Machine learning-based prediction of cardiovascular disease risk: A 5-Year forecast using 22 million data points from Nordic countries and France

Author
Arti Rawat – Team AHeaD BPH Inserm 1219, Univ. Bordeaux, F-33000, Bordeaux, France
Julien Bezin – Team AHeaD BPH Inserm 1219, Univ. Bordeaux, F-33000, Bordeaux, France
Adrian G. Zucco – Copenhagen Health Complexity Center, Department of Public Health, University of Copenhagen
Rachel B. Forster – Department of Health Registry Research and Development, Norwegian Institute of Public Health, Bergen, Norway
Tibor V. Varga – Copenhagen Health Complexity Center, Department of Public Health, University of Copenhagen
Andrea Ganna – Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
Anna-Leena Vuorinen – Unit of Health Sciences, Faculty of Social Sciences, Tampere University, Tampere, Finland; Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
Gayo Diallo – Team AHeaD BPH Inserm 1219, Univ. Bordeaux, F-33000, Bordeaux, France

Citation
Rawat, A., Bezin, J., Zucco, A.G., Forster, R.B., Varga, T.V., Ganna, A., Vuorinen, A-L., Diallo, G. Machine learning-based prediction of cardiovascular disease risk: A 5-Year forecast using 22 million data points from Nordic countries and France.

Introduction

The current study is one of the five use cases selected by the European Health Data Space (EHDS) as a part of a pilot program which will serve research, innovation, policy-making and regulatory purposes, testing the EHDS regulation proposal for secondary use of health data [1]. The EHDS is a regulatory framework designed to standardize the use of health data in the European Union (EU) by establishing a cross-border infrastructure and a secure, collaborative ecosystem where health data from EU member states can be shared for primary and secondary purposes [2].

It focuses on comparing the cumulative incidence and estimating the risk of cardiovascular disease (CVD) using a machine learning model applied to nationwide registry data from **France, Denmark, Finland, and Norway**. These countries were selected due to their **well-established national health registries**, **data availability. Hungary was originally planned to participate**, but due to a **national eHealth system reform and reorganization of responsibilities**, it was not able to join this phase of the study.

Methods

Health records of 22 million individuals aged between 18 to 85 years were extracted from national health registries in France, Denmark, Finland, and Norway between 2010 and 2018. The dataset included demographic variables (age, sex), diagnosis codes (ICD-10), and prescription data (ATC codes). A Gradient Boosted Decision Tree (GBDT) model was trained to predict the **5-year risk (2014–2018)** of a **composite CVD event**, defined as the occurrence of ischemic heart disease, myocardial infarction, angina pectoris, or ischemic stroke.

Input features included binary indicators for the presence or absence of diagnosis and prescription codes during the observation window. Age was treated as a continuous variable, while sex was binary. Features and preprocessing pipelines were harmonized across all countries, ensuring consistency. The same analysis script was executed in secure, federated environments for each country, using identical feature sets.

TABLE 1: Cohort characteristics

	French cohort n=11,744,558	Finnish cohort n=4,145,551	Danish cohort n=4,180,434	Norwegian cohort n=3,636,535
Sex, female, n (%)	6,327,682 (53.9%)	2,108,611 (50.9)	2,115,808 (50.6)	1,813,246 (49.9)
Age, years, mean (sd)	46.95 (17.32)	48.5 (17.6)	47.42 (17.24)	46.74 (17.23)

Model performance was evaluated using the area under the receiver operating characteristic curve (ROC-AUC) and Brier Score. Table 1 presents the cohort characteristics of all countries.

Results

The GBDT model performed similarly across all countries, achieving a ROC-AUC score of 0.83 for Denmark and France, 0.86 for Norway, and 0.87 for Finland. Calibration was also good, with a Brier score of 0.02 for France, 0.04 for both Finland and Norway, and 0.07 for Denmark. To justify the use of GBDT over simpler models, we trained a baseline logistic regression model using only age and sex. This model achieved lower ROC-AUC scores: 0.80 for France and Denmark, 0.85 for Finland, and 0.83 for Norway. These results confirm that GBDT offers improved discriminative performance, particularly in countries with more complex data distributions, supporting its suitability for cross-country CVD risk prediction.

The **CVD incidence per 100,000 person-years** is illustrated in **Figure 1**, which shows that **France** has a **lower incidence rate** than the Nordic countries. This is a significant finding, indicating that France may have **better CVD prevention**, **health services**, or **health-related behaviors** compared to Denmark, Finland, and Norway.

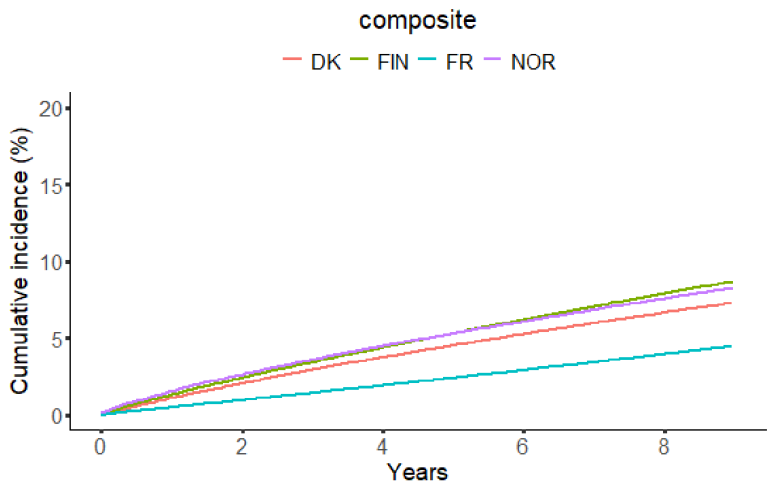

FIGURE 1
Cumulative Incidence curves of composite CVD.

Conclusion

This study provides a preliminary understanding of cross-country comparability, which can guide future research and more targeted investigations. However, future studies should consider the following:

1. Incorporating Time-Varying Covariates: The dynamic nature of medication usage should be taken into account.

2. Addressing Data Imbalance: Sampling techniques should be used to mitigate data imbalance, improving the efficiency of machine learning analysis.

This research lays the foundation for future health data-sharing initiatives and can contribute to better healthcare policy-making in Europe.