MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190508

Towards Structured Data Quality Assessment in the German Medical Informatics Initiative: Initial Approach in the MII Demonstrator Study

Gaetan Kamdje-Wabo[a], Tobias Gradinger[a], Matthias Löbe[b], Robert Lodahl[c], Susanne Andrea Seuchter[d], Ulrich Sax[c], Thomas Ganslandt[a]

[a]*Department of Biomedical Informatics of the Heinrich-Lanz-Center, Mannheim University Medicine, Ruprecht-Karls-University Heidelberg, Mannheim, Germany,*
[b]*Institute of Medical Informatics, Statistics and Epidemiology, University Leipzig, Leipzig, Germany,*
[c]*Department of Medical Informatics, University Medical Center Goettingen, Goettingen, Germany,*
[d]*Medical Center for Information and Communication Technology, University Hospital Erlangen, Erlangen, Germany*

Abstract

The Demonstrator study aims to analyse comorbidities and rare diseases among patients from German university hospitals within the German Medical Informatics Initiative. This work aimed to design and determine the feasibility of a model to assess the quality of the claims data used in the study. Several data quality issues were identified affecting small amounts of cases in one of the participating sites. As a next step an extension to all participating sites is planned.

Keywords:

Data Accuracy; Electronic Health Records; Medical Informatics

Introduction

The German Medical Informatics Initiative (MII) is a long-term nationwide program established to improve research opportunities and patient care through Information Technology (IT) solutions, as well as to facilitate the exchange and use of data across German University Hospitals [1]. It involves four consortia implementing Data Integration Centers (DICs) in every participating university hospitals (MII sites) [1]. The Demonstrator Study was initiated with the intention of a continuous cross-consortia cooperation throughout the project duration. The study carries out a retrospective, descriptive and geo-regionalized exploration of comorbidities and rare diseases of patients in MII sites. The study leverages the standardized "§21" data format mandated by the German Diagnosis Related Group-based billing regulations [2], which enables multicentric analyses even before interoperable data structures and interfaces were fully developed and deployed across the MII sites. The §21-conformant datasets provide a narrow set of data elements for inpatient cases including demographics, diagnoses, procedures and further administrative data. The Demonstrator Study makes use of the Integrated Data Repository Toolkit (IDRT) to extract, transform and load the source §21-conformant data into local i2b2 data repositories (Informatics for Integrating Biology and the Bedside), including a pseudonymization step (see Figure 1). The Demonstrator Study carries out analyses on these repositories to calculate Comorbidity scores and to analyse rare diseases from ICD10 diagnosis data, under consideration of patient ages, encounter duration, hospital admission and discharge reasons, patient zip codes areas and the treating university hospital.

The use of routine clinical data (specifically: claims data) in the Demonstrator Study raises questions concerning the fitness for purpose of the data for the planned analyses. In this regard, the combination of data completeness, data conformance and data plausibility aspects of the data quality (DQ) may considerably determine the validity and veracity of analysis results [3].

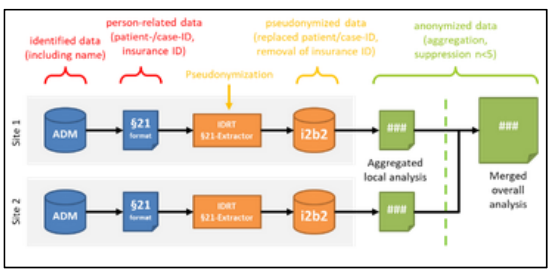

Figure 1 - Data processing pipeline of the MII Demonstrator study

Even though syntactic and coding aspects of the §21 source dataset quality is already tightly defined by billing regulations and controlled yearly by an independent institute in order to ensure their adequacy for billing purposes [2], aspects of data completeness and conformance taking account the aspects of the Demonstrator Study should be additionally addressed, as well as plausibility aspects of the variables in the datasets. The goal of the project was to evaluate the feasibility of constructing and implementing a DQ assessment framework based on a pretest of aggregated §21-conformant data from just one MII site taking part in the Demonstrator Study, before a large scale rollout to all the MII sites is conducted.

Methods

Quality ontology and analysis methods

According to the harmonized terminology of Kahn et al [3], the three quality dimensions completeness, conformance, and plausibility were choosen based on their relevance for the initial DQ approach. Operationalized quality indicators provided by the DQ Assessment guidelines from the German Technology and Method Platform for Networked Medical Research [4] were assigned to each dimensions, and quantitatively assessed using the verification method [3], through seven items formulated in consideration of longitudinal and cross-sectional aspects of the Demonstrator Study (see Table 1).

Implementation of the data quality queries

At the time of the submission, a pre-test was carried out based on the billing data of one of the participating MII sites based on an i2b2 database with data already loaded. We implemented

Table 1- Initial DQA-Framework

DQA-Dimensions	DQA-Indicators	Measurement Items
Completeness	Missing values for data elements (ID-1013)	**1-** Frequency of missing values in data elements
Conformance	Invalid values for qualitative data elements (ID-1021)	**2-** Frequency of invalid values contained in qualitative data elements
	Invalid values for quantitative data elements (ID-1024)	**3-** Frequency of invalid values contained in quantitative data elements
Plausibility	Freedom of Contradiction (ID-1003)	**4-** Men with malignant neoplasms of the female genital organs (C51-C58) as a hospital diagnosis **5-** Women with malignant neoplasms of the male genital organs (C60-C63) as a hospital diagnosis **6-** Men with in-patient delivery (birth) as admission reason **7-** Patients with a difference between the maximum and minimum age within the study period greater than 3 years (study duration)

SQL queries against the i2b2 database resulting in aggregated results. Data use within the project was approved by the ethics commissions of all involved MII sites lead by the Mannheim University Medicine Ethics Board (2018-864R-MA).

Results

Cohort description

The dataset consisted of 152,806 inpatient cases assigned to 94,570 patients. 25,238 patients (26.7%) had more than one case assigned to them (ranging from 1 to 66 cases assigned to a single patient during the observation period: 2015-2017).

Data quality analysis

The DQA results are displayed in

Table 2.

Table 2 - DQA results

DQA-Dimensions	Items	Results
Completeness	1	2532 missing values (1.6%) in the variable “Postal code” 37 missing values (0.02%) in the variable “ICD codes of the principal hospital diagnosis”
Conformance	2	178 invalid codes (0.12%) in the variable “Reason for the hospital admission”
	3	Range of patient ages at the hospital admission: 0-116
Plausibility	4	1 case
	5	5 cases
	6	0 cases
	7	2 patients

Conclusion

The results showed that a subset of 1.6% of cases did not contain a postal code, which, however, can be traced back to foreign patients or homeless patients for which zip codes are either unavailable or are excluded from documentation within the definition of the §21 dataset. In addition a small subset of cases did not contain a diagnosis (0.02%) or contained truncated or invalid codes not listed in the ICD10-GM classification (0.12%). The plausibility analysis revealed only a very small amount (single-digit) of inconsistent cases (see

Table 2). These first results address the feasibility of constructing and implementing a structured DQ analysis model focusing on quality aspects of data completeness, data conformance and data plausibility of the datasets used in the Demonstrator study. In the next steps we plan to implement a visualization of the DQ analysis using the R statistics software. The resulting scripts will be deployed to all MII sites participating in the MII Demonstrator study. This will allow us to carry out and compare DQ analysis based on the datasets from each of the participating MII locations. Further investigation will be perfomed to identify the reasons behind the formation of the problematic cases and to decide about how to handle the given cases before the datasets are used for statistical analysis.

Acknowledgements

This work was funded in part by the German Federal Ministry of Education and Research within the Medical Informatics Initiative, Grant IDs 01ZZ1801A (MIRACUM), 01ZZ1609A (SMITH), 01ZZ1802A (HiGHmed).

References

[1] S. Gehring, and R. Eulenfeld, German Medical Informatics Initiative: Unlocking Data for Research and Health Care. *Methods of information in medicine* 57 (2018), e46-e49.

[2] Institute for Reimbursement System in Hospital. *Data according to §21 KHEntgG - Version 2018 for the data year 2017*, http://www.webcitation.org/74AR3gjC0, Accessed on March 22, 2019.

[3] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, and S.G. Johnson, A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMs 4 (2016).

[4] M. Nonnemacher, D. Nasseh, and J. Stausberg, Data quality in medical research, *Medical Scientific publishing company* (2014).

Address for correspondence

Gaetan Kamdje Wabo,

Department of Biomedical Informatics of the Heinrich-Lanz-Center. Medical Faculty Mannheim of the University of Heidelberg,

Theodor-Kutzer-Ufer 1-3,

68167 Mannheim, Germany.

Email: gaetankamdje.wabo@medma.uni-heidelberg.de