Function, Aim, and Merit of the Personal Health Train for the Secondary Use of Clinical Data in Research

Sascha WELTEN[a], Yongli MOU[a], Yeliz UCER YEDIEL[b], Oya BEYAN[b,c],
Stefan DECKER[a,b], and Toralf KIRSTEN[d]

[a] *Chair Informatik 5, RWTH Aachen University, Germany*
[b] *Fraunhofer Institute for Applied Information Technology (FIT), Germany*
[c] *Institute for Biomedical Informatics, Faculty of Medicine and University Hospital, University of Cologne, Germany*
[d] *Department of Medical Data Science, Leipzig University Medical Center, Germany*

The secondary usage of clinical data in research has been promoted and recommended by several experts and scientists during the last recent years [1]. However, current data protection regulations make seamless data usage difficult and research communities, especially in Germany, run the risk of missing a great opportunity to use this wealth of data [1].

Strongly related to these legal obstacles is the process of data access as it potentially poses threats to privacy and makes the usage of data more difficult [1,2]. Some access policies include either the prior download of the data or restricted access through a cloud, where data holders upload the data [2]. Nevertheless, both inherently contradict the data sovereignty of the data providers as they lose direct control over data, making traceability and provenance challenging. Further, the large data volumes of different formats silo-ed in each institution are not easily transferrable between the involved entities.

To circumvent these shortcomings and enable more responsible processing of data, the Personal Health Train (PHT) constitutes a valuable option to facilitate privacy-preserving data analysis [3].
The PHT follows the paradigm of Federated Learning, which reverts the workflow for the data analysis by bringing the analysis to the data. This methodology empowers data-holding institutions to stay in control over their data but makes data access possible for research. Beyond these advantages, the PHT facilitates the analysis of decentralised data, as it is not dependent on one specific programming language, data source technology, or fixed cryptographic protocols [3,4].

Recent research has shown that the PHT is agnostic to data standards (e.g. FHIR) and is capable of conducting data analysis on sensitive data distributed across multiple institutions [3,5]. In particular, these studies demonstrate that the PHT can manage complex analysis tasks and is a practicable solution for Machine Learning on distributed data.

As the call for the establishment of the secondary usage of data will gain traction during the upcoming years, solutions such as the PHT provide the infrastructure to enable clinical scientists to drive value from data and increase the outcomes of research.

References:

[1] Jungkunz, Martin and Köngeter, Anja and Spitz, Markus and Mehlis, Katja and Cornelius, Kai and Schickhardt, Christoph and Winkler, Eva C.,Forum Marsilius-Kolleg, Bd. 21 (2022): Stellungnahme zur Etablierung der sekundären Forschungsnutzung von Behandlungsdaten in Deutschland,2022, https://doi.org/10.11588/fmk.2022.1.91697

[2] Suver C, Thorogood A, Doerr M, Wilbanks J, Knoppers B, Bringing Code to Data: Do Not Forget Governance, J Med Internet Res 2020;22(7):e18087, URL: https://www.jmir.org/2020/7/e18087, DOI: 10.2196/18087

[3] Oya Beyan, Ananya Choudhury, Johan van Soest, Oliver Kohlbacher, Lukas Zimmermann, Holger Stenzhorn, Md. Rezaul Karim, Michel Dumontier, Stefan Decker, Luiz Olavo Bonino da Silva Santos, Andre Dekker; Distributed Analytics on Sensitive Medical Data: The Personal Health Train. *Data Intelligence* 2020; 2 (1-2): 96–107. doi: https://doi.org/10.1162/dint_a_00032

[4] Wirth, F.N., Kussel, T., Müller, A. *et al.* EasySMPC: a simple but powerful no-code tool for practical secure multiparty computation. *BMC Bioinformatics* 23, 531 (2022). https://doi.org/10.1186/s12859-022-05044-8

[5] Welten, S.; Hempel, L.; Abedi, M.; Mou, Y.; Jaberansary, M.; Neumann, L.; Weber, S.; Tahar, K.; Ucer Yediel, Y.; Löbe, M.; Decker, S.; Beyan, O.; Kirsten, T. Multi-Institutional Breast Cancer Detection Using a Secure On-Boarding Service for Distributed Analytics. *Appl. Sci.* 2022, *12*, 4336. https://doi.org/10.3390/app12094336