Personal Health Train on FHIR

Personal Health Train on FHIR: A Privacy Preserving Federated Approach for Analyzing FAIR Data in Healthcare

Ananya Choudhury(✉), Johan van Soest, Stuti Nayak, and Andre Dekker

Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands
ananya.choudhury@maastro.nl

Abstract. Big data and machine learning applications focus on retrieving data on a central location for analysis. However, healthcare data can be sensitive in nature and as such difficult to share and make use for secondary purposes. Healthcare vendors are restricted to share data without proper consent from the patient. There is a rising awareness among individual patients as well regarding sharing their personal information due to ethical, legal and societal problems. The current data-sharing platforms in healthcare do not sufficiently handle these issues. The rationale of the Personal Health Train (PHT) approach shifts the focus from sharing data to sharing processing/analysis applications and their respective results. A prerequisite of the PHT-infrastructure is that the data is FAIR (findable, accessible, interoperable, reusable). The aim of the paper is to describe a methodology of finding the number of patients diagnosed with hypertension and calculate cohort statistics in a privacy-preserving federated manner. The whole process completes without individual patient data leaving the source. For this, we rely on the Fast Healthcare Interoperability Resources (FHIR) standard.

Keywords: Personal health train · FHIR · FAIR

1 Introduction

We live in an era of information explosion and artificial intelligence. Modern society is creating and making use of data like never before. With new devices and application, the amount of data generated is increasing both in terms of number of individuals an number elements per individual. However, as more and more data are created, people and government are increasingly becoming aware of whether or not it is ethically and legally right to use the data in an unrestricted manner. Of all the information about a person, his medical records are inherently privacy sensitive and confidential. As, such in the healthcare sector, it is important to protect patient privacy. The data protection law in the U.S.A., the HIPAA Act, limits sharing of sensitive data. In the E.U., the General Data Protection Regulation sets a well-formulated directive for securing confidentiality and privacy of citizens so that the data is not available publicly without

A. Bhattacharjee et al. (Eds.): MIND 2020, CCIS 1240, pp. 85–95, 2020.
https://doi.org/10.1007/978-981-15-6315-7_7

explicit, well informed specific consent, and cannot be used to identify a subject without additional information stored separately. PIPEDA in Canada, the Data Protection Act (P.D.A.) in the U.K., the Russian Federal Law on Personal Data, the I.T. Act in India and the China Data Protection Regulations (CDPR), all reflect the increasing global awareness regarding the importance of data privacy and confidentiality [1–4].

Current healthcare data sharing platforms are focused on performing queries on remote data sources and obtaining the results of these data queries. This means that interpretation of this data happens at the receiving end, rather than at the sending end. During this interpretation, issues can occur due to missing provenance or understanding the healthcare process on the sending side. The rationale of the infrastructure (PHT) - is that instead of requesting and receiving data, we are interested in asking a specific question and receiving a corresponding answer. PHT infrastructure is designed to deliver questions and algorithms which can be executed at data source institutes [5]. In this paper, we present a methodology for privacy-preserving analysis of healthcare data using HL7 FHIR standard.

The entire execution is fully controlled by the data source institutes, which means that interpretation and processing will happen at the data source institute as well, rather than at the receiving institute. Hence, we are sharing only the needed knowledge and information about a patient, instead of asking for data. PHT by design is made flexible for existing analysis tools, systems, or configuration in the hospital infrastructures. To perform this interpretation at the source, we need computational resources at the data source institute. More specifically, we need a location within the source institute where applications are received and can be executed in a safe (or sandboxed) environment. This allows the source institute to monitor who is requesting data, for what purpose, and what information is being sent back. This concept of having secure and safe compute resources at the source is not new. Large technology companies are already building towards this view of keeping data sources local, and only centralizing aggregated result statistics. These solutions are marketed using umbrella terms such as the “Intelligent Edge” [6], or and are already implementing similar concepts [7].

From a security perspective, this method reduces the data duplication need. By asking questions (instead of performing data queries), results can be stored, although the original data source is not duplicated. In recurring situations, this means the question needs and can be asked again. Recurrence of any question is an inefficient approach, however, makes the data provenance trail easier. When patients withdraw access to certain kinds of information/data, access management can be effectuated directly and enforced in future questions. In the current situation, all duplications of data need to be accounted for (resulting in a large provenance trail), and the information needs to be removed from these duplications. Second, this approach introduces advantages in terms of audit trails. Next to the regular information stored in audit trails (who requested what data, for which purpose, and stand at what time?), it clarifies the processing for a specific question. For example, an easy request asking for a patient’s body-mass index (BMI) will send an application to the data source, this application retrieves the height and weight of the patient, calculates the BMI. and only sends this BMI value back to the system asking the question. This means that for audit trailing,

we do know which processes and applications processed the data, as well as the exact method of processing. Figure 1 shows the difference between the traditional approach and the PHT approach of knowledge sharing.

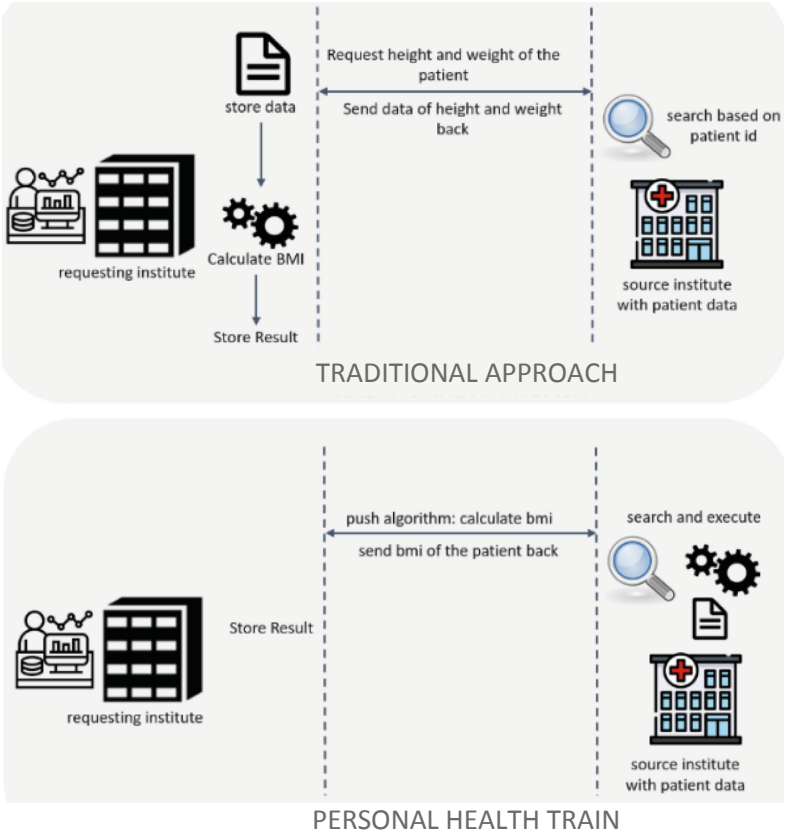


Fig. 1. Traditional approach vs Personal Health Train

Although such an infrastructure would work in an ideal-world situation where there is semantic interoperability, we have to cater for a realistic situation. Hence, such an infrastructure where data stays at the source needs proper definitions of where we can find data (Findable), how we can access this data (Accessible), how we can interpret (Interoperable) the data available, and how we can (Re)use the data. This means that this infrastructure heavily relies on the FAIR (Findable, Accessible, Interoperable, Reusable) principles [8]. This does not mean that every data source is publicly available; however, it should be clear how to use this data. As we are sending applications to the source, it also means that these applications need to be able to interpret FAIR data descriptions. These applications need to act on these descriptions to be able to read the (different) data structures present within the source location. Hence, these trains should be programmed to switch data structures (Interoperability) dynamically and to be able

to access multiple internal data sources (Accessibility). In our approach, this means algorithms and applications need to be able to interpret FAIR data standards, instead of humans.

In recent times, federated machine learning is gaining popularity. Vendors releasing open source packages for federated learning to researchers building algorithms by mathematically splitting the algorithm into distributed and local part, the data science community is increasingly realizing the need to keep data at source. However, many of these approaches do not address data privacy and need data to be stored in the same exact format [9]. PHT leverages containerization technologies for sending applications to the data source and eliminates the need of keeping data in same format by relying on FAIR principles. This leaves less overhead in terms of system requirements at the data source. Also, the computational requirement in case of a centralized approach is higher than that in a distributed setting.

2 FAIR Principles and Data Interoperability

The clinical data stored in the hospitals can be explored for extracting knowledge for both primary and secondary usage. However, for both clinical and research purposes it is important to manage the data in a manner that there is always 'single meaning' of the data no matter where, what and by whom the data is being used. Patient health records contain data captured from many different vendors and healthcare professionals. These may be the general practitioner, hospitals or the patient themselves. Often the data captured by the health care providers or the patients are in their own format and is not understandable once it leaves the domain where data was initially captured. This is largely because of lack of standardization of the data. Also, the data that is identifiable within the realm of one domain may not be uniquely identified across various domains, which complicates analyzing data across domains. An approach to tackle these complexities is by adhering to the FAIR principles.

Healthcare data is also hugely heterogeneous and non-uniform. A person's medical record may consist of data that ranges from radiology images, lab reports, prescriptions, medication list, procedural reports, dietary plans etc. Clinical data are either structured: such as coded data and laboratory results, or unstructured: such as clinical notes and free text comments. Imaging data for those stored in DICOM format contain the structured metadata, whereas the image itself is unstructured [10]. The completeness of scientific and clinical knowledge that can be extracted needs both unstructured and structured data to be harnessed.

Whereas structured data is easier to process, data correctness and completeness becomes a major concern. Unstructured data, on the other hand, contains a detailed description of the clinical condition that is easily understandable by humans but difficult to process by the computer due to lack of standard description and terminology usage. The data landscape in healthcare and the usage of these data for knowledge extraction for better care is hindered by lack of data interoperability and data quality at the source.

Interoperability as defined in the IEEE standard glossary "…is the ability of two or more systems or components to exchange information and use the information that has

been exchanged". Interoperability can be sub divided further as syntactic or structural interoperability and semantic interoperability [11]. Standardization of the data at the source is one way of ensuring interoperability; however, we argue that standardization alone is insufficient for several reasons.

First, the healthcare data exchange standards viz. HL7 Version 2.X and 3.X, OpenEHR and ISO 13606, HL7 CDA, XDS, ODM etc. and more recently the restful HL7 FHIR provide a format for specifying the information so that structure of the information remains same. However, with so many options of standardizing the data, different vendors choose different standards, again raising questions of syntactic interoperability. Thus, structuring data at the source might be a suggested option but not the ultimate solution. Secondly, use of a vocabulary based semantically interoperable system based on terminology and coding standards such as SNOMED CT and WHO Family of classification (ICD, ICF, ICH, ICHI, ICD-O) has already been in practice. However, even though different parties can agree on using the same terminology for their concept representation, understanding the clinical meaning out of it has still been an issue widely unaddressed or minimally addressed. For example, if a patient is admitted in the hospital with fracture in his tibia, this can be coded in an appropriate terminology for example using SNOMED CT. However, the coding terminology is not sufficient to know the following: how the fracture occurred, whether it caused any external injuries and whether the patient suffered from any injury induced diseases and symptoms. As such, keeping semantic consistency of data is more important than structuring the data.

HL7 Version 2.X and 3.X both structures the information and has a rich information model backing it. However, associating the data with terminology services like SNOMED CT, HL7 could only enable semantic interoperability to a certain extant while still failing to communicate the meaning of the clinical context. OpenEHR structures the data in a hierarchical manner based on one or more of the 300 complex archetypes. This emphasizes more on the data persistence rather than the clinical semantic interoperability. It is also common practice to exchange clinical information and documents using XDS. XDS uses a XML-based information representation and a central document registry. Queries to the database is based on health record metadata such as patient id etc.

In the light of all the discussion, HL7 FHIR seems to be a promising solution for achieving interoperability in the simplest possible way. FHIR has a rich information model and structures data in XML and JSON formats. FHIR also provides a RESTful way for querying and exchanging the data. The data elements are encoded in healthcare coding terminologies such as SNOMED CT, ICD and LOINC. All FHIR resources hence can be queried in a uniform way without having to look into the actual data. The FHIR community publishes data structuring guidelines in the form of resources. These resources can be customized for individual requirement or adapted as it is. FHIR extends the capabilities of HL7 version 2.x and version 3.x messaging protocols with a rich information model. Until the advent of HL7 FHIR, ODM was presumed to be the best data exchange standard for clinical research. ODM gained more popularity in clinical research for achieving semantic interoperability [12]. ODM is a cross platform data exchange standard for sharing between heterogeneous systems and allows

integration of multiple data sources. CDISC ODM is a close match for FHIR though; ODM lacks a rich information model. This limits ODM to use its own coding system unlike FHIR where external semantics source is easily incorporated.

One of the primary challenges in achieving semantic interoperability with FHIR will be to make the existing data, modelled in FHIR resources. Many ideas has been proposed in mapping OpenEHR, CDISC ODM, XDS etc. into FHIR [13–15].

3 Methods and Materials

The sensitive nature of patient health records bring challenges surrounding secondary usage of such data. Protecting patient privacy on one hand and making data available for research is a tradeoff faced traditionally in healthcare research. In the previous section we have described how it is possible to share insights from the data, without data having to leave from its origin, hence protecting patient privacy at its core [16, 17].

In this section, we describe the detailed methodology for using distributed health records for secondary usage. We use PHT as the infrastructure for federated data querying and statistical analysis of HL7 FHIR data. One of the pre-requirements for PHT is that the data at the source should be FAIR. We make use of data from two public data repositories and set it up in two data stations [18, 19]. We used the open source implementation of PHT infrastructure [20].

FAIR Data Stations: FAIR data stations are hosted within the organizational and IT system boundary of the hospital. Each data station contains FAIR data and are connected securely to the central server.

Central Server: All communication between the researcher and the data stations occur through the central server. The server acts as a message broker between the data stations and the researcher. The central server also stores the result received from the data stations.

Train: Trains are the containers carrying the algorithms and data query from the researcher to the data station. The researcher designs a train. The most common way to build trains are to wrap the algorithm and scripts in a Docker container.

Private Train Repository: The repository contains algorithms wrapped in containers. The researcher initiates a task by requesting the central server. For this implementation, the train repository is hosted in Docker hub [21].

Track: The track is the metaphor used to describe all communication happening between the researcher, central server and the data stations. The individual components communicate in a RESTful way with each other. Figure 2 shows the schematic diagram of the infrastructure.

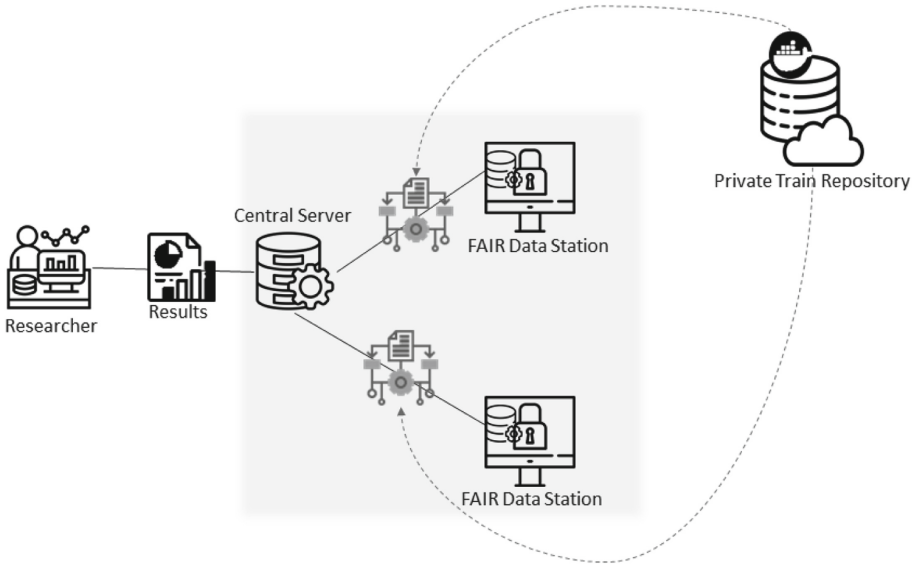


Fig. 2. Schematic diagram of personal health train infrastructure

The prerequisite of PHT is to host data in a FAIR repository. The FHIR resources and coding terminology makes data interoperable syntactically and semantically. All FHIR repositories can be accessed using the <base_url> of the repository. The central resource of FHIR is the patient resource. All other items such as patient observations, diagnostic reports, treatment plans etc. are organized as interlinked XML or JSON files and can be accessed by a unique <uri> assigned to it.

Example: All patients who were born after January 1, 1990

https://example.fhirserver.com:8000/Patient?birthDate=ge01-01-1990

The experiment consists two data stations set up with FHIR resources obtained from two public FHIR data repositories [18, 19]. The two data stations are connected to the PHT infrastructure. We build a train containing the FHIR query and an algorithm to calculate summary statistics. The researcher sends the train to all the data stations to calculate patient cohort statistics. At the data stations, the infrastructure component pulls the specified Docker image from Private Docker registry.

The algorithm for the distributed task consists of a master algorithm running in the researcher's machine. The master algorithm co-ordinates the task among the data stations and the researcher through the central server. The master algorithm also aggregates the results obtained from the individual data stations. The node algorithm wrapped in a Docker container runs at the data station. The node algorithm consists of a data query that loads the data locally and temporarily inside the Docker container. The execution of the node algorithm at the data station is controlled by the infrastructure component running at the data station. The node algorithm computes statistics locally inside the Docker container and sends only the results to the central server. When all the node algorithms complete execution and sends the result back to the server, the master algorithm is notified. The master algorithm retrieves the individual results from the central server and aggregates them to calculate the final output. Figure 3 shows the distribution of algorithms as Master and Node algorithm.

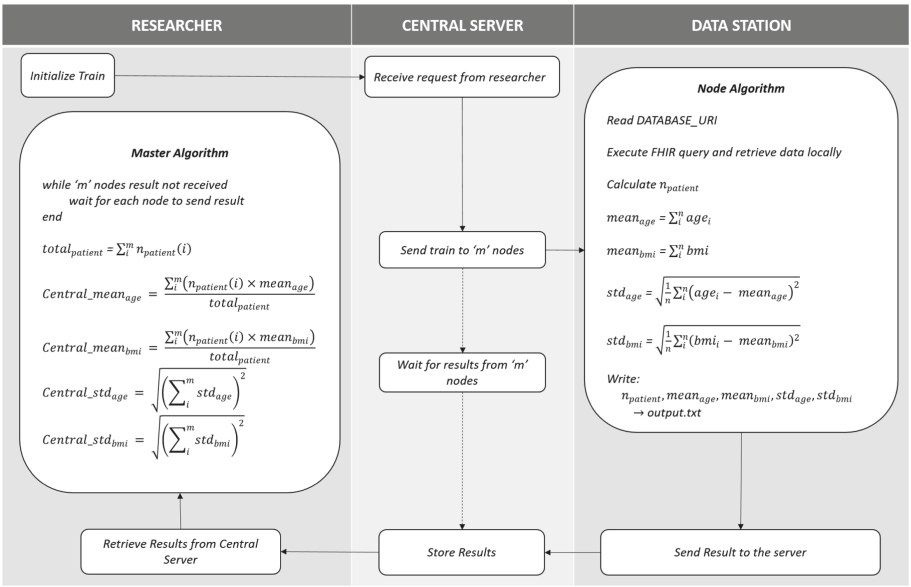


Fig. 3. Algorithm distribution

The master algorithm can also handle complex task such as coordinating and aggregating an iterative distributed machine learning process. Figure 4 shows how the node algorithm executes inside the data station. The node algorithm interacts with the FAIR data repository through an environment variable *DATABASE_URI* set by the infrastructure. The DATABASE_URI, at each data station contains the value of the actual url of the data repository. This keeps the algorithm and the researcher agnostic of the location of actual database. The node algorithm receives input from the master algorithm in the form *JSON* string and writes the result in *output.txt*, which is sent to the central server as *JSON* string.

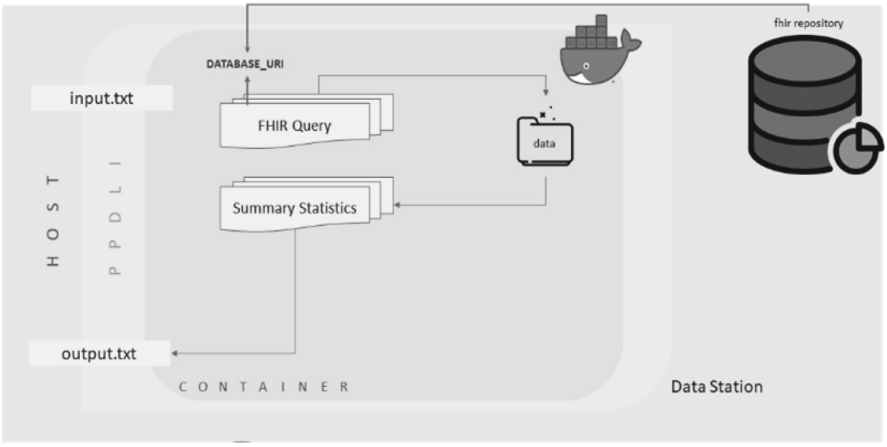


Fig. 4. Node algorithm running inside Docker container.

4 Results

The aim of the experiment was to calculate summary statistics of patient cohorts from distributed FHIR sources without making the data leave the source. The research question was to retrieve *"All matching patients born before 01-01-1990, who are diagnosed with **hypertension** and fetch **age** and **body mass index**."* For the patients diagnosed with hypertension, we send a second query which checks "*Is the hypertension patient also diagnosed with **diabetes**"*. The FHIR queries for fetching data are shown below:

`<base_url>Condition?_include=Condition:patient.birthDate=le1990-01-1&code=`http://snomed.info/sct|38341003

<*base_url*>Observation?subject=<*patient_id*>&category=http://hl7.org/fhir/observation-category|vital-signs&code=http://loinc.org|39156-5

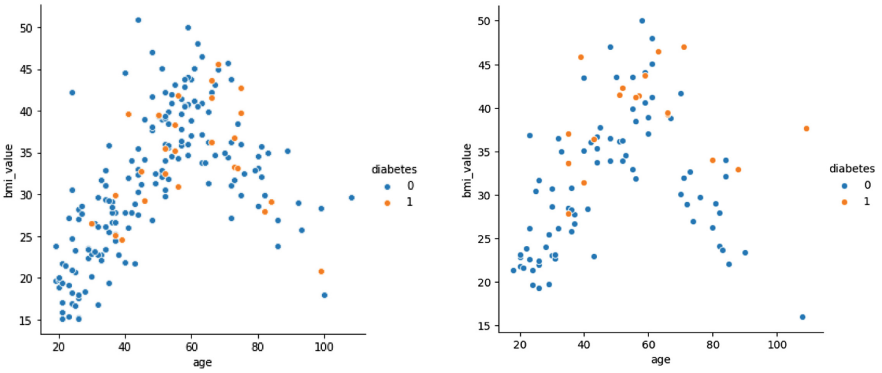
<*base_url*>Condition?subject=<*patient_id*>&code=http://snomed.info/sct|44054006

The first query retrieves all patients who are diagnosed with hypertension specified by SNOMED CT code *38341003* and who are born before *01-01-1990*. The second query fetches the BMI report for each patient. BMI in LOINC is coded as *39156-5*. Finally, the third query retrieves if the patient has also been diagnosed with type 2 diabetes mellitus, identified by SNOMED CT code *44054006*.

Table 1 shows the summary statistics obtained from the distributed cohorts and the aggregated summary statistics. We calculates mean age and BMI. A total of 398 patient information was retrieved from both the sources and mean and standard deviation (std) of age and BMI calculated. Figure 5 shows the age and BMI plot from the two data sources.

Table 1. Summary Statistics

Dataset	Patient Cohort Count	Diabetes	Age		BMI	
			Mean	Std	Mean	Std
FHIR Endpoint 1	199	26	49.72	21.14	32.46	7.94
FHIR Endpoint 2	99	18	50.2	19.63	31.74	8.2
Total	398	44	50.05	28.85	31.98	8.2



(a) FHIR Endpoint 1[19] (b) FHIR Endpoint 2[18]

Fig. 5. Age and BMI plot of hypertension patients.

5 Discussion and Conclusion

In this paper, we showed that PHT, existing healthcare standards and containerization technology can be leveraged for achieving data agnostic, privacy preserving distributed data analytics. Compared to other data sharing platforms or infrastructures, PHT has a significant advantage of scalability and flexibility. Since technologies around the globe are being developed in a very fast speed, it is impossible to ask all hospitals, clinics or other health providers to update their data format, methods of storing data, coding systems in the same speed. With PHT, it is possible to make use of these differently formatted data as long as we have enough metadata description associated with the data. The concept of sending applications and questions instead of requesting data creates many new opportunities both for primary and secondary use of clinical data. It acts as a bridge between the researcher requiring data and healthcare providers containing the data while serving everyone's interests. However, PHT and other existing similar infrastructures do not fix the problems of data preparation and data cleaning, data structure and semantic interoperability.

The paper is part of an ongoing project where we aim to train a machine learning model for predicting diabetes for patients diagnosed with hypertension in a privacy preserving federated manner. The different data sets are geographically and organizationally distributed and are governed by privacy and confidentiality laws. For training a machine learning model it is important for us to know the data distribution at the source. This is essential step before designing a complete machine learning algorithm, as the data distribution at the source impacts the choice of algorithms, hyper-parameter selection and model optimization. Hence, the work presented in this paper is a preliminary but important step. This will be further investigated by including machine learning algorithms in the proposed solution approach for predicting diabetes among patients.

References

1. General Data Protection Regulation (GDPR): Final text neatly arranged. https://gdpr-info.eu/. Accessed 09 July 2019
2. China Data Protection Regulations (CDPR)—China Law Blog. https://www.chinalawblog.com/2018/05/china-data-protection-regulations-cdpr.html. Accessed 26 Mar 2019
3. Data protection - GOV.UK. https://www.gov.uk/data-protection. Accessed 09 July 2019
4. The Personal Information Protection and Electronic Documents Act (PIPEDA) - Office of the Privacy Commissioner of Canada. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/. Accessed 09 July 2019
5. Beyan, O., et al.: Distributed analytics on sensitive medical data: the personal health train. Data Intell. 96–107 (2019). https://doi.org/10.1162/dint_a_00032
6. Intelligent Edge – Future of Cloud Computing—Microsoft Azure, https://azure.microsoft.com/en-us/overview/future-of-cloud/. Accessed 15 Feb 2020
7. Konečný, J., McMahan, H.B., Ramage, D., Richtárik, P.: Federated optimization: distributed machine learning for on-device intelligence. arXiv:1610.02527 [cs] (2016)
8. Hagstrom, S.: The FAIR Data Principles. https://www.force11.org/group/fairgroup/fairprinciples. Accessed 12 Mar 2019
9. Using TFF for Federated Learning Research | TensorFlow Federated. https://www.tensorflow.org/federated/tff_for_research. Accessed 15 Feb 2020
10. DICOM Standard, https://www.dicomstandard.org/. Accessed 15 Feb 2020
11. Oemig, F., Snelick, R.: Healthcare Interoperability Standards Compliance Handbook: Conformance and Testing of Healthcare Data Exchange Standards. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44839-8
12. Tapuria, A., Bruland, P., Delaney, B., Kalra, D., Curcin, V.: Comparison and transformation between CDISC ODM and EN13606 EHR standards in connecting EHR data with clinical trial research data. Digit Health **4** (2018). https://doi.org/10.1177/2055207618777676
13. Leroux, H., Metke-Jimenez, A., Lawley, M.J.: ODM on FHIR: towards achieving semantic interoperability of clinical study data. 10
14. Boussadi, A., Zapletal, E.: A Fast Healthcare Interoperability Resources (FHIR) layer implemented over i2b2. BMC Med. Inf. Decis. Making. **17**, 120 (2017). https://doi.org/10.1186/s12911-017-0513-6
15. Mandel, J.C., Kreda, D.A., Mandl, K.D., Kohane, I.S., Ramoni, R.B.: SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J. Am. Med. Inf. Assoc. **23**, 899–908 (2016). https://doi.org/10.1093/jamia/ocv189
16. Deist, T.M., et al.: Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin. Transl. Radiat. Oncol. **4**, 24–31 (2017). https://doi.org/10.1016/j.ctro.2016.12.004
17. Jochems, A., et al.: Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital – a real life proof of concept. Radiother. Oncol. **121**, 459–467 (2016). https://doi.org/10.1016/j.radonc.2016.10.002
18. HAPI FHIR. http://hapi.fhir.org/. Accessed 16 Feb 2020
19. HL7 FHIR API—Synthea, https://synthea.mitre.org/fhir-api. Accessed 16 Feb 2020
20. IKNL/VANTAGE6. Integraal Kankercentrum, Nederland (2020)
21. Docker Hub. https://hub.Docker.com/. Accessed 16 Feb 2020