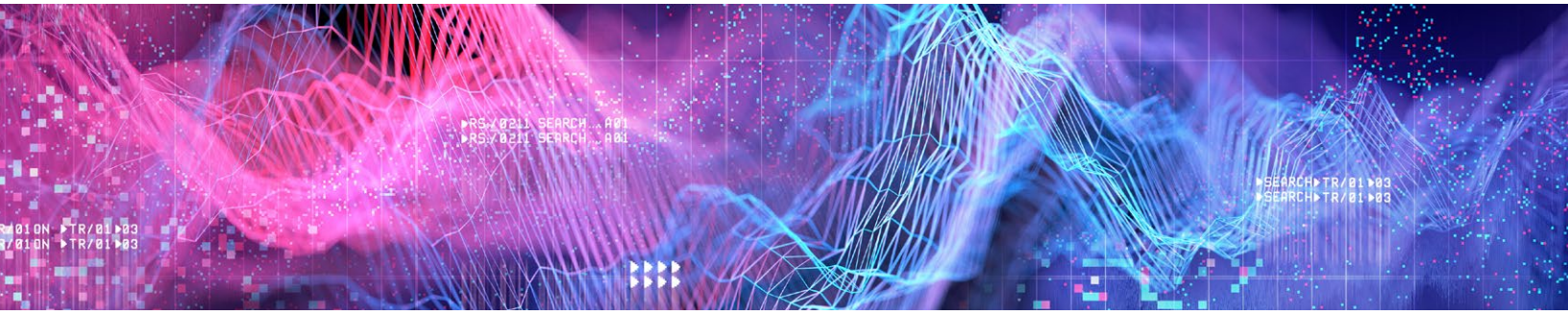

COVID-19 Generates Big Data Worldwide

Personal Health Train, FAIR and FHIR

Digitalization in the healthcare sector has resulted in an explosion of data—known as big data. Recently with the COVID-19 pandemic, nearly 12 million people have been found positive, of which more than 500,000 died.[1]

Ananya Choudhury
Department of Radiation Oncology (MAASTRO), GROW school for Oncology and Developmental Biology, Maastricht University Medical Centre+, The Netherlands

Esther Bloemen-van Gurp, Zuyd Unversity of Applied Sciences and Fontys University of Applied Sciences, The Netherlands, and Board Member, HL7 Netherlands

Johan van Soest, Department of Radiation Oncology (MAASTRO), GROW school for Oncology and Developmental Biology, Maastricht University Medical Centre+, The Netherlands

Andre Dekker, Department of Radiation Oncology (MAASTRO), GROW school for Oncology and Developmental Biology, Maastricht University Medical Centre+, The Netherlands

Big Data and AI in Healthcare

These figures are massive, but what is even more enormous is the amount of data these 12 million patients have created. This is big data and most of the answers that scientist across the globe are looking for are actually hidden in the data itself.

Another example of big data can be seen with cancer. An estimated 18 million new cases and 9.6 million deaths were recorded worldwide in 2018 alone.[2] Considering each patient generates about 1-10 gigabytes of data, the total amount of data generated is about 200 petabytes!

In the last 20 years, electronics and computing devices have decreased in size but have significantly increased in processing power. One of the major benefits of this advancement has come in the form of artificial intelligence (AI) and machine learning technologies.

The healthcare industry is adopting and benefiting from these technologies. From surgery assisting robots, improved and accurate diagnosis in cancer, to personalized treatments and developing new medicines, AI and machine learning is causing a paradigm shift in modern healthcare. With machines that can predict, diagnose, comprehend and learn healthcare sector is empowered like never before.[3]

Data Exchange, Interoperability and Data Protection Laws

Big data and AI are the driving force in modern day healthcare innovation. However, the healthcare sector is far from harnessing the true power of AI. This is because big data is contained in silos that are:

- Located within hospital boundaries and not accessible for research – data exchange and
- Unstructured or poorly structured making them unusable outside the source organization – data interoperability

Lastly, even if we can exchange the data, we may not be allowed to due to data protection laws.

Historically, sharing and exchanging patient data has been guided by the institute which generated the data. In the modern digitalization era, individuals are increasingly becoming aware of the consequences of uncontrolled data sharing. This poses a threat to individual's privacy and confidentiality.

Governments are fast adopting policies and formulating laws that regulate the collection, use and sharing of personal data. Data protection laws in the United States, GDPR in Europe, PIPEDA in Canada, Data Protection Act (DPA) in the UK, China Data Protection Regulation (CDPR) in China, and the IT act in India all reflect the increasing global awareness regarding the importance of preserving data privacy and confidentiality.[4,5,6,7]

In popular discussion, this is often regarded as the *Health Data GoldiLocks Dilemma*—whether to share data or to protect privacy? Or do both? Sharing too little data will prevent care providers from quality clinical decision making. Next generation AI technologies will be starved and promises like personalized medicine will be repressed. Sharing too much data could lead to a possible violation of personal privacy and confidentiality. Trust in healthcare providers would be eroded and value created by healthcare data could be captured by third parties e.g., large technology companies.[8]

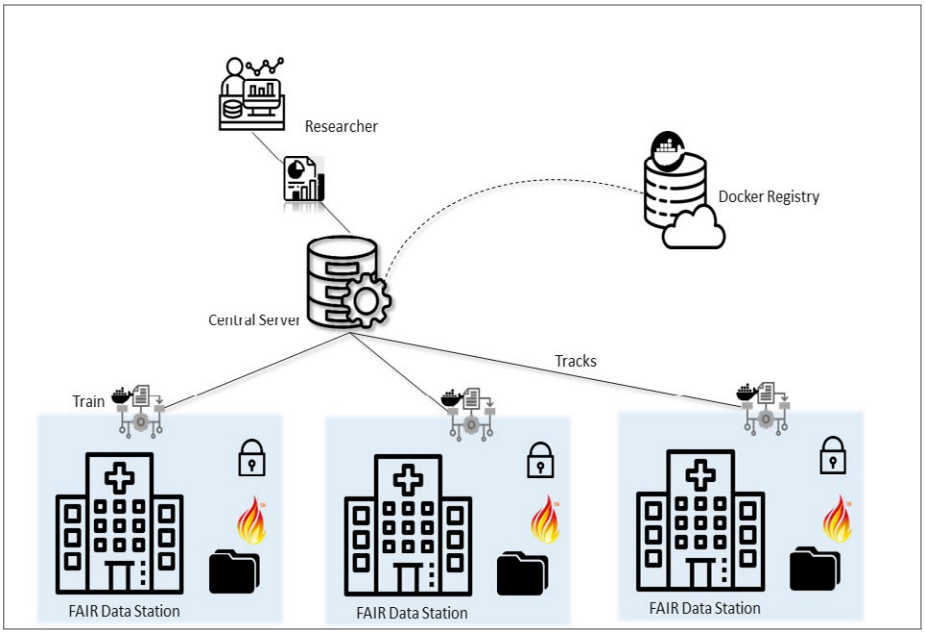

Figure 1: Personal Health Train architecture

Personal Health Train, FAIR and HL7 FHIR

In a world where we are restricted to collect and share data outside the source organization, we can share the analytics to the data. Current healthcare data sharing platforms are focused on performing queries on remote data sources and obtaining the results of these data queries.

The rationale of Personal Health Train (PHT) is that instead of requesting and receiving data, we are interested in asking a specific question and receiving a corresponding answer.[9]

PHT infrastructure is designed to deliver questions and algorithms which can be executed at the data source institutes. 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 side. Hence, we are sharing only the necessary information about a patient instead of asking for data.

The metaphor train in PHT refers to the packaged algorithms and analysis script that are sent to the remote data source. Stations contain the FAIR (Findable, Accessible, Interoperable and Reusable) data and also provide a computation environment for executing the algorithms.

Finally, tracks are the communication channels and mechanism by which the researcher (who initiates the analysis and is looking for answers), the central messaging server and the data stations talk to each other. Figure 1 depicts a schematic diagram of the PHT with three FAIR data stations.

Although such an infrastructure would work in an ideal world scenario where there is semantic interoperability, we have to cater to 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

Continued on page 19

Continued from page 19

Personal Health Train, FAIR and FHIR

interpret (Interoperable) the data available, and how we can (Re) use the data. This means that this infrastructure heavily relies on the FAIR principles.[10]

HL7 Fast Healthcare Interoperability Resources (HL7 FHIR®) as a clinical interoperability standard also establishes a strong relationship and identification with the FAIR data principles. The "I" (interoperability) in FAIR is the core concept in FHIR. FHIR provides a well-defined structure in the form of resources, profiles and extensions, which are the building blocks for ensuring syntactic interoperability.

FHIR also supports all major medical coding terminology standards (e.g., SNOMED CT, ICD, LOINC). Adopting coding terminology in describing health records is a key step in achieving semantic interoperability.

In addition, FHIR is built on top of a rich information model and is supported by rich metadata descriptions in the resources. Furthermore, with the FHIR API, it is possible to find and query patient data from remote servers.

Finally, it has been experimentally shown that PHT and FHIR can go hand-in-hand in achieving privacy preserving federated data analyis in healthcare. As a proof of concept, we designed a patient cohort counter to calculate the number of matching patients from two public FHIR repositories and calculated basic summanry statistics like mean age, mean BMI, standard deviation, age, and BMI relationship in patients diagnosed with both hypertension and diabetes.[11, 12, 13]

The entire process is executed without patient data leaving the source and is completely data agnostic. PHT relies on the metadata information and is independent of the actual data standard. This makes the PHT a generic infrastructure, independent of the (medical) specialty or research domain.

These proof of concept studies show a promising future where large scale clinical data from hospitals can be utilized and machine learning models can be trained for diagnostic as well as predictive analytics. PHT and FAIR data principles using HL7 FHIR as an interoperability solution has the potential to bring ground breaking research in healthcare.

References:

1. COVID-19 Map - Johns Hopkins Coronavirus Resource Center, https://coronavirus.jhu.edu/map.html, last accessed 2020/07/09.
2. Cancer, https://www.who.int/westernpacific/health-topics/cancer, last accessed 2020/07/09.
3. AI In Healthcare: 32 Examples Of Its Growing Impact | Built In, https://builtin.com/artificial intelligence/artificial-intelligence-healthcare, last accessed 2020/07/09.
4. General Data Protection Regulation (GDPR) – Final text neatly arranged, https://gdpr-info.eu/, last accessed 2019/07/09.
5. China Data Protection Regulations (CDPR) | China Law Blog, https://www.chinalawblog.com/2018/05/china-data-protection-regulations-cdpr.html, last accessed 2019/03/26.
6. 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/, last accessed 2019/07/09.
7. Data protection - GOV.UK, https://www.gov.uk/data-protection, last accessed 2019/07/09.
8. says, J.H.: The Health Data Goldilocks Dilemma: Sharing? Privacy? Both?, https://thehealthcareblog.com/the-health-data-dilemma-sharing-privacy-both/, last accessed 2020/07/09.
9. Beyan, O., Choudhury, A., van Soest, J., Kohlbacher, O., Zimmermann, L., Stenzhorn, H., Karim, Md.R., Dumontier, M., Decker, S., da Silva Santos, L.O.B., Dekker, A.: Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intell. 96–107 (2019). https://doi.org/10.1162/dint_a_00032.
10. The FAIR Guiding Principles for scientific data management and stewardship | Scientific Data, https://www.nature.com/articles/sdata201618, last accessed 2019/01/14.
11. Soest, J. van: jvsoest/PHT_on_FHIR_demo. (2019).
12. AnanyaCN: AnanyaCN/PHT_ON_FHIR_HypertensionCohort. (2019).
13. Choudhury, A., van Soest, J., Nayak, S., Dekker, A.: Personal Health Train on FHIR: A Privacy Preserving Federated Approach for Analyzing FAIR Data in Healthcare. In: Bhattacharjee, A., Borgohain, S.Kr., Soni, B., Verma, G., and Gao, X.-Z. (eds.) Machine Learning, Image Processing, Network Security and Data Sciences. pp. 85–95. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-6315-7_7.