Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240735

Leveraging FHIR in Federated Learning Environments: A Data Harmonization Framework for Cohort Studies

Héctor CADAVID[a,1] and Bauke ARENDS[b]
[a] *Netherlands eScience Center, The Netherlands*
[b] *University Medical Center Utrecht, Department of Cardiology, The Netherlands*
ORCiD ID: Héctor Cadavid https://orcid.org/0000-0003-4965-4243
Bauke Arends https://orcid.org/0009-0009-7494-4542

Abstract. MyDigiTwin is a scientific initiative for the development of a platform for the early detection and prevention of cardiovascular diseases. This platform, which is supported by prediction models trained in a federated fashion to preserve data privacy, is expected to be hosted by the Dutch Personal Health Environments (PGOs). Consequently, one of the challenges for this federated learning architecture is ensuring consistency between the PGOs data and the reference datasets that will be part of it. This paper introduces a novel data harmonization framework that streamlines an efficient generation of FHIR-based representations of multiple cohort study data. Furthermore, its applicability in the integration of Lifelines' cohort study data into the MiDigiTwin federated research infrastructure is discussed.

Keywords. Data harmonization, federated learning, FHIR

1. Introduction

The MyDigiTwin is a scientific initiative to develop a platform for the early recognition and prevention of an individual's cardiovascular diseases using big data from multiple reference datasets. To this end, this platform will be supported by the Dutch personal-health environments [1] (*persoonlijke gezondheidsomgevingen*/ PGOs)[2], and models trained with such reference datasets. One of the major challenges of this training process, given the project's unique characteristics, is harmonizing cohort studies' data (e.g., variable definitions, data formats, and measurement scales). On the one hand, this training process must be performed under a privacy-preserving federated architecture, in a way that participant hospitals or research institutions retain control over their data. On the other hand, since the data accessible through PGOs will serve as the input for the trained models, there is a need for consistency (e.g., regarding variable definitions, data formats, and measurement scales) not only between the training datasets, but also between these and the PGO's data. Considering that MedMij [2] is the framework for information exchange used by the PGOs, the project consortium decided to achieve this

[1] Corresponding Author: Héctor Cadavid; E-mail: h.cadavid@esciencecenter.nl

[2] Digital platform where an individual in The Netherlands can collect medical data from multiple healthcare providers and healthcare organizations.

consistency by making the reference datasets compliant with MedMij's underlying standard: the FHIR's Dutch (ZIB) profile [3].

This paper presents a novel data harmonization framework that enables the use of multiple cohort studies on a federated learning architecture by generating a uniform FHIR-based representation of the health data they provide. This framework, by encapsulating all the complexity behind the generation of FHIR resources, aims at making concurrent harmonization processes more efficient, uniform, and less error prone.

2. Methods

Establishing a federated-learning architecture, in the light of the earlier-discussed need for data consistency, requires setting up multiple ETL (Extraction, Transformation and Load) pipelines, one for each institution providing data. As depicted on Fig. 1, the extraction process involves gathering relevant data from the local raw-data sources, which is then transformed into the target format (in this case the MedMij/FHIR-profile) based on the *pairing rules* defined by the domain experts and each dataset expert. In the loading phase, the harmonized data is stored in a way federated algorithms can access it. With FHIR as the target language, the federated learning algorithm designer can then define, with no ambiguity, which variables or features are needed by using FHIRPath expressions (see right side of Fig. 1).

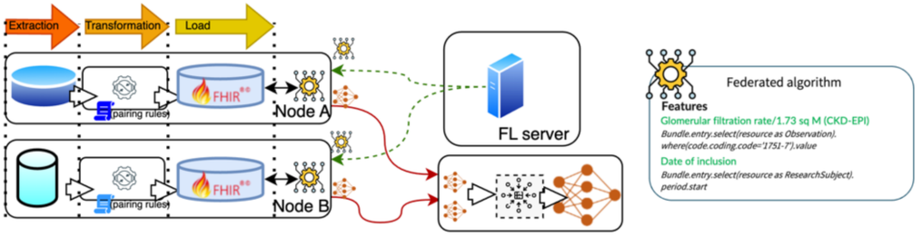

Extraction
Transformation
Load
pairing rules
FHIR
Node A
pairing rules
FHIR
Node B
FL server
Federated algorithm
Features
Glomerular filtration rate/1.73 sq M (CKD-EPI)
Bundle.entry.select(resource as Observation).
where(code.coding.code='1751-7').value
Date of inclusion
Bundle.entry.select(resource as ResearchSubject).
period.start

Figure 1. Federated learning architecture overview. On the left: The nodes where data resides, and the harmonization pipeline required for each one. Center: the federated server that coordinates the overall process, and the Node that aggregates the partial results. Right: an example of how the parameters of a FHIR-compliant algorithm will look like.

Given the intricate nature of implementing multiple ETL pipelines, one of the first milestones of the federated learning architecture was defining a data harmonization pipeline that could be reused, as much as possible, across multiple datasets. We first explored related works in the area, but none fit the project's unique characteristics. In particular, we found that most data harmonization approaches are designed only for data consolidation [4, 5, 6, 7]. Notable exceptions include DataSHaPER [8] and MetisFL [9], which, to the best of the authors' knowledge, are among the very few reported tools for data harmonization within a federated learning framework. However, these solutions were not suitable either as they do not support FHIR as a common data model. Given this, a custom harmonization solution tailored to the MyDigiTwin project needs - including the reusability quality-, was designed in cooperation with domain and data experts through multiple piloting iterations. This pilot, in turn, was guided by a specific task: extracting cardiovascular disease predictors currently explored by MyDigiTwin researchers [10] from Lifelines' cohort study [11] and generating the corresponding MedMij/FHIR compliant resources for each one of the study participants.

3. Results

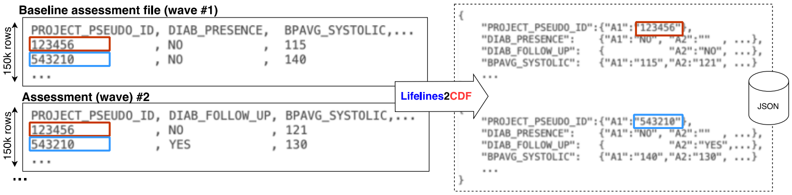
The piloting process outlined above lead to a data harmonization framework with the two core components below described: an intermediate representation format for the data extraction phase, and a reusable FHIR-compliant data transformation engine.

3.1. Intermediate representation for cohort studies (CDF format)

Large-scale cohort studies such as Lifelines [11] and UK Biobank [12] follow different strategies for managing their multiple assessments data, collected over different points in time. For instance, each Lifelines datafile corresponds to an assessment, whereas a UK-Biobank one corresponds to a subset of participants. This variability on the raw data distribution is the first limitation of any data harmonization process reusability. Furthermore, as generating a FHIR-bundle (a FHIR Patient linked to all their corresponding resources) requires the transformations to be performed in a per-patient fashion, some of these management strategies make the process inefficient. To illustrate this, let's consider the two Lifelines data files depicted at the left of Fig. 2, and the following excerpt of a rule that maps Lifelines' data to a Diabetes *clinical status* represented with a FHIR resource[3]:

> *... if DIAB_PRESENCE='NO' and there is a 'YES' on any of the DIAB_FOLLOW_UP follow-up assessments, then the clinical status is Active (SNOMED-55561003).*

As can be seen, evaluating this single condition would require searching, on each assessment datafile, its corresponding value for *DIAB_FOLLOW_UP*. Scaling this to a scenario with multiple variables, and the near 150k participants of Lifelines, gives an idea of the amount of duplicated file scans working with raw data would imply. Given the above, the pilot led to the design of a participant-centered JSON schema for an intermediate representation of cohort-study data depicted on Fig 2. By being participant centered, this intermediate JSON representation enables more efficient computations during the generation of FHIR resources.



```
PROJECT_PSEUDO_ID, DIAB_PRESENCE,  BPAVG_SYSTOLIC,...
123456           , NO           ,  115
543210           , NO           ,  140
...
```

```
PROJECT_PSEUDO_ID, DIAB_FOLLOW_UP, BPAVG_SYSTOLIC,...
123456           , NO            , 121
543210           , YES           , 130
...
```

```
{
   "PROJECT_PSEUDO_ID":{"A1":"123456"},
   "DIAB_PRESENCE":    {"A1":"NO", "A2":""   , ...},
   "DIAB_FOLLOW_UP":   {           "A2":"NO", ...},
   "BPAVG_SYSTOLIC":   {"A1":"115","A2:"121", ...}
   ...

   "PROJECT_PSEUDO_ID":{"A1":"543210"},
   "DIAB_PRESENCE":    {"A1":"NO", "A2":""   , ...},
   "DIAB_FOLLOW_UP":   {           "A2":"YES",...},
   "BPAVG_SYSTOLIC":   {"A1":"140","A2:"130", ...}
   ...
}
```

Figure 2. Comparison of the original 'assessment'-centered Lifelines datafiles, and the proposed CDF format.

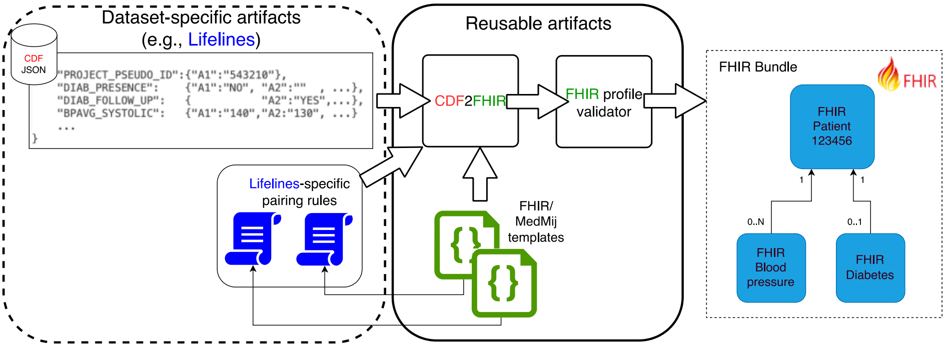
3.2. Generic cohort-data to FHIR transformation engine

Based on the cohort data schema outlined above, which decoupled the harmonization process from the cohort study's underlying management strategy, the reusable transformation engine (CDF2FHIR) depicted on Fig. 3 was developed[4]. It transforms

[3] https://simplifier.net/packages/nictiz.fhir.nl.stu3.zib2017/2.2.12/files/2002573

[4] Available on GitHub at https://github.com/MyDigiTwinNL/CDF2Medmij-Mapping-tool

CDF-based descriptions of cohort studies into FHIR-compliant resources based on two inputs: a set of FHIR templates, which define the structure of the target ZIB-compliant FHIR resources, and the implementation of the pairing rules, which defines how the input (the values given by each CDF) is mapped to a FHIR resource property.



```
"PROJECT_PSEUDO_ID":{"A1":"543210"},
"DIAB_PRESENCE":    {"A1":"NO", "A2":""   , ...},
"DIAB_FOLLOW_UP":   {           "A2":"YES",...},
"BPAVG_SYSTOLIC":   {"A1":"140","A2:"130", ...}
...
}
```

Figure 3. Overview of ETL pipeline. Left: transformation of dataset-specific artifacts to an intermediate representation for cohort studies using dataset-specific pairing rules. Center: a reusable transformation engine (CDF2FHIR) transforms the intermediate representation into a FHIR bundle. Right: example of a FHIR bundle.

The Extraction and Transformation phases of the pipeline were performed for the first node of MyDigiTwin's federated learning architecture. For the Extraction, a tool for transforming a given selection of Lifelines variables into the CDF format was developed[5]. The FHIR templates required for representing the predictors currently explored by the project researchers were implemented following the Dutch-MedMij FHIR profile and tested with the validators provided by HL7[6]. The corresponding pairing rules, on the other hand, were developed using the collaboration features of GitHub for the refinement of their specification and implementation.

4. Conclusions

Federated learning has emerged as a promising element in modern health-data analytics, providing a collaborative approach for model training that directly addresses one of the major concerns in the field: data privacy. This paper introduces a robust data harmonization framework that utilizes FHIR, a standard known for its efficacy in health care data standardization. This framework enables an efficient and consistent workflow for integrating cohort studies within a federated learning environment. It has been successfully applied generating nearly 150.000 FHIR bundles from selected Lifelines variables for use within MyDigiTwin's project federated-learning research infrastructure. At the time of writing, this federated learning infrastructure is under development, and alternatives for enabling federated algorithms to perform queries on the generated FHIR data -i.e., the loading phase of the pipeline- are being evaluated. Future work includes further evaluations and refinements on the framework during the harmonization process on additional cohort studies.

[5] Available on GitHub at https://github.com/MyDigiTwinNL/LifelinesCSV2CDF

[6] https://confluence.hl7.org/display/FHIR/Using+the+FHIR+Validator

Disclaimer and Acknowledgements

This publication is part of the project MyDigiTwin with project number 628.011.213 of the research programme "COMMIT2DATA – Big Data & Health" which is partly financed by the Dutch Research Council (NWO).

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-7675.

References

[1] Brands MR, Gouw SC, Driessens MH. Persoonlijke gezondheidsomgeving. Nederlands Tijdschrift voor Geneeskunde. 2023 Mar 16;167.

[2] van Gorp A. Towards a citizen-centered innovation system for ehealth. IADIS International Journal on Computer Science & Information Systems. 2018 Jan 1;13(1).

[3] HCIM - health and care information models [Internet]. zibs.nl. 2023 [cited 2024 Feb 12]. Available from: https://zibs.nl/wiki/HCIM_Mainpage

[4] Peng Y, Henke E, Reinecke I, Zoch M, Sedlmayr M, Bathelt F. An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. International Journal of Medical Informatics. 2023 Jan 1;169:104925.

[5] Williams E, Kienast M, Medawar E, Reinelt J, Merola A, Ines Klopfenstein SA, Flint AR, Heeren P, Poncette AS, Balzer F, Beimes J. FHIR-DHP: A Standardized Clinical Data Harmonisation Pipeline for scalable AI application deployment. medRxiv. 2022 Nov 13:2022-11.

[6] Adhikari K, Patten SB, Patel AB, Premji S, Tough S, Letourneau N, Giesbrecht G, Metcalfe A. Data harmonization and data pooling from cohort studies: a practical approach for data management. International journal of population data science. 2021;6(1).

[7] Rinaldi E, Stellmach C, Rajkumar NM, Caroccia N, Dellacasa C, Giannella M, Guedes M, Mirandola M, Scipione G, Tacconelli E, Thun S. Harmonization and standardization of data for a pan-European cohort on SARS-CoV-2 pandemic. NPJ Digital Medicine. 2022 Jun 14;5(1):75.

[8] Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BH, Perola M, Stolk RP, Foco L, Minelli C, Waldenberger M, Holle R. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerging themes in epidemiology. 2013 Dec;10:1-8.

[9] Stripelis D, Ambite JL. Federated learning over harmonized data silos. In International Workshop on Health Intelligence 2023 Feb 13 (pp. 27-41). Cham: Springer Nature Switzerland.

[10] Dziopa K, Chaturvedi N, Asselbergs FW, Schmidt AF. Identifying and ranking novel independent features for cardiovascular disease prediction in people with type 2 diabetes. medRxiv. 2023 Oct 24.

[11] Scholtens S, Smidt N, Swertz MA, Bakker SJ, Dotinga A, Vonk JM, Van Dijk F, van Zon SK, Wijmenga C, Wolffenbuttel BH, Stolk RP. Cohort Profile: LifeLines, a three-generation cohort study and biobank. International journal of epidemiology. 2015 Aug 1;44(4):1172-80.

[12] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203-9.