Comparison of Cardiovascular Disease Trajectories: A France Vs Nordic countries' perspective under the HealthData@EU EHDS project*

Gayo Diallo[1,*], Arti Rawat[1,†], Anna-Leena Vuorinen[2,3,†], Rachel B. Forster[4,†], Adrian G. Zucco[5,†], Tibor V. Varga[5,†] and Andrea Ganna[3,†]

[1] *Team AHeaD BPH Inserm 1219, Univ. Bordeaux, F-33000, Bordeaux, France*
[2] *Unit of Health Sciences, Faculty of Social Sciences, Tampere University, Tampere, Finland*
[3] *Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland*
[4] *Department of Health Registry Research and Development, Norwegian Institute of Public Health, Bergen, Norway*
[5] *Copenhagen Health Complexity Center, Department of Public Health, University of Copenhagen*

Keywords
Data Analysis, Cardiovascular Disease, European Health Data Space, Machine Learning

1. Introduction

The HealthData@EU Pilot is a two-year project for building a Pilot version of the European Health Data Space (EHDS) infrastructure for the secondary use of health data (known as the "HealthData@EU") which will serve research, innovation, policy-making and regulatory purposes, testing the EHDS regulation proposal for secondary use [1]. The Pilot project lasted two years (October 2022 - October 2024) with three additional months till December 2024. It brought together 17 partners including health data access bodies, health data sharing infrastructures and European agencies. Further, to show the feasibility and possibility of reuse of data from multiple European countries, the HealthData@EU Pilot project ran five use cases that have been selected upon a recommendation by the European Commission [1]. The aim was to demonstrate the feasibility and potential impact of cross-border exploitation of health data for public health, research, innovation, and improvement of the health care system. Moreover, these use cases were used as a learning source to aid the development of services at both node and central service level.

The current study gives an overall description of the use case entitled "Cardiometabolic diseases" aiming at comparing care pathways for cardiometabolic diseases in European countries and building prediction models using artificial intelligence. More specifically, the use case addresses the issue of using nationwide health registry data in order to predict cardiometabolic disease trajectories using machine learning and make comparisons between three Nordic countries (Finland, Denmark, and Norway) and France. Due to time constraints, the use case focused on cardiovascular diseases (CVD) instead of cardio metabolic diseases.

SWAT4HCLS 2025, 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, June 07–11, 2022, Woodstock, NY

*Corresponding author.
gayo.diallo@u-bordeaux.fr (G. Diallo); arti.rawat@u-bordeaux.fr (A. Rawat); anna-leena.vuorinen@tuni.fi (A. Vuorinen); RachelBedenis.Forster@fhi.no (R. B. Forster); adrigabzu@sund.ku.dk (A. G. Zucco); tibor.varga@sund.ku.dk (T. V. Varga); andrea.ganna@helsinki.fi (A. Ganna)
https://www.gayodiallo.fr/ (G. Diallo)
0000-0002-0877-7063 (G. Diallo); 0000-0001-6597-5720 (A. Rawat); 0000-0002-5658-1305 (A. Vuorinen); 0000-0002-7952-3570 (R. B. Forster); 0000-0002-9827-1559 (A. G. Zucco); 0000-0002-2383-699X (T. V. Varga); 0000-0002-8147-240X (A. Ganna)

[1]https://ehds2pilot.eu/use-cases/

2. Methodology

2.1. Overview of the use case and aim

The overall goal was to use registry data from multiple European countries (Nordic countries, France) to instantiate a dictionary of cardiovascular disease endpoints from international diagnostic codes, evaluate and compare incidence across countries, and to build a trajectory-based prediction model comparing traditional statistical models to machine learning algorithms. In particular, it leveraged a machine-learning approach over longitudinal health trajectories to build prediction models for the 5-year risk of developing severe cardiovascular outcomes based on age, sex, diagnoses, and prescribed medications.

To do so, this use case aimed to address two questions :

- Are health trajectories leading to cardio metabolic diseases (i.e. cardiovascular diseases, type 2 diabetes and stroke) comparable across countries?
- Can we leverage longitudinal disease trajectories to forecast the risk of cardiometabolic diseases?

2.2. Data Sources from the different countries

Each of the four involved countries has individually identified the sources of routinely collected data in order to meet the objectives of the project (2.2). Listed below are the health registry each country identified as a necessary data source. The extracted data concerns all available inpatient WHO International Classification of Disease (10th version) (ICD-10) codes, causes of death, and medications from 01-Jan-2010 to 31-Dec-2018.

Table 1: Data sources and data size used by the different countries for the use case.

	France	**Finland**	**Denmark**	**Norway**
Inpatient encounters	National health data system, hospital discharge database, healthcare insurance system database	Care register for health care (Hilmo)	National Patient Register	Norwegian Patient Registry, Norwegian Cardiovascular Disease Registry
Size of dataset	12M	4M	4.5M	3.5M

2.3. Data Preprocessing

As an initial step in the use case, we developed a disease dictionary to define disease endpoints for a set of cardiovascular diseases based on register data. The purpose was to align the disease definitions across the four countries using ICD-10, while taking into account country-specific coding practices. The disease definitions were based on earlier work with clinical expert groups by communicating directly with clinicians, and based on the predefined analysis protocol inspired by Christensen, et al[2]. An initial data model was described in tabular format, which analysts used to transform their raw data. During this stage, tools were developed to create a common data model (CDM), using a Common Data Model Builder (CDMB) provided by one of the partners of the EHDS pilot project[2].

Quality control was performed per country and by comparing estimates across countries, leading to some reiterations of the data preparation in Finland and Norway. Finally, all countries had transformed their data into the CDM.

[2]https://github.com/cienciadedatosysalud/cdmb

2.4. Data Analysis

All data analysis was conducted using R. However, due to the large dataset for France (which involves data from 12 million individuals), Python (with PySpark) was used for data preprocessing.

We trained independent models in each country considering sex, age and counts of any diagnoses and prescribed medicines recorded in the period 2010 to 2013, inclusive, to predict the risk of a cardiovascular event in the period from 2014 until the end of 2018. The machine learning model used (Gradient boosted trees) was trained based on hyper parameters.

Because data privacy requirements, the same script was adapted and replicated in each node to perform the local analysis without moving the data.

3. Results

All separate models under the same conditions achieved similar performance across all countries (0.83-0.86 ROC AUC) when assessed on a country specific subset of the data kept for validation.

Different variables were relevant in each country, indicating country specific patterns of diagnoses and prescriptions even with comparable performance metrics. This highlights local differences not only in coding practices of the outcomes studied but also on potential indicator variables.

Table 2: Features and performance matrices of cardiovascular prediction models.

	France	**Finland**	**Denmark**	**Norway**
Number of features	1,732	524	493	602
Model performance				
ROC AUC	0.83	0.86	0.84	0.85
Brier Score	0.02	0.04	0.07	0.04

4. Discussion

4.1. Overview of lessons learned

Operating with nationwide registries across four European countries facilitated the study of cardiovascular problems at a population level, reducing the risk of selection bias. The register-based approach to explore disease trajectories and estimate risk enables the assessment of public health risks without requiring collection of clinical variables typically used in risk scores, such as blood pressure. The large sample size of almost 24 million individuals enabled the training of machine learning algorithms. This approach not only allows more precise risk estimation but also helps understand complex interactions and non-linear effects in the whole population. The larger dataset also provided more observations of rare outcomes, such as cardiovascular problems in young individuals.

4.2. Data preparation: computational challenges

One of the major challenges during data preparation and analysis (feature engineering and hyper-parameter optimization for the machine learning model in particular) was computational constraints. This was particularly relevant in France, where due to the large population and initial lack of specific adaptations to handle the volume data, running the scripts on the whole population required too much RAM and could not be achieved, but also in Norway, where extra computational resources were requested. To mitigate these issues, analysts worked with subsets of the data and developed the analysis with mock data created based on their compliance with the predefined data model. Furthermore, hyper-parameter optimization was performed in Denmark and France (e.g. using Spark workers in France) sharing the optimal parameters with the other countries.

4.3. Limitations and future work

Our approach to federated learning involved transferring scripts with a common schema between nodes, allowing for data processing in a standardized manner. This strategy deviates from traditional federated learning, where model training occurs locally at each node, and only model updates are exchanged. This choice was made due to the complexities of implementing true federated learning within the constraints of the European Health Data Space, as highlighted by the following challenges:

Data Access and Variation in Regulation. Major challenges in data access and usage were obtained in various countries. For example, in Norway, delays and uncertainties regarding data access permissions added to strict data security and privacy measures. The pre-existing data was to be reused in the use case. France imposed very strict data minimization rules that reduced the dataset from 66 million entries to 12 million entries. Denmark did struggle in establishing shared data sources, but generally, access to the data was easier than compared to other participating countries. This further reflects the huge differences between data governance and regulatory policies across European nations, a challenge towards joint data efforts.

Data Harmonization and Inpatient/Outpatient Discrepancies. The study initially aimed to include both inpatient and outpatient data for both types from four countries participating in the study. However, the difference in data collection methods for such data created coordination problems. The ambiguity over the very nature of the outpatient data collected in the French dataset added to the dilemma. Consistency and viability ensured that the project ultimately used only the inpatient data. This, therefore, underlines the crucial role that clear data definitions and harmonization procedures play in cross-border projects on health data.

Future works. The following are key areas identified for future development during this pilot project for the European Health Data Space:

- Transitioning to true federated learning: the next generations should consider implementing a true federated learning architecture where model training happens locally, and only model updates are shared. This will increase data privacy and reduce data transfer burdens.
- Improve performance and scalability: The system should be optimized for scalability and performance over numerous nodes and various network conditions. This will ensure robustness and efficiency as the European Health Data Space expands to incorporate more data partners.

Funding

The resources in the cardiometabolic use case are funded by the European Health and Digital Executive Agency (HADEA) ('EU executive agency' or 'granting authority'), under the powers delegated by the European Commission ('European Commission'), and it's relevant national partners in the EHDS2 pilot. Denmark: SUNDHEDSDATASTYRELSEN (DHDA), PIC 888310675, established in ORESTADS BOULEVARD 5, COPENHAGEN 2300, Denmark. Finland: TERVEYDEN JA HYVINVOINNIN LAITOS (Findata/THL), PIC 996697893, established in MANNERHEIMINTIE 166, HELSINKI 00271, Finland. France: PLATEFORME DES DONNEES DE SANTE (Health Data Hub), PIC 895824586, established in 9 RUE GEORGES PITARD, Paris 75015, France.

Tibor V. Varga is supported by the "Data Science Investigator - Emerging 2022" grant from the Novo Nordisk Foundation (NNF22OC0075284).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] European Commission, Communication from the commission - a european health data space: harnessing the power of health data for people, patients and innovation, COM(2022) 196/2, ???? [Internet]. Available from: https://health.ec.europa.eu/publications/communication-commission-european-health-data-space-harnessing-power-health-data-people-patients-and_en.

[2] D. M. Christensen, M. Phelps, T. Gerds, M. Malmborg, A.-M. Schjerning, J. E. Strange, M. El-Chouli, L. B. Larsen, E. Fosbøl, L. Køber, C. Torp-Pedersen, S. Mehta, R. Jackson, G. Gislason, Prediction of first cardiovascular disease event in 2.9 million individuals using danish administrative healthcare data: A nationwide, registry-based derivation and validation study, European Heart Journal Open 1 (2021) oeab015. URL: https://doi.org/10.1093/ehjopen/oeab015. doi:10.1093/ehjopen/oeab015.