RESEARCH REPORT

VTT-R-01225-19

14.2.2020

Data-driven precision medicine ecosystem

PreMed phase 2 report

Authors: Jaakko Lähteenmäki (editor), Richard Fagerström, Mark van Gils, Juha Pajula, Tomi Sorasalmi, Anna-Leena Vuorinen, Peter Ylén

Confidentiality: Public

Report's title Data-driven precision medicine ecosystem - PreMed phase 2 report	
Customer, contact person, address Outi Tuovila Business Finland, Porkkalankatu 1	**Order reference** 3458/31/2017
Project name Business ecosystem for Finnish precision medicine platform	**Project number/Short name** 115511 / PreMed
Author(s) Jaakko Lähteenmäki (editor), Richard Fagerström, Mark van Gils, Juha Pajula, Tomi Sorasalmi, Anna-Leena Vuorinen, Peter Ylén	**Pages** 43/
Keywords precision medicine, personalized medicine, secondary use, artificial intelligence, genome technology, biobank, business ecosystem	**Report identification code** VTT-R-01225-19

Summary

This report summarizes PreMed project activities carried out in phase 2 (1.11.2018 - 31.12.2019). The activities have been focused in three main areas: biobank study (research protocol, data collection, data analysis environment setup), ecosystem simulation model and dissemination (project workshops and other events). Through the retrospective study combining data resources from three biobanks and national registries the PreMed project aims to provide evidence on the benefits of genetic testing in guiding drug therapy. In general, PreMed provides information and experiences of the processes related to exploiting health data resources. This report outlines the research study setting and provides a summary of the study status and experiences from the data collection phase.

The ecosystem model is targeted for simulation of various evolution paths of the data-driven precision medicine ecosystem. In particular, the simulation model aims to provide support for the selection of public financing strategies to boost ecosystem growth. The model development and its parameter setting is still under development. The report provides an overview of the model and some initial simulation results. The project has organized three project workshops for collecting and sharing information between project partners. Collected information on national developments and examples of international activities are summarized in the report.

Confidentiality	Public

Espoo 14.2.2020

Written by	**Reviewed by**	**Accepted by**
Jaakko Lähteenmäki Principal Scientist	Kari Kohtamäki Key Account Manager	Jari Ahola Research Team Leader

VTT's contact address

Jaakko Lähteenmäki, jaakko.lahteenmaki@vtt.fi

Distribution (customer and VTT)

Final version to be available at VTT website (www.vtt.fi/premed).

Contents

1. Introduction

Personal data is increasingly collected and used in the context of digitalized services. Consequently, the amount of personal data stored in information systems is increasing exponentially. Besides the primary purpose, collected personal data is used for so called secondary purposes, such as for monitoring the quality of provided services, in the development of new services or products and for scientific research. In general, secondary use is permitted under certain conditions by the existing privacy legislation, in particular, the General Data Protection Regulation (GDPR). Additionally, specific legislation addressing secondary use of health data (e.g. in the context of biobanks) has been implemented in some countries, including Finland[1,2].

VTT initiated the PreMed project in 2017 with the objective of promoting the development of a data-driven precision medicine ecosystem in Finland. In particular, the project aims at collecting different kinds of companies together with the common objective of exploiting the opportunities provided by health data resources. In the first phase of the project (1.5.2017-31.10.2018) VTT carried out a series of interviews for better understanding of the expectations of companies towards exploitation of data and for identifying the bottlenecks and challenges faced today. The results of the interviews and the analysis of the current status of data-driven precision medicine landscape were compiled into the PreMed phase 1 report[3].

For phase 2 industrial partners joined the project so that the current consortium consists of: Avaintec, FinBB, Crown CRO, Fazer, Medaffcon, Mediconsult, Novartis, Pfizer, Roche Diagnostics and VTT. The project is co-financed by the project partners and Business Finland. The project partners represent different roles of the data-driven precision medicine ecosystem as indicated in Figure 1.

In phase 2 (1.11.2018 - 31.12.2019) the main project activity has been focused in preparing a retrospective cohort study on pharmacogenomics (PreMed PGx study). The objective of the study is to use data from biobanks and national registries to assess the relevance of using genotype data in the context of antithrombotic drug therapy. Besides the scientific objective, the biobank study is expected to provide valuable information and experience on data access processes and related bottlenecks. Additionally, the project has developed a system dynamics model for simulating alternative precision medicine ecosystem development scenarios. This report, summarizes the activities and obtained results of phase 2 of the PreMed project. The third and final phase of the project will be carried out during the year 2020.

[1] Biobank Act, https://www.finlex.fi/fi/laki/kaannokset/2012/en20120688.pdf
[2] Act on the secondary use of health and social data, https://stm.fi/en/secondary-use-of-health-and-social-data
[3] PreMed, phase 1 report, https://www.vtt.fi/sites/premed/deliverables

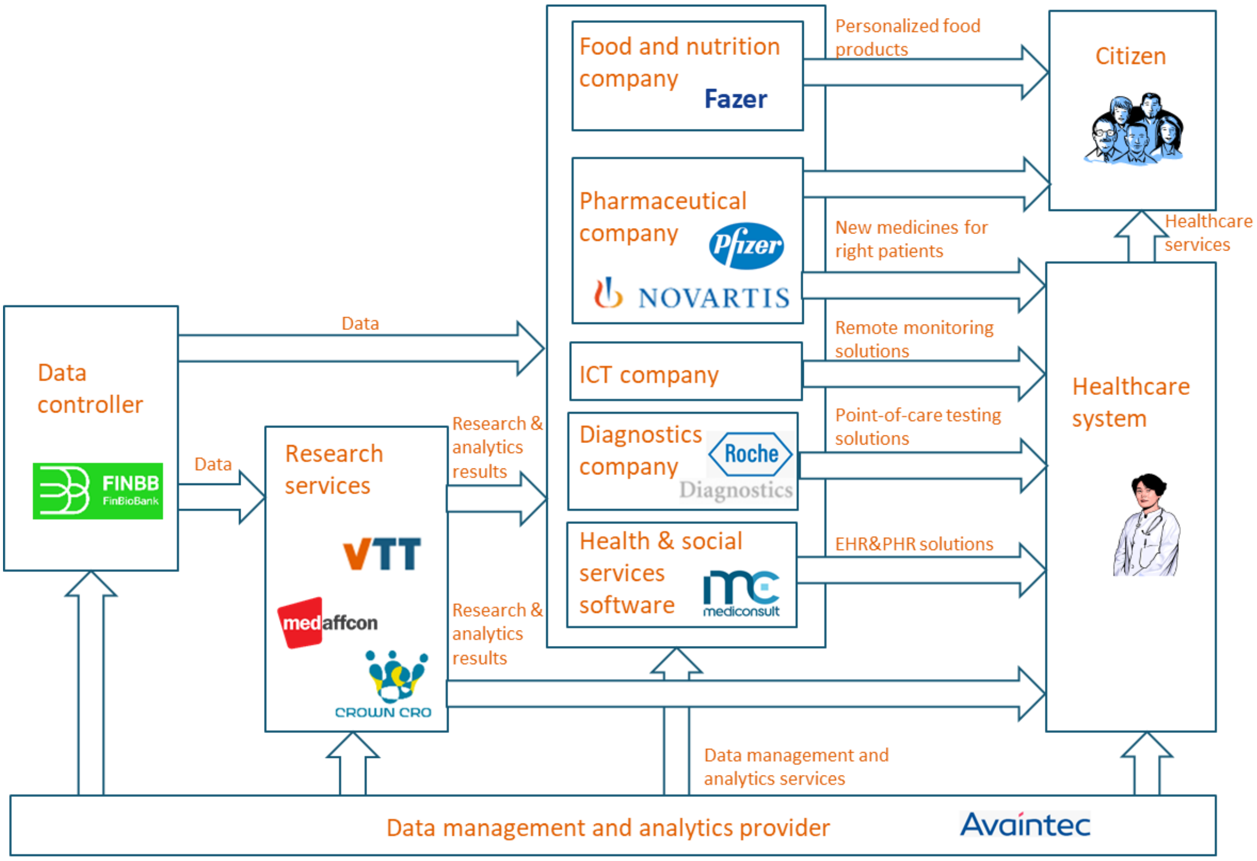
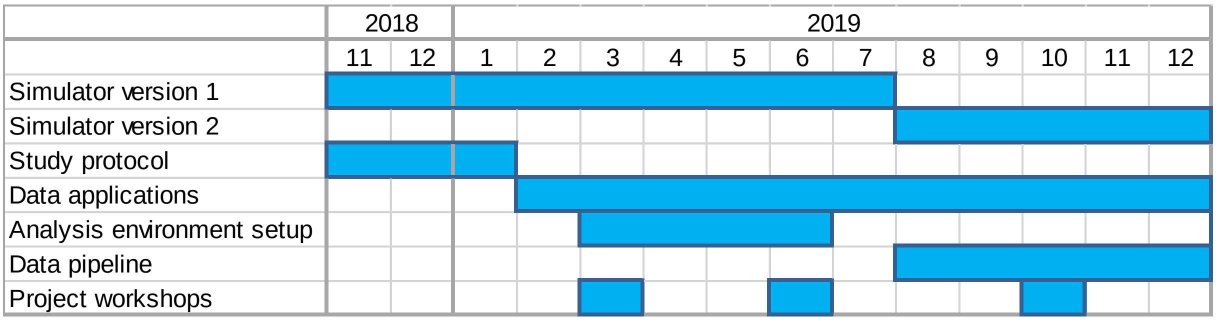

Food and nutrition company
Fazer
Personalized food products
Citizen
Pharmaceutical company
Pfizer
NOVARTIS
New medicines for right patients
Healthcare services
Data
Data controller
FINBB FinBioBank
Data
ICT company
Remote monitoring solutions
Healthcare system
Research services
VTT
medaffcon
CROWN CRO
Research & analytics results
Diagnostics company
Roche Diagnostics
Point-of-care testing solutions
Health & social services software
mediconsult
EHR&PHR solutions
Research & analytics results
Data management and analytics services
Data management and analytics provider
Avaintec

Figure 1. Data-driven precision medicine ecosystem showing roles of PreMed partners.

2. Overview of project activities

Activities carried out during phase 2 of the project are shown in Figure 2. The ecosystem model and its implementation as a simulator was carried out in two parts. The first version with limited functionality was developed and completed by July 2019. Subsequently, the simulator was enhanced by adding several new features during the rest of the year.

The PreMed PGx study protocol was written in the beginning of phase 2 as needed for the data applications to biobanks and national registers. The data application process was a continuing activity throughout the year. Setting up of the analysis environment and data pipeline were carried out in parallel with it. Three project workshops were organized during phase 2. The project activities and obtained results are described in more detail in the following sections.



	2018		2019											
	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Simulator version 1	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦					
Simulator version 2										🟦	🟦	🟦	🟦	🟦
Study protocol	🟦	🟦	🟦											
Data applications				🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Analysis environment setup					🟦	🟦	🟦	🟦						
Data pipeline										🟦	🟦	🟦	🟦	🟦
Project workshops					🟦			🟦				🟦		

Figure 2. PreMed project activities during phase 2.

3. Biobank study - research protocol

3.1 Introduction

The PreMed PGx study is a retrospective cohort study using data from biobanks and national registers. The scientific content for the research protocol was created by VTT researchers in close co-operation with external advisors: Docent Kari Harno, Docent Mika Lehto, Professor Mikko Niemi and Docent Maija Wolf. The research protocol has been updated according to the comments and requests by the biobanks and register controllers during the data application process. A short version of the protocol is presented in the following.

The study is aimed at pharmacogenomics of antithrombotic drugs, which is one of the areas where genome technology is expected to have considerable impact to healthcare and pharmaceutical product development already in a short time frame. The objective of the retrospective cohort study is to assess based on retrospective data the relevance of exploiting the patient's genotype data in the context of antithrombotic drug therapy. The related clinical setting is depicted in Figure 3.

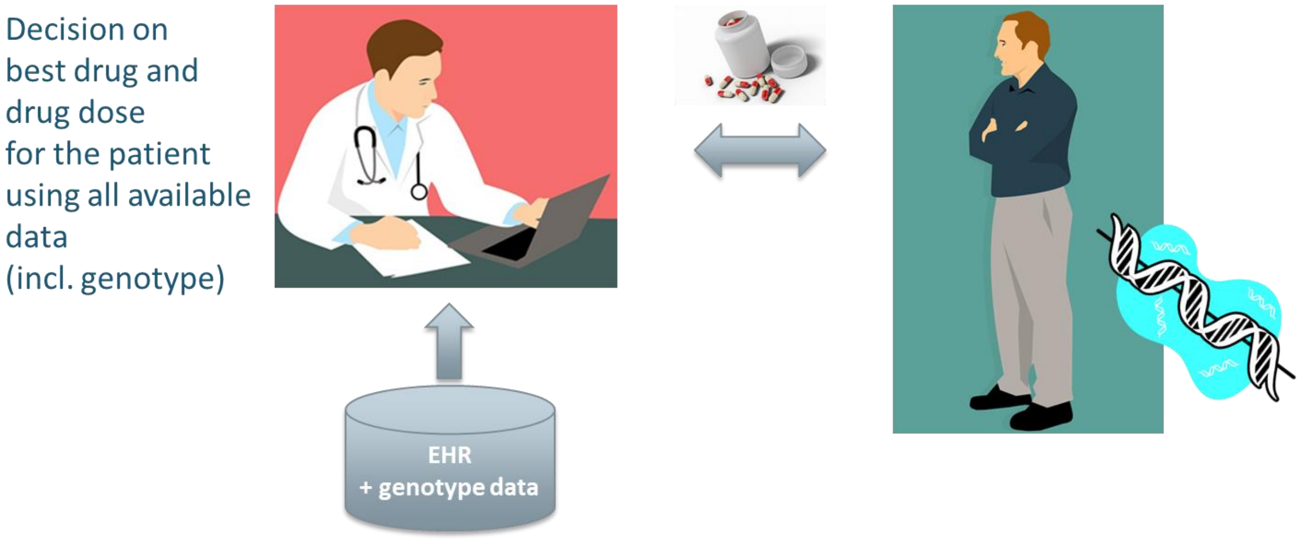

Decision on best drug and drug dose for the patient using all available data (incl. genotype)
EHR
+ genotype data

Figure 3. Clinical setting related to the biobank study.

The study is based on genome data collected by biobanks - especially in the framework of the FinnGen project - combined with patient record and national register data. The results are expected to be beneficial in deciding about adopting pharmacogenomic guidance in antithrombotic drug therapy.

3.2 Clinical background

Antithrombotic (anticoagulant or antiplatelet) therapies are used in the context of various clinical conditions and procedures, including atrial fibrillation (ischemic stroke prevention), deep vein thrombosis, pulmonary embolism, artificial heart valves and prevention of recurrent stroke or heart attack[1]. Careful therapy selection and dosing is of high importance since insufficient prevention of blood clots may lead to life-threatening conditions such as stroke and myocardial infarction. On the other hand, a considerable risk of bleeding is associated with antithrombotic therapies. In a recent study, the annual risk of warfarin users for stroke and bleeding was observed to be 3.1-9.3% and 2.6-7.5% respectively (Lehto et al. 2017). The INR (international normalized ratio) laboratory value and the respective time in therapeutic range (TTR) measure was shown to strongly correlate with improved patient outcomes. A severe adverse effect of heparin is heparin-induced thrombocytopenia, for which prevalence of 0.1-5% and associated mortality of 20-30% has been reported (Salter et al. 2016).

Predicting serious adverse drug reactions (ADRs) is a priority for pharmacogenetic research. Substantial amount of effort has been invested in investigating genotype related factors of drug therapies and a number of pharmacogenetics guidelines have been compiled[2]. Examples of investigated gene variants potentially affecting the response of antithrombotic drugs are listed in Table 1**Error! Reference source not found.**.

[1] NHS Newham Clinical Commissioning Group: http://www.newhamccg.nhs.uk/services/conditions-requiring-anticoagulant-therapy.htm
[2] CPIC: Clinical Pharmacogenetics Implementation Consortium: http://www.pharmgkb.org/page/cpic

Table 1. Antithrombotic drugs and genomic variants to be investigated

Substance	ATC[1] code	Drug consumption DDD[2]/1000 inhab./day[3]	Related gene variants: Single Nucleotide Polymorphism (SNP) or defect	Referred studies
Anticoagulants				
Warfarin	B01AA03	14,8	*CYP2C9* (rs1799853, rs1057910), *VKORC1* (rs9923231), *CYP4F2* (rs2108622)	Monie and DeLoughery 2017
Dabigatran	B01AE07	2,07	*ABCB1* (rs4148738,rs2032582, rs1045642), *CES1* (rs2244613, rs8192935)	Tseng et al. 2018
Rivaroxaban	B01AF01 B01AX06	4,22	*ABCB1* (rs2032582, rs1045642, rs2032582)	Tseng & al. 2018
Apixaban	B01AF02	2,65	*ABCB1* (rs4148738, rs1128503, rs2032582, rs1045642), *ABCG2* (rs2231142), *CYP3A5* (rs776746)	Tseng & al. 2018
Edoxaban	B01AF03	0,09	*ABCB1* (rs1045642), *SCLO1B1* (rs4149056)	Tseng & al. 2018
Heparin	B01AB01	0,08	*IL10* (rs1800896, rs1800871, rs1800872)	Pouplard et al. 2012
Enoxaparin	B01AB05	5,68	*ABCB1* (rs1045642) and *SLCO1B1* (rs4149056)	Vandell et al. 2018
Dalteparin	B01AB04	0,75		
Antiplatelets				
Clopidogrel	B01AC04	7,79	*CYP2C19* (rs4244285, rs4986893), *CES1* (rs71647871)	Tornio et al. 2018, Neuvonen et al. 2018, Kubica et al. 2011
Ticagrelor	B01AC24	0,75	*CYP4F2* (rs3093235, rs3093135), *CYP3A4* (rs35599367)	Tatarunas et al. 2017, Holmberg et al. 2018
Acetylsalicylic acid	B01AC06	69,15	*TXBA2R*(rs1131882), *ADRA2A* (rs4311994), *PLA2G7* (rs7756935) and *CDKN2B-AS1*(rs10120688)	Poistula et al. 2011

Strongest pharmacogenetic (PGx) evidence exists for warfarin, heparin and clopidogrel. For example, in the prospective ENGAGE AF-TIMI 48 trial with a subgroup of 4833 warfarin-treated patients, over-anticoagulation was observed to be correlated with genotype during the first 90 days of treatment (Mega et al. 2015). For direct oral anticoagulants - dabigatran, rivaroxaban, apixaban and edoxaban - potential genome associations have been identified, but evidence is still weak.

The cost-effectiveness of PGx-guided treatments has been evaluated in a number of publications (Verbelen et al. 2017). In 4 (of 6) publications PGx-guided clopidogrel treatment was considered either cost-efficient or cost-saving. For warfarin, 3 (of 12) publications

[1] ATC= Anatomical Therapeutic Chemical
[2] DDD=Defined Daily Dose
[3] Drug consumption 2017 (Fimea) https://www.fimea.fi/web/en/databases_and_registeries/consumption

considered PGx-guided treatment cost-efficient or cost-saving. This number was increased to 7 (of 12) in the case when genotype information was assumed to be already existing without extra cost at the time of prescription. Such situation may increasingly prevail in the future as services (e.g. the Genome Centre[1] in Finland) for secure management and sharing of genome data become available.

Pharmacogenetic tests are largely available by laboratory services. Currently, the use of such tests is still low due to lack of related clinical guidelines. There is a need for electronic health record systems (EHRs) which are capable to connect with genomic information and to provide PGx-related decision support for healthcare professionals (Ji et al. 2016).

3.3 Study objectives

The overall goal is to investigate the feasibility of using genome data in the context of antithrombotic therapy.

The objectives of the study are:

1. **To gain evidence on the association between gene variants and anticoagulation control of warfarin therapy.** Earlier research has already shown associations of *CYP2C9*, *VKORC1* and *CYP4F2* alleles with the efficacy of warfarin. The objective of the present study is to obtain further evidence on these associations in the Finnish population, which has not been reported in the earlier studies.

2. **To assess the clinical and economic impact of using genotype data in guiding warfarin therapy**. The aim is to assess if the knowledge of the genome variants is likely to lead to better anticoagulation control and less adverse drug reactions. Furthermore, the related economic impact will be estimated.

3. **To explore potential genotype-phenotype associations in the context of antithrombotic therapy.** In particular, direct oral anticoagulants will be investigated as their pharmacogenetic properties are not yet well known and their use is rapidly increasing. More detailed studies may be initiated based on the findings.

4. **To assess the current usage of pharmacogenetic information in the context of antithrombotic therapy.** The aim is to increase overall understanding of the total volume and clinical context of using pharmacogenetic tests.

In addition to the clinical objectives listed above the study seeks to collect experience and best practices for carrying out co-operative research activities based on combining clinical data resources (including genome data). Thereby, the study is expected to provide useful information from the perspective of the national activities in setting up the service operator (Findata) for secondary use of data, the genome center and centralized biobank services. Especially in the exploratory part the study aims to deploy and develop new methods needed in processing and data-driven analysis of high-dimension clinical and genome data.

3.4 Study design

The research is conducted as a retrospective cohort study combining genotype data from biobanks with data from national registers and patient records.

[1] Finnish Genome Centre: https://stm.fi/en/artikkeli/-/asset_publisher/genomikeskuksen-perustaminen-etenee-tyoryhman-arviomuistio-lausunnoille

The index date of the study is defined as the date of the patient's first purchase of one of the investigated drugs, in the time frame of 1.1.2007 - 30.6.2018. The follow-up period is defined as:

- start: 2 years before the index date
- end: 6 months after the last purchase of the drug

The inclusion criteria are:

- The patient has had at least one of the inclusion diagnoses in the time frame 1.1.2007 - 30.6.2018 , and
- The patient has used at least one of the investigated drugs in the time frame 1.1.2007 - 30.6.2018 (based on the drug purchase register of Kela), and
- The patient is at least 18 years of age at index date, and
- Genotype data of the patient is available covering at least the variants: *CYP2C9*/rs1799853, *CYP2C9*/rs1057910 and *VKORC1*/rs9923231

The exclusion criteria are:

- Permanent residence in Finland less than 12 months during the follow-up period
- Purchase of any of the investigated drugs in the time frame 1.1.2005 - 31.12.2006

3.5 Methods

The study approach related to the objectives expressed in Section 3.3:

Association between gene variants and anticoagulation control of warfarin therapy (objective 1):

A study setting will be employed where the carriers of *CYP2C9*, *VKORC1* and *CYP4F2* variants are compared with the carriers of the corresponding wild-types. For the analysis a grouping to normal, sensitive and highly sensitive patients will be used (Mega et al. 2015). Standard statistical methods will be used in the analysis to estimate the association between genotypes and phenotypes.

The following phenotypes will be investigated:

- INR (international normalized ratio) laboratory data history will be analysed. The following parameters will be used to evaluate the quality of the anticoagulation control: (1) INR during first month (time-weighted mean), (2) time to reach therapeutic range and (3) time in therapeutic range (TTR) during first three months.
- Bleeding events defined as outcome diagnoses. ISTH[1] criteria will be used in identifying major bleeding (Schulman & Kearon 2005).
- The needed drug dose to maintain anticoagulation at therapeutic level[2]

Assessing the clinical and economic impacts of using genotype data in guiding warfarin therapy (objective 2):

An analysis of the INR history and patient encounters will be performed. Assessment will be made on the clinical and economical impact of PGx-guided treatment, taking into account:

- Potentially improved TTR leading to reduction in adverse effects (bleeding, myocardial infarction or cerebral infarction) and related cost savings
- Potentially reduced need for laboratory tests and related cost savings

Available information on healthcare resource use (HCRU) and associated unit costs will be used in the health economic assessment. Based on the availability, the assessed information will potentially include the following HCRU units, stratified by patient groups with distinct TTR-ranges.

The grand total for each HCRU type will be evaluated, and the estimates will be scaled to “per patient” by dividing the grand total by the number of contributing patients, and to “per patient year” by dividing the grand total by the total follow-up time of the contributing patients.

The potential of the previously reported warfarin dosing algorithms[3] (Gage et al. 2008) will be assessed.

Exploring candidate genotype-phenotype associations in the context of anticoagulation and antiplatelet therapies (objective 3):

A multivariate genotype-phenotype association analysis will be carried out based on data-driven classification methods. The analysis will be separately performed for each investigated drug, for which sufficient amount of data is available. Genome variants to be included in each analysis are listed in Section 3.6. INR laboratory tests are routinely performed only for warfarin users and INR data will only marginally be available for the users of other drugs. For all drugs under investigation the outcomes as listed in Section 3.6. will be used. Also other relevant phenotype data will be derived from the available data sets for this explorative analysis. Possible interactions of identified potentially interactive drugs with the antithrombotic drugs under study will be taken into account in the analysis.

Assessing the current usage of pharmacogenetic information in healthcare (objective 4):

[1] International Society on Thrombosis and Haemostasis, https://www.isth.org/

[2] This can be performed in case warfarin dosing information will be available

[3] CPIC: Clinical Pharmacogenetics Implementation Consortium: http://www.pharmgkb.org/page/cpic

The amount of performed pharmacogenenetic tests will be evaluated based on the available laboratory data. The laboratory codes to be observed are: B -Varfa-D and B -Farma-D (pharmacogenetic panel).

3.6 Data resources

Antithrombotic drugs and genomic variants to be investigated are listed in Table 1 **Error! Reference source not found.**. The objective has been to include all drugs with considerable use for antithrombotic drug therapy in Finland. The genome variants have been selected based on observed or potential association with the investigated drugs as documented in existing literature. The inclusion and outcome diagnoses are listed in Table 2. Additionally the study will use laboratory test results (especially INR-values) and information resulting from healthcare encounters and hospital visits.

Table 2 Diagnoses to be used (inclusion criteria and outcomes)

ICD 10 code	Description	Needed as
I48	Atrial Fibrillation	inclusion criteria
I20-I25, I65-I66, I67.2, I70	Vascular disease	inclusion criteria
I26	Pulmonary embolism	inclusion criteria and outcome
I63, I64, I65, I66, I69.3-I69.8	Stroke / cerebral infarction or atherosclerosis in (pre-)cerebral arteries	inclusion criteria and outcome
I80	Phlebitis and thrombophlebitis	inclusion criteria
I81	Portal vein thrombosis	inclusion criteria
I82	Other venous embolism and thrombosis	inclusion criteria
D50.0, D62, D68.3 I60-I62, I69.0-I69.2, I85.0 J94.2 K22.1, K22.3, K22.6, K25.0, K25.2, K25.4, K25.6, K26.0, K26.2, K26.4, K26.6, K27.0, K27.2, K27.4, K27.6, K28.0, K28.2, K28.4, K28.6, K29.0, K62.5, K63.1, K63.3, K92.0-K92.2 N02 R04, R31, R58 S06.2-S06.6, S06.8	Bleeding events	outcome
D69.6	Heparin-induced thrombocytopenia	outcome (for objective 3)
I24	Other acute ischemic heart diseases	outcome
D68.3	Hemorrhagic disorder due to circulating anticoagulants	outcome
C00-C97	Neoplastic disorders	potential impact to outcome

The patient cohort is formed from three parts provided through the biobanks: Helsinki Biobank, Auria Biobank and THL Biobank. Each biobank will identify the eligible patients based on the inclusion diagnoses and availability of genome data for the subject. The biobanks will be responsible for linking the data for the subject group from the registers (THL, Kela, patient records and laboratory systems). The subjects will be from the following geographical areas:

- **Helsinki Biobank subjects**: Hospital District of Helsinki and Uusimaa, South Karelia Social and Health Care District (Eksote), Kymenlaakso Social and Health Services (Carea)
- **Auria Biobank subjects**: Hospital District of Southwest Finland, Hospital District of Vaasa and Satakunta Hospital District.
- **THL Biobank subjects:** Finland.

3.7 Sample size

The number of patients to be available for the study was estimated based on the initial enquiries to the biobanks as:

- Warfarin pharmacogenetics investigation (objectives 1-2), total: 2678 patients
  - Helsinki Biobank: 735 patients
  - Auria Biobank: 301 patients
  - THL Biobank: 1642 patients[1]
- Explorative study on anticoagulants and antiplatelets (objective 3), total: 5924 patients
  - Helsinki Biobank: 1626 patients[2]
  - Auria Biobank: 666 patients
  - THL Biobank: 3632 patients[3]

Prior studies indicate that the risk rate for bleeding complications in warfarin users is approximately 5% (Lehto et al. 2017) and the prevalence of CYP2C9/rs1799853 or rs1057910 carriers is 35% (Sistonen et al. 2009 and CPIC data[4]). Assuming alpha level of 0.05 and power of 0.9, the estimated number of patients (n=2678) for the warfarin investigation (objective 1) would enable us to detect an increase in the bleeding risk from presumed 5% to 9% (RR=1.8) for the carriers of CYP2C9/rs1799853 or rs1057910.

In earlier studies (Higashi et al. 2002, Aithal et al. 1999) increased bleeding risk of 2.4-3.7 has been observed for CYP2C9/rs1799853 or rs1057910 carriers.

3.8 Limitations of the study

This study is based on real-world evidence data reflecting the actual data available from the selected patient cohort. The data collected for this study is existing standard of care data and may therefore be partially non-standardized and incomplete. Missing values are thus also

[1] Estimated as proportion of warfarin users from the total number of 4734 patients with atrial fibrillation, pulmonary embolism or ischemic stroke.
[2] Estimated number of all anticoagulant/antiplatelet users based on the number of 735 patients with warfarin.
[3] Estimated as proportion of anticoagulant/antiplatelet users from of 4734 patients with atrial fibrillation, pulmonary embolism or ischemic stroke.
[4] CPIC: Clinical Pharmacogenetics Implementation Consortium: http://www.pharmgkb.org/page/cpic

expected. In particular, this limitation applies to THL registry data on primary care (Avohilmo register), which has only been collected since 2011. It is known that the Avohilmo register contents is not complete and uniform across Finland. Missing THL registry data can be partly compensated by patient record data from patient records for the subjects of Helsinki Biobank and Auria Biobank. After the data has been collected the differences between the data sets (Helsinki Biobank, Auria Biobank, THL Biobank) will be assessed and taken into account as appropriate.

For some of the drugs listed in Table 1 the amount of subjects will be low and will not be sufficient for statistically significant conclusions. This is certainly the case for the direct oral anticoagulants which have been in the markets only for a short time. With the resources and timelines of the project, it has not been possible to increase the number of subjects. However, the exploratory part of the project addressing these drugs is still expected to provide valuable outputs indicating relevant topics for further studies.

The drug purchase register of Kela does not include information for over-the-counter drugs. For such drugs, information may be available in the patient record, but will not be complete.

3.9 Significance of results

The results of the project are expected to contribute to improved outcomes of antithrombotic drug therapies.

Concerning warfarin, the study is expected to provide evidence between the association of gene variants, individual responses to drugs and the control of anticoagulation therapy in the Finnish population, which has not been covered by existing research. The study will also assess aspects related to the clinical feasibility and benefits on the use of genome data in the context of warfarin therapy. Thereby, the results of the study may contribute to new care recommendations incorporating genome-dependent factors. Such recommendations are expected to improve therapy outcomes, for example in avoiding adverse effects of warfarin therapy.

Concerning antithrombotic drugs in general, the study will identify candidate genotype associations. Especially interesting drugs are direct oral anticoagulants as their use is rapidly increasing and their pharmacogenetic characteristics are not sufficiently well known yet. This study may identify interesting candidate associations, which may need to be further investigated in follow-on studies. Such research is needed in order to properly understand the genetic associations of the drugs in the Finnish population and to create new clinical guidelines to improve therapy outcomes.

As a whole, the study aims to provide a generic model and best practices for genome data studies exploiting high-dimensional individual-level data mining and modelling approaches. This way the results are intended to be applicable to a wide range of clinical domains and therapies. The study is expected also to provide useful information from the perspective of the national activities in setting up the Findata service operator for secondary use of data, the genome center and centralized biobank services.

4. Biobank study - data collection

As the research protocol was completed the data collection proceeded by writing data applications for the three biobanks (Helsinki Biobank, Auria Biobank and THL Biobank), Kela and THL registry authority. The THL registry application covers the laboratory databases and therefore separate applications to laboratory data register controllers is not needed.

4.1 Data collection approach

The data collection approach of the PreMed study is illustrated in Figure 4. After acceptance of the data application and signing of the Material Transfer Agreement (MTA) each biobank composes their target group (list of person ID's) matching the inclusion criteria concerning diagnosis, age and availability of genotype data. In composing the initial target group Auria Biobank and Helsinki Biobank use diagnosis data from the patient register via corresponding datalakes. THL biobank composes the initial target group based on Finriski and Health 2000/2011 cohort study data. A joint method for pseudonymization and the related secret key was agreed between the biobanks in order to enable detection of overlapping patients from pseudonymized data at a later stage. The initial target groups (person ID lists) are then provided by each biobank to the THL (registry authority), Kela and the controllers (or processors) of the laboratory data registers. Each of them extracts data for the target groups as defined in the study protocol and returns data back to the biobank. The final target groups are formed by applying the drug purchase data of Kela to the initial target groups. Pseudonymized data sets are composed by each biobank and delivered to VTT.

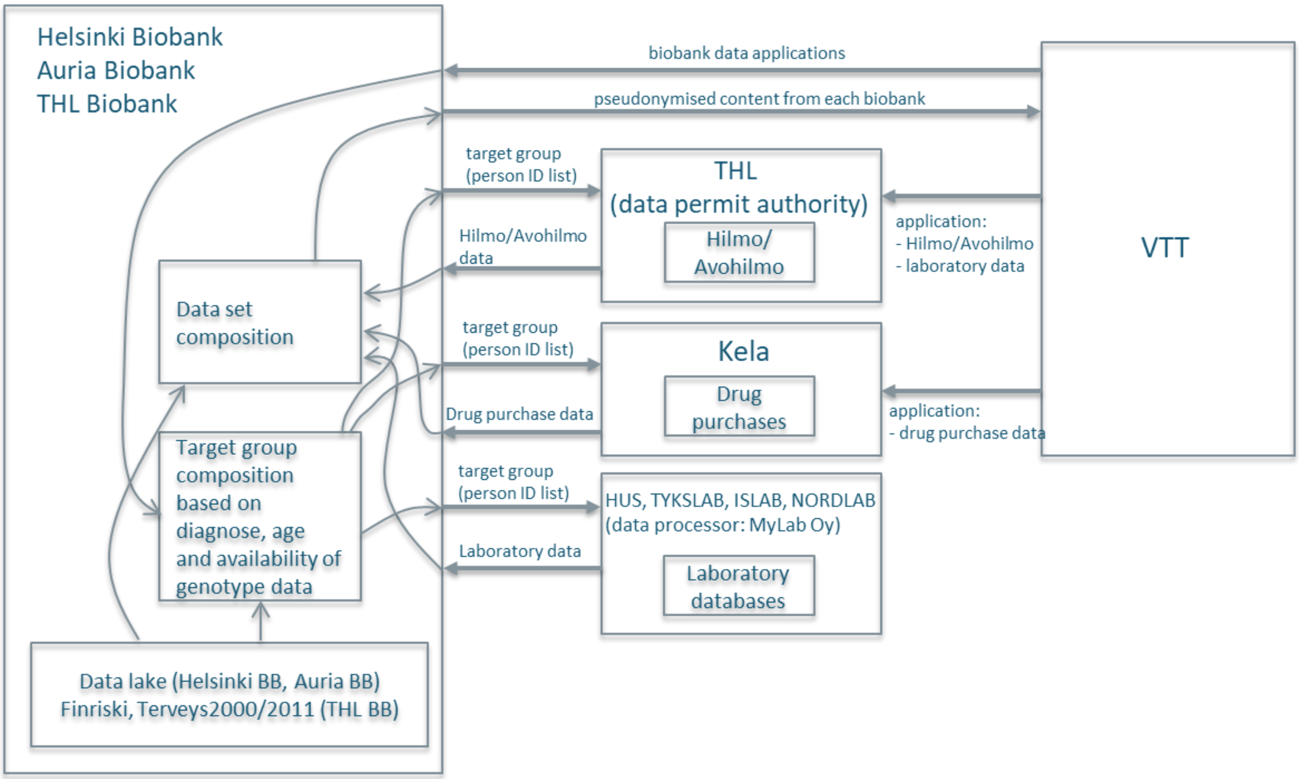

Helsinki Biobank
Auria Biobank
THL Biobank
biobank data applications
pseudonymised content from each biobank
target group (person ID list)
THL
(data permit authority)
Hilmo/ Avohilmo
Hilmo/Avohilmo data
application:
- Hilmo/Avohilmo
- laboratory data
VTT
Data set composition
target group (person ID list)
Kela
Drug purchases
Drug purchase data
application:
- drug purchase data
Target group composition based on diagnose, age and availability of genotype data
target group (person ID list)
HUS, TYKSLAB, ISLAB, NORDLAB
(data processor: MyLab Oy)
Laboratory data
Laboratory databases
Data lake (Helsinki BB, Auria BB)
Finriski, Terveys2000/2011 (THL BB)

Figure 4. Approach for data collection

4.2 Data applications

The data application forms and related processes vary between the data controllers. At the time of PreMed data application preparation Helsinki Biobank and Auria Biobank were using the same online service (Proma), but with different form contents and required annexes. THL Biobank was using the REMS online application system. The application for Kela needed to be filled in a pdf form and submitted by email. THL registry authority's process followed the same pattern but with an application form in RTF format.

The timeline of the data application process is shown in Figure 5. For Helsinki Biobank a review and recommendation by the HUS ethical committee needed to be obtained before the actual biobank application could be submitted. For the other data providers, separate ethical review was optional. The positive recommendation of the HUS ethical committee was annexed to the Helsinki Biobank, THL Biobank, Kela and THL registry applications which were submitted after the ethical review was completed.

As can be shown in Figure 5, the decisions for the applications were mostly received within the time frame initially estimated: Helsinki Biobank and THL biobank in 1,5 months, Auria Biobank in one month and Kela in 3 months. All processes involved some adjustment or clarification to the application before final positive decision was given. No adjustment or clarification was requested by THL registry authority, but the application evaluation process took 8 months. The reason for the excess time was explained to be caused by congestion due to high number of applications and lack of sufficient resources for application processing.

The process of compiling of MTA agreements with the biobanks took time between 1,5 - 3,5 months after acceptance of the application. First data delivery was provided by Auria Biobank in October in line with the MTA. In the case of Helsinki Biobank, THL Biobank and Kela data delivery was bound to the acceptance of the THL registry data application and therefore delayed.

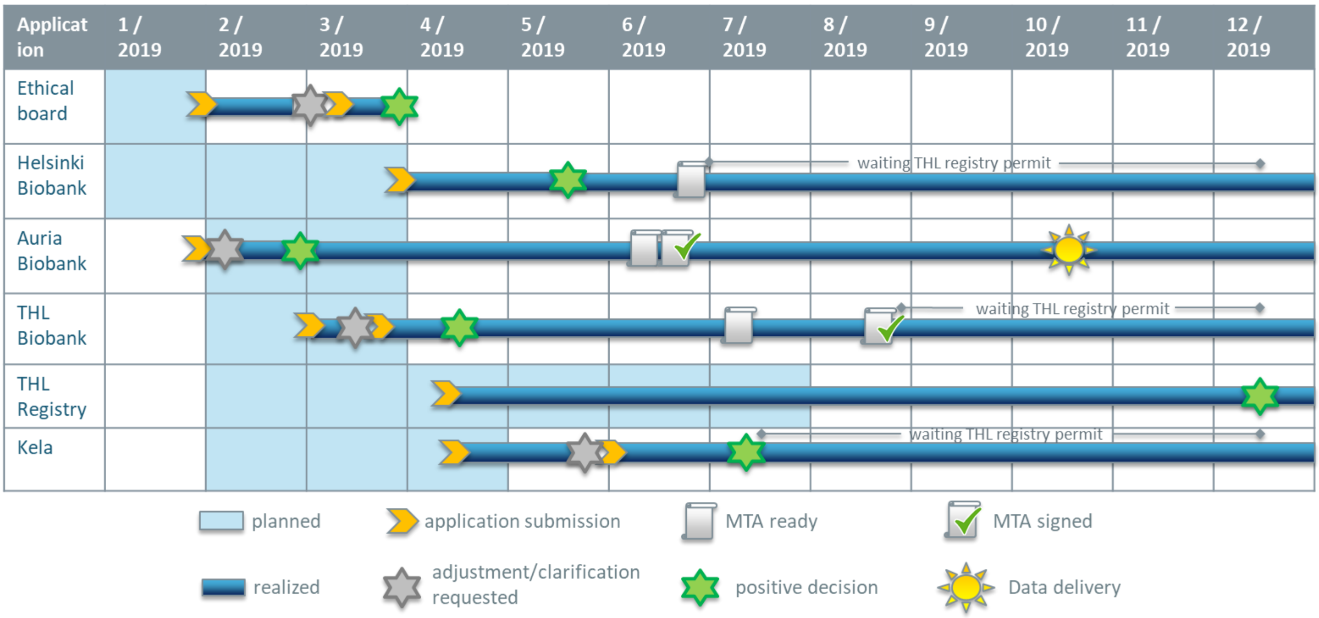


Figure 5. Data application process timeline

4.3 Discussion on data application process

In order to achieve a sufficient amount of patients, the PreMed study needed data from three biobanks. At the time of PreMed study preparation there were no joint biobank services available. Most biobank projects so far have been using data and/or samples from one biobank only - there was not much prior experience about joint biobank projects.

From the biobank customer perspective, there is a need for harmonizing procedures among the biobanks, e.g. in methods for national registries linkage, pseudonymization methods, data delivery mechanisms and data formats. Joint biobank services would be highly appreciated covering at least: data availability requests, biobank applications and agreements. Such services are now gradually evolving. The Fingenious service[1] of FinBB provides now a single channel for finding out the availability of biobank samples and submitting requests for feasibility studies and data/sample access. The services cover six Finnish hospital biobanks and partly also the THL Biobank. During year 2020 the new data permit authority (Findata) starts to provide centralized services for secondary use of health and social services register data. The co-operation model between biobanks and Findata is still under discussion.

5. Biobank study - data analysis environment

5.1 Research server

Data analysis activities are carried out at the research server, which also keeps the original data sets as well as the results of data analyses. The research server resides in a secured physical location at VTT and can be accessed only from computers connected to VTT's intranet. The data is stored in encrypted disks and the server access is only allowed for the study group of VTT's researchers listed in this research plan. The server has 160GB RAM, 20 cpu cores, Nvidia Quadro K5000 GPU and 2TB of storage.

5.2 Data pipeline

The pipeline for producing analysis results from source data is shown in Figure 6. Data is delivered to VTT via the biobanks. In all cases the data is delivered in pseudonymized form (see Section 4.1) over a secure delivery channel, such as secure email or SFTP and organized in data table form, such as CSV. The incoming files are manually loaded in the staging area - a specific folder of the research server.

A set of R scripts has been programmed for making any needed transformations for the data and loading it to the research database (PostgreSQL). By executing the scripts the research database can be re-created from the received data files always when needed.

The research database schema divides into 23 tables collecting data from the various sources. There is one outcome and one event table for the users of each of the 11 investigated drugs. Additionally, there is one outcome table containing key outcomes and background data for all patients.

Based on the drug purchase data the patient's outcome data (diagnoses, laboratory results) is inserted to one or several drug-specific outcome tables. All events (outpatient visits, hospital visits) are inserted to one events table. The outcome tables are used especially in the analysis of the genotype-phenotype associations while the events tables enable the analysis of care paths and related healthcare costs. The summary table enables baseline analysis of the characteristics of the whole cohort. Events table is used to collect information needed in estimating healthcare costs.

Data analysis and visualisation is based on data retrieved from the research database. For example, the following analysis and visualisation scripts will included:

[1] https://finbb.fi/access-finbb-biobanks-fingenious/

- verification of data by comparing to existing statistics (e.g. drug purchase statistics, genotype frequencies, etc.)

- statistical analysis of association between gene variants and anticoagulation control of warfarin therapy as measured by the outcome diagnoses and laboratory results (INR)

- evaluation of healthcare costs based on the events (encounters) data for assessing the clinical and economic impacts

- statistical analysis of candidate genotype-phenotype associations in the context of anticoagulation and antiplatelet therapies

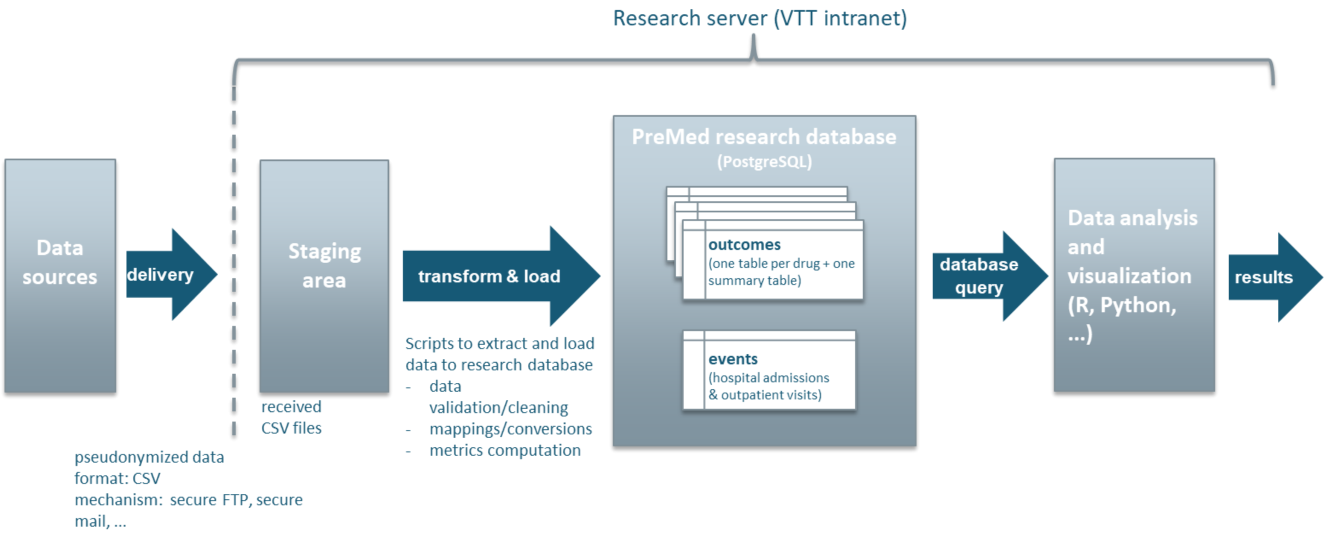

Research server (VTT intranet)
Data sources
delivery
pseudonymized data
format: CSV
mechanism: secure FTP, secure mail, ...
Staging area
received CSV files
transform & load
Scripts to extract and load data to research database
- data validation/cleaning
- mappings/conversions
- metrics computation
PreMed research database
(PostgreSQL)
outcomes
(one table per drug + one summary table)
events
(hospital admissions & outpatient visits)
database query
Data analysis and visualization (R, Python, ...)
results

Figure 6. Data pipeline

As a preparation for the exploratory data analysis the Facets visual data exploration tool[1] has been prepared for the analysis environment. The original version was supporting only Google Chrome web browser but the standalone PreMed implementation is now tested and functional both with Firefox and Google Chrome. The Facets tool includes two applications “Facets Overview” and “Facets Dive”. The first tool calculates the basic statistics from every variable of the given dataset and creates a web application for easy investigation of data. The Dive tool creates a web application which enables to study multiple variables at the same time by grouping them interactively according to the user selections. Maximum of six variables can be compared at a time.

6. Ecosystem simulation

6.1 Overview

The data-driven precision medicine ecosystem is expected to lead to remarkable economic benefits through new products and business opportunities. However, considerable public investments are needed e.g. to establish the required infrastructure (data/samples, biobank

[1] https://pair-code.github.io/facets/

processes), to accelerate the R&D of ecosystem companies and to support related academic research. For decision-makers an important question is how to support the data-driven precision medicine ecosystem growth to achieve maximum benefit for the society. The PreMed project is developing a system dynamics (SD) model for simulation of alternative development paths of the ecosystem. Figure 7 shows the high-level ecosystem diagram indicating alternative targets for public investments. The primary outcome measure is the volume of real world data projects making use of data obtained from biobanks and national registers.

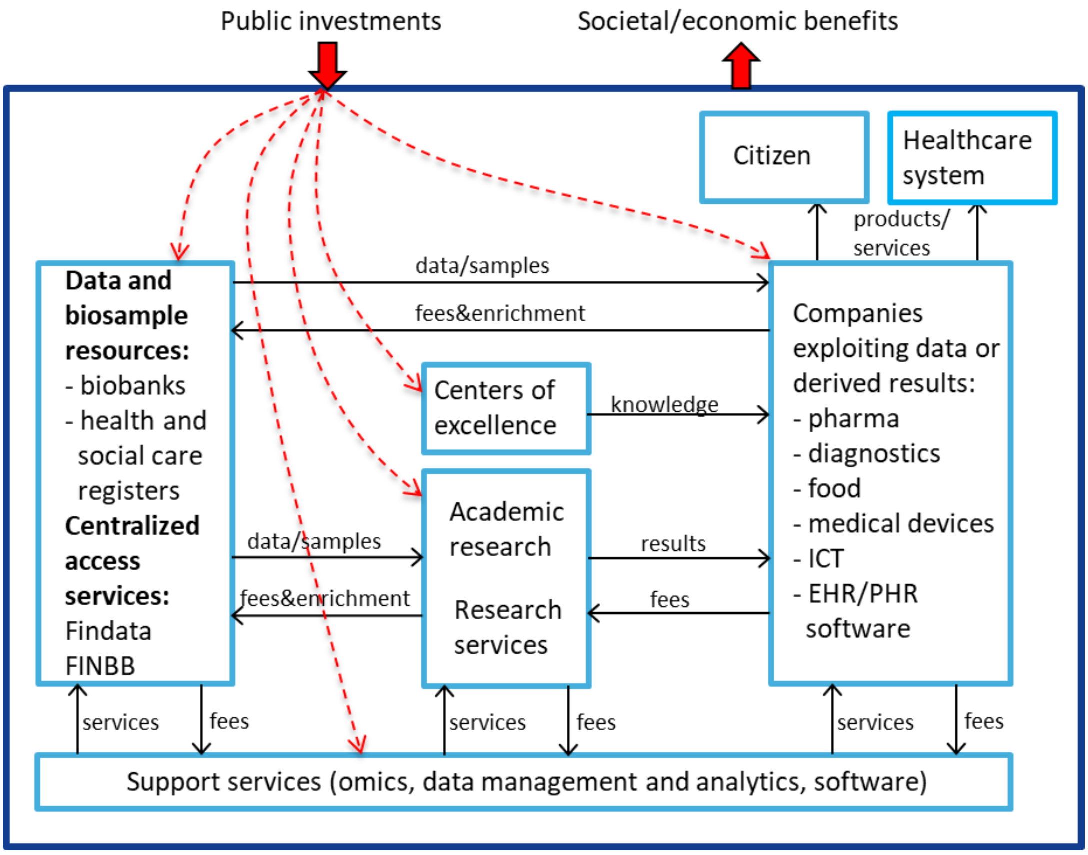


Figure 7. High-level ecosystem model showing main stakeholder groups and dependences affecting the exploitation of health data.

6.2 System dynamics method

System dynamics method is a collection of tools to describe and comprehend how different parts of a system are interconnected and how they create the behaviour of the whole system.

Dynamic complexity (Sterman 2000) does not require a large number of stakeholders in the system, it is caused by interconnections in the system, for example, nonlinear feedback loops and time delays. Dynamic complexity means that the actions and their effects can be far away in time and space, and therefore such systems are in some cases challenging to fully understand without suitable tools, for example, system dynamics.

System dynamics is based on the assumption that the structure of the system causes its behaviour, and thus, the focus is on modelling the structure of the system, i.e. the feedback loops. The structure can be deconstructed into two basic building blocks, 1) positive / reinforcing feedback loop (marked by R in the diagrams) and 2) negative / balancing feedback loop (marked by B in the diagrams). The reinforcing feedback loop tries to cause exponential growth or exponential decay, depending on the current state of the system and the parameters. The balancing feedback loop is a goal seeking behaviour, it tries to drive the state of that loop to match the target value.

When a model is constructed using these two elemental building blocks, it is possible to simulate and study how the different parts of the model affect each other. As the behaviour is caused by the feedback loops, by studying the strengths of the feedback loops it is possible to come to a conclusion which feedback loops are the most important at a given time. Usually some feedback loops dominate in the early phase and then the loop dominance shifts to some other loops. For example, growing ecosystems often grow exponentially in the early phases, the reinforcing feedback loops dominate, and then at some point the growth starts to saturate as the loop dominance shifts to a balancing feedback loop. By understanding the interplay of the feedback loops, it is possible to design system wide strategies in a way that enables the most beneficial behaviour to occur with the least amount of resources.

The benefits of a system dynamics modelling and simulation study consists of the following parts:

- Visualise the system
- Communicate the understanding how the system works
- Challenging assumptions
- Understanding the different effects of different parts of the system on the whole system
- Sensitivity runs to study how the parameter values affect the behaviour
- Trying different strategies and design strategies that produce desired behaviour

We are interested in different behaviour modes (Figure 8), that is, what are the growth drivers in the system (reinforcing feedback loops) or what prevents the system from growing (balancing feedback loops).

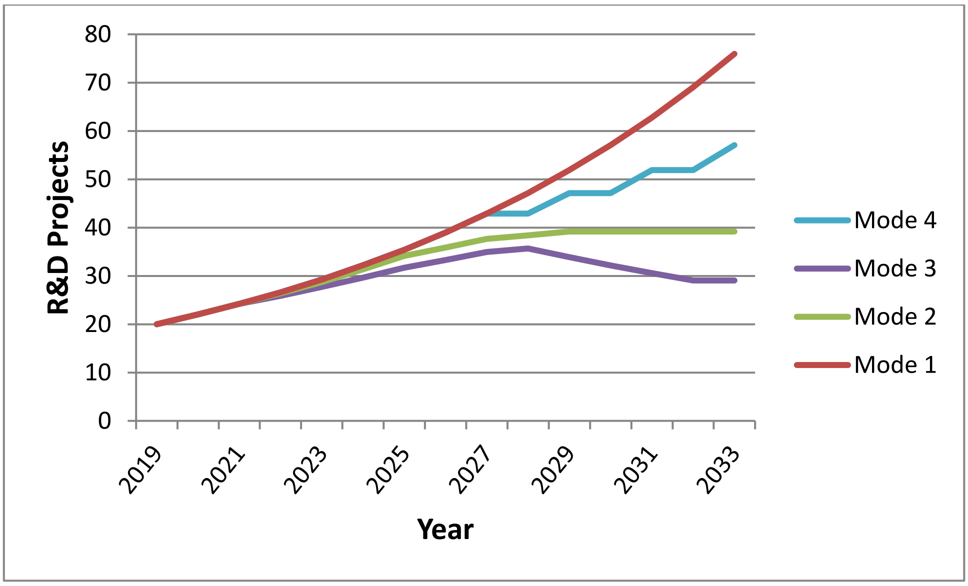


Figure 8: Ecosystem behavioural modes

6.3 Simulation model

As outlined in Figure 7 exploitation of health data involves a wide range of companies and public organizations forming a complex system.

The system model consists of eight sectors / subsystems that interact with each other. Sectors are:

- **Biobank donors and donor services:** Describes the number of persons who have given a consent for the use of their health care data and samples as well as the processes that lead to the increase of the number of given consents, i.e. biobank donors.
- **Biobank samples and data:** Describes the amount of samples and data collected from the biobank donors as well as the processes how the samples are gathered and how the data is generated.
- **Pharmaceutical R&D:** Describes the research and development conducted by the pharmaceutical companies that is specifically related to health care data. Real world evidence (RWE) projects are typical examples of pharmaceutical R&D. The R&D actvitiy is measured in number of projects.
- **Academic research:** Describes the academic research that is specifically based on health care data. Research activity is measured in number of projects.
- **Findata:** Describes the Social and Health Data Permit Authority Findata, which is starting its operation in the beginning of 2020. Findata is designed to be a gate for the secondary use of social and health data, that is, the applications for the use of social and health data will be processed by Findata. The operation of Findata will begin gradually. Before Findata services all applications have been processed by the different data controllers separately.
- **Competence centers / Centers of Excellence (National Cancer Center, Drug Development Centre, Neurocenter Finland):** Describe the operation of different competence centers, respectively National Cancer Center (founded 2019), Drug Development Centre (not yet founded), and Neurocenter Finland (not yet operational), from the point of view of the studied ecosystem, that is, the research coordination and building connections between academic research and industry.
- **Genome Centre:** Describe the operation of National Genome Centre that is under development from the point of view of the studied ecosystem. The Genome Centre is planned to serve as a one-stop shop in all matters related to genomics and to be a population wide genome database. Genome Centre has not yet started its operation.
- **Support services:** Describes the services which the pharmaceutical companies need in their R&D projects and which they are not able (or willing) to do themselves. These services may include different kinds of data analytics, laboratory analyses, software development, etc. Support services consists mainly of private sector companies, but it also includes the analyses done by research centers and other publicly funded organizations.

The main interest of the model, presented in Figure 9, is to study the interconnections between biobank donors, samples and data, and pharmaceutical R&D as well as to study how the public investments (green parameters in the model) or lack of them will affect the evolvement of the ecosystem. In the model public investments are focused to the following parts of the ecosystem in no particular order: 1) Biobanks, 2) Donor services, 3) Findata, 4) Academic research, 5) National Genome Center, 6) Competence centers / Centers of excellence, and 7) Support services.

Investments to biobanks refer to additional funding to individual biobanks or FinBB through the state budget or through targeted R&D projects (such as Finngen). Such investments are targeted to setting up the various biobank processes to build sample and data resources and improve access to them. Donor services are considered as a specific area, where the investments are targeted to bring benefits for the biobank donors with the objective of increasing the number of donors (given biobank consents). Investments to Findata are provided through the state budget and targeted to setting up the basic processes and information systems to handle data permit applications, connect data resources and to support secure processing of data. Investments to academic research (e.g. via the Academy of Finland) refer to public funding of academic research projects based on retrospective data resources. Investments to the competence centers refer to public funding of setting up the national centers of excellence (genome, cancer, neuro and pharmaceutical development). The National Genome Center is modelled as separate entity due to its intended role in maintaining the national genome database. Investments to support services appear as R&D funding (e.g. via Business Finland) to SME companies, which provide various services needed in exploiting data and sample resources.

Notice that pharmaceutical companies is the only major sector in the model that is not receiving public investments, or more accurately, the public investments (e.g. EU-project financing) the pharmaceutical companies may be receiving is included in the R&D budget of the companies and thus not explicitly modelled. The R&D activities are considered mainly from the point of view of the pharmaceutical companies, as it is at the moment the main sector using biobank data and seen as the major sector in the model. However, the model is also applicable to other businesses utilizing health data to support their R&D activities.

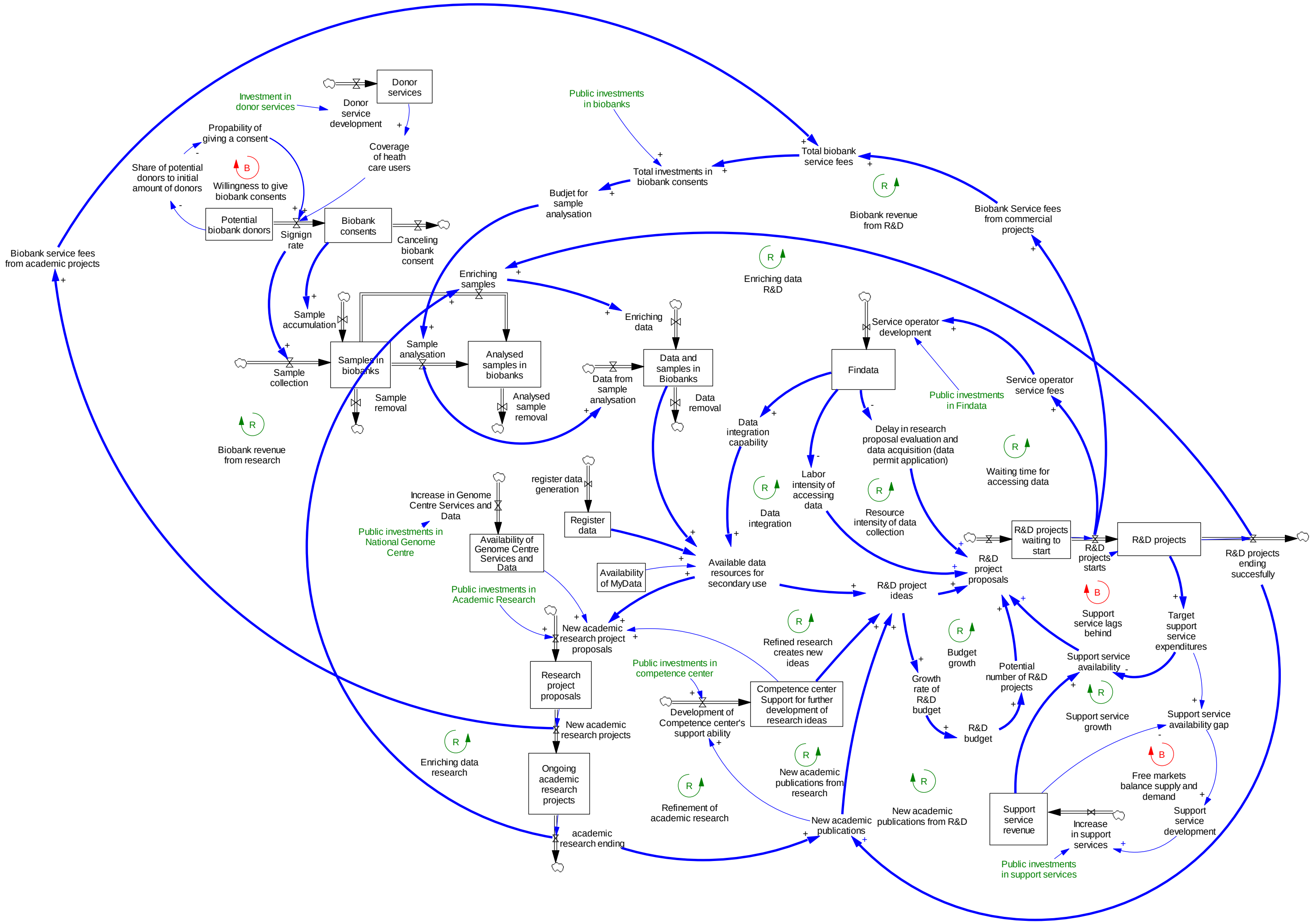

Donor services
Investment in donor services
Donor service development
Coverage of heath care users
Propability of giving a consent
Share of potential donors to initial amount of donors
B
Willingness to give biobank consents
Potential biobank donors
Signign rate
Biobank consents
Canceling biobank consent
Biobank service fees from academic projects
Public investments in biobanks
Total investments in biobank consents
Budjet for sample analysation
Total biobank service fees
R
Biobank revenue from R&D
Biobank Service fees from commercial projects
R
Enriching data R&D
Enriching samples
Enriching data
Sample accumulation
Samples in biobanks
Sample collection
Sample analysation
Analysed samples in biobanks
Sample removal
Analysed sample removal
Data from sample analysation
Data and samples in Biobanks
Data removal
R
Biobank revenue from research
Findata
Service operator development
Public investments in Findata
Service operator service fees
Data integration capability
Delay in research proposal evaluation and data acquisition (data permit application)
Labor intensity of accessing data
R
Waiting time for accessing data
R
Data integration
R
Resource intensity of data collection
register data generation
Register data
Increase in Genome Centre Services and Data
Public investments in National Genome Centre
Availability of Genome Centre Services and Data
Availability of MyData
Available data resources for secondary use
R&D project ideas
R&D project proposals
R&D projects waiting to start
R&D projects starts
R&D projects
R&D projects ending succesfully
Public investments in Academic Research
New academic research project proposals
Research project proposals
New academic research projects
Ongoing academic research projects
academic research ending
R
Enriching data research
Public investments in competence center
Development of Competence center's support ability
Competence center Support for further development of research ideas
R
Refined research creates new ideas
R
Refinement of academic research
R
New academic publications from research
New academic publications
R
New academic publications from R&D
Growth rate of R&D budget
R
Budget growth
R&D budget
Potential number of R&D projects
B
Support service lags behind
Target support service expenditures
Support service availability
R
Support service growth
Support service availability gap
B
Free markets balance supply and demand
Support service development
Support service revenue
Increase in support services
Public investments in support services

Figure 9: Simplified diagram of the simulation model

Analysis of key variables is the model:

- **Effects on Available data resources for secondary use** (see Figure 10)**:**
  - We have assumed that the main driver for the pharmaceutical companies to conduct data driven research is the amount of health data available to use. We have not taken into consideration the possible effects of the number of donors relative to available data, or different kind of health data, but aggregated the effect of these in the single variable *Available data resources for secondary use*.
  - *Available data resources for secondary use* is a combination of the following data resources: *Data and samples in biobanks*, *Register data*, and *Availability of MyData*. Also, the *Data integration capability* affects the amount of data available, as it describes how the different data sources can be combined together, and thus affects the amount of data available, as the combined data can be seen as an own data source.

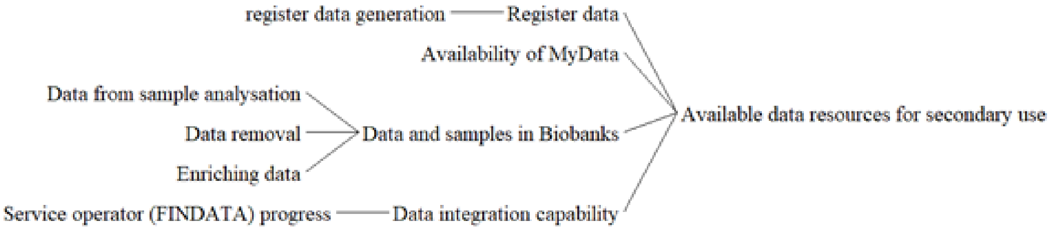

register data generation
Register data
Availability of MyData
Data from sample analysation
Data removal
Enriching data
Data and samples in Biobanks
Available data resources for secondary use
Service operator (FINDATA) progress
Data integration capability

Figure 10: Available data resources for secondary use

- **Effects on R&D project proposals** (see Figure 11)**:**
  - One of the main focal points of the model is to determine what affects the pharmaceutical R&D project starts or the R&D project proposals as the project starts is a delayed function of the project proposals. Figure 11 shows the variables affecting the R&D project proposals.
  - The main effect comes from the *Available data resources for secondary use*, which has been already discussed earlier. The other variables have lesser impact, although, they are still important. *R&D budget* has a major impact on the project proposals in the future if the number of R&D projects increases and the budget becomes a limiting factor for the projects. *Support service availability* has only a negative impact if the support services cannot match the requirements of the ongoing R&D projects. The effects of Findata come through the improved application process reducing the needed effort for accessing data and are a positive factor for project proposals.

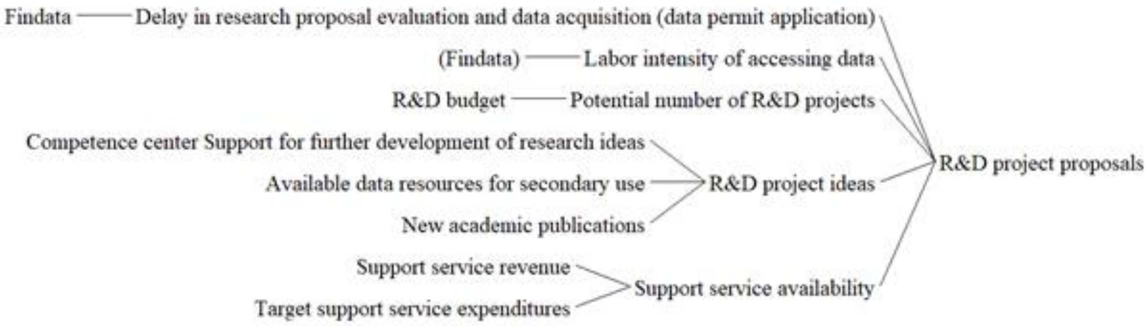

Findata
Delay in research proposal evaluation and data acquisition (data permit application)
(Findata)
Labor intensity of accessing data
R&D budget
Potential number of R&D projects
Competence center Support for further development of research ideas
Available data resources for secondary use
New academic publications
R&D project ideas
R&D project proposals
Support service revenue
Target support service expenditures
Support service availability

Figure 11: R&D project proposals

Loop analysis:

1. **Loop - Willingness to give biobank consents (B):**

    a. Describes how the number of biobank consents will saturate as the number of biobank consents increase, i.e. as the number of potential biobank donors decrease when consents are given and thus the likelihood of finding new biobank donors decreases. The availability of donor services (e.g. generic biobank information for citizens or online services for consent management and return of information) in a sense counterbalances the effect of this loop to some extend as long as the donor services are developed. The donor services block describes the likelihood for potential donors to be exposed to situations for giving consent and to eventually give their consent.

2. **Biobank revenue:**

    a. **Loop - Biobank revenue from R&D (R)**

        i. The revenue from R&D projects enables biobanks to collect, store and analyse more samples and thus there will be more data available for the future R&D projects, which increases the number of R&D projects.

    b. **Loop - Biobank revenue from research (R)**

        i. Same as the *Biobank revenue from R&D* loop but for academic research sector. The revenue from research projects enables biobanks to collect, store and analyse more samples and thus there will be more data available for the future research projects, which increases the number of research projects. However, the fees collected by the biobanks are much smaller compared to the fees collected from the pharmaceutical companies. Thus the overall effect of academic research projects is expected to be smaller.

3. **Data enrichment:**

    a. **Loop - Enriching data R&D (R)**

        i. When samples are analysed in R&D projects according to the biobank sample and data utilization contracts, the enriched data has to be returned back to the biobanks. This in turn increases the amount of available data resources for further use and therefore increases the likelihood of future projects.

        ii. We have assumed that the pharmaceutical companies enrich data to some extent, but at the same time this increases the costs of the R&D projects and thus also affects how many projects can be funded.

    b. **Loop - Enriching data research (R)**

        i. Same loop as the *Enriching data R&D* loop but for the academic research sector. When samples are analysed according to the biobank sample and data utilization contracts, the enriched data has to be returned back to the biobanks. This in turn increases the amount of available data resources for further use and therefore increases the likelihood of future projects.

4. **Findata:** There are three different loops operating through Findata, i.e. *Waiting time for accessing data*, *Resource intensity of data collection*, and *Data integration.* These

three mechanisms all have an effect on the number of project starts and they are all driven by Findata's service fees, and therefore by the number of R&D projects. That is, the more Findata collects service fees, the more it has funds to develop its services. Thus, service fees have an important role, although, the development of Findata is mainly funded by public investments.

a. **Loop - Waiting time for accessing data (R)**

   i. Describes the delay how long it takes for Findata to process the data applications and provide data permits. Findata develops the application process to reduce the evaluation time needed and to streamline the process.

   ii. We have assumed that the application processing delay has an effect on how interested the pharmaceutical companies are to initiate projects based on data.

b. **Loop - Resource intensity of data collection (R)**

   i. Describes the effort in person-months needed by the applicant to meet all the requirements demanded by the data application process. Findata develops the application process to reduce the amount of duplicated work needed and to streamline the process.

   ii. We have assumed that the labor intensity has an effect on how interested the pharmaceutical companies are to utilize the data.

c. **Loop - Data integration (R)**

   Describes the possibilities and efficiency of integration of different data sources. The ability to integrate data from different sources, namely from different national registers, healthcare service providers' registers and from biobanks, enhances the utility of the data. Data integration capability has therefore effect on the amount of available data resources for secondary use. The amount of data for secondary use affects in turn how interested the pharmaceutical companies are to utilize the data. Findata develops the integration capability of the different data sources.

5. **Academic publications:**

a. **Loop - New academic publications from research (R)**

   i. Describes the effect of publications resulting from academic research on R&D project ideas. Publications may lead to new R&D projects and thereby more service fees and returned data to biobanks and back to academic research. Biobank *revenue from R&D* and the loops containing Findata, i.e. *Waiting time for accessing* data, *Resource intensity of data collection*, and *Data integration*.

b. **Loop - New academic publications from R&D (R)**

   i. Describes the effect of publication from pharmaceutical R&D projects on project ideas. The project ideas affect the number of R&D projects and as ending R&D projects publish a certain amount of academic publications on average, then the ending projects feed back to the academic publications. The amount of academic publications of an

R&D project is expected to be lower compared an academic research project.

6. **The role of competence centers:**
   a. **Loop - Refinement of academic research (R)**
      i. Describes the effect of competence centers (centers of excellence) on academic research. It is assumed in this model that the establishment of competence centers lead to better coordinated research and increase of quality and volume of academic publications. This loop enhances directly the *Loop - New academic publications from research*
   b. **Loop - Refined research creates new ideas (R)**
      i. Describes the effect of competence centers on R&D projects. Otherwise works the same way as the *Loop - Refinement of academic research*.
7. **Loop - Budget growth (R)**
   a. *Describes the growth of the R&D budget of pharmaceutical companies.* This loop is one of the most important loops when considering the long term growth of the pharmaceutical companies' R&D expenditures, as the R&D budget is at some point in time going to saturate the growth of the number of R&D projects if it does not grow. The model assumes that the number of R&D project ideas affects the growth rate of R&D budget, i.e. if there is enough research ideas, then the budget is increased annually, thus causing an exponential growth defined by the annual growth rate.
8. **Support service:**
   a. **Loop - Support service lags behind (B)**
      i. Describes how the demand and supply of support services affects the project proposals. As the number of R&D projects is growing relatively fast, it is likely that the supply of services lags behind the demand, depending on the growth strategies (i.e. the *Loop - Free markets balance demand and supply*) of the companies comprising the support services. If the support services cannot meet the demands of the ongoing R&D projects, then it is likely that this affects the future project proposals, as the pharmaceutical companies have difficulties in accessing the services needed for successful projects. Public investment can be used to enhance the growth of the support services to level off the discrepancy of the demanded and offered services.
   b. **Loop - Free markets balance supply and demand (B)**
      i. Describes how the supply and demand is balanced, i.e. how the support services react to the growing number of R&D projects. The faster this loop operates, that is, the smaller the delay for service development is, the faster the service offering of the support service matches the demand and the less problems this loop causes. Public investment can be used to enhance the growth of the support services and thus balance the supply and demand. This loop, if working slowly, can cause the supply to lag far behind the demand, and thus, affect the loop *Support service lags behind* cause to hinder the growth of the pharmaceutical R&D.

c. **Loop - Support service growth (R)**

   i. Describes the growth of the support services in the long term. As the number of R&D projects increase, the support services try to match this growth by the *Loop - Free markets balance supply and demand*.

6.4 Simulations

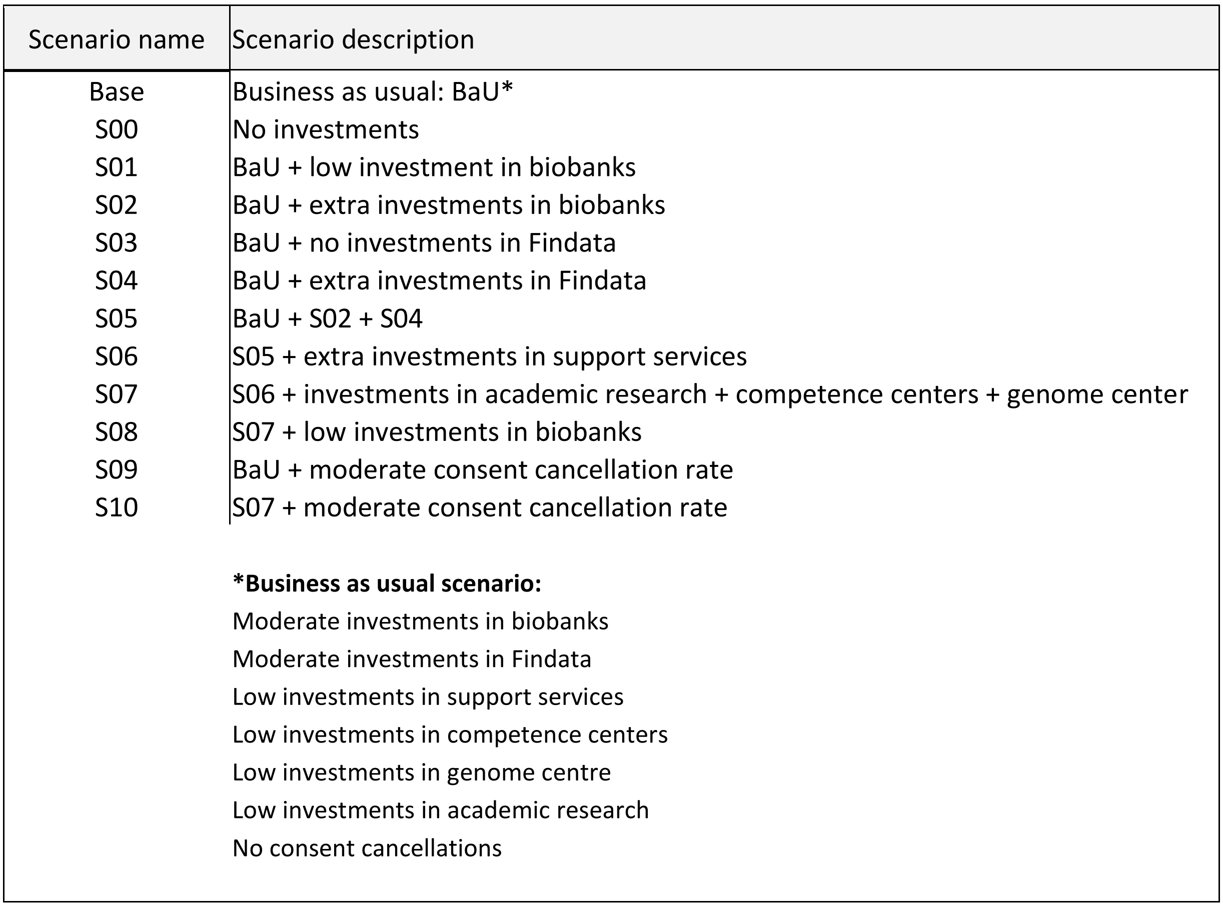
Simulation results of the number of R&D project starts is shown in Figure 12. The main parameter values are presented in Table 3. Table 4 contains an explanation for the scenarios presented.

In Figure 12 the preliminary simulation results are presented. As can be seen, the different scenarios show a huge variability in the number of pharmaceutical R&D project starts, from 50 to 150 projects per year. Our aim is to study what combination of public investment would be the most suitable strategy for the development of this business ecosystem. In here “most suitable” does not only mean optimal in terms of what amount of invested euros produces the highest amount of X (e.g. R&D project starts) in the system, but also the robustness of the scenarios under uncertainty. Currently, we haven’t been able to specify the exact parameter values of all the parameters or the exact formulation of cause and effect relationships, and therefore there is certain amount of uncertainty regarding these. More background information is being collected in order to improve the model and its input parameters.

Table 3: Main parameter values

Parameters	Value	Units
Biobanks and data		
Biobank unit price per project	0,05	MEuro/Project
Units of samples per person	3	Sample/Person
Sample accumulation per person	0,2	Sample/(Person*Year)
Units of data per analysed sample	5	Data/Sample
Average number of persons in a biobank project	5000	Person
Unit price of sample collection	5	Euro/Sample
Unit price of sample analysation	30	Euro/Sample
Unit price of data for storage and administration	0,5	Euro/Sample/Year
Share of analysed samples removed	0,5	Dmnl
max data growth rate	0	1/Year
Ref share of projects that need enriching	0,25	Dmnl
R&D projects (pharmaceutical companies)		
R&D project planning delay	1	Year
Ref R&D expenditures per project	1	MEuro/Project
Ref annual growth rate of R&D budget	0,03	1/Year
R&D budget growth starting year	0	Year
Share of total R&D budget to data projects	0,5	Dmnl
r&d project duration	2	Year
Time for R&D to perceive available data	0,2	Year
Time for R&D to perceive benefits of publications	1	Year
Time for R&D to perceive benefits of competence centers	1	Year
Time for R&D to perceive support service availability	1	Year
ref biobank service cost	0,05	MEuro/Project
Service operator (Findata)		
Service operator unit price per project	0,015	MEuro/Project
Service operator development start time	0	Year
Init labor intensity	0,8	Year/Project
Init evaluation delay	1	Year
Init data integration capability	0,5	Dmnl
min labor intensity	0,1	Year/Project
min evaluation delay	0,2	Year
Research (universities and research institutions)		
Ref potential number of research project starts per year	50	Project/Year
Average expenditure of research project	1	MEuro/Project
Avg duration of academic research project	2	Year
Support service		
Adjustment time of support service development	2	Year
support service unit price per project	0,025	MEuro/Project/Year
Publications		
Delay for academic publishing	1	Year
publications per R&D project	0,5	Publication/Project
publications per research project	2	Publication/Project

Table 4: Simulation scenarios



Scenario name	Scenario description
Base	Business as usual: BaU*
S00	No investments
S01	BaU + low investment in biobanks
S02	BaU + extra investments in biobanks
S03	BaU + no investments in Findata
S04	BaU + extra investments in Findata
S05	BaU + S02 + S04
S06	S05 + extra investments in support services
S07	S06 + investments in academic research + competence centers + genome center
S08	S07 + low investments in biobanks
S09	BaU + moderate consent cancellation rate
S10	S07 + moderate consent cancellation rate

***Business as usual scenario:**
Moderate investments in biobanks
Moderate investments in Findata
Low investments in support services
Low investments in competence centers
Low investments in genome centre
Low investments in academic research
No consent cancellations

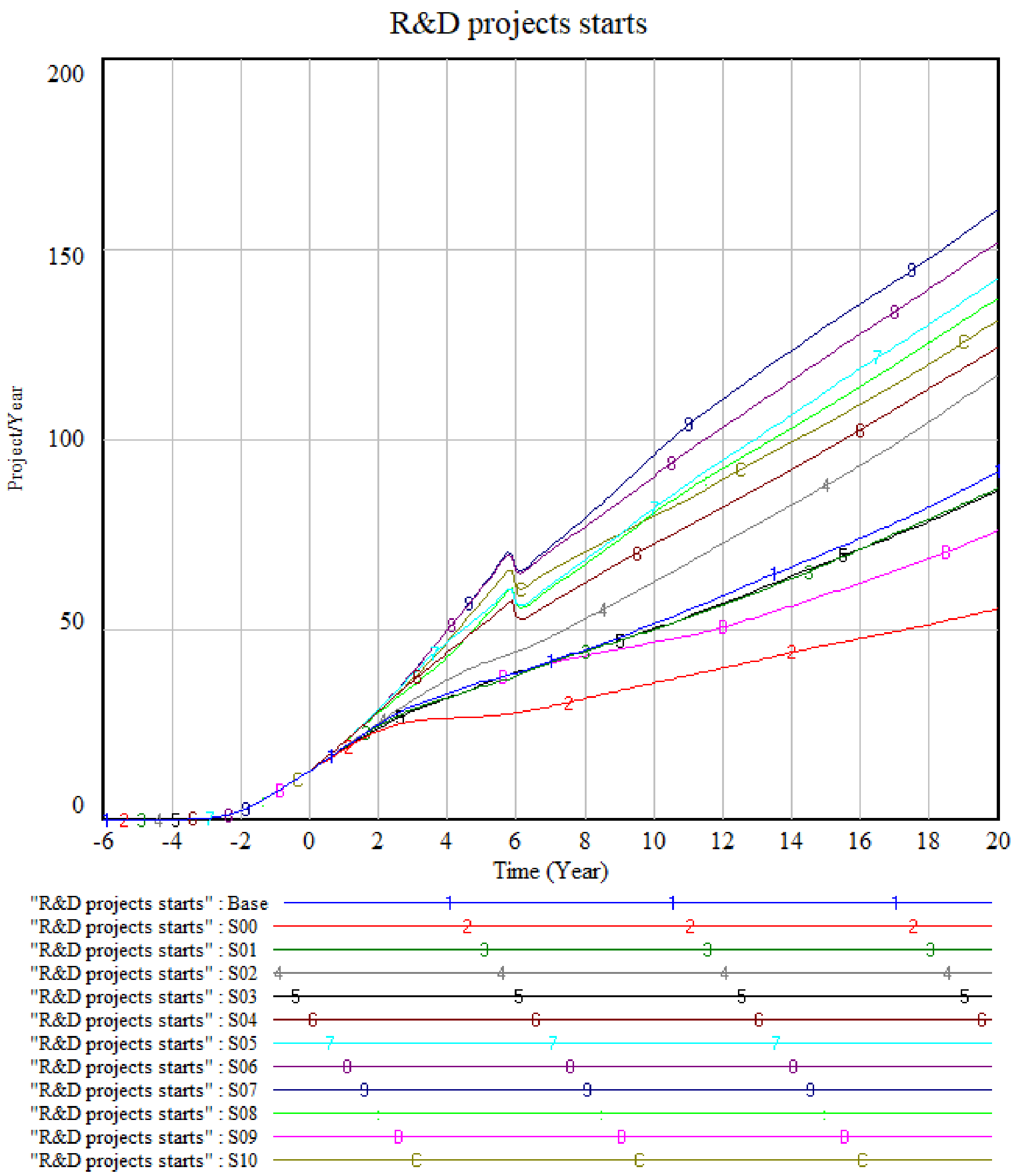


Figure 12: Preliminary simulation results.

It is difficult to estimate the effect of some of the interconnections described in the model as there is no data available in some of the cases. For example, what is the effect of the variable *Labor intensity of accessing data* on R&D project proposals? At the moment, we have estimated the effect as well as possible, and as the project continues, we are going improve the estimation based on comments from PreMed project participants and external experts. Also, we have conducted extensive sensitivity analyses to study the effects of different interconnections in the model and to identify the key variables and feedback loops in the system. As, even though all feedback loops have an effect on the system, some are significantly more important when considering the future development of the ecosystem.

6.5 Results

Preliminary analysis of the model suggests that public investments in biobank operations (e.g. for sample analysation) are very important at this point of the development of the ecosystem. As the revenue stream from data utilization service fees is not nearly enough to cover the expenses of the biobanks. Also, in order for the ecosystem to be healthy in the long run, the biobanks have to find different ways of monetarizing the samples and data they base their business.

The results given in Figure 12 are only preliminary and we continue to specify the parameter values used as well as to improve the model structure. The model and parameter settings will be adjusted during the third phase of the project. Updated results along with analysis and discussion are expected during first half of the year 2020.

7. Follow-up of the PreMed project domain

7.1 Dissemination and networking activities

During its second phase the project has organised three workshops where topical presentations related to secondary use of data have been given and the progress of the PreMed project has been reported. In addition to project participants, also external experts have been participated as invited guests. Other dissemination and networking activities include: (1) two poster presentations in scientific conferences (NBCC 2019, EMBC 2019), (2) other conference presentations (HIMSS, DHN, ECHA, Smart Health for Europe), (3) FinnGen and Health Tuesday events, (4) continuously updated project website, (5) news release (12.3.2019) giving rise to two network magazine articles and (6) VTT blog (21.3.2019). Additionally, several presentations on PreMed have been given in the context of visits of international groups in Finland (Korea, China/Sichuan, Singapore).

7.2 Roadmap

The PreMed project has provided a compact roadmap table highlighting important developments towards the data-driven precision medicine ecosystem and related infrastructures (Table 5). The national infrastructure for secondary use of data took a step forward in May 2019 as the act on secondary use of health and social services data came into effect. Consequently, the new data permit authority (Findata) was established, with first services gradually starting during year 2020. Another important step towards centralised services was the opening of the Fingenious service of the Finnish Biobank (FinBB).

Fingenious enables feasibility and access requests to be done for all hospital biobanks and the THL biobank via one joint service. The relation between Findata and FinBB remains still unclear, but it is expected that a "one-stop shop" covering also services to link biobank data and samples with national registry data is targeted and could be available by 2022. Based on information from Findata web site, the Findata services will gradually start during 2020 and will be extended to Kanta services by 2021. Progress of the national centers of excellence has been slower than expected. However, it seems that the remaining three planned centers (neuro, pharma development and genome) could finally be established in year 2020. Development of additional centers covering other areas of healthcare might start in 2021. The enforcement of the genome law will be followed by the development of the genome database. We expect the first version of the genome database to be in operation in 2022 in the most positive case. The genome center web page does not provide any estimate on the time schedule[1]. The on-going FinnGen project has exceeded expectations in collecting sample donors. It is expected that the project is able to achieve the targeted amount (500 000) of genotyped participants by 2023 as planned. The project attracted two new pharma company partners (GSK and Sanofi) in 2019. Negotiation with additional partners is on-going and are expected to be published in 2020. Open API interfaces provided by new EHR systems will improve the possibilities to integrate AI-based and data-driven tools with healthcare processes. The usage of the new Apotti system will considerably increase after the second deployment in May 2020.

Kanta services are the core of the national infrastructure enabling access to health data and since year 2019 also for social data. The MyKanta (OmaKanta) services will include social services data in the beginning of 2020. Prominent enhancements of the Kanta system also include the introduction of the Kanta PHR service ("Omatietovaranto") which enables health and wellness applications to be connected via an open API interface. Value of the service is expected to increase later (expected 2022), when data can be exchanged between health or social care organization and the citizen via the PHR. Currently, the PHR does not enable such data sharing. The Kanta PHR enables the citizens to manage and share their health data. Thereby, it is a potential solution for sharing personal data also for secondary use.

The reform of the Finnish social and health services system has been delayed, but the expectation is that the work proceeds and leads into improved service delivery system. Efficient exploitation of data (e.g. in management and value-based healthcare) is expected to be an important element of the new service system.

New EU regulation - clinical trial regulation (CTR) and medical device regulation (MDR) - will be enforced in 2020. The new CTR brings a major change to the way clinical trials are conducted in the EU. In particular, it will harmonize practices and improve collaboration among the member states. The new MDR will significantly expand the definition of medical devices. In particular, software used for a medical purpose (e.g. a decision support system) will fall under class IIa or higher and require certification by a notified body. The new in-vitro diagnostics regulation (IVDR) will be enforced in 2022.

As indicated in Table 6 the services for biobank donors are still scarce. However, in line with the GDPR, register controllers would need to provide information for donors concerning the use of biobank samples and patient data. Consequently, there is currently strong interest towards setting up services for biobank donors, and we believe that new services will gradually become available in 2021.

As precision medicine projects are consuming substantial public investments, they are expected to provide direct benefits in healthcare. We expect direct impacts to be seen e.g. through the adoption of pharmacogenomics and polygenic risk scores in healthcare.

Finland will host the HIMSS Europe and Health 2.0 international conference during years 2019, 2020 and 2021. Already the first conference of the series was a success and enabled Finland to promote its capabilities and infrastructures for data-driven precision medicine. The events of 2020 and 2021 provide an opportunity to attract further interest towards Finland.

The PreMed project ends in the end of 2020. The PreMed PGx study is expected to provide a useful model for new industry-driven biobank projects and valuable experience for biobanks, national registries, FinBB and Findata to improve data access processes. During year 2020 a follow-on project will be prepared with the objective to start PreMed 2.0 in 2021. Concrete impact, e.g. new business opportunities and products based on data are expected

[1] http://www.genomikeskus.fi/en/q-a.html

in year 2022. A permanent organisation to coordinate data-driven precision medicine ecosystem may turn out to be needed and could be established in 2023.

Table 5. Data-driven precision medicine roadmap

	2019	2020	2021	2022	2023 →
Biobanks	• FinBB / Fingenious: joint feasibility and access requests [3] [12]	• Biobank Act renewal in force		• one-stop shop for joint biobank studies	
Secondary use of data	• Legislation on Secondary use of social and health data in force [4]	• FinData services start gradually [2]	• Findata response times comply with secondary use legislation and apply also to Kanta [4]		• services cover hospital data lakes and all Kanta data [1]
National centers of excellence (genome, cancer, neuro, pharma development)	• draft genome law released for comments [8] • establishment of the national cancer center [6]	• establishment of other national centers of excellence [5][7][8]: neuro, pharma development, cancer, genome • genome law [13]	• preparation of additional national centers starts	• first version of genome database in operation	• genome database integrated with patient information systems
Projects	• FinnGen data available: 180k [10], new partners join • Apotti in limited use	• FinnGen data available for 250k individuals, new partners join, second phase starts • Apotti 2nd deployment phase [9]	• FinnGen data available: 320k • business arising from FinnGen (in Finland): 15 M€ [11] • BF: PH and AI programs completed	• FinnGen data available: 390k	• FinnGen data available: 500k • business arising from FinnGen (in Finland): 48 M€ [11]
Finnish healthcare system	• centralized service for social data in use (Kansa) •	• more connected Kanta PHR apps (in addition to Terveyskylä) [14] • social data in MyKanta		• healthcare professionals' access to personal data in Kanta PHR	• Kanta data in secondary use • Clinical data accessible in Kanta PHR • SOTE reform in force
International	• large number of precision medicine initiatives ongoing • cross-border ePrescription with Estonia • Finnish EU presidency • HIMSS Europe and Health 2.0 2019 in Finland	• EU clinical trial regulation enforced [15] • EU medical device regulation (MDR) enforced [16] • UK Brexit • HIMSS Europe and Health 2.0 2020 in Finland	• Donor's access services of biobanks becoming increasingly available as a consequence of legislation • HIMSS Europe and Health 2.0 2021 in Finland	• Impact of precision medicine in healthcare (pharmacogenomics, polygenic risk scores, etc.) • EU in vitro diagnostics regulation (IVDR) enforced [16]	•
Business ecosystem	• PreMed: biobank study execution and ecosystem simulation tool development • discussion on parallel collaboration projects	• PreMed: system dynamics model for data exploitation • recommendations for public bodies to boost growth • preparation of follow-on project (PreMed 2.0)	• start of PreMed 2.0 targeting at building international co-operation and business opportunities (content and target country to be defined)	• PreMed 2.0 continued. Project impact in opening business opportunities for data-driven products and services	• PreMed 2.0: establish permanent organisation for data-driven ecosystem coordination

7.3 International developments

In phase 1 of the PreMed project a review of international activities in precision medicine was carried out[1]. During phase 2 we have followed up some of the ongoing major international initiatives. Table 6 shows examples of major initiatives with information about study approach, clinical focus, number of participants, and funding. The table also summarizes the policies of the initiatives concerning access to data and/or samples by researchers, industry and the donors themselves.

The initiatives are based on different strategies in enrolling participants. Some of the initiatives (e.g. All of Us and UK Biobank) try to attract a wide variety of citizens as participants and a substantial proportion of their participants are healthy volunteers. Following up such populations enables risk factors of diseases of common diseases to be detected and evaluated. Other initiatives (such as FinnGen and 100k genomes project) recruit most or all of the participants in healthcare settings. These cohorts are dominated by individuals with existing health conditions. Thereby, the collected data and sample resources are highly valuable in detailed studies on individual diseases (including rare diseases) and related therapies. National Human Genome Research Institute (NHGRI) and Australian Genomics are examples of programmes enabling collaboration and funding of research, forming a basis for a wide range of individual research projects. In terms of absolute number of participants the All of Us project is highest with targeted cohort of one million individuals. FinnGen with the targeted 500000 participants is comparable to corresponding initiatives of UK and China. When measured by participants per population or by invested funding per population Estonian biobank and FinnGen are unique in size.

Concerning the objectives of the PreMed project, it is of interest to observe how the data and sample collections gathered in the initiatives are available for use for the different stakeholder groups: external researchers, companies and donors. Most of the initiatives enable access to data and/or samples for external researchers by default. The access is always subject to a research plan with valid scientific objectives. There is more variability in the policies of granting access for companies. Most of the initiatives (e.g. FinnGen, UK Biobank, 100k Genomes Project and All Of Us) enable the use of data and samples by companies as long as the use is based on a valid research plan with scientific objectives.

The FinnGen the project group includes researchers from both universities and pharma companies. After the project, the genome data is returned to Finnish biobanks, where it is available for scientific research (including research carried out by companies). Some of the initiatives (e.g. Estonian Biobank and the EPIC study) have not explicitly communicated at their web pages their policy towards industry driven studies. Kadoorie biobank limits the access to public research organizations and health service organisations. None of the initiatives listed Table 6 disclose data for industrial R&D without a valid research plan with scientific research setting and objectives.

Accessing own data is increasingly seen as a fundamental right of the donor and also an important opportunity for the biobanks and research projects to attract volunteers to participate. As seen in Table *6*, not much have been done yet to enable the donors to access data retrieved from their samples (e.g. genomic variants) or even to access general information about the studies where their data or samples have been used. The All of Us project has recently announced a genetic counselling service, which would enable to share

[1] PreMed Phase 1 report: https://cris.vtt.fi/ws/portalfiles/portal/20131333/PreMed_ecosystem_stakeholder_needs_and_opportunities.pdf

genotyping and sequencing results with the donors along with appropriate counselling[1]. The Estonian Biobank has started providing data for the donors. So far, 2000 donors (1% of total participants) have received their polygenic risk score (PRS) along with counselling[2]. Helsinki Biobank has carried out a study concerning the return of results to the donors[3]. The study provides a proposal for a joint donor portal for the Finnish biobanks[4].

The 100k Genomes Project provides clear guidance for the participants concerning return of findings[5]. Findings related to the original disease (participant's reason to participate) are always returned to the participant. The participant needs to accept this in order to be included in the study. Additionally, the participant may choose to allow "additional findings" to be looked for and returned - e.g. genomic variants. Any findings will be first fed back to the NHS, to confirm the result. A clinician then gives the results to the participant along with discussion on their meaning. The 100k Genomes Project web page does not provide information on how extensively findings have been returned in practice.

[1] https://allofus.nih.gov/news-events-and-media/announcements/nih-funds-genetic-counseling-resource-ahead-million-person-sequencing-effort
[2] https://www.slideshare.net/THLfi/andres-metspalu-the-estonian-genome-project
[3] https://www.finngen.fi/sites/default/files/inline-files/Return%20of%20results_250419.pdf
[4] https://www.vtt.fi/sites/premed/Documents/Vaatimusmaarittely%20osallistamisportaalille.pdf
[5] https://www.genomicsengland.co.uk/information-for-participants/findings/

Table 6. Examples of precision medicine initiatives.

Initiative	Approach	Clinical focus	N / share of population (thousands / %)	Yearly funding total (M€)	Yearly funding per population (€)	Access to external researchers	Access to industry	Donor’s services
All of Us / NIH (USA) [18]	prospective cohort	all diseases	1000 / 0.30%	258[1]	0,8	yes	yes	Genetic counselling service being developed
UK Biobank [19]	prospective cohort	all diseases	500 / 0.75%	15[2]	0,2	yes	yes	n/a
China Kadoorie biobank (China / UK) [20]	prospective cohort	chronic diseases	500 / 0.04%	n/a	n/a	yes	no (only HC organizations)	n/a
Estonian Biobank [17]	prospective cohort	all diseases	200 / 15%	2,3[3]	1,7	yes	yes (not ruled out)	Yes, partial (PRS + counselling)
EPIC study / IARC-WHO [21]	prospective cohort	chronic diseases	520 / 0.14%	n/a	n/a	yes	yes (not ruled out)	n/a
100k Genomes Project (UK) [23]	research project	rare diseases, cancer	85 / 0.13%	45[4]	0,7	yes	yes	Results return process defined and documented (no online service)
FinnGen (Finland) [24]	research project	all diseases	500 / 9.1%	10[5]	1,8	yes	yes	Right to get information on findings (no online service)
NHGRI / NIH (USA) [25]	research/collaboration/ funding programme	genetic diseases, cancer	n/a	325[6]	1,0	yes	depending on project	n/a
Australian Genomics [26]	research/collaboration/ funding programme	rare diseases, cancer	n/a	18[7]	0,73	n/a	n/a	Online service for consent giving

[1] US congress approval for year 2018: https://allofus.nih.gov/news-events-media/media-toolkit/all-us-research-program-backgrounder
[2] MRC and Wellcome Trust funding: https://www.ukbiobank.ac.uk/wp-content/uploads/2018/10/Funding-UK-Biobank-summary.pdf
[3] Based on year 2019 investment: https://www.sm.ee/en/news/genome-project-100000-samples-collected-2019-least-50000-more-people-can-join
[4] Based on public funding decision for 2015-2020: https://www.pharmaceutical-journal.com/news-and-analysis/news-in-brief/funding-for-genomics-england-to-reduce-by-40m-under-2019/2020-dhsc-spending-plans/20206548.article?firstPass=false
[5] Based on budgeted funding of 59 M€ for 2017 - 2022: https://www.finngen.fi/en/node/38
[6]US government funding budgeted for 2019, https://www.genome.gov/Pages/About/Budget/NHGRIFY2019CJ.pdf
[7] Based on announced funding round, https://www.australiangenomics.org.au/grants-round-opens-for-65-million-genomics-research-fund/

8. Conclusions

PreMed project activities carried out in the phase 2 (1.11.2018 - 31.12.2019) have been reported. The activities have been focused in three main areas: biobank study (research protocol, data collection, data analysis environment setup), ecosystem simulation model and dissemination (project workshops and other events).

The biobank study carried out in the PreMed project aims to provide evidence on the benefits of genome-based tests in the context of drug therapy. In this way, it is expected to promote the use of genome tests and related services in healthcare. At more general level, PreMed provides information and experiences of the processes related to healthcare data access. Thereby, it is expected to lower the threshold for companies to exploit healthcare data and results from retrospective studies in their business.

The biobank study lags behind the original time schedule by six months due to delays in the data collection process. The six months extension from original project ending (30.6.2020) has been accepted by the project partners and Business Finland. Despite of the delay, the objectives of the study have not been changed. All needed data resources seem to be accessible and all needed administrative decisions concerning data disclosure have been made.

The project has already been of benefit in pointing out bottlenecks in accessing data resources especially related to combining data from different registers. These experiences are expected to be valuable in the further development of services for data access, in particular those of Findata and FinBB.

The data-driven precision medicine ecosystem model and its implementation as a simulator was carried out in two parts. The first version with limited functionality was developed and completed by July 2019. Subsequently, the simulator was enhanced by adding several new features during the rest of the year. The primary outcome measure of the simulator is the volume of real world data projects making use of data obtained from biobanks and national registers. The ecosystem model enables the simulation of various ecosystem evolution paths affected by different public investment strategies. The effect of investments to biobanks, Findata, academic research, national centers of excellence and support services provided by SME companies can be estimated. The model and its parameters are still under development. Initial results have been provided in the report. The simulator is expected to be valuable for public entities and authorities in providing support for the selection of financing strategies to boost ecosystem growth. Also other stakeholders can find the simulator useful in increasing the understanding of the ecosystem dynamics.

During its second phase the project has organised three workshops where topical presentations related to secondary use of data have been given and the progress of the PreMed project has been reported. The Finnish health data resources are globally unique. They provide an excellent basis for world class products and services in the health and wellness domain. One important objective of the project is to ensure that opportunities for companies to exploit data in compliance with GDPR will be taken into account. PreMed is in a good position to achieve this objective thanks to the strong participation of companies.

References

Aithal, Guruprasad P., Christopher P. Day, Patrick JL Kesteven, and Ann K. Daly. 1999. “Association of Polymorphisms in the Cytochrome P450 CYP2C9 with Warfarin Dose Requirement and Risk of Bleeding Complications.” *The Lancet* 353(9154):717–19.

Gage, BF, C. Eby, JA Johnson, E. Deych, MJ Rieder, PM Ridker, PE Milligan, G. Grice, P. Lenzini, AE Rettie, CL Aquilante, L. Grosso, S. Marsh, T. Langaee, LE Farnett, D. Voora, DL Veenstra, RJ Glynn, A. Barrett, and HL McLeod. 2008. “Use of Pharmacogenetic and Clinical Factors to Predict the Therapeutic Dose of Warfarin.” *Clinical Pharmacology & Therapeutics* 84(3):326–31.

Holmberg, Mikko T., Aleksi Tornio, Maria Paile-Hyvärinen, E. Katriina Tarkiainen, Mikko Neuvonen, Pertti J. Neuvonen, Janne T. Backman, and Mikko Niemi. 2018. “*CYP3A4*22* Impairs the Elimination of Ticagrelor, But Has No Significant Effect on the Bioactivation of Clopidogrel or Prasugrel.” *Clinical Pharmacology & Therapeutics*.

Ji, Yuan, Jennifer M. Skierka, Joseph H. Blommel, Brenda E. Moore, Douglas L. VanCuyk, Jamie K. Bruflat, Lisa M. Peterson, Tamra L. Veldhuizen, Numrah Fadra, Sandra E. Peterson, Susan A. Lagerstedt, Laura J. Train, Linnea M. Baudhuin, Eric W. Klee, Matthew J. Ferber, Suzette J. Bielinski, Pedro J. Caraballo, Richard M. Weinshilboum, and John L. Black. 2016. “Preemptive Pharmacogenomic Testing for Precision Medicine.” *The Journal of Molecular Diagnostics* 18(3):438–45.

Kubica, Aldona, Marek Kozinski, Grzegorz Grzesk, Tomasz Fabiszak, Eliano Pio Navarese, and Aleksander Goch. 2011. “Genetic Determinants of Platelet Response to Clopidogrel.” *Journal of Thrombosis and Thrombolysis* 32(4):459–66.

Lehto, Mika, Jussi Niiranen, Pasi Korhonen, Juha Mehtälä, Houssem Khanfir, Fabian Hoti, Riitta Lassila, and Pekka Raatikainen. 2017. “Quality of Warfarin Therapy and Risk of Stroke, Bleeding, and Mortality among Patients with Atrial Fibrillation: Results from the Nationwide FinWAF Registry.” *Pharmacoepidemiology and Drug Safety* 26(6):657–65.

Mega, Jessica L., Joseph R. Walker, Christian T. Ruff, Alexander G. Vandell, Francesco Nordio, Naveen Deenadayalu, Sabina A. Murphy, James Lee, Michele F. Mercuri, Robert P. Giugliano, Elliott M. Antman, Eugene Braunwald, and Marc S. Sabatine. 2015. “Genetics and the Clinical Response to Warfarin and Edoxaban: Findings from the Randomised, Double-Blind ENGAGE AF-TIMI 48 Trial.” *The Lancet* 385(9984):2280–87.

Monie, Dileep D. and Emma P. DeLoughery. 2017. “Pathogenesis of Thrombosis: Cellular and Pharmacogenetic Contributions.” *Cardiovascular Diagnosis and Therapy* 7(Suppl 3):S291–98.

Neuvonen, Mikko, E. Katriina Tarkiainen, Aleksi Tornio, Päivi Hirvensalo, Tuija Tapaninen, Maria Paile-Hyvärinen, Matti K. Itkonen, Mikko T. Holmberg, Vesa Kärjä, Ville T. Männistö, Pertti J. Neuvonen, Jussi Pihlajamäki, Janne T. Backman, and Mikko Niemi. 2018. “Effects of Genetic Variants on Carboxylesterase 1 Gene Expression, and Clopidogrel Pharmacokinetics and Antiplatelet Effects.” *Basic & Clinical Pharmacology & Toxicology* 122(3):341–45.

Poistula, M., A. Kaplon-Cieslicka, M. Rosiak, A. Kondracka, A. Serafin, K. J. Filipiak, A. Czlonkowski, G. Opolski, and P. K. Janicki. 2011. “Genetic Determinants of Platelet Reactivity during Acetylsalicylic Acid Therapy in Diabetic Patients: Evaluation of 27 Polymorphisms within Candidate Genes.” *Journal of Thrombosis and Haemostasis* 9(11):2291–2301.

Pouplard, Claire, Pascale Cornillet-Lefebvre, Redha Attaoua, Dorothée Leroux, Carinne Lecocq-Lafon, Jérôme Rollin, Florin Grigorescu, Philippe Nguyen, and Yves Gruel. 2012. “Interleukin-10 Promoter Microsatellite Polymorphisms Influence the Immune Response to Heparin and the Risk of Heparin-Induced Thrombocytopenia.” *Thrombosis Research* 129(4):465–69.

Salter, Benjamin S., Menachem M. Weiner, Muoi A. Trinh, Joshua Heller, Adam S. Evans, David H. Adams, and Gregory W. Fischer. 2016. “Heparin-Induced Thrombocytopenia.” *Journal of the American College of Cardiology* 67(21):2519–32.

Schulman, S., C. Kearon, and Subcommittee on Control of Anticoagulation of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis. 2005. “Definition of Major Bleeding in Clinical Investigations of Antihemostatic Medicinal Products in Non-Surgical Patients.” *Journal of Thrombosis and Haemostasis* 3(4):692–94.

Sistonen, Johanna, Silvia Fuselli, Jukka U. Palo, Neelam Chauhan, Harish Padh, and Antti Sajantila. 2009. “Pharmacogenetic Variation at CYP2C9, CYP2C19, and CYP2D6 at Global and Microgeographic Scales.” *Pharmacogenetics and Genomics* 19(2):170–79.

Sterman, J. D. Business Dynamics: Systems Thinking and Modeling for a Complex World. New York, NY, USA: McGraw-Hill Companies, 2000,. p. 21-23

Tatarunas, Vacis, Nora Kupstyte, Remigijus Zaliunas, Agne Giedraitiene, and Vaiva Lesauskaite. 2017. “The Impact of Clinical and Genetic Factors on Ticagrelor and Clopidogrel Antiplatelet Therapy.” *Pharmacogenomics* 18(10):969–79.

Tornio, Aleksi, Rob Flynn, Steve Morant, Elena Velten, Colin N. A. Palmer, Thomas M. MacDonald, and Alex S. F. Doney. 2018. “Investigating Real-World Clopidogrel Pharmacogenetics in Stroke Using a Bioresource Linked to Electronic Medical Records.” *Clinical Pharmacology & Therapeutics* 103(2):281–86.

Tseng, Andrew S., Reema D. Patel, Heidi E. Quist, Adrijana Kekic, Jacob T. Maddux, Christopher B. Grilli, and Fadi E. Shamoun. 2018. “Clinical Review of the Pharmacogenomics of Direct Oral Anticoagulants.” *Cardiovascular Drugs and Therapy* 32(1):121–26.

Vandell, A. G., J. Lee, M. Shi, I. Rubets, K. S. Brown, and J. R. Walker. 2018. “An Integrated Pharmacokinetic/Pharmacogenomic Analysis of ABCB1 and SLCO1B1 Polymorphisms on Edoxaban Exposure.” *The Pharmacogenomics Journal* 18(1):153–59.

Verbelen, M., M. E. Weale, and C. M. Lewis. 2017. “Cost-Effectiveness of Pharmacogenetic-Guided Treatment: Are We There Yet?” *The Pharmacogenomics Journal* 17(5):395–402.

Web links

[1] Kanta secondary use, https://thl.fi/web/tiedonhallinta-sosiaali-ja-terveysalalla/mita-tiedonhallinta-on-/asiakas-ja-potilastietojen-harmonisointi-toisiokayttoon

[2] FinData services, https://www.findata.fi/palvelut/palvelut-asiakkaille/

[3] FinBB Fingenious, https://finbb.fi/fi/etusivu/projektit/

[4] Secondary use legislation, https://www.finlex.fi/fi/laki/alkup/2019/20190552

[5] https://stm.fi/laakekehityskeskus

[6] https://stm.fi/syopakeskus,https://stm.fi/artikkeli/-/asset_publisher/kansallinen-syopakeskus-on-perustettu

[7] https://stm.fi/neurokeskus

[8] https://stm.fi/genomikeskus

[9] Apotti deployment, http://www.logy.fi/media/seminaarikuvat/julkishallinnon-hankinnan-ja-logistiikan-paiva/05_eija-isolahti_apotti-hanke-husin-nakokulmasta22052019.pdf

[10] FinnGen status update (Aarno Palotie), https://www.finngen.fi/sites/default/files/inline-files/FinnGen_ecosyst_day_Palotie.pdf

[11] FinnGen business estimate (Health Tuesday, 4.9.2018)

[12] IHAN webinar 16.9.2019, https://www.slideshare.net/SitraHyvinvointi/ihan-make-data-work-for-peoples-health

[13] Genomics to Healthcare event (Sandra Liede), https://www.slideshare.net/THLfi/sandra-liede-the-finnish-genome-strategy-and-the-genome-law

[14] Terveyskylä/Omapolku, https://www.terveyskyla.fi/palvelut/omapolku-palvelukanava-ja-digihoitopolut/omaseurantalaitteet-omapolulla

[15] EU Clinical Trials Regulation status, https://ehaweb.org/organization/newsroom/news-and-updates/current-status-of-the-clinical-trials-regulation/

[16] EU MDR and IVDR status, https://www.ema.europa.eu/en/human-regulatory/overview/medical-devices

[17] Anders Metspalu: Estonian Genome Project, https://www.slideshare.net/THLfi/andres-metspalu-the-estonian-genome-project

[18] All of Us programme, https://allofus.nih.gov/

[19] UK Biobank, https://www.ukbiobank.ac.uk/

[20] China Kadoorie Biobank, https://www.ckbiobank.org/site/

[21] EPIC Study, https://epic.iarc.fr/

[22] Estonian Biobank, https://genomics.ut.ee/en

[23] 100k Genomes Project, https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/

[24] FinnGen, https://www.finngen.fi/en

[25] National Human Genome Research Institute, https://www.genome.gov/

[26] Australian Genomics, https://www.australiangenomics.org.au/

Terms and acronyms

Term / acronym	Description
ADR	adverse drug reaction
AI	artificial intelligence
API	application programming interface
ATC	Anatomical Therapeutic Chemical
B2B	business-to-business
B2C	business-to-consumer
BBMRI	Biobanking and BioMolecular resources Research and Infrastructure
CDSS	clinical decision support system
CSV	comma separated file
DDD	Defined Daily Dose
DHN	Digital Health Nordic
ECHA	European Connected Health Alliance
EHR	electronic health record
FDA	Food and Drug Administration
FINBB	Finnish Biobank co-operative
FTP	file transfer protocol
GDPR	general data protection regulation
HCRU	healthcare resource use
HIMSS	Healthcare Information and Management Systems Society
HUS	The Hospital District of Helsinki and Uusimaa
ICT	information and communications technology
INR	international normalized ratio
ISTH	International Society on Thrombosis and Haemostasis
IVDR	in-vitro diagnosis regulation
Kela	The Social Insurance Institution of Finland
MDR	medical device regulation
MTA	Material Transfer Agreement
NBCC	Nordic-Baltic Cardiology Conference
NHGRI	National Human Genome Research Institute
NIH	National Institute of Health
PGx	pharmacogenomics
PHR	personal health record
PM	precision medicine
PRS	polygenic risk score
RTF	rich text format
SD	system dynamics

SFTP	secure file transfer protocol
SMEs	small and medium-sized enterprises
STM	Ministry of Social Affairs and Health
THL	National Institute for Health and Welfare
TTR	time in therapeutic range