Check for updates

*Scandinavian Journal of Public Health*, 1–8

**ORIGINAL ARTICLE**

# Integrating data from multiple Finnish biobanks and national health-care registers for retrospective studies: Practical experiences

JAAKKO LÄHTEENMÄKI[1], ANNA-LEENA VUORINEN[2], JUHA PAJULA[2], KARI HARNO[3], MIKA LEHTO[4,5], MIKKO NIEMI[6,7] & MARK VAN GILS[2,8]

*[1]VTT Technical Research Centre of Finland Ltd (Espoo), Finland, [2]VTT Technical Research Centre of Finland Ltd (Tampere), Finland, [3]Department of Health and Social Management, University of Eastern Finland, Finland, [4]Heart and Lung Centre, Helsinki University Hospital, Finland, [5]University of Helsinki, Finland, [6]Department of Clinical Pharmacology and Individualised Drug Therapy Research Programme, University of Helsinki, Finland, [7]HUS Diagnostic Centre, Helsinki University Hospital, Finland, and [8]Tampere University, Finland*

**Abstract**

*Aim:* This case study aimed to investigate the process of integrating resources of multiple biobanks and health-care registers, especially addressing data permit application, time schedules, co-operation of stakeholders, data exchange and data quality. *Methods:* We investigated the process in the context of a retrospective study: Pharmacogenomics of antithrombotic drugs (PreMed study). The study involved linking the genotype data of three Finnish biobanks (Auria Biobank, Helsinki Biobank and THL Biobank) with register data on medicine dispensations, health-care encounters and laboratory results. *Results:* We managed to collect a cohort of 7005 genotyped individuals, thereby achieving the statistical power requirements of the study. The data collection process took 16 months, exceeding our original estimate by seven months. The main delays were caused by the congested data permit approval service to access national register data on health-care encounters. Comparison of hospital data lakes and national registers revealed differences, especially concerning medication data. Genetic variant frequencies were in line with earlier data reported for the European population. The yearly number of international normalised ratio (INR) tests showed stable behaviour over time. ***Conclusions:*** **A large cohort, consisting of versatile individual-level phenotype and genotype data, can be constructed by integrating data from several biobanks and health data registers in Finland. Co-operation with biobanks is straightforward. However, long time periods need to be reserved when biobank resources are linked with national register data. There is a need for efforts to define general, harmonised co-operation practices and data exchange methods for enabling efficient collection of data from multiple sources.**

**Keywords:** *Biobank, health-care data, register data, real world data, genotype, secondary use, pharmacogenetics, precision medicine*

## Introduction

Exploitation of health data resources is essential to increase the efficiency and quality of health-care and to facilitate development of innovative therapies and pharmaceutical products. Considerable efforts are ongoing globally to enable efficient and secure use of health data [1–6]. Finland is well-positioned to benefit from health data resources. The nationwide centralised Kanta Services provide a standard-based architecture for electronic prescriptions and storage of patient records and social services data [7]. Data on service episodes are systematically collected in national registers for statistical and research purposes [8,9]. Biological samples and related data are collected by 10 public biobanks and one private biobank [10]. Finnish national legislation includes a unique framework facilitating secondary use of data for legitimate purposes ranging from basic research to R&D activities of companies and monitoring of service processes and quality. The Biobank Act, in force

Correspondence: Jaakko Lähteenmäki, VTT Technical Research Centre of Finland Ltd, PO Box 1000, VTT, 02044 Espoo, Finland.
E-mail: jaakko.lahteenmaki@vtt.fi

*Date received 1 February 2021; reviewed 1 February 2021; accepted 1 March 2021*

© Author(s) 2021
Article reuse guidelines: sagepub.com/journals-permissions
DOI: 10.1177/14034948211004421
journals.sagepub.com/home/sjp
SAGE

since 2013, defines the conditions for using biological samples and related data [11] and provides a coherent basis for biobank services. The Act on the Secondary Use of Health and Social data [12], in force since 2019, defines a new organisation (Findata) as responsible for national services, including data access and permit services, as well as infrastructures for safe processing of data in compliance with the General Data Protection Regulation [13]. Additionally, the Finnish Biobank co-operative (FinBB) provides joint services targeted at helping to access the resources of Finnish biobanks.

The usage of longitudinal real-world health data has been shown to be a powerful tool to identify clinical associations and biomarkers [14,15]. However, several challenges of data usage have been identified in relation to ethical issues [16], data accuracy [17] and combining data across national borders [18]. In particular, the performance of the Finnish biobanks has recently been evaluated in the context of cancer research and challenges related to time schedules and the availability of competent medical and technical resources have been identified [10]. The need for wider and more comprehensive use of health and social data has been expressed [19].

The Biobank Act enables the biobanks to link their materials (including genome data) with phenotype data in patient records and national registers. Consequently, biobanks have a central role in the formation of research cohorts linking data from multiple sources. The purpose of this case study was to investigate the process of combining data resources, especially addressing data permit application, time schedules, co-operation of stakeholders, data exchange and data quality. The Findata and FinBB services were not yet available during the data collection phase. Thus, the results of this case study provide baseline information for future evaluation of these new services.

## Methods

### *Study case*

We explored the process of exploiting data from several sources, that is, multiple biobanks and national registers, in the context of a retrospective study – Pharmacogenomics of antithrombotic drugs (hereafter 'PreMed study'), registered at clinicaltrials.gov (NCT04001166) [20]. The objective of the PreMed study is to assess the relevance of using genotype data in the context of antithrombotic drug therapy by investigating genotype associations of antithrombotic drugs in the Finnish population. The study focuses on warfarin, as a significant proportion of the population (36% for *CYP2C9* alleles in the Finnish population [21]) carries genetic variants which impair the metabolism of the drug, potentially causing an elevated risk for bleeding complications. Additionally, the PreMed study includes explorative investigation of genotype–phenotype associations for a larger group of antithrombotic drugs. The PreMed study has been approved by the ethics committee of the Hospital District of Helsinki and Uusimaa. Results of the PreMed study on genotype–phenotype associations have been published separately [20].

Table I. Data sources and corresponding data types in the PreMed study.

| Data source | Data type |
|---|---|
| Auria Biobank | Genotype data, clinical data |
| Helsinki Biobank | Genotype data, clinical data |
| THL (Hilmo/Avohilmo registers) | Data on health-care encounters |
| THL Biobank | Genotype data, demographic data |
| Kela (Finnish Prescription Register) | Data on medicine dispensations |
| Hospital districts and municipalities | Laboratory data |

### *Data sources*

The initial analysis revealed that with the contribution of three Finnish biobanks (Auria Biobank, Helsinki Biobank and THL Biobank), enough genotyped subjects could be included in the study to meet the required statistical power. The biobanks provided genotype data, and these were combined with corresponding phenotype data retrieved from different health-care registers. The data sources used are listed in Table I. Data were collected for the period from 1 January 2005 to 31 December 2018.

Helsinki Biobank and Auria Biobank are public, hospital-based biobanks jointly covering the area of seven hospital districts, with approximately three million residents in southern and western Finland. Both biobanks have access to patients' clinical data, including diagnoses, laboratory results, procedures and medications documented in the context of specialised care encounters in various information systems (e.g. electronic health record systems (EHRs)) in the hospital districts. These data, referred to as 'hospital data', are accessible via hospital data lakes and can be linked with the biobanks' sample and data resources.

The Finnish Institute for Health and Welfare (THL) has a statutory role in carrying out studies and developing instruments to monitor the well-being and health of the Finnish population. The PreMed study used the Register of Primary Health

Care Visits (Avohilmo) and the Care Register for Health Care (Hilmo) of THL, which contain healthcare encounter data from primary and specialised care episodes, respectively. The THL Biobank is an administrative part of THL, with responsibility for collecting and storing biological sample collections and survey data for research. Cohorts used in this study from the THL Biobank were the National FINRISK Study [22] and the Health 2000 and Health 2011 surveys [23].

The Social Insurance Institution of Finland (Kela) is a government agency that provides basic economic security for people living in Finland. Kela maintains several national health and social data registers, which are available for administrative and scientific research purposes.

Laboratory results are stored in laboratory systems of hospital districts and municipalities from where they can be retrieved for scientific research with the permission of THL. The processor of the laboratory databases was Mylab, Inc., on behalf of the hospital districts and municipalities as data controllers.

At the time of the PreMed data collection, the permits to access national register data needed separate requests for each register controller. As stipulated by the new legislation [12], these permits are currently granted by Findata.

### *Data collection process*

The process adopted for data collection is presented in Figure 1. To minimise the need for exchanging identifiable health data between the biobanks, each biobank created their own sub-cohort independently and delivered it in a jointly pseudonymised form to VTT Technical Research Centre of Finland. We further combined the sub-cohorts into the final cohort.

## Results and discussion

### *Project timeline and resources*

We started the data collection process in January 2019 by submitting the ethical review application to the ethics committee of the Hospital District of Helsinki and Uusimaa, followed by data requests to the three biobanks as indicated in Figure 1. We then submitted the data permit applications to the national registers (THL and Kela) in April 2019. After required iterations to complete the applications, all three biobank data requests were accepted in May, and the respective Material Transfer Agreements (MTAs) were iterated during summer 2019. After the THL register approval in December 2019, the data were delivered by THL and Kela to biobanks in February–March 2020. The formation and delivery of the sub-cohorts was completed by the end of April, and we were able to complete the preparation of the final study cohort by the end of May 2020. Altogether, the data collection phase took 16 months. The data costs, including applications and data extraction of biobank and national register data, were approximately €80,000. The required effort from VTT researchers was approximately five person-months invested in data permit applications, contracts and curation (e.g. data collecting, integrating and verification).

Our pre-estimate for the duration of the data collection phase was nine months based on the information available in the planning phase. Substantial delay to the data collection was caused by the congested register data permit approval process of THL, which took about eight months. As the acceptance of all data applications was set as a precondition in the MTAs of THL and Helsinki biobanks, the completion of the full cohort was directly affected by this delay. There was no such precondition included in the Auria Biobank MTA, enabling earlier delivery of partial data.

### *Co-operation of stakeholders*

As illustrated in Figure 1, each biobank formed their own pseudonymised sub-cohort and delivered it to VTT. None of the biobanks had the leading role during the process. Our observation from the process is that established mechanisms to execute projects with multiple biobanks are only just emerging because most biobank projects are currently still limited to only a single biobank. Nevertheless, co-operation with the biobanks was good throughout the data collection, and co-operation practices were successfully agreed between the parties during the process.

Joint meetings were organised with the biobanks to monitor the progress, to define data formats and to share methods and tools between the biobanks for cohort formation. The pseudonymisation service first implemented by Auria Biobank was also used by the other two biobanks. The shared pseudonymisation algorithm enabled data delivery to VTT without direct personal identifiers while maintaining the possibility of detecting if the same individuals were present in more than one sub-cohort. However, no such overlap was found.

To ensure compatibility between the sub-cohorts, common data structures for genome and hospital data were agreed upon between the biobanks following the existing templates of THL Biobank and Auria Biobank, respectively. Each biobank received register data from THL and Kela in the same formats, which

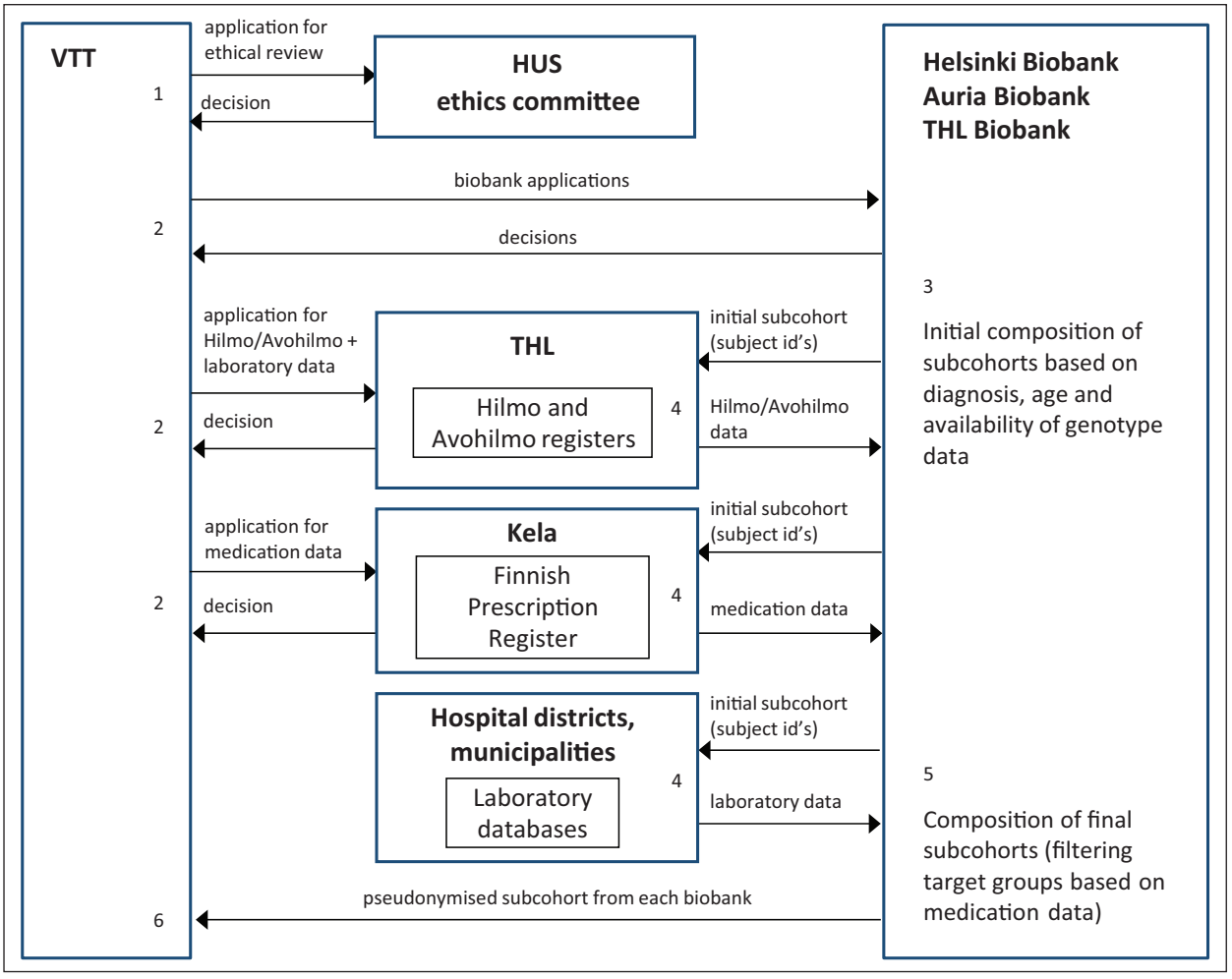


Figure 1. Pharmacogenomics of antithrombotic drugs (PreMed) study data collection process with phases: (1) ethical review, (2) applications to data controllers, (3) initial sub-cohort composition, (4) composition of national register and laboratory data sets based on initial sub-cohorts, (5) composition of final sub-cohorts and (6) composition of the final cohort.
VTT: Technical Research Centre of Finland; HUS: Hospital District of Helsinki and Uusimaa; THL: Finnish Institute of Health and Welfare; Kela: The Social Insurance Institution of Finland.

eliminated the need for further harmonisation between the sub-cohorts.

Based on the experience, the process could be made more efficient if data formats and related sample files were publicly available and could be accessed in the study design phase. This would enable the biobank customer to design and implement the data pipeline to be ready while waiting for the approvals of the data permit applications. The complexity of the data pipeline was increased by the fact that the formats between hospital data and national registers are different. Additionally, there are differences between the formats of Hilmo and Avohilmo, as well as between the laboratory test codes between the hospital districts. The overall observation is that there is a clear need for harmonisation of data formats for secondary use of health data and biobanking.

Table II. Estimated and actual number of subjects in the PreMed study.

| | Pre-estimate | | Actual | |
|---|---|---|---|---|
| | Warfarin sub-study | Full study cohort | Warfarin sub-study | Full study cohort |
| Auria Biobank | 301 | 666 | 353 | 839 |
| Helsinki Biobank | 735 | 1626 | 636 | 1521 |
| THL Biobank | 1642 | 3632 | 2171 | 4645 |
| total | 2678 | 5924 | 3160 | 7005 |

### *Cohort size*

Table II includes the initially estimated and actual number of subjects obtained from the three biobanks. The cohort sizes are separately shown for the two PreMed objectives: the sub-study that included warfarin-treated patients only and the explorative

Table III. Comparison of hospital data (hosp.) with national register data (reg.) of medicine dispensations from Kela and diagnoses from THL register for the period 2007–2018.

| | Auria Biobank sub-cohort | | | Helsinki Biobank sub-cohort | | |
|---|---|---|---|---|---|---|
| | Patients with data hosp./ reg. | Data non-existent hosp./reg. | Data unmatched | Patients with data hosp./ reg. | Data non-existent hosp./reg. | Data unmatched |
| *Drug or diagnosis* | | | | | | |
| Warfarin | 316/356 | 13.8%/2.8% | 46.9% | 403/656 | 40.2%/2.7% | 70.9% |
| Dabigatran | 46/61 | 34.4%/13.0% | 62.5% | 69/79 | 40.5%/31.9% | 76.6% |
| Rivaroxaban | 192/210 | 17.6%/9.9% | 38.7% | 162/221 | 34.4%/10.5% | 53.8% |
| Apixaban | 84/90 | 15.6%/9.5% | 64.5% | 168/164 | 15.2%/17.3% | 52.5% |
| Enoxaparin | 592/327 | 11.6%/51.2% | 50.9% | 1080/722 | 25.5%/50.2% | 55.6% |
| Dalteparin | 110/39 | 15.4%/70.0% | 12.1% | 43/99 | 74.7%/41.9% | 40.0% |
| Clopidogrel | 253/262 | 8.8%/5.5% | 15.9% | 291/432 | 36.6%/5.8% | 31.4% |
| Ticagrelor | 20/18 | 5.6%/15.0% | 11.8% | 151/142 | 20.4%/25.2% | 20.4% |
| All drugs (mean) | 202/170 | 13.3%/19.0% | 41.1% | 295.9/314.4 | 30.3%/22.2% | 53.0% |
| *Diagnosis group* | | | | | | |
| Bleeding outcome | 139/163 | 16.0%/1.4% | 7.3% | 252/308 | 19.8%/2.0% | 6.5% |
| Thromboembolic outcome | 237/236 | 1.7%/2.1% | 6.0% | 463/483 | 5.8%/1.7% | 7.7% |
| Atrial fibrillation | 395/399 | 2.8%/1.8% | 15.5% | 617/670 | 8.8%/1.0% | 23.2% |
| Vascular disease | 377/395 | 5.1%/0.5% | 16.3% | 728/732 | 3.8%/3.3% | 20.2% |
| Pulmonary embolism | 74/73 | 0.0%/1.4% | 1.4% | 150/157 | 6.4%/2.0% | 8.8% |
| Stroke | 164/164 | 3.0%/3.0% | 7.5% | 315/326 | 5.2%/1.9% | 6.5% |
| Phlebitis | 93/97 | 7.2%/3.2% | 4.4% | 172/171 | 5.8%/6.4% | 6.2% |
| Neoplastic disorder | 270/278 | 3.2%/0.4% | 16.4% | 446/479 | 7.9%/1.1% | 22.2% |
| All diagnoses (mean) | 219/226 | 4.5%/1.5% | 12.0% | 392.9/415.8 | 7.5%/2.1% | 15.5% |

investigation of a larger group of antithrombotic drugs (full cohort). The estimated numbers were based on information received from the biobanks before data applications were submitted. In the end, the actual number of subjects was higher than was initially predicted, so that targeted statistical power was achieved. The underestimation of the cohort sizes may be explained by the limited coverage of drug data in the hospital data lakes (see next section).

### *Medication and diagnosis data*

As the Auria Biobank and Helsinki Biobank sub-cohorts included overlapping data from hospital data lakes and national registers for the same subjects, it was possible to compare these data sources. Table III shows comparisons on data coverage for eight antithrombotic drugs and eight diagnosis groups. The table indicates the number of patients for whom medication data and diagnosis data were available in both hospital data lakes and national registers and the proportion of patients with non-existent or unmatched medication data and diagnosis data. Data were deemed non-existent in the hospital data if the information was available in the register data but was not documented or was missing in the hospital data. Data were deemed non-existent in the register data if the information was available in the hospital data but was not documented or was missing in the register data. Data were defined as unmatched if they were documented in both data sets (hospital and register) but there was a difference of more 30 days in the dates of the first documentation entry (the date of medication initiation or the date of the first diagnosis).

Limited coverage of hospital medication data was expected based on the a priori information communicated by the biobanks to us. The expected reason is that medication data have been stored in various systems and formats during the years, and it is difficult to retrieve data for use via the data lake systems. Also, the practices in documenting medication use in EHRs vary. Especially for the early phase of the study period (2007–2014), a large part of medication data was not documented in the hospital data sets. Restricting the data sets to the years 2015–2018 reduces the share of non-existent data from 13.3% to 10.2% and from 30.3% to 8.7% for Auria and Helsinki sub-cohorts, respectively. Similarly, this limitation reduces the share of unmatched medication data from 41.1% to 37.3% and from 53.0% to 29.1% for Auria and Helsinki sub-cohorts, respectively. Non-existent hospital medication data may also partly result from antithrombotic drug therapy provided solely in the context of primary care episodes, which are not covered by hospital data but are covered by Kela medication data. A high number of dalteparin and enoxaparin users were not identified

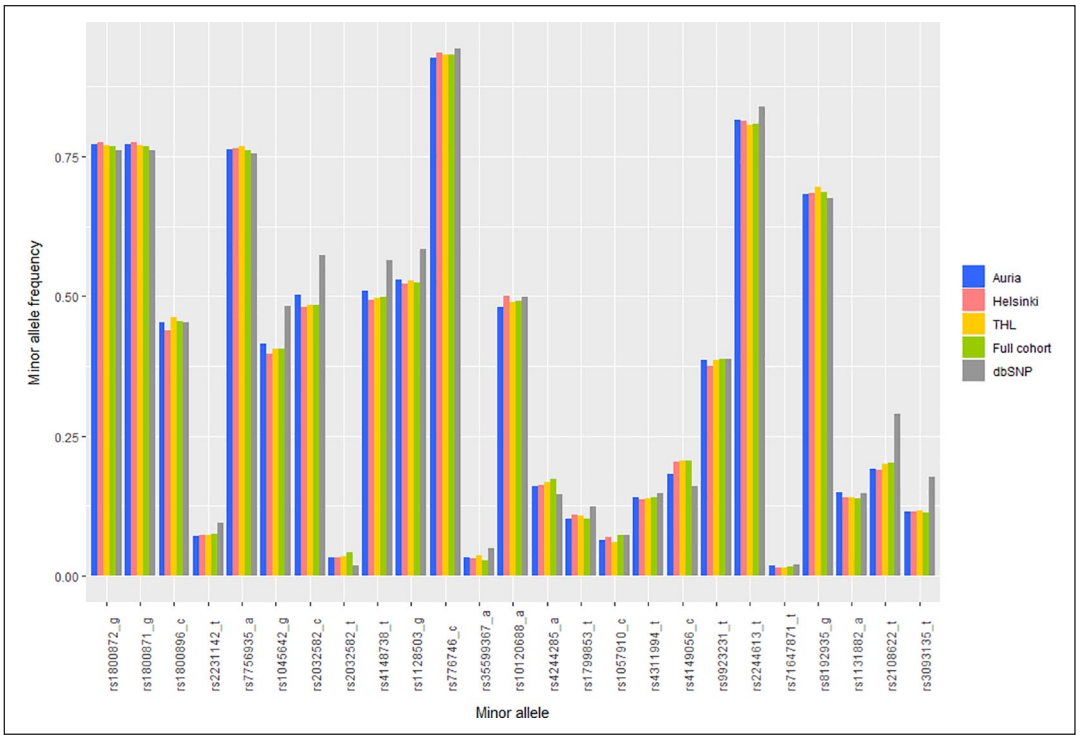


Figure 2. Minor allele frequencies for the three sub-cohorts (Auria Biobank, Helsinki Biobank and THL Biobank), the combined full cohort and earlier reported data for the European population (dbSNP).

from the Kela registers because the drugs are administered by subcutaneous injection and are often delivered during hospitalisation, in which case they do not appear as medicine dispensations in the Kela register.

Concerning both diagnosis and medication data, differences may also result from cases when the individual has moved in or out of the hospital district during the study period. Data from those subjects may not appear in the hospital data but are covered by the national registers – information on place of residence changes was not available for the study. In general, the differences between hospital data and national register data were smaller for diagnoses, which might follow from the fact that both data sources cover specialised care episodes only.

### *Genotype data*

Resulting from the ongoing FinnGen project [24] and previous THL population study projects, a large set of harmonised genome data (population of 500,000 genotyped donors to be covered by 2023) is currently available for research through the biobanks. Based on existing research, we selected 26 single nucleotide polymorphisms (SNPs) to be investigated in the study. We defined the availability of *CYP2C9* (rs1799853, rs1057910) and *VKORC1* (rs9923231) alleles as an inclusion criterion to ensure coverage of the most relevant variants related to warfarin. All but two of the 26 SNPs were available for the study from the biobanks. The two missing SNPs (rs4986893 and rs3093235) were planned to be investigated in the explorative part of the study.

The genotype data were imputed. Imputation quality score varied in the range 0.898–1.000, with a mean value of 0.995. The allele frequencies for the full cohort and the three sub-cohorts are shown in Figure 2. Allele frequencies appear to be closely similar in all sub-cohorts, and there are no major differences from the European population [25]. The genotype data were observed to be in Hardy–Weinberg equilibrium based on a chi-square test ($p$-values for SNPs varying in the range 0.07–0.99).

### *Laboratory data*

We included results of 12 laboratory tests in the PreMed study protocol. Test results were available for a subset of 92.5% of the full cohort. Most importantly, international normalised ratio (INR) tests are used to monitor the effectiveness and safety of the warfarin therapy. The laboratory tests are also important indicators of actual drug use. We assessed the completeness of the INR data by investigating the number of tests per patient per year for the three sub-cohorts

during the period 1 January 2007–31 December 2018. The medians (interquartile ranges) of the annual number of tests were 24.5 (18.4–35.8), 25.4 (19.2–35.6) and 23.0 (17.1–33.4) for the Auria, Helsinki and THL Biobank sub-cohorts, respectively. The number of tests is in line with the current care guidelines for warfarin therapy and previous literature [26–28]. Linear regression analysis showed a decreasing trend of −0.48 (95% confidence interval (CI) −1.35 to 0.38) and −0.46 (95% CI −0.93 to 0.02) for the annual number of tests per patient in the Helsinki and THL Biobank sub-cohorts, respectively, while the Auria Biobank cohort showed an increasing trend of 0.26 (95% CI −1.34 to 1.85) tests per patient. Based on the regression analysis, the yearly number of INR tests has not considerably changed during the observation period, suggesting that there was no severe distortion in the laboratory INR data sets.

The use of specific pharmacogenetic tests is still negligible in clinical practice. Pharmacogenetic test results were only available for seven of the warfarin users.

## Conclusions

Our case study shows that a large cohort consisting of versatile individual-level phenotype and genotype data, for the purpose of a retrospective pharmacogenetic study, can be constructed by integrating data from several biobanks and health data resources in Finland.

The main challenge was the duration of the data collection process, which was 16 months in total. Especially from the point of view of industry-driven research, the process should be shorter, less labour intensive and more predictable. Most of the delays in the process can be attributed to the extraction of data from the national registers and related to data permit applications. These delays may be shorter in the future once the recently established Findata services are in full operation. Co-operation with the biobanks during the data collection process was good, and the project benefitted from open sharing of tools, methods and information between the parties.

The characteristics of the obtained study cohort were well in line with the data specification in our original research protocol, and the number of subjects exceeded the expected number of participants. We performed tests on the data sets to indicate possible errors or inconsistency. In general, we observed the data quality to be good. As expected, differences between hospital data lake records and national register data existed, especially for medication. The results suggest an improving trend in the medication data accuracy towards the end of the study period.

The study experiences indicate a need for a predefined general model for executing projects with contributions from multiple biobanks and national register controllers. This model should cover practices for co-operation and harmonised data structures for data access and exchange [29]. Initiation and acceleration of international level efforts, such as the creation of the European Health Data Space [30], is important, as there is also a rapidly growing need to facilitate data collection across country borders.

## Acknowledgements

The study used data obtained from Auria Biobank (study number: AB19-9833), Helsinki Biobank (study number: HBP20190038) and THL Biobank (study number: BB2019_6). We thank all study participants for their generous participation in the biobank research. We also thank Merja Perälä, Perttu Terho and Mikko Tukiainen of Auria Biobank; Theresa Knopp, Miika Koskinen and Otto Manninen from Helsinki Biobank; and Niina Eklund, Anni Joensuu and Katariina Peltonen from THL Biobank for their contribution in defining the sub-cohorts for the study. We also thank the contribution of Medaffcon Oy researchers, Tanja Nieminen and Maija Wolf, for support in defining the data needs for health-care resource usage evaluation.

## Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This study was funded by Business Finland, VTT Technical Research Centre of Finland Ltd, Karl Fazer AB, Novartis Finland Oy, Pfizer Oy, Roche Diagnostics Oy, Avaintec Oy, Crown CRO Oy, Mediconsult Oy and Biobank Cooperative Finland.

## ORCID iDs

Jaakko Lähteenmäki https://orcid.org/0000-0002-9358-6332

Mika Lehto https://orcid.org/0000-0002-8691-5142

## References

[1] Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform* 2014;52:28–35.

[2] Vayena E, Dzenowagis J, Brownstein JS, et al. Policy implications of big data in the health sector. *Bull World Health Organ* 2018;96:66–8.

[3] Nalin M, Baroni I, Faiella G, et al. The European cross-border health data exchange roadmap: case study in the Italian setting. *J Biomed Inform* 2019;94:103183.
[4] Argudo-Portal V and Domènech M. The reconfiguration of biobanks in Europe under the BBMRI-ERIC framework: towards global sharing nodes? *Life Sci Soc Policy* 2020;16:1–15.
[5] Pastorino R, De Vito C, Migliara G, et al. Benefits and challenges of big data in healthcare: an overview of the European initiatives. *Eur J Public Health* 2019;29:23–27.
[6] Hopp WJ, Li J and Wang G. Big data and the precision medicine revolution. *Prod Oper Manag* 2018;27:1647–64.
[7] Aarnio E, Huupponen R, Martikainen JE, et al. First insight to the Finnish nationwide electronic prescription database as a data source for pharmacoepidemiology research. *Res Soc Adm Pharm* 2020;16:553–9.
[8] Gissler M and Haukka J. Finnish health and social welfare registers in epidemiological research. *Nor Epidemiol* 2004;14:113–20.
[9] Sund R. Quality of the Finnish Hospital Discharge Register: a systematic review. *Scand J Public Health* 2012;40:505–15.
[10] Vesterinen T, Salmenkivi K, Mustonen H, et al. Performance of Finnish biobanks in nationwide pulmonary carcinoid tumour research. *Virchows Arch* 2020;476:273–83.
[11] Finnish Biobank Act, https://www.finlex.fi/en/laki/kaannokset/2012/en20120688.pdf (accessed 27 January 2021).
[12] Secondary use of health and social data – The Ministry of Social Affairs and Health, https://stm.fi/en/secondary-use-of-health-and-social-data (accessed 20 January 2021).
[13] General Data Protection Regulation, GDPR (EU) 2016/679 https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed 27 January 2021).
[14] Peippo MH, Kurki S, Lassila R, et al. Real-world features associated with cancer-related venous thromboembolic events. *ESMO Open* 2018;3:363.
[15] Hernesniemi JA, Mahdiani S, Tynkkynen JA, et al. Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome – the MADDEC study. *Ann Med* 2019;51:156–63.
[16] Andreassen OA. eHealth provides a novel opportunity to exploit the advantages of the Nordic countries in psychiatric genetic research, building on the public health care system, biobanks, and registries. *Am J Med Genet B: Neuropsychiatr Genet* 2018;177:625–9.
[17] Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. *Am J Med Genet B Neuropsychiatr Genet* 2018;177:601–12.
[18] Maret-Ouda J, Tao W, Wahlin K, et al. Nordic registry-based cohort studies: possibilities and pitfalls when combining Nordic registry data. *Scand J Public Health* 2017;45:14–9.
[19] Kilpeläinen K, Parikka S, Koponen P, et al. Finnish experiences of health monitoring: local, regional, and national data sources for policy evaluation. *Glob Health Action* 2016;9:28824.
[20] Vuorinen A-L, Lehto M, Niemi M, et al. Pharmacogenetics of anticoagulation and clinical events in warfarin-treated patients: A register-based cohort study with biobank data and national health registries in Finland. *Clin Epidemiol* 2021;13:183–195.
[21] Sistonen J, Fuselli S, Palo JU, et al. Pharmacogenetic variation at CYP2C9, CYP2C19, and CYP2D6 at global and microgeographic scales. *Pharmacogenet Genomics* 2009;19:170–9.
[22] Borodulin K, Vartiainen E, Peltonen M, et al. Forty-year trends in cardiovascular risk factors in Finland. *Eur J Public Health* 2015;25:539–46.
[23] Aromaa A and Koskinen S. Health and functional capacity in Finland: baseline results of the Health 2000 health examination survey, www.julkari.fi/handle/10024/78534 (2004, accessed 27 January 2021).
[24] FinnGen. FinnGen research project, https://www.finngen.fi/en (accessed 27 January 2021).
[25] National Center for Biotechnology Information – dbSNP, https://www.ncbi.nlm.nih.gov/snp/ (accessed 20 January 2021).
[26] January CT, Wann LS, Calkins H, et al. 2019 AHA/ACC/HRS focused update of the 2014 AHA/ACC/HRS Guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society in collaboration with the Society of Thoracic Surgeons. *Circulation* 2019;140:e125–51.
[27] Dlott JS, George RA, Huang X, et al. National assessment of warfarin anticoagulation therapy for stroke prevention in atrial fibrillation. *Circulation* 2014;129:1407–14.
[28] Lehto M, Niiranen J, Korhonen P, et al. Quality of warfarin therapy and risk of stroke, bleeding, and mortality among patients with atrial fibrillation: results from the nationwide FinWAF Registry. *Pharmacoepidemiol Drug Saf* 2017;26:657–65.
[29] Klann JG, Joss MAH, Embree K, et al. Data model harmonization for the All Of Us Research Program: transforming i2b2 data into the OMOP common data model. *PLoS One* 2019;14:e0212463.
[30] European Health Data Space. Public health, https://ec.europa.eu/health/ehealth/dataspace_en (accessed 27 January 2021).