Health Informatics Meets eHealth
D. Hayn and G. Schreier (Eds.)

doi:10.3233/978-1-61499-759-7-204

A Web-Based Tool to Evaluate Data Quality of Reused Health Data Assets

Christopher WENDL[1,a], Georg DUFTSCHMID[a], Deniz GEZGIN[a], Niki POPPER[c], Florian MIKSCH[b] and Christoph RINNER[a]
[a] *Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna* [b]*dwh Simulation Services, Austria*
[c] *COCOS – Computational Complexe Systems, Technical University of Vienna, Austria*

Abstract. Background: Data from the health care domain is often reused to create and parameterize simulation models for example to support life science business in the evaluation of new products. Data quality assessments play an important part to help model users in interpreting simulation results by showing deficiencies in the raw data used in the model building and offers model builders a comparison of data quality amongst the used data assets. Objectives: Assess data quality in raw data prior to creating simulation models and prepare results for model users. Methods: Using a literature review and documentation of previous models created, we searched data quality criteria. For eligible criteria we formulated questions and viable answers to be used in a questionnaire to assess data quality of a data asset. Results: We developed a web tool to evaluate data assets using a generic data model. Percentage results are visualized using a radar chart. Conclusion: Data quality assessment with questionnaires offers model builders a framework to critically analyse raw data and to detect deficiencies early in the modelling process. The summarized results can help model users to better interpret simulation results.

Keywords. Data Quality, Computer Simulation, secondary use

1. Introduction

The development of pharmaceuticals and medical devices poses various challenges to life science companies not only from a technical point of view but also in the context of process management and evaluation of new products [1]. The goal of the imProve project [2] is to support life science businesses in bringing innovation to market using simulations of the Austrian health landscape.

Simulation models are models of reality which are created and parameterized by the model builders using information resources describing reality. In the health care domain such data is often obtained through reuse [3] i.e. the intelligent reutilization of patient data obtained in routine care or clinical trials for medical research. Clinical trials and prospective studies are conducted based on a study protocol where all needed data points and data requirements are listed beforehand and can be considered during data acquisition. When data is reused, i.e. was collected for other purposes, not the same standards were applied in the acquisition nor can be expected [4]. Therefore data quality assessments are necessary. Especially model users interpreting or parameterizing simulation models need to be aware of limitations in the data disguised by the models.

[1] Corresponding Author: Christopher Wendl, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria, E-Mail: christopher.wendl@meduniwien.ac.at.

Wand and Wang developed an ontological base to assess data quality [5]. They listed 26 data quality dimensions and categorized them into internal and external views. In [6] these dimensions are further categorized hierarchically into intrinsic, contextual representational and accessibility data quality dimensions. In [7] data assets are evaluated in respect to their reusability by consulting data quality experts and comparing these assets to an asset with a very high quality. Results are presented using radar charts.

We developed a web based questionnaire for model builders to evaluate data quality criteria without the need of consulting any data quality experts and allow model users to get a quick overview of the underlying data used in the modelling process. This work presents the preliminary results of an ongoing bachelor thesis.

2. Methods

In the first step, criteria to categorize data assets and data quality were searched. Precedence was given to criteria with special focus on the medical domain. Beside a literature review we searched for Austrian data assets suitable for the model creation and parameterization process by looking at previous models and simulations of the Austrian health care landscape. The found criteria were textually evaluated with respect to their eligibility for reused health data. For each eligible criterion a question with permitted answers was formulated.

A generic database based on the Entity-Attribute-Value (EAV) design [8] was used to persist the criteria categories, criteria with corresponding questions, permitted answers as well as the results of the data asset evaluation in a MariaDB [9] database. In Figure 1 the four tables of the EAV design are shown. The EAV design allows adding new criteria and categories without changes in the database schema.

For the web tool the PHP framework Laravel [10] was used. The model view controller (MVC) design pattern was applied, separate forms to enter new categories and corresponding criteria with questions and answers were implemented. The evaluation of a data asset was performed using a questionnaire displaying all the questions from the criteria table. The result of a data quality assessment was visualized using the JavaScript framework D3[11] and the radar charts plugin.

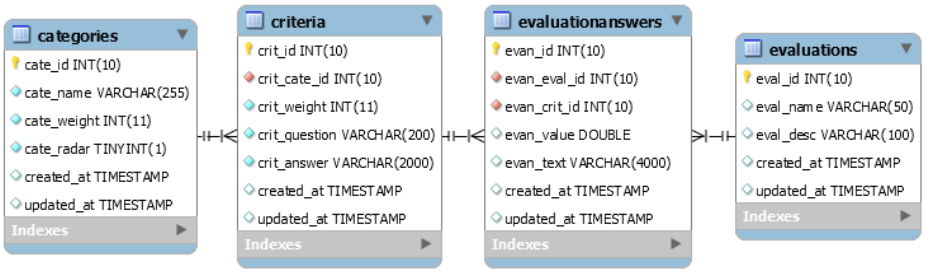

Figure 1. Entity–relationship diagram showing the EAV design used in our evaluation tool. A criterion (i.e the attribute) belongs to one category. In an evaluation (i.e the entity described) many evaluation answers (i.e the values) are assessed. Each evaluation answer corresponds to one criterion.

3. Results

The 15 quality criteria from Wang and Strong [6] were selected as base for the quality criteria. The main reason for selectin the criteria of Wang and Strong were the existing interpretation for the medical domain published in [4] as well as existing definitions for the criteria and the many citations in literature. We found some very similar criteria [12, 13], but due to the missing definitions and missing definitions in a clinical context, Wang and Strong's criteria were selected. The 15 criteria were reduced to 10 criteria by combining and omitting as following. The reputation criterion was omitted since the definition of reputation is the source of data [6], which does not influence the quality of a data asset, especially in a clinical context. Furthermore we combined relevance and value of the data (i.e. the definitions are very similar and data with higher value would be more relevant), amount of data and completeness (i.e. in a clinical context a complete data asset needs to have the proper amount of data), interpretability and ease of understanding (i.e. those definitions are practically the same) and representational consistency and concise representation (i.e. consistent representation in a data asset means that the data is well formatted so those two criteria definitions are very similar) by considering the definition of these criteria and aiming for similarities. The hierarchical categories (i.e. intrinsic data quality, contextual data quality, representational data quality, accessibility data quality) were adopted directly.

By analysing previous simulations we found that in most cases the raw data was collected for other purposes (e.g. routine care, reimbursement purposes, legal purposes, health survey, studies, registries, etc.) and was reused in the model creation and parameterization process. We found that the reason why information is documented influences what and how it is documented and has to be considered during the model creation and interpretation phase (e.g. when using claims data, diagnoses not relevant for reimbursement are not documented). Further, the different data assets are either available in aggregated form (e.g. literature, health surveys, etc.) stratified aggregations (e.g. Statistic Austria offering population information stratified by age-group, place of residence, year, etc.), individual record per activity (e.g. reimbursement data one record per hospital stay) or data assets with direct (e.g. electronic health records) and indirect (e.g. pseudonymous data) person identifiers. We added additional criteria to assess the initial purpose the reused data was collected for and the granularity of the raw data. However, those two criteria do not reflect the quality of a data asset and only offer additional information to model builders

To help model builders understand the purpose of a criterion, for each criterion one or more questions were formulated and to each question a set of permitted answers was attached. For each answer we defined either a percentage corresponding to the amount how good or bad data quality is reflected or a textual value. We were using categories to group criteria and reduce the complexity of the result visualisation. As shown in Table 1 the categories were directly related to those in [6]. The contextual data quality is a measure that describes the quality of the asset in its entirety. A low contextual data quality could be due to an out of date data set or many missing values. The intrinsic data quality category lets model users know how good the specific entries of a data asset are. A low intrinsic data quality could indicate many typos in a dataset. The representational data quality shows consistency in the representation of the data (e.g. entries following a standard) and indicate to the model user how much pre-processing and cleansing of the raw data was applied or necessary. The last two categories are the granularity and the source of data, they tell the model user in what granularity the data was available to the

Table 1. Overview of data quality criteria selected for questionnaire with corresponding category and number of questions per criteria

Category criterion	Number of questions
Intrinsic data qualityAccuracy	3
Believability	2
Objectivity	2
Contextual data qualityRelevance	3
Timeliness	3
Completeness	3
Representational data qualityInterpretability	2
Representational consistency	2
Quality of accessibilityAccessibility	2
Access security	1
Other relevant criteriaData Source	1
Granularity	1

model builder and indicates if fine grained parameters could be implemented. The question to granularity has answers from coarse grained aggregated data assets to fine grained individual level data assets. Since the source of data is only for informational purposes there will only be a text field where users can enter a source

The definitions in [4] were used to guide us in the formulation of our questions. For example accuracy is defined as "the extent to which data are correct, reliable, and free of error, or in a clinical context, data values should represent the true state of a patient within the limitations of the measurement methods" [4]. Using this information the example question in Figure 2 was formulated.

During the visualization we distinguished between percentage and textual answers. For percentage answers, the arithmetic mean of all answers from the criteria in one category was calculated and displayed as a spoke in the radar chart. For each category a spoke was displayed (see Figure 2 showing 5 spokes). Textual answers were displayed below the radar chart on the result page. Each evaluation resulted in a distinct URI and can be made available together with the model. This allows model builders to evaluate data assets by answering the questionnaire and to quickly get an overview of the quality of evaluated assets by considering the radar chart. It also allows them to compare the quality of evaluated data assets amongst each other.

4. Discussion

The categorization of data is a multidimensional problem with different aspects in focus depending on the use case and priorities. In our initial design we focused on criteria from Wang and Strong [6] to evaluate data quality and added two criteria to better reflect the aspect of data granularity and origin of data. Similarities and differences of our approach to the ones of other researchers are discussed in the following.

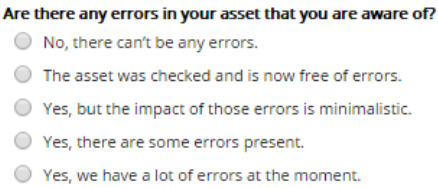

Figure 2. Question to assess accuracy focusing on the "free of error" part.

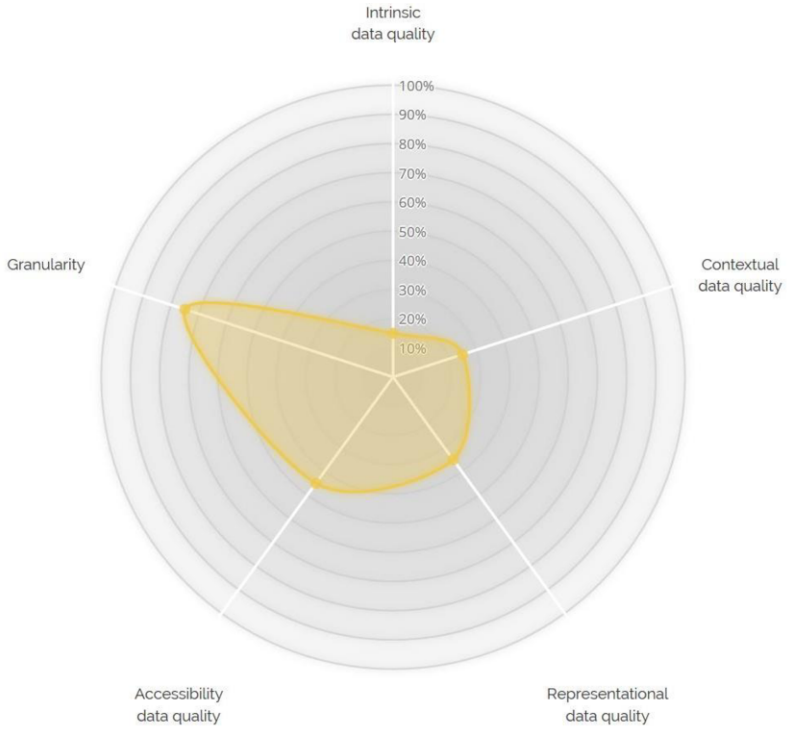

Figure 3. Sample radar chart showing the result of a data asset questionnaire of a specific data asset.

In the interoperable asset registry for asset discovery and assessment presented in [7] quality metrics were categorized into eight domains (development process, maturity level, trustworthiness, support and skills, sustainability, semantic interoperability cost and effort, maintenance) and were evaluated by 20 experts. Similar to their approach we use a radar chart to visualize results of a questionnaire. Our assessment focuses on data quality criteria of the underlying raw data used in the model creation process in contrast to their criteria focused on assessing the suitability for reuse.

The Health Data Navigator [14] is a tool to assess the performance and evaluate the quality of data sources used for comparative evaluation of health systems. Data sources are split into aggregated data or individual level data and vary from health status data to efficiency, cost and expenditure data. Quality criteria are split into entry errors, breaks and consistency of terminology and are documented in textual form. Also the other criteria (i.e. coverage, linkage, access, strength and weaknesses) are documented as free text. We used the quality criteria from Wang and Strong [6] in our assessment since they covered more than these three aspects.

The EMIF catalogue [15] was developed as part of an IMI European project and allows researchers to find databases which fulfil their particular research study requirements. Using 12 categories and 208 questions, amongst others general information (contact information to access the data sources, database populations), Database characteristics (start date, etc.), linkage data set description, examples of covered data elements, publication and comments can be assessed. The EMIF catalogue offers a very detailed description of a data asset and can be a valuable source to select raw data for creating and parameterizing simulation models.

In [4] a framework for data quality assessment in electronic health record is presented. The quality criteria from Wand et al. were adapted to requirements in comparative effectiveness research. The framework can be applied for single site data but mainly focuses on multi-site data to detect inconsistencies between the sites. We reused their interpretation of quality criteria for the medical domain to formulate our question.

OMOP CROUCH [16] consists of a set of scripts that can be applied to a consolidated data source to evaluate 35 different data quality criteria for all input data sources. If the raw data used in the modelling process is available in a consolidated standardized form, a similar generic approach to automatically evaluate data quality criteria is feasible for our use case.

In a next step we plan to evaluate the suitability of our questionnaire approach with predefined answers to evaluate data quality of data assets. The suitability of qualitative versus quantitative criteria will be analysed and compared to data quality assessment with other methods. We will review our quality criteria (i.e. add new criteria and tweak questions, reconsider merging of criteria) by applying the questionnaire to real data assets and evaluate it with model users.

Data quality assessment with questionnaires offers model builders a framework to critically analyse raw data and to detect deficiencies early in the modelling process. The generic design of the web tool allows us to easily delete or add criteria or categories and tweak the questions and answers. Our questionnaire currently offers a quick overview of the data quality of a specific data asset, for in depth analysis additional questions and more categories could be added. The summarized results can help model users to better interpret simulation results.

Acknowledgement

This research was supported by Vienna Business Agency.

References

[1] Markiewicz K, van Til JA, IJzerman MJ. Medical devices early assessment methods: systematic literature review. International journal of technology assessment in health care. 2014;30(2):137-46.

[2] dwh GmbH. imProve - Managing the Health Product Development. Available from: http://www.dwh.at/de/expertise/projekte/improve/ (accessed March 2017).

[3] Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association. 2007;14(1):1-9.

[4] Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Medical care. 2012;50.

[5] Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. Commun ACM. 1996;39(11):86-95.

[6] Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst. 1996;12(4):5-33.

[7] Moreno-Conde A, Thienpont G, Lamote I, Coorevits P, Parra C, Kalra D. European Interoperability Assets Register and Quality Framework Implementation. Studies in health technology and informatics. 2016;228:690-4.

[8] Friedman C, Hripcsak G, Johnson SB, Cimino JJ, Clayton PD, editors. A generalized relational schema for an integrated clinical patient database. Proceedings of the Annual Symposium on Computer Application in Medical Care; 1990: American Medical Informatics Association.

[9] MariaDB. Available from: https://mariadb.org/ (accessed March 2017).
[10] Laravel. Available from: https://laravel.com/ (accessed March 2017).
[11] D3. Available from: https://d3js.org/ (accessed March 2017).
[12] Stvilia B, Mon L, Yi YJ. A model for online consumer health information quality. Journal of the American Society for Information Science and Technology. 2009;60(9):1781-91.
[13] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association. 2013;20(1):144-51.
[14] Hofmarcher MM, Smith PC. The Health Data Navigator. Your toolkit for comparative performance analysis. A EuroREACH product. Vienna: European Centre for Social Welfare Policy and Research, 2013.
[15] European Medical Information Framework - (EMIF). EMIF Catalogue 2017. Available from: https://emif-catalogue.eu/ (accessed March 2017).
[16] Observational Medical Outcomes Partnership. OSCAR-Observational Source Characteristics Analysis Report (OSCAR) Design Specification and Feasibility Assessment 2011. Available from: http://omop.fnih.org/OSCAR.