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

doi:10.3233/SHTI230706

Pitfalls in Analyzing FHIR Data from Different University Hospitals

Matthias LÖBE[a,b,1], Christian DRAEGER[a,b], Alexander STRÜBING[a,b], Julia PALM[c], Frank A. MEINEKE[a,b] and Alfred WINTER[a]
[a] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig*
[b] *LIFE – Leipziger Forschungszentrum für Zivilisationserkrankungen, University of Leipzig*
[c] *Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital, Friedrich-Schiller University, Jena, Germany*

Abstract. The German Medical Informatics Initiative has agreed on a HL7 FHIR-based core data set as the common data model that all 37 university hospitals use for their patient's data. These data are stored locally at the site but are centrally queryable for researchers and accessible upon request. This infrastructure is currently under construction, and its functionality is being tested by so-called *Projectathons*. In the 6th Projectathon, a clinical hypothesis was formulated, executed in a multicenter scenario, and its results were analyzed. A number of oddities emerged in the analysis of data from different sites. Biometricians, who had previously performed analyses in prospective data collection settings such as clinical trials or cohorts, were not consistently aware of these idiosyncrasies. This field report describes data quality problems that have occurred, although not all are genuine errors. The aim is to point out such circumstances of data generation that may affect statistical analysis.

Keywords. Data Quality, HL7 FHIR, EHR data

1. Introduction

Ensuring data quality in the sense of the suitability of medical data for a defined application purpose is a challenge that is particularly significant in the secondary use of health care data for research [1]. In addition to "real" data quality problems, such as incorrect entries or systemic errors, misinterpretations arise due to special circumstances of data collection and data processing [2]. While experts from the field of health care data integration are aware of these limitations, in complex multicenter data use chains, a direct communicative link between data user (scientific research) and data collector (health care) often does not exist. At the same time, the data bodies of Hospital Information Systems are extremely complex, and the same data may be curated, transformed, or optimized for different uses. For this reason, deep domain knowledge is necessary to detect problems originating in transformation and selection processes.

[1] Corresponding Author: Matthias Löbe, Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany; E-mail: matthias.loebe@imise.uni-leipzig.de.

The German Medical Informatics Initiative (MII) has set itself the goal of transforming the patient data of university hospitals into a common model with an agreed format. This model consists of consented data elements and is divided into modules such as demographic data, patient encounters, diagnosis, procedures, laboratory values or medications. HL7 FHIR was chosen as the exchange format, FHIR profiles were specified, and a governance established. In addition, the mandatory use of international terminologies such as ICD-10, LOINC, or SNOMED CT, which have not been used in primary systems to date, was agreed upon. The technical challenges involved are enormous, and aspects of data quality have so far played a subordinate role - also for reasons of lack of availability of real data.

2. Methods

Proof of the functionality of the solutions and processes developed as part of the MII is ensured by cross-site test cases. These test cases are called Projectathons and have an ascending complexity. The goal is to find weaknesses and bugs, improve the data sharing infrastructure, and share solution approaches in the community. The first Projectathons focused on testing the FHIR core data set modules on local data. This was followed by testing the queryability of the data structures with FHIR Search and the integration of the central application and registry portal (FDPG) and dashboard functions. In the 6th Projectathon, the application, contract, and data provision process were run manually for a real clinical question for the first time. This involved first submitting a feasibility query to sites to get a case number estimate considering graded patient consent. Then, a data use request was submitted, data selection scripts were developed, distributed, and executed locally. De-identified data were forwarded to a central site for analysis.

To avoid overwhelming the sites with an overly complex data structure, a rather simple hypothesis was selected: which value of the laboratory parameter NT-proBNP is a suitable marker for diagnosis for cardiological diseases such as atrial fibrillation, taking age and gender into account? This so-called “atrial fibrillation” use case was first addressed in the 6th Projectathon and is currently being continued in the 7th Projectathon, where the same research question is addressed in multiple use cases with different approaches towards data aggregation and analysis. To do so, a search for cardiology diagnoses for patients in whom NTproBNP level was measured during hospitalization was developed. Accordingly, data from the demographics, case, diagnosis, and laboratory modules were necessary for the query. Data extraction and analysis scripts were written using R statistical software, and the R package dataquieR [3] from the cohort study context was used for data quality reports. Using R as common ground for all implementations allowed the generation of those data quality reports both at the point of data extraction in the data integration center as well as before the central analysis, by the researchers. Thereby, the data quality reports formed the basis of discussion with the experts from the data integration centers with detailed knowledge of their ETL process, the researchers with insights towards the use case specific requirements, the technicians responsible for script implementations as well as experts with detailed knowledge of MII specific FHIR profiles.

3. Results

As previously stated, biometricians do not know the idiosyncrasies of data origins in detail and are often not experts in FHIR or coding with medical terminologies. In clinical research, biometricians typically work close to the data collection. They are involved in creating the study protocol and the data catalog that is to be collected. Data capture forms and surveys are tested on realistic data. Exports from a clinical data management system are similar across studies and are analyzed using standardized routines. Below, we report some of the specific challenges and pitfalls encountered during statistical analysis of FHIR-based EHR data from multiple sites.

3.1. FHIR isn't tabular

The first problem is that usually tabular data structures are analyzed, and external data is provided in CSV format. FHIR data, on the other hand, comes as XML or JSON bundles. Additionally, the data structure included is tree-like and varies in detail depending on the origin location and data point. The reason for this is to retain maximum information from the primary systems. In this vein, some concepts must be annotated with one or more terminology code. Making imports into statistical software complex. Consequently, a tool named fhircrackr[2] [4] for flattening the tree structure has been implemented.

3.2. Temporal anomalies

Temporal implausibilities such as laboratory value measurements after the end of the stay are often addressed in prospective data collection at the time of data entry or during query management. For hospital data, automated verification of inputs generally does not exist. However, a deeper investigation showed that these are not necessarily misentries. Instead, the concept of a medical encounter is very complex; there are billing cases, medical visits, and point-of-care contacts that are mapped very differently in EHR systems. In addition, the case type can also change at the time of stay (from outpatient to inpatient), which leads to such results.

3.3. Coding of missing information with valid values

A frequent problem is to mark up missing values for mandatory fields. In research, missing codes are often used for this purpose. They are supposed to map exceptions without leaving the value range of the data type and requiring an additional form field, for example '99' for 'unknown' as an answer to the question about the number of biological children. In our case, we found birthdays such as 1900-01-01 or 1111-11-11 to be more common than one would expect. This information can be mapped in FHIR in a much more standards-compliant way (e.g., with the construct `dataAbsentReason`), and additionally poses a potential threat to analyses if it is not detected correctly.

[2] https://github.com/POLAR-fhiR/fhircrackr

3.4. Implausible distributions

When analyzing individual sites, it became apparent that the characteristics or distribution functions of individual characteristics did not meet expectations and differed significantly from the majority of sites. The reasons for this lay in the still prototypical character of the data catalogs. In some cases, for example, data from different years or from different wards (ICU, pediatric clinic) were included and it was possible that these also overlapped unfavorably (laboratory values 2020-2022, medications 2019-2020 and then not again until July 2022). Although this will reduce with productive operation, one can never completely exclude such gaps in data series. There is a lack of metadata describing data sets to better detect this, otherwise studies of longitudinal trends in incidence and prevalence could be compromised.

3.5. Misinterpretations due to underspecified queries

The FHIR model of the MII core data set is very expressive and the informative value should be used in the transformation process from the primary systems. However, an initial version of the query scripts simply queried all diagnoses or laboratory parameters of a type, without considering that exclusion diagnoses or incorrect entries would thus be counted as diagnoses. Moreover, preliminary laboratory values would be counted in addition to the desired parameters. Care must be taken that meaning-modifying constructs such as FHIR `Condition.clinicalStatus.code` or `verificationStatus.code` are understood correctly. However, there was a lack of knowledge that such statements would also be included, as this may not typically be documented in prospective data collections.

3.6. Scale limits/comperators are not respected correctly

A similar problem lies behind the observation that some sites have no (or only a few) very high (or low) values for certain laboratory parameters. The reason is that measuring devices and analysis methods have different scale limits and values exceeding these limits can no longer be measured absolutely. In FHIR there is an attribute `comparator` for this as an addition to numerical values, in order not to complicate the calculation by strings like ">2000". Especially for questions like the NT-proBNP example, where extreme values are interesting, a disregard of a comparator can be fatal.

3.7. Multiple terminology codes for the same concept

Some sites returned no results at all for NT-proBNP measurements, even though the parameter is collected at virtually all university laboratories. This was explained by missing codes in the FHIR Search query. LOINC is the standard terminology for laboratory values and NT-proBNP has seven different codes depending on the kind of property, specimen type, etc. For laboratory values, all eligible codes must be searched and queried, otherwise not all results will be returned.

3.8. Many units of measurement

When querying across many sites, many different units of measurement are to be expected. In our example, there were 9 different units of measure for NT-proBNP values, even the same LOINC codes can have different units of measure such as picograms per milliliter [pg/mL] = nanograms per liter [ng/L]. This requires parsing and conversion.

3.9. Missing or multiple values

Exceptions such as no measured value for existing measurements of NT-proBNP or multiple measured values present a further challenge. In research studies, exactly one value is often expected. In clinical reality, preliminary laboratory reports or issues such as "not enough material" may result in a measurement having no value. On the other hand, high-resolution data can result in many widely varying readings of NT-proBNP. This is an expected behavior e.g., with drug administration, but must be considered statistically.

3.10. No values at all are returned or results that do not meet the query criteria

FHIR can be customized for one's own needs through the mechanism of *profiling*. When no encounter data were found that included an NT-proBNP value, it was because the MII profile called for was "MII facility contacts." However, it turned out that some sites were using the generic FHIR `Encounter` resource instead. On the other hand, there was also the case where all diagnoses were returned even though only certain diagnoses were wanted. This was due to the fact that "wildcards" are not possible in queries for diagnoses. When looking for "atrial fibrillation" (code I48 in ICD-10), one must include all subcodes. Some FHIR servers cannot handle many parameters properly and simply return all diagnoses. Such bugs can only be detected exploratively.

4. Discussion and Conclusions

FHIR is a relatively young standard and information systems with native FHIR support are still rare. In this respect, the MII plays a pioneering role, since FHIR data for all major data kinds and from all university hospitals are available in a standardized form for the period starting around 2020. This allows many research hypotheses to be tested quicker, with larger case numbers and lower costs. Nevertheless, there are challenges to the structure and content of the data that do not occur in this way in traditional data collection and are currently unknown to many researchers. First, FHIR was primarily designed as a data exchange format. Many resources are deliberately kept generic and have only a few mandatory fields and alternative modeling options (e.g. Encounter towards Condition of a diagnosis or the other way around). One might be tempted to model very specific profiles that restrict such alternatives. However, defining very strict constraints on FHIR profiles will lead to having a suite of profiles for every single use case. That would be easier for analysis, but much harder for data providers.

Second, aspects of intrinsic data quality such as plausibility limits for laboratory values or phenotypes have not even been addressed here. Some MII sites have developed or adapted their own solutions for data quality assessments [5], but none operate on FHIR

data. For future data repositories such as the envisioned European Health Data Space (EHDS) [6], vocabularies or even semantic richer ontologies to describe the content of datasets, their quality, and their provenance still need to be developed to enable trust. One approach could be the development of FAIR data quality indicators that are managed in a technology-neutral way in a central repository.

Declarations

Conflict of Interest: The corresponding author declares that there is no conflict of interest from authors according to II.B from the recommendations of ICMJE.

Contributions of the Authors: ML wrote the initial manuscript. AS, CD, JP implemented scripts for the MII Projectathons. All authors reviewed the manuscript.

Acknowledgements: Research was supported by BMBF grants SMITH (01ZZ1803A, 01ZZ1803C), POLAR (01ZZ1910A, 01ZZ1910C) and DFG grant WI 1605/10-2.

References

[1] Kahn MG, Callahan TJ, Barnard J et.al. (2016) A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC) 4, 1244. https://doi.org/10.13063/2327-9214.1244

[2] Kohane IS, Aronow BJ, Avillach P et.al. (2021) What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res 23, e22219. https://doi.org/10.2196/22219

[3] Schmidt CO, Struckmann S, Enzenbach C et.al. (2021) Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 21, 63. https://doi.org/10.1186/s12874-021-01252-7

[4] Palm J, Meineke FA, Przybilla J, Peschel T. "fhircrackr": An R package unlocking FAST Healthcare Interoperability Resources for statistical analysis. Applied Clinical Informatics. 2023;14(01):054–64. https://doi.org/:10.1055/s-0042-1760436

[5] Löbe M, Kamdje-Wabo G, Sinza AC, Spengler H, Strobel M, Tute E (2022) Towards Harmonized Data Quality in the Medical Informatics Initiative - Current State and Future Directions. Studies in health technology and informatics 289, 240–243. https://doi.org/10.3233/SHTI210904

[6] Bernal-Delgado E, Craig S, Engsig-Karup T et. al. (2022) European Health Data Space Data Quality Framework. TEHDAS Deliverable 6.1. https://tehdas.eu/app/uploads/2022/05/tehdas-european-health-data-space-data-quality-framework-2022-05-18.pdf