ICT for Health Science Research
A. Shabo (Shvo) et al. (Eds.)

doi:10.3233/978-1-61499-959-1-80

Clinical Information Model Based Data Quality Checks:Theory and Example

Erik TUTE[a,1], Antje WULFF[a], Michael MARSCHOLLEK[a], Matthias GIETZELT[a]
[a] *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School*

Abstract. Introduction: We describe principles of leveraging clinical information models (CIMs) for data quality (DQ) checks and present the exemplary application of these principles. Methods: openEHR compliant CIMs are used to express DQ-checks as constraints. Test setting is the process of extracting, transforming and loading (ETL) assisted ventilation data from two patient data management systems (PDMS) of a pediatric intensive care unit into a local openEHR-based data repository. Results: A generic component logs aggregated DQ-check results for ~28 million entries. DQ-issue types in the presented results are range-, format- and value set violations. Discussion: CIMs are suitable means to define DQ-checks for range-, format-, value set and cardinality constraints. However, they cannot constitute a complete solution for standardized DQ-assessment.

Keywords. Data quality, quality assessment, reuse, interoperability

1. Introduction

Reusing electronic patient data originating from the routine care processes for medical research is an important research topic in medical informatics. A challenge in reusing data originating from different source systems or even organizations is semantic interoperability. One approach to tackle this challenge is the integration of data into so-called clinical information model (CIM) based data repositories sharing common CIMs [1]. A CIM is a shared implementable definition of the clinical concepts represented by the data. The medical data integration centers (MeDIC) of the HiGHmed [2] consortium employ this approach to reach semantic interoperability in multi-centric data reuse scenarios across eight German university hospitals.

Data quality (DQ) is the suitability of a dataset for a given task (cf. [3]). Therefore, before reusing data for another task, assessing its DQ is advisable. Literature proposes plenty of different DQ measurement methods (MM) for DQ-assessment (e.g. [3]). Many of the already applied MMs deal with checking the data for range-, format- or value set violations, missing mandatory values or other cardinality constraints. CIMs complying with the standard openEHR [4] can express such constraints, which suggests leveraging them for this kind of DQ-checks.

The objective of this contribution is to present the principles of leveraging openEHR-CIMs for DQ-checks (precisely: range-, format-, value set and cardinality

[1] Corresponding Author: Erik Tute, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany; E-mail: erik.tute@plri.de.

checks). The presented test of these principles lays the foundation for more comprehensive model-based DQ-assessment procedures in a local MeDIC.

2. Methods

2.1. Theory

In openEHR, the CIMs are called archetypes and templates (cf. [4]). Archetypes are reusable models defining the clinical data items for a clinical concept, e.g. a blood pressure and its associated values like the cuff size, without any specific use case in mind. Templates are definitions of content for a specific use case. Templates can be thought of as definitions of specific documents or messages. Defining a Template usually comprises combining Archetypes, further constraining them as well as defining terminology bindings. The modeler defines the constraints on data fields while defining the CIM, usually using a modeling tool. The resulting model is expressed in Archetype Definition Language (ADL). Table 1 lists examples for cardinality-, range-, format- and value set constraints in ADL 1.4.

Table 1. Examples for cardinality-, range-, format- and value set constraints in ADL 1.4.

Cardinality	**Range**	**Format**	**Value set**
<existence> … <lower>0</lower> <upper>1</upper> </existence>	<magnitude … <lower>0</lower> <upper>1000</upper> </magnitude>	<pattern> openEHR-EHR-CLUSTER\.device(-[a-zA-Z0-9_]+)*\.v1 </pattern>	<terminology_id> <value>local</value> </terminology_id> <code_list>at007</code_list> … <code_list>at0.12</code_list>

2.2. Example Setting

The test setting for CIM-based DQ-checks is the process of extracting, transforming and loading (ETL) assisted ventilation data from two different patient data management systems (PDMS) of a pediatric intensive care unit into a local openEHR-based data repository. We describe DQ analysis results for the data from one PDMS system to give an impression about the capabilities and limitations of the presented method. The PDMS stores the data in an entity-attribute-value database model distinguishing only numeric and string values without setting further constraints for the values. The openEHR-based data repository checks if the incoming data complies with the constraints defined in the CIM.

3. Results

3.1. Implementation Result

An ETL-process implemented using SQL Server Integration Services (SSIS) extracts the data from the relational source tables, transforms it into openEHR compositions (i.e. an openEHR compliant serialized object) and sends the compositions as HTTP-requests to the standardized openEHR REST-API of our local openEHR repository (Think!EHR Platform). More details on the ETL-process implementation are described in [5].

The openEHR repository checks if the composition complies with the CIM-definition and sends either a success or an error code as response. The "bad request" response means problematic content. One of the possible content problems is "invalid composition data", which means that the data sent does not comply with the specified CIM definition. A generic C#-component implemented within SSIS parses these bad responses and logs information aggregated per constraint violation type over one ETL-process execution. This kind of information consists of an issue description, the number of issue occurrences and a list of problematic entries. Eq. (1) shows an example log-entry.

73 x Invalid value at /content[...]/items[at0031,'Inspired tidal volume']/value,
expected: M:[0.0..2000.0], P:[1..1], U:ml, actual:aValue
Extra Infos: actual: quantity M:11337.0, P:1, U:ml --- [...]

(1)

Based on the logged information we noticed and resolved a mapping error and iteratively implemented data cleansing. For systematic and definitely identifiable badly formatted values in the source data, data cleansing derives correctly formatted values. It simply removes remaining problematic entries if these make up less than 1% of the entries in one composition and additionally stores problematic values in an error table.

3.2. Observations for Example Dataset

The ETL-process took a running time of 72 minutes for processing 12 141 784 records from the source table, creating 9 850 contributions consisting of 28 632 794 entries. The stated time includes extracting data from source tables, transforming it into contributions, sending the contributions, performing the DQ-checks, processing the responses and performing data cleansing (for 354 contributions 1147 problematic entries were removed). Table 2 lists the counts and types of problematic entries removed.

Table 2. Counts and types of remaining DQ-issues leading to removed entries in last data cleansing step.

DQ-issue description	Count
Text value instead of numeric value	25
Unexpected unit	2
Unexpected value (probably systematic and resolvable, but not without medical expert knowledge, which was not available)	209
Value far out of range (implausible value)	908
Value marginally out of range (probably value constraint to restrictive)	3

The first iterations of the ETL-job revealed that the number of problematic entries in a composition is the most influential factor regarding the time needed to process a composition. Composition processing and response generation time on the openEHR data repository increased significantly with the number of problematic entries.

4. Discussion

Our main finding is that CIMs are suitable means to define DQ-checks for range-, format-, value set and cardinality constraints. CIM-based data repositories can realize the constraint-application on electronic patient data. Information generated about problematic entries proved to be useful to notice mapping errors in the ETL-process as well as to implement simple data cleansing.

Plausibility checks often depend on values of other attributes for the same patient, e.g. a male patient should not be pregnant or a death date should not precede a treatment begin. OpenEHR-CIMs cannot express these kind of checks. Anani et al. [6] describe these kind of DQ-checks using openEHR Guideline Definition Language (GDL). GDL is a rule language, which closely integrates openEHR-CIMs. That way also more advanced DQ-checks are possible with the limitation, that each rule can just check constraints within one patient's data. Thus, this method still cannot express MMs computing measures over multiple patients, e.g. distributions for comparing temporal or multi-site differences (cf. [7]).

A central aspect of DQ is its task dependence. Johnson et al. emphasize this aspect [8] when proposing DQ-assessment based on separate task and domain ontologies. CIM-based definition of DQ-checks suits well to express constraints derived from different task and domain ontologies. One CIM can define constraints based on the original medical concept (e.g. a human age should be in the range 0 - 150) and a modified CIM can define constraints based on the task (e.g. for a certain study patient age should be in the range 65-85).

Although we described and tested CIM-based DQ-assessment this work is not intended to give an assessment about the DQ in our exemplary setting. In our case, this would not be reasonable without involving medical experts knowing and making use of the data. However, more interesting than local data quality values would be, what kind of DQ-issues can be found, and cannot be found, using our method and if our method found all issues it was supposed to find. However, because of limited resources for creating a gold standard, which labels all DQ-issues existing in the dataset, we did not compare our findings to one. Therefore, the biggest limitation of this work is that we only approximate these questions. We describe what kind of DQ-issues we found using our method (Table 2) and we point out that literature on DQ describes plenty of other MMs (a few discussed above) having a right to exist because they find different kinds of DQ-issues than our CIM-based method. Although, the numbers in Table 2 may seem small for ~12 million records, especially since data from routine care is often reported to have bad DQ, we still think the numbers are reasonable, because medical devices automatically created most of those values, which increases their conformance with range-, format-, value set and cardinality constraints.

Another limitation of our work is that we did not attempt to optimize our implementation for performance nor did we perform multiple runs of our ETL-process.

Thus, the only statement regarding performance of our approach is that the observed running time of 72 minutes for the whole ETL-process is all right for our purposes.

An unsolved problem stated in DQ-literature is that there is no agreed upon standard for DQ-MMs and –assessments and consequently these are not comparable (cf. [9]). While basing checks on standardized CIMs is a step forward, there is still a need to overcome the lack of a standard result format.

The presented approach for definition of range-, format-, value set and cardinality constraints based on openEHR-CIMs will be a building block for DQ-assessment in our local HiGHmed MeDIC. CIM-based DQ-assessment methods going beyond that are planned as future work as outlined in [10].

Acknowledgements

We want to thank our colleagues from the Central Information Management department for Educational and Scientific IT Systems at Hannover Medical School for supporting us with data access and data integration matters.

The research for this paper was done within the project „HiGHmed“ as part of the German Medical Informatics-Initiative (MI-I). The project is funded by the German Federal Ministry of Education and Research (BMBF, grant id: 01ZZ1802C).

References

[1] Haarbrandt B, Gerbel S, Marschollek M, Einbindung von openEHR Archetypen in den ETL-Prozess eines klinischen Data Warehouse, 59. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS). Göttingen, 07.-10.09.2014, German Medical Science GMS Publishing House, Düsseldorf, 2014. DocAbstr. 230; 2014.

[2] Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al., HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries, Methods of information in medicine **57** (2018), 66-81.

[3] Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, et al., Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature, International Journal of Medical Informatics **82** (2013), 10–24.

[4] openEHR Foundation, Welcome to openEHR, available from: https://www.openEHR.org, last access: 12.10.2018.

[5] Tute E, Haarbrandt B, Integrating relational data into clinical information model based data repositories, 62. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS). Oldenburg, 17.-21.09.2017, German Medical Science GMS Publishing House, Düsseldorf, 2017, available from: https://www.egms.de/static/en/meetings/gmds2017/17gmds145.shtml

[6] Anani N, Mazya MV, Chen R, Moreira TP, Bill O, Ahmed N et al., Applying openEHR ' s Guideline Definition Language to the SITS international stroke treatment registry: a European retrospective observational study, BMC Medical Informatics and Decision Making **7** (2017).

[7] Saéz C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez J, Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories, J Am Med Inform Assoc **23** (2016), 1085–1095.

[8] Johnson SG, Speedie S, Simon G, Kumar V, Westra BL, Application of An Ontology for Characterizing Data Quality For a Secondary Use of EHR Data, Applied clinical informatics **7** (2016); 69–88.

[9] 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 **4** (2016).

[10] Tute E, Striving for Use Case Specific Optimization of Data Quality Assessment for Health Data, Studies in health technology and informatics **251** (2018), 113-116.