1st Conference on Research Data Infrastructure

Securing RDM

https://doi.org/10.52825/CoRDI.v1i.282

Published: 07 Sept. 2023

Distributed Privacy-Preserving Data Analysis in NFDI4Health with the Personal Health Train

Yongli Mou[1] [https://orcid.org/0000-0002-2064-0107], Feifei Li[2] [https://orcid.org/0000-0003-4815-8547], Sven Weber[2] [https://orcid.org/0000-0002-8518-9097], Sabith Haneef[3], Hans Meine[https://orcid.org/0000-0002-7557-5007], Liliana Caldeira[https://orcid.org/0000-0002-9530-5899], Mehrshad Jaberansary[2] [https://orcid.org/0000-0003-3407-1387], Sascha Welten[1] [https://orcid.org/0000-0001-5570-9672], Yeliz Yediel Ucer[13] [https://orcid.org/0000-0002-6845-7774], Guido Prause[https://orcid.org/0009-0008-4273-4957], Stefan Decker[13] [https://orcid.org/0000-0001-6324-7164], Oya Beyan[23] [https://orcid.org/0000-0001-7611-3501], and Toralf Kirsten[45] [https://orcid.org/0000-0001-7117-4268]

[1]Chair of Computer Science 5, RWTH Aachen University, Germany

[2]Institute for Medical Informatics, Faculty of Medicine, University Hospital Cologne, University of Cologne, Germany

[3]Fraunhofer FIT, Sankt Augustin, Germany

[4]Department of Medical Data Science, University Medical Center Leipzig, Germany

[5]Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Germany

Fraunhofer MEVIS, Bremen, Germany

Department of Diagnostic and Interventional Radiology, Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany

Abstract: Data sharing is often met with resistance in medicine and healthcare, due to the sensitive nature and heterogeneous characteristics of health data. The lack of standardization and semantics further exacerbate the problems of data fragments and data silos, which makes data analytics challenging. NFDI4Health aims to develop a data infrastructure for personalized medicine and health research and to make data generated in clinical trials, epidemiological, and public health studies FAIR (Findable, Accessible, Interoperable, and Reusable). Since this research data infrastructure is distributed over various partners contributing their data, the Personal Health Train (PHT) complements this infrastructure by providing a required analytics infrastructure considering the distribution of data collections. Our research has demonstrated the capability of conducting data analysis on sensitive data in various formats distributed across multiple institutions and shown great potential to facilitate medical and health research.

Keywords: NFDI4Health, distributed data analytics, personal health train

1 Introduction

In the medical sciences, data is continuously collected from patient care services, registries, clinical trials, epidemiologic studies, and other research projects. Data collections are managed in a highly fragmented manner. Only a few of them are web-based accessible, while most of them are not findable and often there is no widespread knowledge about them - even within the same institution where the study was conducted. The NFDI4Health consortium, as part of the German National Research Data Infrastructure (NFDI) Initiative, aims at bridging the highly fragmented data collections generated by clinical trials and epidemiological and public health studies. To overcome the current limitations resulting from this fragmentation, NFDI4Health establishes an infrastructure allowing each data owner to continually manage its data collection locally and, thus, keep the sovereignty about the data, but need to make the data collection FAIR (Findable, Accessible, Interoperable, Reusable) and, hence, register it at a so-called Local Data Hub (LDH). There are LDHs all over Germany that are connected with a German Central Health Study Hub (CSH) which is the central access point for scientists to search and request data of interest. To complement the distributed data management, NFDI4Health contributes with infrastructures allowing to analyse data in a distributed mode, such as the Personal Health Train (PHT). In this paper, we sketch how the PHT is used within NFDI4Health to share medical study data for a common analysis and to simultaneously preserve the privacy of personalized medical data.

2 Methods and Materials

The predominantly used centralized analysis requires all requested data to be collected from partners who want to contribute to the medical research. On the contrary, the PHT constitutes a paradigm shift, i.e., bringing the algorithm to the data, and provides a novel distributed flexible approach that enables the use of sensitive personal data for privacy-preserving data analysis in a network of participants, while data owners stay in control of their own data [1]. Incorporating the FAIR principles, the PHT aims to facilitate medical and health research that applies not only to populations but also can be tailored to individuals.

The PHT has two main concepts, namely the Stations and Trains [2]. The stations are nodes that provide computational resources and execute analytic tasks in a secure environment in a way that it can run without any further installation. Specific analysis tasks are encapsulated and executed at stations. A station will be attached to an LDH, and data will be attached (connected to) at each station. One of the primary motivations behind the PHT is to empower data owners to take control of their health data. Therefore data remain in their original location attached at an LDH, and there is no automatic execution of the overall train keeping the opportunity for each data provider to check the source code before and the obtained results from their own station afterward.

Beyond these advantages, the most important feature of the PHT is the flexibility of the choice of data source technologies, as it can deal with different data types (e.g., radiology or genomics) and data formats or standards (e.g., FHIR or DICOM) [3], [4]. Besides, the PHT is agnostic to the code languages such that researchers can develop their own algorithms in different programming languages, including R and Python.

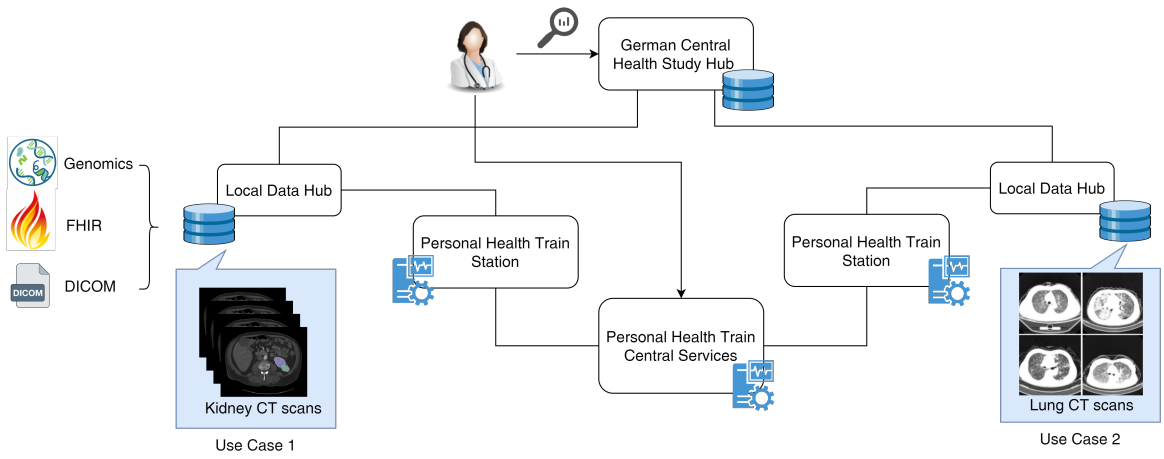

Figure 1. In NFDI4Health, the decentralized architecture comprises the German Central Health Study Hub (CSH) and a collection of Local Data Hubs (LDHs) and the Personal Health Train (PHT) provides the infrastructure for distributed analytics. LDHs are connected with the CSH and a PHT station is attached to an LDH. Researchers search and request data of interest through CSH and request data analytics through the PHT central services.

3 Results

While different groups in NFDI4Health work highly parallel on several infrastructural components and organizational procedures (e.g., governance acts) for the overall architecture, it is challenging to provide infrastructural components without precise requirements and interface definitions. Therefore, we are in contact with some groups within NFDI4Health while improving, testing, and adapting the PHT analysis infrastructure to new requirements and interface designs. While the first set of LDHs is currently established and filled with first content, i.e., metadata about clinical and epidemiological studies, there is currently a low but growing number of studies accessible in this early project stage that can be included in typical data analysis. However, at this project stage, data that can be used to solve medical research problems using complex algorithms, such as from the artificial intelligence method spectrum, or of more complex types, such as image and genomics data, is still missing. Therefore, we created use cases to show that the PHT infrastructure is working in principle, as shown in Figure 1.

A first use case focuses on the recognition of kidney tumors in patients to characterize the current stage of the disease and the location of the tumor within the kidney in order to provide therapy recommendations. The basis for this study is computer tomography (CT) image collection of known tumor patients in two locations. While these patients have already been treated in hospitals, the idea is to include them in a multi-center study to evaluate the overall segmentation, which can later be used for assessing the outcome of different therapy approaches. Recognizing the tumor stage and location requires segmenting the available CT images and identifying the tumor portion(s) within the images. We apply methods from the deep learning spectrum, in particular, the nnUNet [5] for 3D segmentation and then use this to extract the radiomics features. Instead of moving CT images for a centralized analysis, the PHT circulates the trained nnUNet model from server to clients and then transfers the extracted features for a federated analysis.

A second use case focuses on recognizing lung cancer in patients available at multiple sites. However, image data need to be harmonized when they have been taken by different devices and/or by different protocols. The amount of images produced by each device and protocol is different and sometimes small. Therefore, the approach is to produce synthetic data taking the available CT images into account in a way the synthetically generated data amount is equally distributed over devices and protocols and large enough for learning the differences from each subtype. The synthesized data is generated based on the model CycleGAN [6] at each LDH. It is used for style transfer and synthetic data generation to mitigate the distribution shifts from different devices and protocols. Based on this generated synthetic data set, we will apply a 3D pretrained segmentation model [7], allowing us to recognize and extract the anatomical structures within the lung.

4 Discussion

In response to the growing demand for more extensive knowledge acquisition through improved utilization of research data and the ensuing social advantages, the PHT offers the necessary infrastructure to facilitate secure, privacy-preserving, and standardized distributed data analytics across various medical and health data providers and researchers. Our studies have made significant progress in integrating federated and incremental learning using the PHT infrastructure. Taking both use cases into account, we will study future challenges of distributed analysis, in particular, when complex analysis methods are applied such as from machine learning method spectrum or by using complex data types including large-scale images (e.g., magnetic resonance images, CT) and genetics data. Moreover, we will align the requirements obtained from the intended use cases with those from other NFDI4Health groups in terms of interoperability, interrelating PHT stations to LDHs and the CSH as well as governmental procedures.

Funding

This work was done as part of the NFDI4Health Consortium (www.nfdi4health.de). We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 442326535.

References

[1] O. Beyan, A. Choudhury, J. Van Soest, *et al.*, "Distributed analytics on sensitive medical data: The personal health train," *Data Intelligence*, vol. 2, no. 1-2, pp. 96–107, 2020.

[2] S. Welten, Y. Mou, L. Neumann, *et al.*, "A privacy-preserving distributed analytics platform for health care data," *Methods of information in medicine*, vol. 61, no. S 01, e1–e11, 2022.

[3] Y. Mou, S. Welten, M. Jaberansary, *et al.*, "Distributed skin lesion analysis across decentralised data sources," in *Public Health and Informatics*, IOS Press, 2021, pp. 352–356.

[4] S. Welten, L. Hempel, M. Abedi, *et al.*, "Multi-institutional breast cancer detection using a secure on-boarding service for distributed analytics," *Applied Sciences*, vol. 12, no. 9, p. 4336, 2022.

[5] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[7] J. Wasserthal, M. Meyer, H.-C. Breit, J. Cyriac, S. Yang, and M. Segeroth, "Totalsegmentator: Robust segmentation of 104 anatomical structures in ct images," *arXiv preprint arXiv:2208.05868*, 2022.