FinOMOP - a population-based data network

Javier Gracia-Tabuenca[1*], Perttu Koskenvesa[2*], Pia Tajanen[3*], Sampo Kukkurainen[4*], Gustav Klingstedt[5*], Anna Hammais[5], Persephone Doupi[5], Oscar Brück[2], Leena Hakkarainen[4], Annu Kaila[2], Marco Hautalahti[8], Toni Mikkola[4], Marianna Niemi[4], Pasi Rikala[7], Simo Ryhänen[6], Anna Virtanen[7], Arto Mannermaa[6], Arto Vuori[5], Joanne Demmler[5], Eric Fey[2], Terhi Kilpi[5*], Arho Virkki[3*], Tarja Laitinen[4*], Kimmo Porkka[2*]

[1]FinnGen, Helsinki University, Tampere, Finland, [2]Hospital District of Helsinki and Uusimaa, Helsinki, Finland, [3]Hospital District of Southwest Finland, Auria Clinical Informatics, Turku, Finland, [4]Wellbeing Services County of Pirkanmaa and Tampere University Hospital, Tampere, Finland, [5]Finnish Institute for Health and Welfare (THL), Helsinki, Finland, [6]North Savo Hospital District, Kuopio, Finland, [7]Productivity Leap, Joensuu, Finland, [8] Finnish Biobanks-FINBB

* equal contribution

Background

Secondary healthcare in Finland has been centralized to 5 university and 16 central hospital districts. Although national health data guidelines exist, in practice, most hospital districts have poorly interoperable IT solutions.

Hospital-level raw data residing in vendor-independent, secure data lakes provides a robust and granular source of clinical data for research and development. In addition, Finnish Institute for Health and Welfare (THL) governs several, legislation-mandated population registries. FinnGen, a national public-private partnership genome project, has converted the clinical annotation data to the OMOP CDM, with germline genomic data collected from 500,000 biobank participants (2).

The FinOMOP project aims at high-quality, granular mapping of key medical data sources (EMRs, registries; primary and secondary health care) to the OMOP CDM, for a comprehensive, population-based health data network in Finland.

Methods

FinOMOP consortium is a tight network of administrative and IT personnel, data scientists and clinical domain experts committed to shared, high-quality and -granularity OMOP processes. The consortium is administered by a steering group responsible for funding, developmental prioritization and legal/ethical issues. FinOMOP also operates vocabulary, ETL, and analytics groups. FinOMOP consortium is legally represented by FinBB (Finnish Biobank Cooperative), a co-operative owned by the Finnish Universities, University Hospitals and THL. FinBB serves as a one-stop shop for administering external collaborations (national/international; academic and commercial) via a single

contract or DTA, with focus on federated data and expertise sharing and analysis.

OMOP CDM conversion projects were started at the university hospitals (secondary care), and FinnGen (genomic data) in 2020, and in THL in 2022. Part of the funding was derived from the EU IMI2 EHDEN data partnership for all data sites. Currently, the hospital districts cover 70% (3.9/5.5 million) of the Finnish population with the aim for a complete coverage by 2025. The THL registries, including FinRegistry, already are population-based and cover the whole country.

OMOP mapping to ver. 5.4 is performed by conforming to the OHDSI guidelines and tooling by a multidisciplinary team at each site. National vocabulary mappings are maintained and distributed through a private Github repository, supported by regular meetings and workshops. This repository version-controls an USAGI file for each national vocabulary. FinOMOP members can continuously add new mappings, new vocabularies, or propose changes. Periodically, the USAGI files are transformed into concept and concept-relationship OMOP vocabulary tables and added to the FinOMOP databases. Thus, the FinOMOP databases not only are mapped to the same standard OMOP concepts, but also share the same custom (2 billionaires) non-standard concept IDs. ETL coding has been outsourced to EHDEN-certified SMEs.

Results

Currently, the primary OMOP mapping of 3.8M Finnish patients has been completed, with a granular and comprehensive population of all the key OMOP clinical data tables. OHDSI guidelines and tooling were shown to be compatible with the varying data infrastructures at each site. The ETL pipelines have included extensive quality control, including the use of the OHDSI Data Quality Dashboards.

Federated analysis pilots have already been performed using acute leukemia as a granular use case of a rare disease. The OMOP CDM facilitated the construction of a deep learning model to predict acute leukemia patient survival using federated and swarm learning (HPE). Several proof-of-concept studies using different condition domains will be performed by the end of 2023, including international collaborative efforts.

As an example of a regulatory use case, the consortium is currently running a comprehensive, fully federated post-launch launch evidence generation (PLEG) pilot program in three therapeutic areas for the competent authorities (Finnish Medicines Agency FIMEA), sponsored by the The Finnish Innovation Fund Sitra.

FinnGen projects regularly use ATLAS to build case-control cohorts for running genome and phenome wide association studies (GWAS and PheWAS) through custom tools. The OMOP CDM has also facilitated GWAS meta-analyses.

Although our vocabulary repository is private, we have built a public status dashboard that shows the vocabularies mapped [https://finngen.github.io/FinOMOP_OMOP_vocabulary/StatusReport/dashboard.html#databases-coverage]. Some of these are shared by nordic countries and we encourage them to contact us to contribute. To share our process and tools we have shared an example repository [https://github.com/FinOMOP/FinOMOP_OMOP_vocabulary_test], which can be cloned and used by

other members of the OHDSI community. The vocabulary maintenance is continuous and driven by the FinOMOP vocabulary working group.

Conclusion

The FinOMOP network provides a unique population-based resource for national and international collaborative work (research, HTA, competent authorities, health and social system performance assessment). Conforming to the OHDSI OMOP data model and infrastructure has provided a robust roadmap on the path to reliable, privacy-preserving and fit-for-purpose evidence generation, using federated analytics of well-curated data.

References

1. Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, et al. Swarm Learning for decentralized and confidential clinical machine learning. Nature. 2021;594(7862):265–70.
2. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023;613(7944):508–18.