FinOMOP Swarm Learning – Distributed Deep Learning for Patient-Specific Predictive Modelling of Acute Myeloid Leukemia

Eric Fey[1,8], Valtteri Nieminen [1,8], Salma Rachidi[1], Hartmut Schultze[2], Vytis Vadoklis[3], Perre Gustafsson[4], Johansson Markus[5], Kauko Tommi[4], Anna Hammais[6], Kukkurainen Sampo[7], Niemelä Sami[7], Tuomas Hakala[3], Alexey Ryzhenkov[1], Tomi Mäkelä [1], Oscar Brück[8], Joachim Schultze[9], Tarja Laitinen[2], Arho Virkki[3], Kimmo Porkka[1,8]

1 iCAN Digital Precision Medicine, University of Helsinki, 2 Hewlett Packard Enterprise, 3 Tietoevry Oy, 4 Hospital District of Southwest Finland, VARHA, Auria Clinical Informatics, 5 Istekki Oy, 6 THL Finnish Institute for Health and Welfare, 7 Tampere University Hospital, PIRHA, 8 HUS Helsinki University Hospital, 9 German Center for Neurodegenerative Diseases

Background

Deep learning to construct predictive models for precision medicine has great potential for precision medicine. The predictive power of deep learning is theoretically only limited by the amount and information content of the data. However, unlocking the full potential of deep learning faces two key challenges; sample size and availability of high-quality data. Local datasets of a single institution will always be limited and somewhat biased, and this is also true for centralized datasets of consortia. To truly scale, deep learning needs to take advantage of distributed datasets from hundreds of data nodes around the globe. Additionally, data diversity and quality are crucial. To deploy federated and distributed learning at scale, data need to be harmonized, and quality controlled to make sure that the right data are available for the model.

The FinOMOP Swarm Leaning Pilot addresses these challenges through three main aims:

1. Setting up a comprehensive, transparent, harmonized swarm-learning network between the key public health data owners in Finland, in particular the university hospitals.
2. Evaluate the technical and legal framework for performing swarm-learning[1] analytics in a secure and privacy-preserving manner within the swarm network (Figure 1.)
3. Perform technical and clinical proof-of-concept studies utilizing swarm learning.

Results show that it is possible i) to establish a highly-secure, privacy-preserving, fully-transparent, and auditable swarm-learning network between health data holders (here: university hospitals) ii) train meaningful predictive models using longitudinal EHR data collected in clinical routine (here: acute myeloid leukemia), iii) obtain certification for such this system from a national data authority (here: Findata).

The FinOMOP SL pilot started with a proof of concept utilizing the Finnish data nodes but is now entering the next phase to scale to OMOP-ready data partners in Europe and globally.

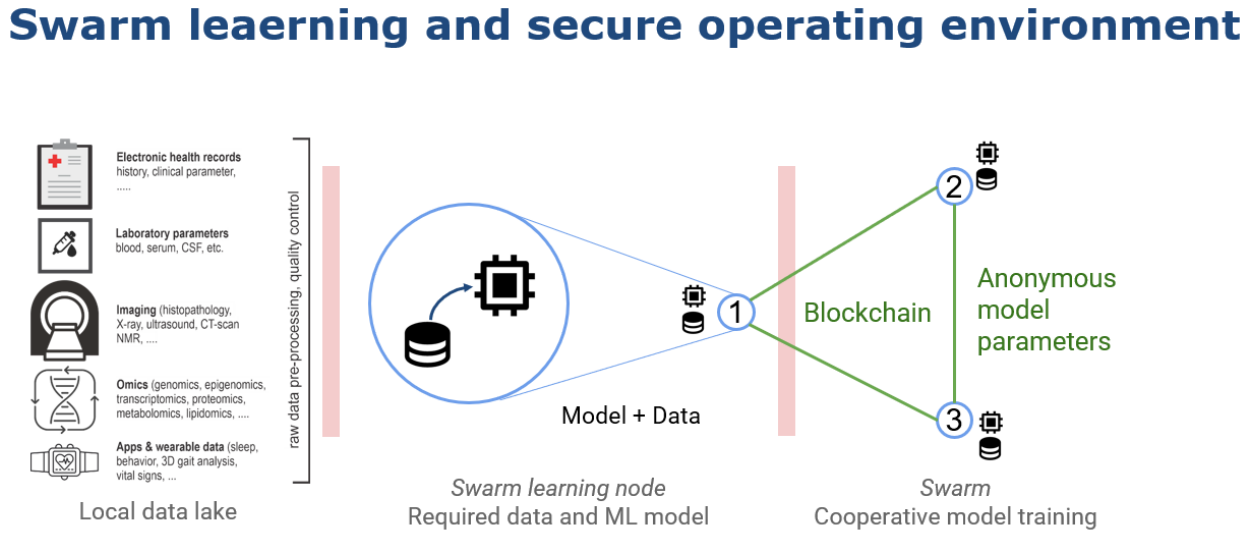

Figure 1. Swarm learning concept. Both potentially sensitive training data and personalized model predictions stay local in the data controllers' secure processing environment. Training of a joint model is facilitated through the encrypted, secure exchange of anonymous model parameters and model performance statistics based on zero trust technology. All transactions (requests and information exchanges) are logged in an immutable blockchain.

Methods

Overall approach (Figure 2):

1. Start with the three largest university hospitals in Finland, HUS in Helsinki, VARHA in Turku, PIRHA in Tampere, who together cover about 70% of the Finnish population.
2. Use a predictive model for acute myeloid leukemia (AML) as the first use case[2] and conduct a federated feasibility analysis leveraging the OMOP CDM to define the scope of the first model.
3. Develop the mathematical and computational framework to integrate time-to-event (survival) modelling into swarm learning, using OMOP as the data backend, and that can handle longitudinal data and missing values.
4. Implement and iteratively refine the swarm-learning technology in all three university hospitals and seek certification from the national data authority (Findata).
5. Develop a model locally, then train and validate this model in the swarm using OMOP data from all data partners.

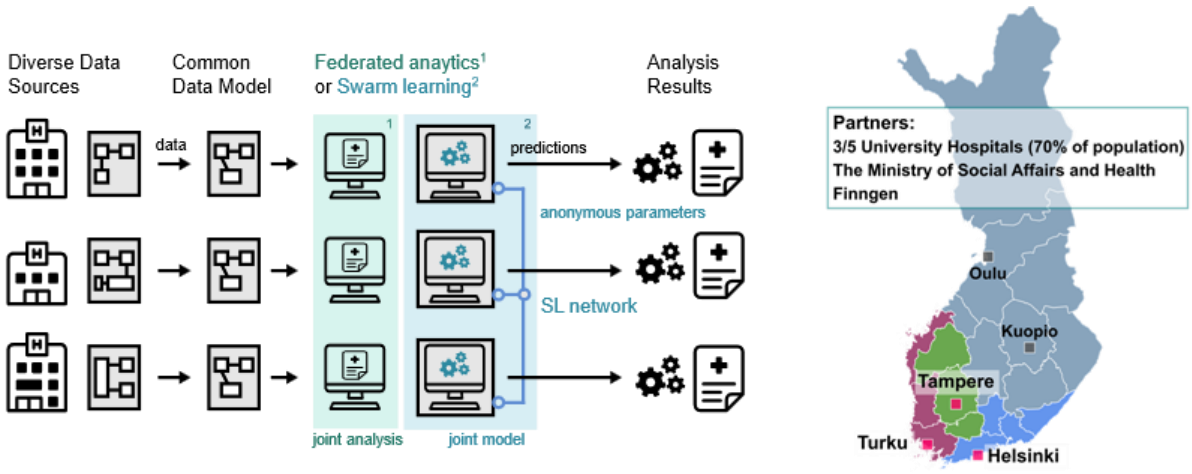

Figure 2. *Study concept.* ***Left:*** *EHR data from the three largest university hospitals in Finland were harmonized to OMOP enabling federates analytics (joint analysis) and swarm learning (joint model), a form of federated machine learning (ML). In contrast to federated ML, swarm learning does not require a central orchestrator. In contrast to [1]federated analytics where scripts are run locally and results can be shared asynchronously afterwards, [2]training the joint model (federated ML, swarm learning) requires the coordinated exchange of model parameter in real-time during training. Both training data and personalized model predictions stay local but can be shared afterwards if so desired. In contrast to federated learning, swarm learning logs all transactions, such as parameter updates during training, in a blockchain, thereby providing a secure, immutable audit log.* ***Right:*** *Map of Finland illustrating the catchment areas of the three participating university hospitals, covering about 70% of the Finnish population.*

Results

1. **SL has been implemented in all three participating hospitals.** For this implementation, we designed a solution that meets the requirements of our secure processing environments in the hospitals as mandated by Findata. The architecture of this solution separates the communication with the external swarm network from the internal traffic within the secure processing environment (machine learning nodes and data-access) using a reverse proxy setup.
2. **Federated analyses using OMOP scripts scoping the available data** across the three hospitals, identified a set of measurements that are routinely taken for AML patients in all three hospitals. Initial results of locally build models suggest that longitudinal blood count measurements for up to 21 days post diagnosis are highly predictive of relapse-free and overall survival, and that good performance can already be achieved with 15 days of longitudinal data.
3. **A transformer-based modelling framework for longitudinal data was developed** that integrates deep learning and survival modelling into SL. Taking advantage of recent advances in deep learning, we developed a transformer module specifically tailored for processing longitudinal data based on a dual attention mechanisms across time and features and residual (skip) connections that facilitate training of deep multi-layer models.
4. **A test run of the 3-node SL network has been performed** with the developed AML model architecture but using artificially generated “fake” data in the first instance. “Real” training of the AML across all three participating hospitals using real-world data is currently ongoing, and we hope to present these results at the symposium.
5. **The implemented swarm learning system has been audited** by the Finnish National Data

Authority (Findata, https://findata.fi/en/) providing a first-of-its-kind certification for such federated system for the use of real-world hospital data.

Conclusion

Building joint predictive models from distributed longitudinal, real-word data spread across multiple institutions is feasible and has great potential. OMOP and SL facilitate this model building. The framework is highly secure, as certified by Finland's National Data Authority, and generally applicable to predictive problems, including predicting treatment responses, patient survival, adverse events, and optimal treatments. Further, our first use-case demonstrates that using dynamic, short term follow up data only - blood count measurements up to 21 days post diagnosis - holds valuable information for predicting long-term prognosis, including overall survival. These models will be further refined using more traditional baseline features (cytogenetics, mutation profiles, clinical characteristics). OMOP Genomics mapping is ongoing. In conclusion, our OMOP-based swarm-learning framework facilitates constructing, training, and iteratively maturing predictive models for precision medicine in multicenter network settings.

References

1. Warnat-Herresthal, S. *et al.* Swarm Learning for decentralized and confidential clinical machine learning. *Nature* **594**, (2021).
2. Malani, D. *et al.* Implementing a Functional Precision Medicine Tumor Board for Acute Myeloid Leukemia. *Cancer Discov* **12**, (2022).
3. Lee, C., Zame, W. R., Yoon, J. & Van Der Schaar, M. *DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks*. www.aaai.org.