

2.3. Evaluation

For evaluation of the configuration of the EDC system, we introduce a holistic level for rm_ca: depending on the degree of presence of the respective metadata, compliance is rated from 0 to 3 based on gaps, i.e. the number of questions that cannot be sufficiently answered to assess the contradictions conclusively. The different levels indicate: 0: not present, 1: present with 2 gaps, 2: present with one gap, 3: present without gaps. A scenario for level 1 is the case where a timestamp answers the question about the date of the data capture without any reference to the onset of a disease or diagnosis timepoint. To evaluate the EDC system support to assess contradictions, the implemented checks for the contradiction assessment (ic_ca) are examined. The check may consist of one implemented rule or a rule-set. Again, we define a scale that ranges from 0 to 2, depending on the completeness of the implemented rules. The different levels indicate 0: not established, 1: partially established, 2: fully established. A certain check for a scenario i (ic_ca_i) is considered partial if not all interdependencies are included. For each scenario i, a total maximum score (ca_tms(i)) is determined by the sum of 3 points for each required metadata type in the respective scenario (3-9 points) and always 2 points for ic_ca. For each study s, the total score (ca_ts(i,s)) is determined by the sum of the individual scores for the different metadata and the implemented check.

Table 1. Elicitation of required metadata (rm_ca) and implemented checks (ic_ca) for conclusive contradiction assessment. rm_ts=timestamps, rm_mm=measurement methods, rm_mu=measurement units.

Additional Requirements	Parameters	Scale
Required Metadata (rm_ca)	rm_ts,rm_mm,rm_mu	0-3
Implemented check within eCRF (ic_ca)	ic_ca	0-2

3. Results

3.1. Blood Pressure and Hypertension Assessment

The three blood pressure items DP, SP, and HT are all related to a measurement method. SP and DP have measurement units. All three items have two interdependency rules that depend on the timepoints. Therefore, the total maximum score is 11: ca_tms(I1) = 3+3+3+2 = 11. The total scores of SÜP and SAF are given in Table 2, together with the available metadata items that contain the required information. SÜP is rated 7, SAF has a total score of 8. Both studies captured required timestamps (visit_date, examination_date, & event_start_date for SÜP, monitor_date, home_time, & visit_label for SAF) for a conclusive contradiction assessment. However, the presence of "event_start_date" in SÜP would validate the comparison of hypertension onset against the DP and SP examination date while its absence in SAF hinders similar comparison, therefore rm_ts(I1,SAF) is level 2. The methods for DP and SP measurements and regional HT rules are missing for SÜP, while explicitly captured in SAF. As a consequence, contradictory findings of blood pressure measurements with respect to HT indication are inconclusive. Both studies embedded measurement units in the respective fields for DP and SP data entry, resulting in full score in rm_mu. SÜP has implemented a check between DP and SP but no check for the relation to HT, resulting in ic_ca(I1,SÜP) = 1. SAF has no checks implemented at all on this topic, resulting in ic_ca(I1,SAF) =0. We observe that both studies, although with similar total scores, have different flaws in the implementation that make contradiction assessment difficult.

Table 2. Grading of required metadata rm_ca and implemented checks ic_ca of contradiction assessment of blood pressure (DP, SP) and hypertension (HT) on the two studies s: SÜP and SAF. ca_tms = total maximum score, ca_ts = total score.

Study s	metadata items rm_ts	rm_ts (I1,s)	metadata items rm_mm	rm_mm (I1,s)	rm_mu (I1,s)	ic_ca (I1,s)	ca_ts(I1,s) /ca_tms(I1)
SÜP	*Visit_date,* *Exam_date,* *Event_start*	3	none	0	3	1	7/11
SAF	*Monitor_date,* *home_time,* *visit_label*	2	home_evening, home_morning	3	3	0	8/11

3.2. Diabetes and Insulin Medication Assessment

The two interdependent items DM and INS are linked by one interdependency rule, that requires the timepoint of insulin medication is not earlier than the assessment time of diabetes mellitus. As measurement methods and units are not relevant here, the total maximum score is 5: ca_tms(I2) = 3+2 = 5. From Table 3, it can be inferred that SÜP and SAF captured the necessary timestamps to ensure insulin medication did not precede diabetes mellitus diagnosis, but in different ways: SÜP documents explicitly the examination, event and medication time, while SAF documents the timestamp of the visit and asks explicitly for current medication. Therefore, both studies reached level 3 in rm_ts. In SÜP, insulin medication can only be selected if diabetes mellitus is confirmed. However, there is no implemented check between diabetes mellitus and insulin medication catalogue which makes it vulnerable to contradictory entries. Therefore, we rated ic_ca(I2,SÜP) =1. In SAF, no implemented check is found. Accordingly, ca_ts(I2,SÜP) = 4, so SÜP fulfilled 80% of the requirements, while ca_ts(I2,SAF) = 3, so SAF fulfilled 60% of the requirements.

Table 3. Grading of additionally required metadata and interdependency rules in the assessment of diabetes mellitus (DM) and Insulin (INS) medication. % = not relevant

Study s	metadata items rm_ts	rm_ts (I2,s)	rm_mm (I2,s)	rm_mu (I2,s)	ic_ca (I2,s)	ca_ts(I2,s) /ca_tms(I2)
SÜP	*Visit_date, Exam_date, Event_start, Med_start*	3	%	%	1	4/5
SAF	*Visit_label, current_med*	3	%	%	0	3/5

4. Discussion

Our results show that information required for conclusive contradiction assessment is not fully documented as (meta-)data items in the investigated studies. Some information such as the blood pressure measurement method is typically defined in the SOPs and might even be displayed as informative text in the form, but are not available as structured data. While implementation of checks during setup of the study database is still possible, later automatic contradiction assessment is not possible. Availability of the discussed metadata will be in particular crucial for contradiction assessments of anamnestic questions in routine clinical data [8], as no common SOPs are implemented in health care. While missing values within any of the interdependent data items will hinder contradiction assessment as noted by Schmidt et al. [5], the focus in this work is on the completeness of additional information required to validate suspected contradictions. We scored the fulfillment level of the different requirements however, the individual levels and ca_ts should be considered as qualitative gradings rather than quantitative measures - while full assessment of the described interdependencies is only supported if all requirements are fulfilled (ca_ts(i,s) == ca_tms(i)), one cannot deduce from ca_ts(i,s) < ca_tms(i), if one will be able to conduct a contradiction assessment with information e.g. from the SOPs or further non-structured study documentation, or if the metadata is complete but no checks are implemented. However, we think ca_ts is in particular useful during implementation of the study database. We observed the study protocol as an important factor for the collection of required items – for example SAF captured the home location and the daytime of the blood pressure measurement, because it is designed as a home monitoring study. To prevent entry of contradictory values as reported for different studies including SÜP [9,10], it is encouraged to establish and enforce contradiction checks within the eCRFs. However, it should be noted that the user experience during data entry suggests a cautious approach when enforcing complex interdependency rules to avoid difficulties in data capture. Capabilities of different EDC systems can help mitigate this effect as we observed in secuTrial® where rules can be enforced as either soft or hard rules during data entry. Study endpoints may help in estimating the relevance of a certain contradiction and decision on the level of enforcement of targeted rules during data capture. As an illustration, the current EDC design does not support the enforcement of contradiction checks between an anamnesis form and the medication catalogue. If a diabetes definition is of greater relevance to a study, it would be encouraged to establish an interdependency rule between diabetes mellitus diagnosis and the varieties of insulin medications to prevent contradictory information.

5. Conclusion

Required metadata may impede conclusive assessment of contradictions in health data sets. An initial analysis of the study infrastructure is highly recommended at the onset of data collection to early address potential gaps. Our analysis points to the need for improved collaboration between principal investigators, data managers, and biometricians involved in the design of clinical studies. The findings in this work will be considered during the schema implementation of future studies to ensure a robust definition of items.

Acknowledgement: The study was carried out using the clinical-scientific infrastructure of NUKLEUS of the Network University Medicine (NUM 2.0, grant number 01KX2121), and the German Center for cardiovascular research (DZHK, grant number 81Z0300108) both funded by the Federal Ministry of Education and Research (BMBF)

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

Ethical approval and consent: NAPKON and DZHK approved the use of the publicly available dataset tables.

Abbreviations: EDC - electronic data capture, eCRF - electronic case report form, SOP - Standard Operating Procedures, NAPKON - National Pandemic Cohorts Network, DZHK - German Centre for Cardiovascular Research, SÜP – Cross-sectoral platform, SAF and SCREEN-AF - Home-Based Screening for Early Detection of Atrial Fibrillation in Primary Care Patients Aged 75 Years and Older, DP - Diastolic blood pressure, SP - Systolic blood pressure, HT - Hypertension, DM - diabetes mellitus, INS – insulin medication.

References

[1] Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. JNCI: Journal of the National Cancer Institute. 2017 Nov 1;109(11). doi:10.1093/jnci/djx187/4157738

[2] Lockery JE, Collyer TA, Reid CM, Ernst ME, Gilbertson D, Hay N, et al. Overcoming challenges to data quality in the ASPREE clinical trial. Trials. 2019 Dec;20(1):686. doi:10.1186/s13063-019-3789-2

[3] Miksad RA, Abernethy AP. Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. Clin Pharmacol Ther. 2018 Feb;103(2):202–5.

[4] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs. 2016 Sep 11;4(1):18. doi:10.13063/2327-9214.1244

[5] Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, 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. 2021 Dec;21(1):63.

[6] Yusuf KO, Hanss S, Krefting D. Towards a Consistent Representation of Contradictions Within Health Data for Efficient Implementation of Data Quality Assessments. In: Hägglund M, Blusi M, Bonacina S, Nilsson L, Cort Madsen I, Pelayo S, et al., editors. Studies in Health Technology and Informatics [Internet]. IOS Press; 2023. doi: 10.3233/SHTI230123

[7] Schwaneberg T, Weitmann K, Dösch A, Seyler C, Bahls T, Geidel L, 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 Failure. 2017 Nov;4(4):440–7

[8] McGuckin T, Crick K, Myroniuk TW, Setchell B, Yeung RO, Campbell-Scherer D. Understanding challenges of using routinely collected health data to address clinical care gaps: a case study in Alberta, Canada. BMJ Open Qual. 2022 Jan;11(1):e001491.

[9] Yusuf KO, Miljukov O, Hanß S, Schoneberg A, Wiesenfeldt M, Stecher M, et al. Consistency as a Data Quality Measure for German Corona Consensus items mapped from National Pandemic Cohort Network data collections. Methods Inf Med. 2023 Jan 3;a-2006-1086.

[10] Yusuf K, Tahar K, Sax U, Hoffmann W, Krefting D. Assessment of the Consistency of Categorical Features Within the DZHK Biobanking Basic Set. In: Röhrig R, Grabe N, Hoffmann VS, Hübner U, König J, Sax U, et al., editors. Studies in Health Technology and Informatics [Internet]. IOS Press; 2022. doi: 10.3233/SHTI220809.

[11] Schons M, Pilgram L, Reese JP, Stecher M, Anton G, Appel KS, et al. The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. Eur J Epidemiol. 2022 Aug;37(8):849–70

[12] de la Sierra A. New American and European Hypertension Guidelines, Reconciling the Differences. Cardiol Ther. 2019 Dec;8(2):157–66