Secure Distributed Medical Analytics for German Healthcare institutions using the Personal Health Train (PHT-meDIC)

Marius de Arruda Botelho Herr[1,2,4], Michael Graf[4], Peter Placzek[4], David Hieber[4], Felix Bötte[2], Stephanie Biergans[1], Mete Akgün[2], Nico Pfeifer[2], Oliver Kohlbacher[1,2,3,4]

[1]Medical Data Integration Center, University Hospital Tübingen; [2]Institute of Bioinformatics and Medical Informatics (IBMI), University Tübingen; [3]Applied Bioinformatics, Dept. of Computer Science, University of Tübingen; [4]Translational Bioinformatics, University Hospital Tübingen

Abstract

Transferring data between different hospitals is often restricted, and federated analysis of clinical data is a viable alternative. Current federated analytics frameworks (e.g., DataSHIELD[1] or MedCO[2]) are often limited in the type of input data or analysis that can be performed. In the Personal Health Train (PHT) paradigm, the analysis algorithm (wrapped in a 'train') travels between multiple sites (e.g., hospitals - so-called 'train stations'), hosting the data in their protected infrastructure, and only transfers results rather than the data. Structured pseudonymized clinical data is stored in FHIR servers at Data Integration Center's (DIC's) based on the HL7/FHIR profiles of the German National Core Data Set[3]. Implementing trains as secured containers enables even complex data analysis workflows to travel between sites, i.e., genomics pipelines or deep-learning algorithms - analytic methods that are generally not easily amenable. We present PHT-meDIC[4], a productively deployed, interoperable, open-source implementation of the Personal Health Train paradigm. The scope of applications for this platform ranges from machine learning algorithms to sophisticated omics and image analysis with arbitrary input data. Light-weight virtualization permits the automated deployment of complex data analysis pipelines (e.g., genomics, image analysis) across multiple hospitals in a secure and scalable manner. We combine different open-source third-party services with several custom-developed services. A separation into various services allows flexible adaption and extension in a scalable form. We achieved constant monitoring and persistent execution of trains and are providing governance template documents for deployment. Hospitals have pseudo-identifiers within the infrastructure and can only access their repository, and such inference attacks are less likely. Results are always encrypted at rest. Only participating sites and the submitting user can access them. Manipulation of trains will be detected at any stage. Furthermore, researchers can use additional privacy mechanisms (e.g., Paillier cryptosystem). The execution is within an encapsulated environment using study specific FHIR servers or data warehouses. We successfully deployed the implementation for distributed analyses of large-scale data. Our platform has been extended for interoperability in the Leuko-Expert project with other Medical Informatics Initiative partners' PHT[5] architecture. Documentation and source code is accessible at https://github.com/PHT-meDIC.

References

1. Marcon Y, Bishop T, Avraam D, Escriba-Montagut X, Ryser-Welch P, Wheater S, Burton PB, González JR (2021) *Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD*. PLOS Computational Biology 17(3):e1008880. March 30, 2021 https://doi.org/10.1371/journal.pcbi.1008880
2. Jean Louis Raisaro, Juan Ramon Troncoso-Pastoriza, Mickael Misbach, Joao Sa Sousa, Sylvain Pradervand, Edoardo Missiaglia, Olivier Michielin, Bryan Ford, and Jean-Pierre Hubaux (2019) *MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data.* IEEE/ACM Trans. Comput. Biol. Bioinformatics 16, 4 (July 2019), 1328–1341. https://doi.org/10.1109/TCBB.2018.2854776
3. Bild, R., Bialke, M., Buckow, K., Ganslandt, T., Ihrig, K., Jahns, R., … Prasser, F. (2020). Towards a comprehensive and interoperable representation of consent-based data usage permissions in the German medical informatics initiative. *BMC Medical Informatics and Decision Making*, *20*(1), 103. doi:10.1186/s12911-020-01138-6
4. de A. B. Herr, M., Graf, M., Placzek, P., König, F., Bötte, F., Stickel, T., … Kohlbacher, O. (2022). *Bringing the Algorithms to the Data - Secure Distributed Medical Analytics using the Personal Health Train (PHT-meDIC)*. doi:10.48550/ARXIV.2212.03481
5. Sascha Welten, Yongli Mou, Laurenz Neumann, Mehrshad Jaberansary, Yeliz Ucer Yediel, Toralf Kirsten, Stefan Decker, Oya Beyan; Privacy-Preserving Distributed Analytics Platform for Health Care Data. *Methods of Information in Medicine* 2022; (DOI:10.1055/s-0041-1740564)