Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230123

Towards a Consistent Representation of Contradictions Within Health Data for Efficient Implementation of Data Quality Assessments

Khalid O. YUSUF[a], Sabine HANSS[a,c] 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] *German Center for Cardiovascular Research, Partner Site Göttingen, Germany*
ORCiD ID: Khalid O. Yusuf https://orcid.org/0000-0003-2556-6898

Abstract. Contradictions as a data quality indicator are typically understood as impossible combinations of values in interdependent data items. While the handling of a single dependency between two data items is well established, for more complex interdependencies, there is not yet a common notation or structured evaluation method established to our knowledge. For the definition of such contradictions, specific biomedical domain knowledge is required, while informatics domain knowledge is responsible for the efficient implementation in assessment tools. We propose a notation of contradiction patterns that reflects the provided and required information by the different domains. We consider three parameters (α, β, θ): the number of interdependent items as α, the number of contradictory dependencies defined by domain experts as β, and the minimal number of required Boolean rules to assess these contradictions as θ. Inspection of the contradiction patterns in existing R packages for data quality assessments shows that all six examined packages implement the (2,1,1) class. We investigate more complex contradiction patterns in the biobank and COVID-19 domains showing that the minimum number of Boolean rules might be significantly lower than the number of described contradictions. While there might be a different number of contradictions formulated by the domain experts, we are confident that such a notation and structured analysis of the contradiction patterns helps to handle the complexity of multidimensional interdependencies within health data sets. A structured classification of contradiction checks will allow scoping of different contradiction patterns across multiple domains and effectively support the implementation of a generalized contradiction assessment framework.

Keywords. Data quality, health data, boolean minimization, metadata

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

1. Introduction

Contradictions in data quality (DQ) assessments are typically understood as (nearly) impossible combinations of data values of interdependent data items within a data set [1]. In health data, a typical example is that diastolic blood pressure (DP) must not be higher than systolic blood pressure (SP). Handling of a single interdependency between two data items is well established and implemented in most DQ assessment frameworks, as we show below. There are several taxonomies used in literature to describe the semantic nature (e.g. logical or empirical) of contradictions [2–5]. This is useful for the definition of contradictions on the biomedical domain expert level, while the implementation within a DQ assessment tool is typically realized by Boolean rules that are somehow agnostic to the semantic nature of a contradiction. For example, a conditional expression ($x > y$) might be related (a) to DP and BP as well as (b) age at two different time points. While these examples might be assigned to different semantic groups, for example (a) as atemporal and (b) as temporal plausibility according to [4], from the implementation point of view, a clear and simple implementation of contradiction rules is desired [6]. Therefore, we propose a structural representation of contradiction patterns that is useful to simplify increasing complexity of multidimensional interdependencies, as it describes the dimensionality of the contradictory dependency and the minimum of derived Boolean rules. For the aforementioned blood pressure case, there is only one contradictory dependency between the two data items. Another pair of data items however may have more interdependencies. An example is the notation of fever and the body temperature: We have to distinguish the two cases where the body temperature is below or above a certain temperature, where one case requires the fever_item set to *yes*, while the other requires the fever_item set to *no* or at least *not set*. This would result in two Boolean rules in the implementation.

2. Methods

We consider three parameters (α, β, θ) for the proposed structural representation of contradictions i.e. the number of: 1) interdependent data items α, 2) contradictory dependencies β, and 3) minimal Boolean rules θ. While β represents the number of distinct contradictions defined by biomedical domain experts, θ is derived by grouping multiple similar contradictions using all plausible common denominators (CD). A CD could either be a conditional expression, an item, a value or their combination. As a rule, any defined rule that evaluates to multiple numerical or categorical values is set to its value-range or value-set respectively—provided it represents a distinct contradiction. Also, each minimal Boolean rule within θ must be bounded unambiguously such that it is independent of other rules. We examined six R-packages that support contradiction assessment *(assertive, dataquierR, DQAstats, pointblank, testdat, and validate)* on what Boolean rules are implemented in these packages [7]. Based on recent quality assessments on biobank and COVID-19 data [8,9], we defined different classes of contradiction patterns on the sets of interdependent data items.

3. Results

3.1. Contradiction pattern implemented in R-packages

As shown in table 1, all R-packages implemented contradiction checks on $\alpha = 2$ with $\beta = 1$ which translates to $\theta = 1$ in each case. These scenarios represent the simplest form of contradiction patterns.

Table 1. Contradiction pattern in existing R-packages (c.f. [7] for information on R-packages). α = number of interdependent items, β = number of contradictory dependencies, θ = number of Boolean rules

R-Package	Notation (α, β, θ)	α	β	θ
assertive	*(2,1,1)*	x,y	Is_less_than(x,y)	1
dqastats	*(2,1,1)*	bank_balance(bb), credit_worthiness(cw)	is_negative(bb) & cw == yes	1
pointblank	*(2,1,1)*	x,y	col_val_lt (vars(x), y)	1
testdat	*(2,1,1)*	x,y	expect_cond(x, y.length>=1)	1
validate	*(2,1,1)*	staff, staff_cost	validator(if (staff >=1) staff_cost >= 1)	1
con_contradictions	*(2,1,1)*	Age_followup, Age_baseline	A_less_than_B(Age_1, Age_0)	1

3.2. Contradiction pattern in the biobank domain

Three interdependent data items used in storing information about the collection of citrate samples were investigated, i.e. primary receptables (pr), desired aliquots filled (af), and actual aliquots count (ac) [8]. From table 2, a contradiction pattern was established where $\alpha = 3$, $\beta = 5$, and $\theta = 3$. Transforming β to θ resulted in the reduction of 4 distinct contradictory rules to 2 Boolean rules using their respective CD (pr==0) and (pr > 0 & af == no & ac ==). Another contradiction pattern refers to contradictions in items related to pre-analytic states of blood samples, please refer to [8] for further information.

Table 2. Contradiction pattern between biosample associated data items

α	β	θ
Number of primary receptable (pr)	pr == 0 & af == yes	pr == 0 & isTrue(af == yes | ac > 0)
All aliquots filled (af)	pr == 0 & ac > 0	pr > 0 & af == no & ac in (0,4)
Aliquot count (ac)	pr > 0 & af == no & ac == 0 pr > 0 & af == yes & ac < 4 pr > 0 & af == no & ac == 4	pr > 0 & af == yes & ac < 4

3.3. Contradiction pattern in the COVID-19 domain

We considered the consistency of different groups of interdependent data items that were mapped from different cohorts to the German Corona Consensus (GECCO) dataset [9]. Table 3 shows the aforementioned contradiction pattern between fever and body temperature (T_b) items: a normal T_b ($35 \leq T_b < 38.3$) should not evaluate to fever_item set to *yes* and an elevated T_b ($38.3 \leq T_b \leq 45$) should not evaluate to fever_item set to *no* or *not set*. There are no plausible CD in this case, hence, β equals θ and is assigned to

class (2,2,2). The T_b ranges and fever_item being compared to the elevated T_b evaluate to multiple values pointing to distinct contradictions.

Table 3. Contradiction pattern between body_temperature (T_b) and fever

α	β	θ
fever body_temperature (T_b)	$35 \leq T_b < 38.3$ & fever == yes $38.3 \leq T_b \leq 45$ & fever in (no, notset)	$T_b >= 35$ & $T_b < 38.3$ & fever == yes $T_b >= 38.3$ & $T_b <= 45$ & fever in (no, notset)

A more complicated example is the *(10,10,2)* class for a comparison on the presence of pulmonary disease (PD) anamnesis with its nine documented indicators. PD is a dependent variable on its set of indicators such that when a study participant answers the question about the presence of PD affirmative, this has to be followed-up with at least a specific disease that belongs to the PD family. As presented in table 4, β is derived from the combination of values of PD and their implausible indicators. By exploiting the CD, β is reduced to θ = 2. While β preserve the atomicity of the rules provided by domain experts, θ is implemented in the assessment tool by grouping multiple similar β within θ. Further examples are the comparison of COVID-19 severity and 19 severity indicators, resulting in a (20,38,4) pattern (c.f. [9] for further information).

Table 4. Contradiction pattern between branch question of pulmonary disease anamnesis and documented indicators.

α	β	θ
Chronic Lung Disease (pd)?	pd == no & asthma == yes	pd == no & any_of(asthma, copd, fibr, ph, ohs, apn, osas, cf, others) == yes
Asthma	pd == no & copd == yes	pd == yes & all_of(asthma, copd, fibr, ph, ohs, apn, osas, cf, others) == no
Chronic obstructive pulmonary disease (copd)	pd == no & fibr == yes	
Lung fibrosis (fibr)	pd == no & ph == yes	
Pulmonary hypertension (ph)	pd == no & ohs == yes	
Obesity hypoventilation syndrome (ohs)	pd == no & apn == yes	
Sleep apnoea (apn)	pd == no & osas == yes	
Obstructuve sleep apnoea (osas)	pd == no & cf == yes	
Cystic fibrosis (cf)	pd == no & others == yes	
Other lung disease (others)	pd == yes & all_pd_indicators == no	

4. Discussion

We demonstrate an efficient way of representing the dimensionality of contradictions where the underlying structure of multiple rule combinations is described. Our results indicate θ may be significantly lower than β in multidimensional interdependencies. While the preservation of β will aid the explanation and traceability of identified contradictions, θ ensures an efficient implementation of Boolean rules within DQ

assessment tools—in particular for large datasets. Though different contradiction indicators are useful for the semantic description of contradictions, a structural classification simplifies the varying dimensionality of contradictions for ease of evaluation. A holistic approach to contradiction assessment is a step towards building a DQ assessment tool that would be applicable across multiple domains. Uniform representation of contradiction patterns will ensure a harmonized way of comparing contradiction patterns when considering the fitness of existing assessment tools for internal use. Conclusively, a structured classification of contradiction checks will support the implementation of a generalized contradiction assessment framework effectively and may help researchers to identify Boolean rules in a structured way. While the evaluations in the described cases are performed manually, we envision a tool that helps to translate the domain specific definition of contradictory rules to a normalized form of the informatics domain.

Acknowledgement: The work is jointly funded by NUM, grant number 01KX2121 and DZHK, grant number 81Z0300108

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

Ethical approval and consent: DZHK and TORCH approved the use of the TORCH dataset. Also, use of the data from the cohorts was approved by the NAPKON use and access committee.

References

[1] De Marneffe MC, Rafferty AN, Manning CD. Finding contradictions in text. InProceedings of acl-08: Hlt 2008 Jun (pp. 1039-1047).

[2] 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. doi: 10.1186/s12874-021-01252-7

[3] Johnson SG, Pruinelli L, Hoff A, Kumar V, Simon GJ, Steinbach M, et al. A Framework for Visualizing Data Quality for Predictive Models and Clinical Quality Measures. AMIA Jt Summits Transl Sci Proc. 2019;2019:630–8

[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] Nonnemacher M, Nasseh D, Stausberg J, Bauer U. 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. 230 p. (Schriftenreihe der TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V).

[6] Duşa A. A mathematical approach to the boolean minimization problem. Qual Quant. 2010 Jan;44(1):99–113. doi:10.1007/s11135-008-9183-x

[7] Mariño J, Kasbohm E, Struckmann S, Kapsner LA, Schmidt CO. R Packages for Data Quality Assessments and Data Monitoring: A Software Scoping Review with Recommendations for Future Developments. Applied Sciences. 2022 Apr 22;12(9):4238. doi:10.3390/app12094238

[8] 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 [cited 2022 Sep 10]. Available from: https://ebooks.iospress.nl/doi/10.3233/SHTI220809

[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. doi:10.1055/a-2006-1086