DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
12

SULEV REISBERG

Developing Computational Solutions for Personalized Medicine

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

12

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

12

SULEV REISBERG

Developing Computational Solutions for Personalized Medicine

1632

UNIVERSITY OF TARTU
Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in informatics on June 17, 2019 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor

Prof. PhD. Jaak Vilo
University of Tartu
Tartu, Estonia

Opponents

Prof. MD. Dan Mark Roden
Vanderbilt University School of Medicine
Nashville, Tennessee, USA

Dr. PhD. Patrick Kemmeren
Princess Máxima Center for Pediatric Oncology
Utrecht, Netherlands

The public defense will take place on August 29, 2019 at 15:15 in J. Liivi 2-404, Tartu.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906
ISBN 978-9949-03-091-0 (print)
ISBN 978-9949-03-092-7 (PDF)

University of Tartu Press
http://www.tyk.ee/

To my lovely family,
supporting parents
and inspiring colleagues

ABSTRACT

The general idea of personalized medicine is to provide more effective clinical care and prevention of diseases by utilizing individual differences mostly in genetics, but also in detailed electronic health records (EHR) and other data. I provide an overview of the definition and its elements of personalized medicine, also the current state of the field. To date, personalized medicine is used in oncology and for testing developmental diseases in children. For more broader use, several challenges need to be dealt with. Some of these are addressed in this thesis. We show, by using genetic data from Estonian Biobank and 1000 Genomes Project, that polygenic risk score models are biased towards Europeans and should not be used for people from other populations. Similarly, frequencies of single nucleotide variants associated with asthma and liver diseases among Estonians are close to Europeans but different from the others. To bring personalized medicine into state-level clinical use, one has to integrate them to the workflows of the EHR systems. By combining genetic data and EHR of the participants of Estonian Biobank, we conducted a phenome-wide association study, by utilizing genetic variants related to asthma and liver diseases in order to find new gene-disease associations. Although we did not identify novel associations, we showed that this data could be effectively used for validation studies. Finally, we describe a pharmacogenomics recommendation pipeline for producing individual pharmacogenomic recommendations for 44,000 gene donors. We show that genotyping arrays with imputation can be used as cost-effective alternatives for whole genome sequencing in pharmacogenomic testing.

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
EHR	Electronic health record
Genome	Complete set of DNA of an individual organism
Genotype	Genetic data of an individual, usually used in a context of data from genotyping arrays (covering only parts of the full genome) or single position of the genome
GWAS	Genome-wide association study
FDA	U.S. Food and Drug Administration
Phenome	Physical characteristics (as a whole) of an individual
Phenotype	Physical characteristics of an individual
PheWAS	Phenome-wide association study
SNP	Single nucleotide polymorphism; same as *SNV*, but has to be frequent enough (usually considered 1%) worldwide
SNV	Single nucleotide variant (in a genome)

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

Ref. I

Reisberg S, Iljasenko T, Läll K, Fischer K, Vilo J. **Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations.** PloS one. 2017 Jul 5;12(7):e0179238. Full text given on page 65.

Ref. II

Reisberg S, Galwey N, Avillach P, Sahlqvist AS, Kolberg L, Mägi R, Esko T, Vilo J, James G. **Comparison of variation in frequency for SNPs associated with asthma or liver disease between Estonia, HapMap populations and the 1000 genome project populations.** International journal of immunogenetics. 2019 Jan 19. Full text given on page 66.

Ref. III

James G[1], **Reisberg S[1]**, Lepik K, Galwey N, Avillach P, Kolberg L, Mägi R, Esko T, Alexander M, Waterworth D, Loomis AK, Vilo J. **An exploratory phenome wide association study linking asthma and liver disease genetic variants to electronic health records from the Estonian Biobank.** PloS one. 2019 Apr 12;14(4):e0215026. Full text given on page 67.

[1] These authors contributed equally to this work.

Ref. IV

Reisberg S, Krebs K, Lepamets M, Kals M, Mägi R, Metsalu K, Lauschke VM, Vilo J, Milani L. **Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions.** Genetics in Medicine. 2018 Oct 16:1. Full text given on page 68.

Author's contribution to these articles

Ref. I – The author of this thesis led the study, extracted the data, created necessary analysis scripts and conducted the analysis, wrote the major part of the manuscript.

Ref. II – The author of this thesis extracted the data, was one of two authors who made the calculations and wrote the manuscript.

Ref. III – The author of this thesis led the study, extracted the genomic data, filtered and prepared diagnosis and laboratory measurement data from electronic health records, was one of three authors who wrote the necessary analysis scripts and one of three authors who wrote the main part of the manuscript.

Ref. IV – The author of this thesis participated preparing and extracting the necessary genomic data, wrote all the scripts and pipeline to analyze the data and made necessary calculations for the manuscript, was one of three authors who led the project and wrote the manuscript.

Publications not included in the thesis

Ref. V

Liivlaid H, Eigo N, **Reisberg S**. **Eriarstiabi haigestumusstatistika võrdlus Tervise Arengu Instituudi ja Eesti Haigekassa andmetel.** ***Comparing specialist care morbidity statistics on two datasets: National Institute for Health Development and Estonian Health Insurance Fund.*** Eesti Arst. 2019 Jan;98(1):17–26 Accessed on 20 March 2019: `https://eestiarst.ee/eriarstiabi-haigestumusstatistika-vordlus-tervise-arengu-instituudi-ja-eesti-haigekassa-andmetel` (in Estonian)

Ref. VI

Reisberg S, Sirel R, Kalda R, Merzin M, Pruulmann J, Vilo J. **Elektrooniliste terviselugude analüüsimise võimalused Tartu perearstide infosüsteemi näitel.** ***Analysis of the electronic health record dataset of the general practitioners of Tartu.*** Eesti Arst. 2013 Aug;92(8):452-459 Accessed on 20 March 2019: `https://doi.org/10.15157/ea.v0i0.11413` (in Estonian)

Ref. VII

Mooses K, Oja M, **Reisberg S**, Vilo J, Kull M. **Validating Fitbit Zip for monitoring physical activity of children in school: a cross-sectional study**. BMC public health. 2018 Dec;18(1):858.

Ref. VIII

Roberto G, Leal I, Sattar N, Loomis AK, Avillach P, Egger P, Van Wijngaarden R, Ansell D, **Reisberg S**, Tammesoo ML, Alavere H. **Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the EMIF project**. PLoS One. 2016 Aug 31;11(8):e0160648.

Ref. IX

Reimand J, Arak T, Adler P, Kolberg L, **Reisberg S**, Peterson H, Vilo J. **g: Profiler – a web server for functional interpretation of gene lists (2016 update)**. Nucleic acids research. 2016 Apr 20;44(W1):W83-9.

Other published work of the author

Ref. X

Reisberg S, Talvik HA, Koppel K, Laur S, Vilo J. **Description of the current status and future needs of the information architecture and data management solutions for the national personalised medicine pilot project**. A report for Ministry of Social Affairs of Estonia. 2015. Accessed 21 March 2019: `https://www.sm.ee/sites/default/files/content-editors/eesmargid_ja_tegevused/Personaalmeditsiin/description_of_the_current_status_and_future_needs_of_the_information_architecture_and_data_management_solutions_for_the_national_personalised_medicine_pilot_project.pdf`

1. INTRODUCTION

Traditional medicine is mostly relying on the symptoms-based disease diagnosing, and selected therapies have evidence of working on large sets of people (Chen and Snyder, 2013; Fröhlich et al., 2018). Indeed, doctors also take into account individual factors such as gender, age, blood type, *etc.*, however, the therapies are still inefficient for significant proportions of the patient population (Spear et al., 2001). This is mostly due to the over-simplification of the complex nature of most diseases and not considering the differences in the individual genetic background of the patients (Loscalzo and Barabasi, 2011) – the problem often addressed as *"one size does not fit all"* (PMC, 2017). Personalized medicine is a new approach that utilizes individual genetic information in combination with detailed medical records and other data to provide more targeted – and therefore more effective – therapies based on detailed subgrouping of the patients.

A lot of research has been done in this field already. Several research findings could be potentially taken into use in routine medical practice as of today. However, there is a myriad of challenges to tackle beforehand, including regulatory, technical, ethical, educational and financial obstacles. This thesis is mainly about technical challenges, particularly related to large-scale computations, and is motivated by Estonian state-level ambition to bring personalized medicine into national clinical practice.

In Chapter 2, I provide a general overview of the concept of personalized medicine. In Chapter 3, I elaborate on the current state of the field in more detail to give a better understanding of the data and approaches behind the publications of this thesis. Chapter 4 shows that one cannot apply the same polygenic risk estimation models blindly to all people because it could lead to incorrect estimations. Chapter 5 reveals how personalized medicine is influencing electronic health records. Chapter 6 shows how to combine genetics and health records to discover new associations which could lead to better options for disease prevention in the future. Finally, we describe the developed pipeline that can be used for adjusting drug dosage for each patient based on his/her genome in Chapter 7.

2. PERSONALIZED MEDICINE – PROMISES BUILT ON DATA

This chapter gives an overview of the concept of personalized medicine. I first describe its different definitions and later explain its core elements.

2.1. Variety of definitions

Although medicine has always been an individualized interaction between a patient and a doctor, the term *personalized medicine* has entered into common use and research publications during recent decades (Pokorska-Bocci et al., 2014). Having several aliases like *personalized healthcare*, *stratified medicine*, *precision medicine*, *systems medicine*, and *P4 medicine* (predictive, personalized, preventive and participatory) (Duffy, 2015), the general idea behind these has mainly been the same: to provide more effective clinical care and prevention of the diseases by more finely dividing patients and diseases into subgroups (Katsnelson, 2013; Pokorska-Bocci et al., 2014).

New subgrouping options rely mostly on the vast amount of additional data about the patient that has become available during recent decades. First, due to the widespread use of electronic health records instead of notes on paper, the whole medical history of the patient is now easily accessible and can be shared between physicians. Second, huge amounts of data are also produced by machines. Machines can produce large quantities of high-density data (*e.g.* digital images, genetic information) in a very short period of time. This has led to the rapid growth of any digital data. It is estimated that digital data doubles in size every two years, and it is even faster (48% growth annually) in healthcare (Gantz and Reinsel, 2012).

Personalized medicine has been mostly associated with molecular data, especially genetic data. Due to the rapid decline in costs of genetic testing, more and more genetic data have become available in order to finely characterize the patients and diseases to discover the true causes of the illness and therefore to select the best treatment. However, the attention has turned to other data sources besides genetics also (Duffy, 2015; Fröhlich et al., 2018). For instance, a lot of information can now be collected by the patient – such as heart rate and atrial fibrillation measured by smart-watches or smart-phones (Galloway et al., 2013), or physical activity measured by step counters (Mooses et al., 2018). Consequently, the collection of the information can start long before a person goes to the doctor and even before any disease has manifested. That provides a great opportunity for clinicians to monitor even healthy individuals and deal more thoroughly with the prevention of the diseases.

However, there is another important enabler that has empowered the analysis capabilities of these vasts amount of data – increased computational power. It is already almost impossible to analyze all the available information – patient data and clinical knowledge – thoroughly by a single human and therefore, computers are needed to assist physicians in their clinical decision-making. Computer's role would be to do the necessary data filtering and analysis, present the results to the doctor and let them make the final treatment decision (Welch and Kawamoto, 2012).

In 2015, the Ministry of Social Affairs of Estonia conducted a feasibility study for bringing personalized medicine into healthcare on a national scale (Reisberg et al., 2015). After thorough discussions, the definition of personalized medicine was put as follows: "Personalized medicine refers to prevention, diagnosis and treatment of health disorders, based on individual risk-tailored approach using computational decision support analysis of person's phenotype and genotype data. The goal of personalized medicine is to contribute towards preventive, predictive and participatory health system." This nicely summarizes what we said above – personalized medicine uses both genetic and phenotype (physical characteristics, clinical) data, computers are involved in the process of making recommendations not only to treat the disorders but also to prevent healthy people from getting ill.

We will describe each of these components in more detail in the following sections.

2.2. Genetic data

In this Section, we briefly explain what biological DNA is and how it is turned to digital format.

2.2.1. The essence of DNA

As mentioned above, genetic data has been the most important driving force of personalized medicine. This thesis also relies on the analysis of human genetic data. Humans, like most of the living organisms, have their genetic information stored in a molecule called **DNA** (deoxyribonucleic acid) located in the nuclei of the cells. All DNA molecules have the same repeating building blocks – four types of **nucleotides**, denoted by a single letter depending on what type of nucleobase they contain – **C** (cytosine), **G** (guanine), **A** (adenine) or **T** (thymine). These nucleotides join to one another and compose long chains (millions or even billions of elements). For each individual, the exact ordering of the nucleotides – called **genome** – is unique.

For replication purposes, every DNA chain has a complementary bound parallel chain that carries exactly the same information. These chains are called **strands** and therefore, DNA is called a double-stranded molecule.

The ordering of the nucleotides carries essential information for each cell – how to grow, function, reproduce, *etc*. Particularly, there are approximately 20,000 regions in the genomes – called **genes** – each containing subparts called **exons**, that cover approximately 3% of the whole genome (Braschi et al., 2018).

Exons carry the information for making **proteins** – the building blocks of the whole organism that regulate body processes, transport materials within the body, protect against foreign substances, *etc.* Therefore, genes determine most of the physical characteristics (**phenotype**, also **phenome**) of an individual. Each gene has a symbol and name, *e.g. INS* stands for insulin. All exons together are called an **exome**.

The genomes of the organisms of a particular species are very similar to each other – for instance, every two people share approximately 99.9% of their genomes (1000Genomes, 2015; HapMap, 2003). However, as the total length of the human genome is nearly 3.2 billion elements, the remaining 0.1% still leaves room for millions of differences in total (1000Genomes, 2015; Hinds et al., 2005). What is more, as humans are diploid organisms, they inherit their genome as two "copies"[1] from their parents, both split into 23 slices – physically arranged as **chromosomes**. This means that we have two "copies" of each gene in our genome – called **alleles** of the gene. One is inherited from the mother, one from the father. Due to slight differences of the genomes, we may have different versions of the same gene, leading to different proteins as a result. This might play an important role if one of the versions is "faulty" and not functioning properly. We will explore this further in Section 3.4.

As mentioned, all human genomes have their unique differences. Majority of the differences (also called **mutations**) are single nucleotide variants, shortly called **SNVs**, where there is a single nucleotide replacement in the genome (*e.g.* A is replaced by G, denoted as A>G). Different versions of the same variant (here, A and G) are called **alleles**. If SNV is frequent enough (usually >1% in the population), it is often called a **SNP** (single nucleotide polymorphism) instead. Other changes are structural, including insertions (for instance, there is G**CA**CC instead of GCC), deletions (*vice versa*) or repeated parts of the genome where the exact number of repetitions varies across people, called copy number variations (CNV).

In this thesis, genetic data of approximately 50 thousand Estonian gene donors are used.

2.2.2. Turning biological DNA into digital format

The modern technology can read the whole human genome and turn it into a digitally analyzable format. The work of this thesis is conducted purely on digital genetic data. **Whole genome sequencing** that detects the full nucleotide sequence of the genome or **exome sequencing** that analyses the coding part are the most thorough ones. A cost-effective alternative to them is **genotyping** – a method that detects only specific nucleotides of the genome (**genotype**). Modern genotyping arrays can detect up to one million genetic variants, that can be effectively used for

[1] Although half of the chromosomes come from the mother and half from the father and they are very similar to each other, there are always slight differences between them and therefore a widely used word "copy" is not fully correct in this context as they are not identical nor coming from the same source.

predicting the remaining nucleotides between them – a process called **imputation**. For a good imputation, a representative reference genome, preferably from the same population, is needed.

For imputation, it is necessary to distinguish which nucleotides are read from which pairing chromosome. It is accomplished through a process called **phasing** before a proper imputation can be applied. A problem arises from the fact that common sequencing and genotyping methods lose the source chromosome information and it is not known which nucleotides are read from maternal and which from paternal chromosome. Therefore, in case this kind of information is needed, a phasing process that tries to reconstruct the two DNA chains of both pairing chromosomes has to be conducted.

From personalized medicine perspective, the question of which sequencing/ genotyping methods could be utilized for clinical use provides a great scientific and practical interest. In Chapter 7, we analyze how they perform as providers of pharmacogenomics information – to estimate, how fast the drugs are metabolized in our body.

For each genome, it is usually important to know how it differs from the average genome – this is why a comparison with the **reference genome** is used. Reference genomes are averaged genomes of a number of people and there are several widely used reference genomes available. Perhaps the most commonly utilized is a reference genome called **GRCh37** (Consortium et al., 2001) that is used in this thesis also. In a reference genome, each nucleotide has a specific position, which is a distance from the beginning of the chromosome. If in some position the alternative nucleotide is common, usually a special identifier for the SNV is assigned, *e.g. rs12979860* denotes a C>T nucleotide change in chromosome 19 at position 39738787. An actual genotype for a particular sample for the same position can be given as 'C/C' (meaning that the sample does not have the variant) or 'C/T' (the sample has the variant in one chromosome, but not in the other) or 'T/T' (the sample has the variant in both chromosomes).

There are several file formats for expressing the particular genetic data of a sample – SAM, CRAM, VCF, *etc.* In this thesis, *de facto* standard Variant Call Format **VCF** (Danecek et al., 2011; Rehm et al., 2013) is used. In Figure 1, there is an example of VCF data given. While VCF format represents the genetic differences between samples, it is more suitable for scientific use when one investigates the genetic differences of two sets of people (cases and controls). In order to utilize VCF data for individual use in personalized medicine, there is a set of limitations that need to be taken into account.

2.2.3. Limitations of VCF format

The notation of the same genetic variant in VCF format can vary, especially for structural variants. For instance, deletion GTTTTT**TT**A>GTTTTTA on chromo-

```
#CHROM  POS       ID          REF ALT SAMPLE
...
2       234668879 rs57191451  C   CAT 0/1
7       117188682 .           GTT G   0/0
10      96521657  rs12248560  C   T   0/1
16      31107689  rs9923231   C   T   0/0
16      31107927  rs17878544  T   C   0/0
16      31110501  rs17880887  G   T   0/0
...
```

Figure 1. Example extraction of genotype data of the author of this thesis. The data is given in VCF format, but some columns are removed for saving space. “0” denotes the nucleotide sequence of the reference genome and “1” an alternative sequence. Thus, “0/1” for rs12248560 means that from one parent the author has inherited nucleotide C on chromosome 10 at position 96521657 and nucleotide T from the other.

some 7 in position 117188682 in the *CFTR* gene which is used for drug metabolization (**Ref. IV**), is normalized and represented as GTT>G in whole genome VCF files (chromosome 7 line in Figure 1) and I>D in genotyping array data. This also results in the chromosomal position difference of the variant (the position is 117188688 for genotyping array), which in turn creates challenges in automating the detection of the variant in different genotype files.

Many genetic variants have great importance in how an individual metabolizes medications. For instance, position 234668879 on chromosome 2 can have different mutations in gene *UGT1A1* – instead of a genetic sequence CAT, an individual might have CATAT, C, or CATATAT, each having a different effect on the functional characteristics of a protein. Although all these variants were detected by whole-genome sequencing of 2,400 gene donors in Estonian Biobank, some of these were detected with poor quality and only variant notation of CAT>CATAT remains in the files after a quality check. Due to the variant normalization process, the variant labels are trimmed (variant normalization, `https://genome.sph.umich.edu/wiki/Variant_Normalization`) to the shortest possible notation C>CAT in the VCF file (chromosome 2 line in Figure 1). Therefore, these variants can be mistakenly interpreted as C and CAT versions of the variant, as they actually correspond to CAT and CATAT. For providing personal pharmacogenomic information about the speed of drug metabolism, there is a major difference.

Perhaps the most challenging task regarding genetic data in the light of personalized medicine is to interpret variants that are missing in the VCF files. It is crucial to distinguish between whether the presence of a variant can be ruled out or if it simply remained undetected (Rehm et al., 2013). To improve future genetic tests, it is also important to investigate the causes of undetected variants.

The clearest cause of undetected variants is restrictions of the platforms used for genotyping. For instance, whole exome sequencing rarely detects variants outside coding regions of the genome, *e.g. rs9923231*, that are crucial for pharmacogenomics testing. Genotyping arrays, on the other hand, only cover a limited set of predefined DNA positions, particularly restricting the analysis of rare variants. Although imputation helps in increasing the number of variants detected, it is, in turn, restricted by variants that are present in the respective reference genomes used for phasing and imputation.

The second cause of unclarity in undetected data lies in the limitations set by VCF format itself. It was developed to reduce the number of variants from genome sequencing by only including positions that differ from the reference genome. However, when one is making a complete survey of variants for allele calling, this creates a challenge in determining whether variants not present in the VCF were covered at sufficient quality during the sequencing to rule out the presence of the variant in the person's genome. Cases where all sequenced individuals carry the alternative allele also create challenges, as such monomorphic variants are not included in the VCF file. Therefore, awareness of these issues and manual inspection of all input variants is important.

A final cause of missing data is due to the filtering of variants during the quality control (QC) process. Importantly, several genetic variants related to drug metabolism were present in our data before QC in **Ref. IV**. For instance, *rs1985842* T>C was present in both whole-genome and exome sequencing data. However, after the QC step, these variants were removed. In total, 20 variants (4.9%) were removed during the QC to keep confident variant calls only. Again, when calling alleles, it is essential to flag variants that have been removed during QC to ensure that these are not automatically assumed to be alleles of the reference genome.

The issues raised above highlight the importance of inspecting the original sequencing and genotype files as well as QC logs to determine which positions can be called as reference allele and where we do not have sufficient information to make this call.

2.2.4. Increasing volume of genetic data

Due to the rapid developments of gene technology, the speed and amounts of producing digital genetic data have increased exponentially in recent years. Full human DNA sequence can now be determined in one day, producing up to 100 GB of data per sample (He et al., 2017). It also costs less than ever before. While the first human whole genome sequence cost 0.5-1 million dollars in 2003, the price has now dropped to nearly 1,000 dollars (Reuter et al., 2015). Several studies are currently ongoing to sequence or genotype thousands of new individuals and it is estimated that genomes of 60 million Americans will be fully sequenced by 2025 (Khan and Mittelman, 2018). In Estonian Biobank, approximately 2,400 samples have been fully sequenced to date. Using genotyping arrays is even more popular. They can nowadays detect approximately 1 million genetic positions and are cheaper than sequencing by an order of magnitude (Martin et al., 2019). This has led to tremendous amounts of genetic data available in different databases (*e.g.* HapMap project (HapMap, 2005), followed by 1000 Genomes project (1000GenomesProject, 2015)) and biobanks worldwide and these keep growing rapidly. Being so far mainly used for scientific purposes, their use for returning research results and providing personal genetic counselling to participants of the biobanks is more and more being discussed (Gottesman et al., 2013).

Genetic data is collected also by private enterprises. It is estimated that 50% of all clinical trials led by pharmaceutical companies collect DNA from patients to aid in drug development (for the Study of Drug Development, 2011). In addition, hundreds of companies are also providing direct-to-consumer genetic testing online, including health-testing (Phillips, 2016). The first officially approved direct-to-consumer tests got their approval from U.S. Food and Drug Administration (FDA) in 2015 (Curnutte, 2017). One can collect a DNA sample at home and receive results of the genetic test online. The consumer genetics sector has grown exponentially since 2016, reaching to 10 million genotyped individuals worldwide by mid-2018, and it is estimated to increase 10-fold by 2021 (Khan and Mittelman, 2018).

The enormous volume of the data poses challenges for its use in clinical practice. It has become apparent that the bottleneck of personalized medicine has shifted from data generation to data management and interpretation (Alyass et al., 2015). Data storing requires not only large quantities of disk space but also secure environment, procedures and infrastructure for accessing the data (Evans, 2016). In order to analyze the data, sufficient amounts of computational power, memory and data science skills are required. Finally, the major value of the genetic data reveals itself in integration and interpretation with other types of data like electronic health records (He et al., 2017) that will be described in the next section.

2.3. Electronic health records

Electronic health record (**EHR**), also called electronic medical record (EMR), is the longitudinal record of patient health information generated by multiple encounters (physicians) in any care delivery setting (De Moor et al., 2015). Usually, these records contain all or some of the following attributes: diagnoses, drugs, treatments, procedures, surgeries, laboratory measurements, complaints, healthcare service bills and other clinical notes. This kind of information has been recorded in a paper form for decades, but since around 2002-2010 these data have been moved to electronic medium (Charles et al., 2015; Schade et al., 2006) and presently majority of the physicians use them (PMC, 2017). There is a good reason behind this – it enables faster access to the patient information and simplifies sharing the data with other physicians, reduces redundant data capture and medical errors, and also provides faster statistical reporting options (Schade et al., 2006).

However, EHR is not only a replacement for the old paper records. New types of high-resolution data have also become part of EHR: computer tomography (CT) images, magnetic resonance images (MRI), real-time monitoring data through medical sensors (electrocardiogram data, measurements for sleep apnea, *etc.*) are just some examples (Fröhlich et al., 2018; He et al., 2017).

Therefore, the volume of EHR data is increasing worldwide. For instance, in the Personalized Medicine Research Project at Marshfield Clinic, the database size for 20,000 patients is approximately 3.3 GB (mean 165 KB per patient, He et al. (2017)). In comparison, the author of this thesis has had access to Estonian central e-health database *Digilugu* which contains 21 million EHR documents of 1.4 million patients from 2012-2016 and takes 620 GB of disk space (440 KB/patient on average). The continuous growth of the volume of EHR data can be also seen from Figure 2.

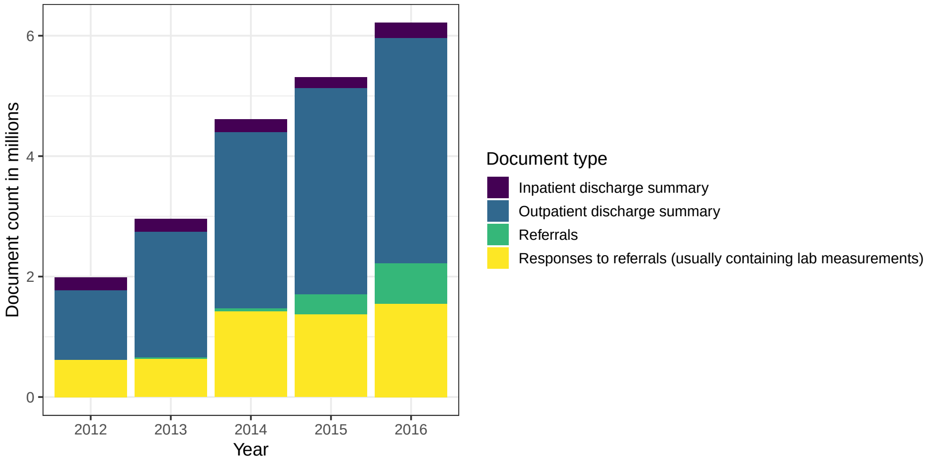


Figure 2. Counts of different types of EHR documents in Estonian central e-Health database in 2012-2016

During the current decade, the high value of the **secondary use of EHR data** for clinical research has also been acknowledged (Botsis et al., 2010). For instance, EHR can be used for building risk prediction models as they contain all patients who are in touch with the medical system in contrast to cohort-based models (Goldstein et al., 2017). It is not only cheaper to use EHR data instead of building a new study cohort for the specific purpose for every study, but also more effective – the volume of the data is much larger, it better reflects the complexity of the medicine and can be used for investigating multiple outcomes. During the years 2011-2016, an ambitious project EHR4CR was conducted. This brought together 10 pharmaceutical companies and 34 academic partners to utilize data from hospital EHR systems for clinical research (De Moor et al., 2015). They envisioned that detailed EHR data could be used for the feasibility assessment of the study and patient recruitment – to connect patients to the right clinical trial.

Another, even larger initiative – European Medical Infrastructure Framework[2] (EMIF, 2013-2018) – contained 58 teams from academia and pharmaceutical field who worked together to improve access and use of health data, specifically focusing on Alzheimer's disease and metabolic disorders. **Ref. III** of this thesis was written during EMIF project. I am also a member in an ongoing project EHDEN[3] that implements a federated health data network in Europe for scientific purposes and aims to harmonize health records of 100 million Europeans.

Despite a number of terminology and documentation standards used in healthcare (*e.g.* ICD-10[4] for diseases, ATC[5] for drug substances, LOINC[6] for health measurements, SNOMED CT[7] for different kind of clinical data; HL7[8] for transferring clinical data between applications, *etc.*), different countries use different standards with various local modifications in them. Importantly, these standards also change in time. Therefore, the constant work of harmonizing EHR databases is highly needed.

EHR data are also important enablers for genetic research. Therefore, a number of projects have already linked them with genomic data (He et al., 2017). For instance, to discover genomic variants associated with clinical conditions identified using EHR data (Gottesman et al., 2013). For clinical trials, case and control groups can be finely defined by using comprehensive health records together with genetic information. No less important, the results of the studies can be more easily integrated back to EHR systems.

While having lots of benefits, EHR data also have their shortcomings. Perhaps the biggest challenge is the data quality (De Moor et al., 2015) as EHR data tend to be very "messy" – contain lots of missing data, repeated measurements, loss of follow up, input errors, *etc.* (Gottesman et al., 2013). It often lacks standards and some information (*e.g.* symptoms) might be given only as free text notes. Therefore, reconstructing the true patient state from the EHR is a challenge on its own (Hripcsak and Albers, 2012).

In chapter 5, I describe the research that I have done in the EHR field in more details.

2.4. Self-collected data (health wearables)

As it was mentioned above, one of the keywords in personalized medicine is *participatory*, which means that individuals are expected to take more responsibility for managing their health condition. For instance, by monitoring and responding

[2] http://emif.eu/

[3] http://ehden.eu/

[4] https://icd.who.int/browse10/2016/en

[5] https://www.whocc.no/atc/

[6] https://loinc.org/

[7] http://www.snomed.org/

[8] http://www.hl7.org/

to fluctuations in the data generated by their wearables (McLean, 2013). These wearable technologies include smart-watches, wristbands, subcutaneous sensors, *etc.*, equipped with gyroscopes, accelerometers, optical sensors, cameras, temperature sensors and many more (Yetisen et al., 2018), capable of monitoring a range of medical risk factors.

Hundreds of millions of wearable devices have already been sold worldwide (Yetisen et al., 2018). However, the potential of the wearables is still in infancy and has been studied mostly within academic research rather than for medical use in a real-world context. Perhaps the main problem has been the questionable accuracy and reliability of these devices. Yetisen et al. (2018) summarize that many evaluation studies of widely used wearables have shown up to 25-30% inaccuracies of measuring physical activity, heart rate, and burnt calories. Therefore, additional clinical studies are needed to validate the accuracy of wearable devices before integrating them into healthcare systems.

I have also helped to carry out one of such study. In **Ref. VII**, we used FitBit activity trackers worn by third-grade students to measure their physical activity at school setting and compared the measured accuracy with more expensive research-grade accelerometers. We found that though a few FitBit trackers failed during the data collection and some inaccuracies were observed, the overall results were relatively similar to the accelerometers. Thus, FitBit devices could be used as cost-effective alternatives for similar studies. Interestingly, one of the major challenges of this study was to build the IT infrastructure for distributing the trackers among students every morning, collecting and synchronizing the data after the last lesson and create the aggregated data instead of high-resolution timepoints (Figure 3). This highlights the need to have a proper IT infrastructure in place in order to use any of such solutions on a large scale – either for research or medical use.

2.5. Linking different databases

In order to take into account all information about the patient – one of the central concepts of personalized medicine –, there has to be a way to link all pieces of data together (Duffy, 2015; Wu et al., 2017).

For new discoveries, scientists have to combine different databases not only to put together different types of information (*e.g.* genomic data and EHR like we discussed in Section 2.3) but also to increase statistical power by collecting more cases to detect the associations. For instance, in eMERGE project about discovering genetic variants associated with clinical conditions identified using EHR, the power increased when studies were deployed across a network where cases and controls were shared (Gottesman et al., 2013).

Another example is UK Biobank, one of the largest open resource for studies in the personalized medicine field. It contains genomic data and deep phenotyping data for half a million volunteers, including biological measurements, lifestyle

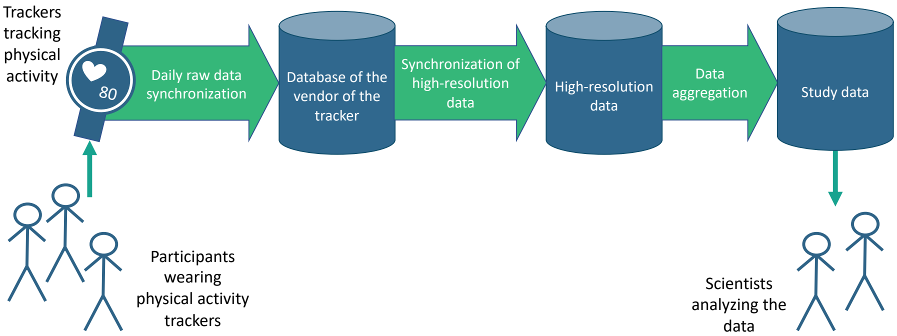

Trackers tracking physical activity
80
Daily raw data synchronization
Database of the vendor of the tracker
Synchronization of high-resolution data
High-resolution data
Data aggregation
Study data
Participants wearing physical activity trackers
Scientists analyzing the data

Figure 3. Data flow of studying physical activity in Mooses et al. (2018)

indicators, markers in blood and urine, and imaging of the body and brain (Bycroft et al., 2018). Gathering the follow-up information is currently underway by linking health and medical records from several databases.

However, to move from the discovery phase to clinical implementation, such database integrations should be made instantly and online. As emphasized by Personalized Medicine Coalition, "all of this requires providers to adopt powerful health information technology (IT) platforms that enable instant connections between real-world clinical results and molecular data so that providers can make clinical decisions based on a body of scientific knowledge that exceeds the training, experience, or memory of any single practitioner" (PMC, 2017). This requires at least two things – a unique **personal identifier** for each individual across all databases, and an IT infrastructure that enables online interoperability between databases. Estonia is in a good situation here as all residents have a personal ID and there is also a national level secure data exchange layer **X-Road**[9] (Figure 4). As of today, 99% of Estonian state level services are conducted online and using X-Road for them is mandatory. X-Road is also implemented in Finland, Kyrgyzstan, Namibia, Faroe Islands, Iceland and Ukraine.

2.6. From computerized analysis towards clinical decision support systems

Not only the amount of patient-level data is increasing. The growth of medical knowledge is also accelerating. At the same time, physicians who are overwhelmed by the vast amount of information, are under the pressure for taking

[9] https://e-estonia.com/solutions/interoperability-services/x-road/

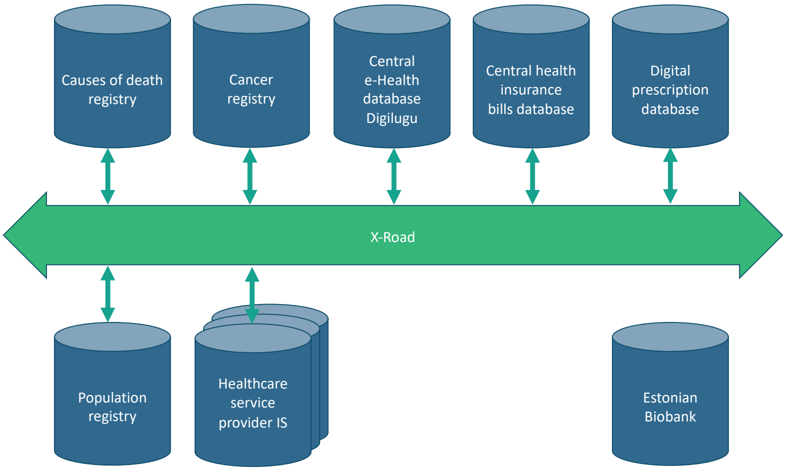

Causes of death registry
Cancer registry
Central e-Health database Digilugu
Central health insurance bills database
Digital prescription database
X-Road
Population registry
Healthcare service provider IS
Estonian Biobank

Figure 4. X-Road in Estonia – an example of IT infrastructure for linking databases. Most of the health-related databases are exchanging the data online via X-Road. The personal ID for each individual is provided by the population registry. Hospital IT systems are sending discharge summaries to central e-health database and healthcare bills to insurance bills database, the majority of the prescriptions are made through a central prescription database. Also, national level registries like causes of death registry and cancer registry are using X-Road. Note that Estonian Biobank is currently not linked to X-Road, mostly due to regulatory reasons.

all these details into account for high-quality clinical decisions (Obermeyer and Emanuel, 2016). Unfortunately, due to the volume and complexity of the data, physicians might not have proper data science skills to do such analysis and this is where **bioinformatics** – an essential enabler for personalized, predictive, and preventive medicine – is brought into play, as each of these medical applications requires data from multiple sources and multiple scales to be integrated (Duffy, 2015; Fernald et al., 2011; He et al., 2017; Phan et al., 2012).

However, bioinformaticians doing the analysis on demand in the back-office would not be a scalable solution for the nation-wide implementation of personalized medicine. There is a need for appropriate **clinical decision support (CDS)** tools for prompting their use at the point of care and delivering results in an easily interpretable format (McCarthy et al., 2013).

According to Aleksovska-Stojkovska and Loskovska (2010), typical CDS contains three main components (see Figure 5):

- Patient-related data, such as EHR (medical history, diseases, symptoms, laboratory results, diagnostic images, treatment plans) and genetics;
- Medical knowledge base – given in a format that computers can use;
- Inference (reasoning) mechanism – empirically validated computer algorithms for combining patient data and medical knowledge base to generate

conclusions and recommendations.

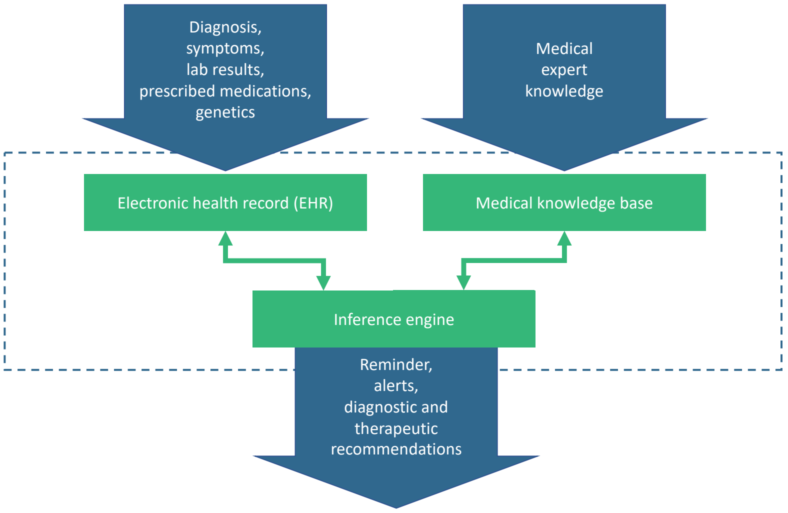


Figure 5. A general model of clinical decision support system. Figure adapted from Aleksovska-Stojkovska and Loskovska (2010).

This is the way how the integration of personalized medicine into clinical practice is also seen – the actionable genomic information needs to be matched with the knowledge-based CDS systems and deployed through EHRs (Evans, 2016; Gottesman et al., 2013). While such integration indeed poses several challenges, it also provides a feasible opportunity to identify clinically actionable genetic variants for individualized diagnosis and therapy (Fernald et al., 2011; He et al., 2017) which is the core of personalized medicine.

In this chapter, the concept and key components behind personalized medicine approach were described. In the next chapter, we elaborate more closely on the current state of this field.

3. CURRENT STATE OF PERSONALIZED MEDICINE IN THE WORLD AND ESTONIA

In this chapter, we focus on specific clinical applications where the personalized medicine approach is already utilized or has a high potential to be taken into use in the near future.

3.1. Targeted therapies and carrier screening in oncology

In this thesis, **germline DNA** – DNA that is inherited from mother and father – is analyzed. However, one may be interested in DNA changes within an individual – called **somatic mutations**. For instance, DNA is continuously altered in cancer cells and exploring somatic changes in them provide a great interest as they make the tumour cells different from normal cells and help to select better treatment targeted for that type of cancer.

A classical example of using DNA data in oncology starts with the discovery in the mid-1980s that in about 30% of breast tumours a cell-surface protein HER2 (human epidermal growth factor receptor) is overexpressed (Slamon et al., 1987), causing tumour cells to grow and spread faster than the ones with normal levels of the protein. That led to the development of new drug trastuzumab in 1998, which binds to HER2 receptors and blocks them from receiving growth signals. HER2 testing has thereafter been a standard procedure for breast cancer patients (Rüschoff et al., 2017; Tannock et al., 2016).

Another example is drug gefitinib, which was approved for non-small cell lung cancer by U.S. Food and Drug Administration (FDA) in 2003. However, after a few months on the market, it became clear that the drug did not work significantly better than placebo, and the approval was withdrawn (Kazandjian et al., 2016). During the followed investigation, it was found that the drug was effective only among patients having certain mutations in *EGFR* (epidermal growth factor receptor) gene. Therefore, FDA re-approved the drug in 2015, requesting genetic testing before prescribing the medication. This is a good example of how *one size does not fit all* and identifying the target population that most likely benefit from the drug has great importance.

There are many other examples of using genetic profiling of the tumour tissue before making decisions for therapy. As of today (27 March 2019), FDA lists 86 drugs used in oncology that have some genetic information in the drug labelling (FDA, 2019). Out of all oncology drugs currently in development, 73% are personalized medicines (PMC, 2017). There is also a paradigm shift taking place in cancer care as instead of defining disease by anatomical location, the therapy can be selected by genetic profiling of the tumour cells. Since 2017, FDA has twice approved a drug to treat tumours with a specific genetic change regardless of the type of cancer (Challener, 2019).

Unfortunately, cancer cells are highly adaptable and can often develop a resistance to single-target-drugs after some time. Using a combination of targeted therapies is troublesome due to toxic effects, and, therefore, it has been difficult to show significant long-term effect for tailored cancer drugs (Nawrocki, 2018). Additionally, further limiting the target population makes the market size smaller, drugs more expensive and less cost-effective (Tannock et al., 2016). These are the big challenges that pharmaceutical companies in cancer care have to face as of today.

Besides treatment of cancer, genetic testing is also used for prevention and early detection of the disease – this is called **carrier screening**. The most widely used example is the use case related to hereditary breast (5% of all breast cancers, Key et al. (2001)) and ovarian cancer. Women with certain germline mutations in *BRCA1* and *BRCA2* genes (responsible for producing proteins that repair damaged DNA) have a very high risk (up to 87%) of developing breast cancer (Petrucelli et al., 1998) and these mutations also increase the risk of ovarian (Antoniou et al., 2003) and prostate cancer (Edwards et al., 2003). Therefore, knowing that individual is carrying pathogenic variants in *BRCA1* or *BRCA2* is a crucial piece of information from prevention (more frequent screening, mastectomy suggested) and better survival perspective.

In Estonian Biobank, genetic data of approximately 5 thousand gene donors have been screened as of today, and nearly 50 individuals out of them carry high-risk *BRCA1/2* mutations. For almost half of them, personal genetic counselling has been provided.

3.2. Testing for developmental diseases in children

Personalized medicine is trying to move towards the earliest possible point in the course of the disease. It is extremely valuable for detecting risks of developmental disorders in children, where early genetic testing can treat or prevent diagnosis at birth (newborn screening), before birth (in utero) or sometimes even before conception (carrier testing) (McCarthy et al., 2013).

Some severe diseases are caused by a single gene, called **Mendelian diseases**. For instance, the parents might both be the carriers of an "affected" gene causing cystic fibrosis (approximately every 1:30 people are, De Boeck et al. (2014); de Vries et al. (1996)), but the disease has never manifested on them as they have another working "copy" of that gene also. However, there is a 25% chance that their child gets two affected genes (one from each parent), resulting in a severe life-threatening disease which affects lungs and other organs. Therefore, especially when there is a family history of such conditions, genetic **carrier testing** is suggested for parents before planning a family.

Sometimes the condition of the patient is so rare (occurs in less than 1 out of 1000-2000 people, called **rare diseases**) that it is extremely hard to detect the true cause of the condition. It is estimated that approximately 80% of such cases have a genetic origin and a substantial part of them are found in children (Boat et al., 2011). Therefore, whole genome and exome sequencing technologies provide an opportunity to conduct a thorough search over the patient's genome to find the causal mutation and obtain an accurate diagnosis (Yang et al., 2013). However, this can be an extremely challenging task, like searching for "the needle in a haystack" as the causal genetic mutation is rare, possibly novel, and it might require an enormous effort to find a patient with a similar rare condition and genetic background to build evidence for causality (Philippakis et al., 2015).

Chromosomal disorders (having a missing or an extra copy of a chromosome) usually start with the cell division error in the development of egg or sperm cells. Particularly, when the division of the cell does not split chromosomes equally. It can also occur after fertilization in the developing embryo where new cells of the baby are built through continuous cell division process. Perhaps the most known disease of chromosomal disorders is Down syndrome that has three copies of chromosome 21. Genetic testing can help to detect such abnormalities early and nowadays it is possible to do these test non-invasively (non-invasive prenatal diagnostics, NIPD, tested substance is taken from mother's blood instead of inserting a needle into the uterus).

Some diseases, such as phenylketonuria (Blau et al., 2010), might not have any visible symptoms in the beginning, but can lead to severe disability later without prompt intervention. Therefore, in some countries, **newborn screening** is mandatory and state-supported to protect children for rare, treatable disorders at birth (McCarthy et al., 2013). In Estonia, all newborns are tested against 20 treatable inborn errors of metabolism since 2015 (coverage >99.5% of newborns) (Reinson, 2018).

3.3. Polygenic risk scores

In contrast to Mendelian diseases, most of the common diseases have a complex origin and are caused by a number of different factors including gender, age, lifestyle, and hundreds of genetic variants in the patient's DNA. Type 2 diabetes and coronary artery disease are some examples of such diseases. Although it has been shown that several genetic variants are associated with them, each variant has only a tiny effect when taken separately. However, if a person has a high proportion of such mutations, and combined with other risk factors, the cumulative effect is large enough to end up with the disease (Martin et al., 2019).

For assessing the genetic risks of getting such complex diseases (and other traits), **polygenic risk scores** (**PRS**, also known as *genetic risk scores*, *GRS*) are built. They are mathematical models containing a set of genetic variants, each having a certain weight and based on the actual DNA of the patient these weights are summed up so that the resulting score reflects the estimated risk of getting the disease. Higher score usually indicates higher genetic risk and *vice versa* (Figure 6).

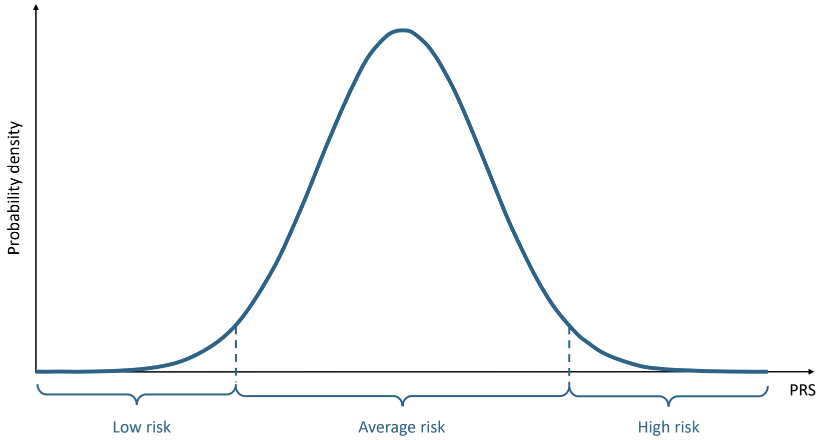


Figure 6. Distribution of polygenic risk score (PRS). A low score for an individual indicates a low genetic risk of the disease, a high score indicates a high risk.

Although several PRS models have been built and their estimation capabilities demonstrated (Aly et al., 2011; Dudbridge, 2013; Euesden et al., 2014; Khera et al., 2019; Läll et al., 2017), there is a problem that they mostly work only in European populations, making it difficult to use them in admixed populations (Khera et al., 2019; Martin et al., 2019), where multiple divergent genetic lineages have interbred (Rius and Darling, 2014).

Therefore, additional investigation in this field is required and the problem still remains largely unresolved. **Ref. I** was one of the first studies to demonstrate this problem and showed how this could dramatically affect the accuracy of the risk estimation if the tested individual is from a different population. We elaborate on this more thoroughly in Chapter 4.

3.4. Pharmacogenomics

Pharmacogenomics, an important part of personalized medicine, is a study of how genetic variation influences responses to drugs. It has been shown that the same drug may work differently among patients – for some individuals the drug may have a lower (insufficient) effect while for the others it may cause toxicity and other side effects (Wilkinson, 2005). It largely depends on how many "working" copies of the genes responsible for drug metabolism a particular patient has, leading to either poor, intermediate, normal, or in some cases also rapid and ultrarapid metabolism of a drug (this response to a drug is called **pharmacogenomic phenotype**) (Caudle et al., 2017; Nebert et al., 2003). Therefore, a clinical decision about the most appropriate drug and its dosage should be made by taking into account the genetic background of a particular patient.

There are several genetic tests available for detecting pharmacogenomic phenotype, which is the basis for making pharmacogenomic recommendations for an individual. However, these tests are usually very limited, targeting a very small group of drugs and testing only against very common phenotypes (Chua and Kennedy, 2012). Taken also into account the debate over cost-effectiveness of such tests, pharmacogenomic testing has not become into common practice yet (Ashley, 2016). In **Ref. IV**, for the first time, we developed a pipeline for estimating pharmacogenomic phenotype for 11 genes by testing against a wide range of actionable phenotypes and by utilizing cost-effective genotyping arrays. This will be further described in Chapter 7.

3.5. Integrating personalized medicine to (state-level) routine care

As it could be seen from above, there is enough evidence in several clinical applications to start integrating the methods of personalized medicine into routine practice. However, there are also several barriers to tackle. In this section, we briefly describe these obstacles and give a short overview of the current state of integration of nation-wide personalized medicine approach in Estonia.

3.5.1. Challenges to tackle

Several challenges need to be addressed before personalized medicine could become an acceptable part of clinical practice.

Evidence. Showing the clinical and economic evidence of the impact to convince authorities to approve, insurance companies to cover and physicians to use genomic-based therapies has probably been the biggest challenge of personalized medicine (McCarthy et al., 2013; PMC, 2017). Here are some examples of the questions that need answers: is there enough evidence that genetic testing provides accurate results and reflects the cause of diseases correctly?; is there enough evidence of clinical benefit?; does the acting based on genetic testing improve the outcome – quality of life and medical care (Ginsburg and Phillips, 2018)? These questions are related to other types of data also, not only to genetics – EHR, laboratory measurements, lifestyle, self-collected data, *etc.* (Chow et al., 2018).

Actionability. Even if there is enough evidence of the causality, an intervention mechanism has to be available also (Severin et al., 2015). There is little help if nothing can be done to prevent the predicted outcome, or there are no guidelines for intervention. For instance, some genetic variants may alter the speed of drug metabolism, but as long there is no action plan how to react if any of these variants are found in a patient's genome, it cannot be used in a clinical decision making (Relling and Evans, 2015). Some markers (for instance, PRS) might indicate a high risk of a disease, but there has to be an action plan also in place to react to these findings so that clinical benefit could be obtained.

Cost-effectiveness. Even if there is a concrete action plan for intervention, testing and acting according to the plan might not always be cost-effective. Although screening whole genomes of all people would allow detecting individuals with high risks of specific diseases, this method is still considered not optimized for cost. It has been shown that although many personalized medicine tests provide better health, it tends to come at a higher price, highlighting the need for further debate over the cost-effectiveness of personalized medicine (Buchanan et al., 2013; Phillips et al., 2014).

Computational feasibility. In order to utilize computers for assisting physicians in their decision making, fully specified (proven and accepted by clinical committees) computer algorithms are needed to rely on. A number of such diagnostic (answering to question "is a disease or condition present?") and prognostic ("will the disease or condition occur in the future?") algorithms have been published (Obermeyer and Emanuel, 2016). However, the reporting of such algorithms has been mostly poor, usually containing insufficient information not only about algorithm development and validation but also for reproduction, making them difficult to use in clinical practice (Collins et al., 2015).

Additionally, even if the algorithms are well defined, it does not mean that the necessary input data exists and has the right format at the place of care (Obermeyer and Emanuel, 2016). For instance, obesity, defined via high body mass index, has been considered an important risk factor for several diseases and mortality for a long time (Calle et al., 1999), but its use (representation) in Estonian national health records is still not standardized, meaning that it is almost impossible to use this information automatically for computer algorithms in clinical practice as of today. This also holds for smoking status, blood pressure, heart rate, *etc.*

Finally, in order to provide clinical guidance in real time, the algorithms should be able to run in real time. This is not only related to computational power but also to have a convenient access to the required data, especially to genetic information. As the volume of the genetic data can be large, it is technically challenging to exchange these amounts of information over the computer network quickly, meaning that the data have to be processed close to the storage. This highlights the need for high performance storing and computing platforms to facilitate personalized medicine in clinical practice (Alyass et al., 2015).

Integration into clinical workflow. Integrating the concepts of personalized medicine into routine clinical workflows has been one of the biggest challenges in this field. This means – to present relevant information to clinicians at the point of care (Gottesman et al., 2013) so that the right treatment for the right patient at the right time could be provided (Ginsburg and McCarthy, 2001). As most of the doctors work with EHR systems, these computerized assistants have to function within electronic health records as a part of the clinical workflow to automatically alert treating physicians about relevant information that could help inform treatment decisions (PMC, 2017; Welch and Kawamoto, 2012). This applies not only to genetic but also to other types of data, *e.g.* history of the diagnoses or the therapy currently under consideration (Sittig et al., 2008). For instance, alerting a physician about the high risk of diabetes is not justified in case the disease has already been diagnosed based on his/her EHR data. Alerts about drug dosage adjustments should not be made when prescribing the drug has never been under consideration.

3.5.2. Estonian personalized medicine program

It should be clear from the above that personalized medicine is an interdisciplinary field where experts from several domains – clinicians, bioinformaticians, statisticians, biologists – are needed (Wolkenhauer et al., 2013). Therefore, universities, research funding agencies, and governments should support connecting researchers from diverse scientific backgrounds (Alyass et al., 2015).

In several countries, state-level personalized medicine programs have been recently launched for that purpose – 100,000 Genomes project[1] in United Kingdom (in 2012), Precision Medicine Initiative[2] in the USA (2015), Genomic Medicine 2025[3] in France (2016), a similar project in Denmark[4] (2017), *etc.* They largely share a common objective – to build a dedicated inter-disciplinary team of healthcare experts, researchers, industry participants, and government representatives with sufficient funding to improve the diagnosis and prevention of the diseases by utilizing genetic information. Most of them are focusing on targeted therapies in oncology and rare diseases as a first step.

In Estonia, a feasibility study for nation-wide piloting of personalized medicine was carried out in 2015. Four topics were thoroughly investigated to identify the most promising clinical fields to start with, describe what kind of information architecture is needed, what kind of clinical decision support solutions are available, and how should the management organization of the program look like[5]. After identifying the main shortcomings in the field, several projects have been started afterwards to deal with them specifically. In 2018, a 3-year-long project[6] was launched to develop and validate new genetics-based disease risk assessment methods for preventing breast cancer and cardiovascular diseases. This year (2019), a decision support project was started to provide a set of validated deci-

[1] https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/ (accessed on 3 April 2019)

[2] https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative (accessed on 3 April 2019)

[3] https://aviesan.fr/fr/aviesan/accueil/toute-l-actualite/plan-france-medecine-genomique-2025 (accessed on 3 April 2019)

[4] https://www.sum.dk/~/media/Filer%20-%20Publikationer_i_pdf/2017/Personalised-Medicine-Summary/SUM_klar_diagnose_summary_UK_web.ashx (accessed on 3 April 2019)

[5] All reports are available at https://www.sm.ee/et/personaalmeditsiini-juhtprojekti-eeluuring (accessed on 3 April 2019)

[6] https://www.etag.ee/uuring-kaardistab-rinnavahi-ja-sudameveresoonkonna-haiguste-ennetuse-voimalusi/ (in Estonian, accessed on 3 April 2019)

sion support algorithms for general practitioners. Besides them, another project was started in 2019 to develop a comprehensive state-level IT infrastructure for connecting genomic data from Estonian Biobank, Estonian EHR databases, health insurance data, and digital prescription system, in order to provide pharmacogenomics recommendations through clinical software[7]. It will be largely based on pharmacogenomics pipeline described in **Ref. IV** and will be further explained in Chapter 7. The key databases involved in this project will be briefly described in the next sections.

3.5.3. Estonian Biobank

Estonian Biobank[8] is a national population-based biobank, established in 1999 mostly by private partners (original name EGeen). After the end of the collaboration, the University of Tartu took control of the biobank and it is now led by Institute of Genomics. It acts within Human Genes Research Act[9] of Estonia and is purposed for research activities. It aims to collect biological samples and genetic material of Estonian gene donors and link them to EHR data to investigate the genetic, environmental, and behavioural background of common diseases in the Estonian population (James et al., 2019).

After a decade of operation, nearly 52 thousand voluntary participants (approximately 5% of Estonian adult population) had joined Estonian Biobank by 2012 and signed a broad informed consent (Leitsalu et al., 2014). Each of them had donated a blood sample, and a thorough questionnaire had been filled in during standardized health examination by trained medical personnel. The survey contained more than 600 questions about nationality, education, family tree, lifestyle (including smoking and alcohol consumption), diseases, and medications. Additionally, height, weight, blood pressure and other objective measurements were recorded during the examination visit.

During these years, several specific studies have been conducted on these data, and many specific modules have been added for diabetes, psychiatry, cardiovascular diseases, physical activity, *etc.* Also, different types of analyses of biological samples have been conducted (Leitsalu et al., 2014).

For most of the participants, their genetic data is also available in digital format. Particularly, whole genome sequencing data is available for 2,400 donors and exome sequencing data for another 2,400 donors as of the present date. DNA of 41,000 donors has been genotyped using either Illumina Global Screening Array (GSA, n=33,000) or Illumina OmniExpress array (OMNI, n=8,000). In **Ref. II** and **Ref. III**, data from GSA array were used. Whole genome data is used in **Ref. I** and all sources in **Ref. IV**.

[7] `https://www.openaccessgovernment.org/personalised-medicine-estonia/55550/` (accessed 3 Apr 2019)

[8] `https://www.geenivaramu.ee/en` (accessed 4 April 2019)

[9] `https://www.riigiteataja.ee/en/eli/ee/531102013003/consolide/current` (accessed on 4 April 2019)

One of the strengths of Estonian Biobank is the permission to collect health-related data from other sources. Therefore, additional data about the gene donors have been continuously updated from hospitals, state-level EHR databases, and registries.

A detailed overview of the cohort of Estonian Biobank as of 2014 is given in the article by Leitsalu et al. (2014). However, things have rapidly changed in recent years. In the light of national personalized medicine program, the Estonian government provided funding for recruiting additional 100,000 gene donors in 2018[10] which was completed in only nine months by the end of the year. Only informed consent together with the blood sample was collected this time, and the samples will be genotyped by the end of 2019. Health data of new participants will be acquired entirely from EHR databases. Due to the enormous public interest, the recruitment was extended to additional 50,000 gene donors in 2019. Therefore, it is expected that Estonian Biobank will have 200,000 participants (15% of the total population) by the end of 2019.

It has been shown that even if the data were collected for research purposes, the participants are likely willing to receive personal genetic risk information, especially when there is an effective preventive measure (Johansson et al., 2019). To pilot individual genetic counselling, Estonian Biobank has provided personal feedback for nearly 2,000 participants as of today. This includes reporting the genetic risk of diseases such as type 2 diabetes, cardiovascular diseases, and other conditions such as early menopause, carrier status of several diseases (*e.g.* breast cancer, cystic fibrosis), and pharmacogenomics information about drug metabolism.

This thesis has a strong connection to these developments. **Ref. I** and **Ref. II** are studying the frequencies of genetic variants across different populations, including Estonians, in order to highlight the differences of the frequencies, which should be taken into account when making genetics-based decisions. **Ref. IV** describes a pipeline for producing pharmacogenetic recommendations, used when providing personal counselling for gene donors. **Ref. III** explains how genetics data linked to EHR could be a valuable resource for studying new genetics-disease associations.

3.5.4. Estonian central EHR databases

Besides hospital EHR systems, there is a number of central health databases in Estonia (see Figure 4 on page 26). In this section, we briefly introduce the most notable ones.

Central e-Health database Digilugu. Central database of in- and outpatient discharge summaries, referrals and responses to referrals (mostly laboratory measurements), called *Digilugu*, is the largest central health database in Estonia (see

[10] `https://www.tai.ee/et/instituut/koostooprojektid/100-000-uut-geenidoonorit` (accessed on 4 April 2019)

also Figure 2 on page 22). It was originally used as a data sharing platform between physicians as sending case summaries to the database has been mandatory, and all physicians can see the patient records added by everyone else. However, there is also an online patient portal available where patient himself/herself can see his/her documents. Data collection began in 2009.

All Digilugu documents have Health Level 7 version 3 format (Dolin et al., 2001) where part of the information is strictly structured (diagnoses, medications, patient and clinician data), but the rest (complaints, laboratory measurements) less strictly, even containing sections of free text. Loose format restrictions increase the variation of data quality and from a scientific perspective poses several challenges to deal with. This will be further elucidated in Chapter 5.

Digital prescription. As of today, 99% of all prescriptions in Estonia are made via digital prescription system called *e-Prescription*[11], launched in 2009. The doctors fill in the online form and patients can immediately purchase the medications from any pharmacy on their will. One of the strengths of e-Prescription in Estonia, when compared to other similar systems, is that it also contains fill data (shows whether the medicine was actually purchased or not).

Insurance bills. There is a single health insurance broker in Estonia – Health Insurance Fund – covering all people having state-level health insurance (94% of the population, Liivlaid et al. (2019)). Reimbursement of health expenditures is strictly based on health service bills, which are reported electronically to the central insurance bills database by health care providers. Therefore, clinicians are motivated to send the reports in a timely and well-formatted manner, which is a reason why this database is perhaps the most complete EHR database in Estonia. Data collection was started in 2002.

[11] `https://e-estonia.com/solutions/healthcare/e-prescription` (accessed on 4 April 2019)

4. DIFFERENCES BETWEEN POPULATIONS HAMPERS USING POLYGENIC RISK SCORES BLINDLY (REF. I-II)

Polygenic risk scores (PRS) are prediction metrics (models) used for assessing the genetic risk of developing certain conditions/diseases in the future, *e.g.* type 2 diabetes. They are composed of a large set of disease-associated genetic variants which are weighted so that people having a high probability of developing the disease will get a high score, and people with the low probability get a low score (see Figure 6 on page 31) (Belsky and Israel, 2014). However, for complex diseases, the exact proportion of genetic risk to total risk is usually unknown – for instance, estimated to be around 20-80% for type 2 diabetes (Ali, 2013). Therefore, PRS is not a highly accurate prediction model without taking into account additional risk factors such as lifestyle, age, and environment. Nonetheless, it would still be a reasonable option for selecting people for targeted lifestyle modification programs for disease prevention (Schellenberg et al., 2013).

As PRS is a prediction model (answering to a question "what happens to me?"), it does not explain causality ("why it happens?"). Generally, genetic variants are chosen to PRS model from genome-wide association studies (GWAS), where certain alleles of the variants are found to be associated with the disease (Belsky and Israel, 2014), but it does not necessarily mean causality. Several variants in human genome are correlated with each other, and, therefore, for strongly correlated variants with similar frequency, similar association with the disease can be seen for all of them. Usually, only one of these variants that shows the strongest association, is included in PRS model. Instead of being a causal variant, this could also be non-causal (Carlson et al., 2013). As long as the correlations between the causal and non-causal variant remain similar, this does not affect the PRS result. In case the correlations change, the result of the PRS model could be dramatically altered. The problem is that correlations between variants vary across different populations, and, therefore, a model that works well for one population, might not work for the others. In addition, there might be additional causal variants in other populations. When the model is built on data from one population, it may miss several causal variants from the others, and, therefore, not reflect the true genetic risk of these people. Finally, as the frequencies of variant alleles vary among populations, the mean of PRS differs as well, affecting the result of PRS model in other populations.

Most of the currently available genomic data are from European populations. Therefore, also PRS models are biased towards Europeans (Martin et al., 2019), especially when the PRS contains a large number of genetic variants. **Ref. I** was one of the first studies to indicate this problem. In this study, we investigated two PRS models, developed mostly on European genomic data and using a very large number of variants, probably the largest PRSs by the time of the publication. The first of them was developed by Läll et al. (2017) for assessing the genetic risk of type 2 diabetes, the other by Abraham et al. (2016) for evaluating the risk of coronary heart disease. In our paper, we showed that the resulting risk scores are so different across different populations that applying the same model blindly to all people would lead not only to slightly biased but to incorrect disease risk estimations.

Several studies have reached the same conclusion afterwards, and the problem has remained largely unresolved as of now (Martin et al., 2017, 2019).

In **Ref. II**, we investigated allele frequencies of the variants that had been associated with asthma and liver disease in the literature. Specifically, we compared allele frequencies in Estonia and other populations and tested, in how many of them the difference was significant. Similarly to **Ref. I**, we observed that allele frequencies of these disease-associated variants among Estonians are close to other Europeans but more distant to other populations, especially Africans.

Taking this into consideration, one has to be cautious when estimating genetic risk in the clinical setting. Differences in the genetic background of the patients may cause biases in the models and lead to wrong conclusions as a result. Even in relatively homogeneous populations, there might be individuals from different ancestries. Therefore, when using genetics-based risk models, we emphasize the importance of assessing the genetic suitability of the patient and avoid providing risk estimations to anyone else.

These publications have had a direct impact on Estonian Biobank participants. During genetic counselling, genetic risks for type 2 diabetes and coronary heart disease are not provided if the ethnicity of the gene donor is not stated as Estonian.

5. TURNING NATIONAL ELECTRONIC HEALTH RECORDS INTO BETTER FORMAT

Estonian EHR data is exceptional from several aspects. First, they are mostly written in Estonian, requiring Estonian natural language processing tools and skills to analyze them on a large scale. Second, central EHR databases (discharge summaries, prescriptions, bills) cover almost the whole population and can be relatively easily linked together electronically. The latter provides a unique opportunity not only to do research on national data but also to build new personalized medicine services for the entire country.

The author of this thesis has a long experience of working with Estonian national EHR data, especially with facilitating scientific use of these data. If one wants to integrate new evidence-based personalized medicine approaches to national-level clinical workflows and EHR systems, the underlying data and systems must be well understood. Due to local nuances, the publications in this topic (**Ref. V**, **Ref. VI**, **Ref. X**) have been mostly targeted to local researchers and authorities, written mainly in Estonian, and are therefore not included in this thesis. However, in this chapter, a brief overview of these studies is provided.

5.1. High-quality decisions need high-quality information

Across the world, EHR data have been mainly collected for the treatment of a particular patient rather than scientific purposes. These documents can be seen as collections of information about a patient, stored in a format which allows easy manual recording and reading them through later (*e.g.* by a clinician) if needed. Therefore, the format restrictions of such documents are usually relatively loose in order to not limit adding any kind of relevant information (Scholte et al., 2016). However, as soon as one wants to apply some automatic analyses on these data – either for research or for clinical decision support systems – such format is not suitable anymore. Well-defined format and structure are essential for machine readability (Hripcsak and Albers, 2012; Obermeyer and Emanuel, 2016). For this reason, converters are needed to extract the necessary information from EHR documents and transform into better format.

In Estonia, the author of this thesis is a member of the health informaticians team (a joint effort by the University of Tartu and private companies STACC and Quretec) who have developed a data extraction and filtering pipeline for Estonian central discharge summary documents. It starts with exploding summary documents into smaller pieces and adding medical facts from the fragments to the relational database, followed by fact duplication filtering (sometimes new summary documents are started with duplicating some old ones, so that the old facts are also present in the new document) and fact extraction from structured and free text parts, including identifying laboratory measurements and unifying their units. Development of the pipeline has taken years by now, and the work continues. However, this has been a vital prerequisite for any further analysis of these data – either for assessing the data quality or building new services on top of them. Note that in order to use EHR data as input for personalized medicine algorithms, the data has to be extracted with high quality.

5.2. Need for automatic patient-level data summarization

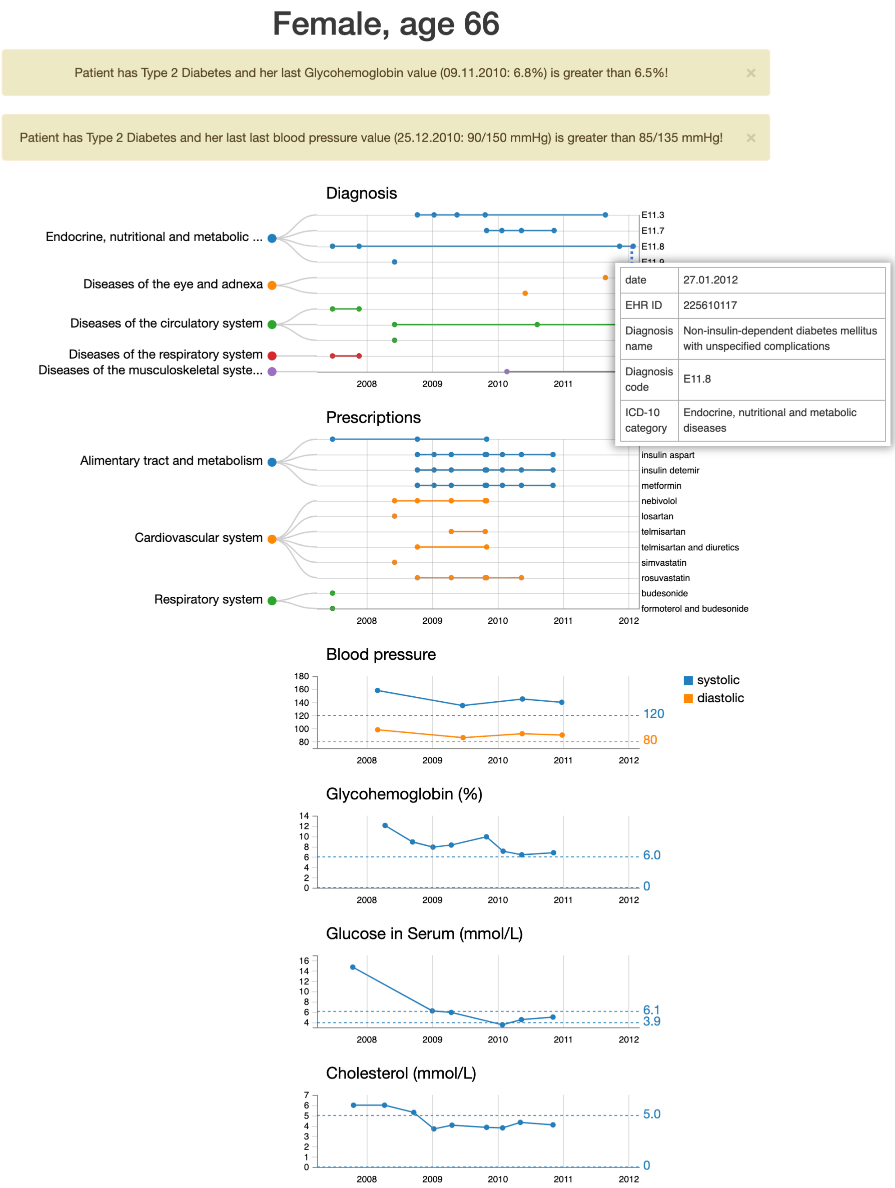
Recently, Estonian clinicians raised the problem in the Parliament that due to a large number of EHR documents per patient, it takes an unreasonable amount of time for a physician to look them through (Koppel, 2018). This is an indication of a problem that researches foresaw more than a decade ago – that automatic summaries of patient-level information are needed (Sittig et al., 2008). Although several research groups have developed patient-level summarization tools and prototypes to date and investigated different summarization techniques (Feblowitz et al., 2011; Hsu et al., 2012; Laxmisan et al., 2012; Pivovarov and Elhadad, 2015; Rind et al., 2013; West et al., 2014), we have also build our own version of it (Figure 7). It was designed for dealing with large amounts of data to get a quick overview of what has happened to a particular patient. Practising physicians in Estonia who have consulted us have found it very useful to overcome the information overload problem. Additionally, we have demonstrated its interaction with clinical guidelines of type 2 diabetes so that alerts about undone procedures are shown to the doctor.

5.3. Moving towards common data models

Using common health data structures is important for several reasons. Firstly, different coding systems and free text information make it extremely difficult to directly replicate any external study or apply a computer algorithm to the target environment. Therefore, it is hard to estimate their applicability in the target population. For instance, in **Ref. VIII** we presented how difficult it is to find a common “defining criterion” for patients of type 2 diabetes in different European databases. Secondly, data are more and more spread over several databases, and it is required to gather them at the time of care. Therefore, the semantic interoperability – the ability to exchange data with unambiguous, shared meaning – is necessary. Thirdly, as in personalized medicine the target patient groups for different treatment options are getting narrower, there is a need to combine several datasets to achieve a large enough number of a specific type of patients for a proper research study. This emphasizes the necessity of harmonization of the standards and data formats of health data cross-border. Recently launched EHDEN project is a great example of this – aiming to harmonize health records of 100 million Europeans into a common data format in 4 years for enabling reproducible science. They provide funding for observational data sources to map their data to OMOP common data model (Hripcsak et al., 2015) so that these data could be more easily used for research studies. I have already conducted an initial mapping of two central EHR databases to OMOP common data model (Digilugu, health insurance bills database). However, this effort is ongoing.

5.4. Well formatted EHR data opens new opportunities

When successful, having EHR data in a proper format opens new opportunities for investigating disease trajectories (Figure 8), associations between diseases and lab measurements or, in combination with genetic data, discover new associations between genetic variants and diseases (Evans, 2016). **Ref. III** (further explained in Chapter 6) is a perfect example of this – EHR diagnoses and laboratory measurements were combined with genetic data in order to validate the current DNA-disease associations and discover new ones. If not applicable to date, this can lead to better personalized treatments or prevention in the future.



Female, age 66

Patient has Type 2 Diabetes and her last Glycohemoglobin value (09.11.2010: 6.8%) is greater than 6.5%! ×

Patient has Type 2 Diabetes and her last last blood pressure value (25.12.2010: 90/150 mmHg) is greater than 85/135 mmHg! ×

date	27.01.2012
EHR ID	225610117
Diagnosis name	Non-insulin-dependent diabetes mellitus with unspecified complications
Diagnosis code	E11.8
ICD-10 category	Endocrine, nutritional and metabolic diseases

Figure 7. Example of the automatic patient-level summary report for Estonian discharge summary data. Diagnoses, prescriptions, blood pressure and laboratory measurements are displayed in a zoomable timeline graph. Above them, alerts about undone procedures, required by clinical guidelines for type 2 diabetes, are shown. From each data point, a tooltip with additional information can be opened.

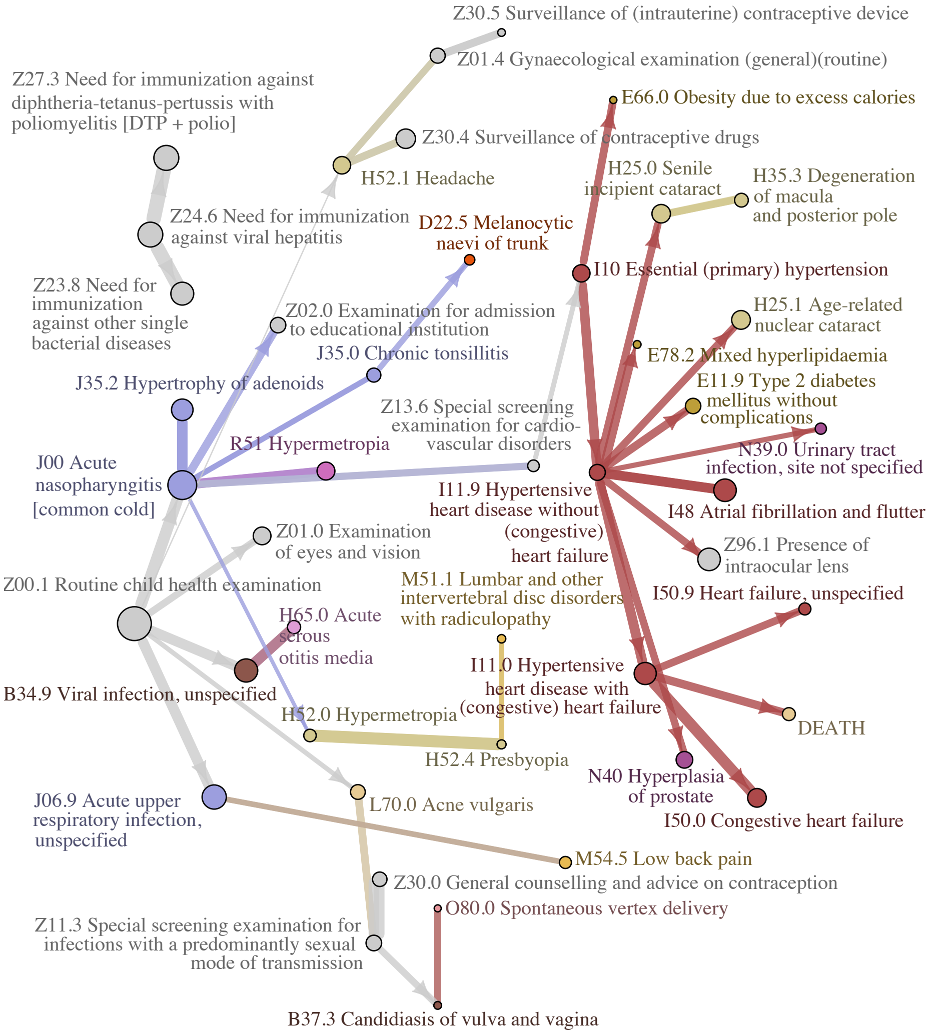

Z30.5 Surveillance of (intrauterine) contraceptive device
Z01.4 Gynaecological examination (general)(routine)
Z27.3 Need for immunization against diphtheria-tetanus-pertussis with poliomyelitis [DTP + polio]
E66.0 Obesity due to excess calories
Z30.4 Surveillance of contraceptive drugs
H52.1 Headache
H25.0 Senile incipient cataract
H35.3 Degeneration of macula and posterior pole
Z24.6 Need for immunization against viral hepatitis
D22.5 Melanocytic naevi of trunk
I10 Essential (primary) hypertension
Z23.8 Need for immunization against other single bacterial diseases
Z02.0 Examination for admission to educational institution
H25.1 Age-related nuclear cataract
J35.0 Chronic tonsillitis
E78.2 Mixed hyperlipidaemia
J35.2 Hypertrophy of adenoids
E11.9 Type 2 diabetes mellitus without complications
Z13.6 Special screening examination for cardio-vascular disorders
R51 Hypermetropia
N39.0 Urinary tract infection, site not specified
J00 Acute nasopharyngitis [common cold]
I11.9 Hypertensive heart disease without (congestive) heart failure
I48 Atrial fibrillation and flutter
Z01.0 Examination of eyes and vision
Z96.1 Presence of intraocular lens
Z00.1 Routine child health examination
M51.1 Lumbar and other intervertebral disc disorders with radiculopathy
I50.9 Heart failure, unspecified
H65.0 Acute serous otitis media
I11.0 Hypertensive heart disease with (congestive) heart failure
B34.9 Viral infection, unspecified
H52.0 Hypermetropia
DEATH
H52.4 Presbyopia
N40 Hyperplasia of prostate
J06.9 Acute upper respiratory infection, unspecified
L70.0 Acne vulgaris
I50.0 Congestive heart failure
M54.5 Low back pain
Z30.0 General counselling and advice on contraception
O80.0 Spontaneous vertex delivery
Z11.3 Special screening examination for infections with a predominantly sexual mode of transmission
B37.3 Candidiasis of vulva and vagina

Figure 8. Most common diagnoses and their order throughout life, based on EHR data of 1.3 million patient in central e-Health database of Estonia. Each node denotes one ICD-10 code, the size of the node indicates the number of patients having that diagnosis, the width of the arrow indicates the proportion of patients having both diagnoses. The number of incoming arrows for each node is limited to 1.

6. COMBINING ELECTRONIC HEALTH RECORDS AND GENETIC DATA ALLOWS INVESTIGATION OF NEW ASSOCIATIONS FOR FUTURE CARE (REF. III)

A number of genome-wide association studies (GWAS) where correlations between a genome-wide set of genetic variants and specific diseases are searched for in order to find disease-associated (and possibly causal) genetic variants, has grown rapidly after 2002 when the first study of that kind was published (Ozaki et al., 2002). In 2010, a reverse type of study was introduced – **phenome-wide association study** (**PheWAS**). In this approach, a small set of genetic variants, usually associated with some disease, are tested towards a large set of other diseases (or traits) to find out whether any previously unknown correlations peak up, potentially indicating a shared causal pathway (Denny et al., 2010). Uncovering genetic markers that signal potential disease could help to focus on disease prevention and early intervention.

PheWAS is heavily dependent on EHR data as it requires a diverse set of clinical events to test against. However, settings where biobank information could be linked to EHR are not very common to date. Estonian Biobank is one of those. It has not only genetic information available for nearly 5% of the Estonian population but also EHR data from state-level databases and laboratory measurements from main hospitals. Therefore, with the motivation to support new discoveries in clinical science, we utilized this valuable data and conducted a PheWAS on them (**Ref. III**). Particularly, we studied genetic variants that had been previously associated with asthma and liver disease, and tested in this computationally extensive study, whether there are other associations between these variants and health events. Genetic data of 26 thousand samples and EHR data of over 2,000 different diagnosis codes and 25 types of laboratory measurements were used. We confirmed 7 asthma and liver disease associations and found 2 phenome-wide significant associations with other diseases (type 1 diabetes, autoimmune thyroiditis). Although these particular associations were not completely novel, we believe that there was just not enough statistical power (too few cases) to detect all associations. However, as the EHR data of the gene donors grows in time, the power will also increase. Additionally, we showed that such integrated database settings could be effectively used for validation studies. Note that in order to utilize some previously published association in clinical practice, one should always validate the association in the target population – not only to verify its validity but also to assess the potential impact of the intervention.

This study has great importance in the light of a rapid expansion of Estonian Biobank. It is expected that there will be 200 thousand participants in the Biobank by the end of 2019. Therefore, it is possible to rerun the analysis on a much larger dataset in the near future, which could increase the statistical power and possibly allow detecting many more associations. This is the reason why Estonian Biobank and PheWASs are seen as valuable resources for discovering common disease pathways which would possibly lead to better treatment of these diseases in the future.

7. BUILDING PHARMACOGENOMICS RECOMMENDATION PIPELINE FOR 44,000 GENE DONORS (REF. IV)

The pharmacogenomics field has developed fast during the last decade. More than 250 drugs have pharmacogenomic information available in their information leaflets as of today, 69% of them are non-oncology drugs (FDA, 2019). They describe how a pharmacogenomic phenotype of a patient should be taken into account when selecting the right dosage or drug.

The problem is that detecting pharmacogenomic phenotype is not a trivial task. Although several genetic tests are available for that, they are usually quite limited – using only a small set of genetic variants and single genes. Therefore, it is not easy to decide in what cases which tests to use and at what costs.

At the same time, for a large amount of people, their genetic data already exists in various biobanks. In Estonia, they are stored in Estonian Biobank. Although thorough public mapping tables exist for linking genetic variants to pharmacogenomic phenotypes (Whirl-Carrillo et al., 2012), they are hard to use by medical doctors due to missing or ambiguous guidelines on how to interpret these tables.

Therefore, we saw great potential in utilizing existing genetic information and pharmacogenomic knowledge for comprehensive pharmacogenomic testing. In **Ref. IV**, we combined two pieces of information – the existing definition tables of pharmacogenomic alleles and the individual genomes – in order to provide personal pharmacogenomic recommendations for each gene donor (see also Figure 5 on page 27). We built a software pipeline for providing such pharmacogenomic recommendations for all donors who had their genetic data digitally available in the Biobank at that time (44 thousand participants). However, it turned out that building a universal pipeline requires several hurdles to overcome: non-unified notation of genomic data (see also Section 2.2.3), missing machine-readable pharmacogenomic knowledgebase, variety in the amount of evidence, missing actionability for some associations, *etc.* We describe some of these challenges in **Ref. IV**. Despite these issues, we were able to calculate the pharmacogenomics report for all 44 thousand people, and this is now used as a standard part of personal genetic counselling of Biobank participants.

We showed that 99.8% of the donors need a dosage adjustment for at least one of the medications (otherwise possibly suffering from side effects or having no effect of the drug), which is larger proportion than estimated before (98.5% by Dunnenberger et al. (2015)), highlighting the enormous potential and need for using the pipeline for the whole nation also. Moreover, we compared the results derived from different sequencing/genotyping methods and showed that the genotyping array – a cost-effective alternative for full sequencing (Martin et al., 2019) – could be successfully used for pharmacogenomic testing.

This study has had a direct impact. In spring 2019, a national-level project was launched to build the necessary IT infrastructure to bring the developed pharmacogenomics pipeline into state-level use. As using offline report is relatively inconvenient, the true value would reveal when the pipeline and alert mechanism were integrated into clinical workflows. That is, a pharmacogenomic alert should be shown to the physicians at the time of prescribing a drug (Gottesman et al., 2013). This also requires that the genetic information exists in some database (Ashley, 2016) and is securely maintained but accessible by EHR systems online when needed (Evans, 2016).

8. CONCLUSION

The driving force behind this thesis has been Estonian state level desire to bring personalized medicine into routine care. However, there is a number of challenges to overcome. This thesis is investigating some of these, mainly computational ones, by combining several datasets, conducting large scale calculations and developing computer algorithms for large scale use.

The results of this thesis can be summarized as follows.

- Due to differences in genetic background, polygenic risk scores used for assessing genetic risks of the diseases can produce incorrect risk estimations when applied to individuals from different populations. In admixed populations, these scores should be used with caution. It is suggested to verify the ancestry of an individual before using a polygenic risk score for disease risk estimation.
- Genomic data together with electronic health records in Estonian Biobank can be effectively used for phenome-wide association studies to discover new associations between DNA and diseases and thereby increase our understanding of these diseases. It can also be used for validation studies.
- Almost everyone needs a dosage adjustment for some medications. We describe a pipeline for producing individual pharmacogenomics recommendations for 44 thousand biobank participants.

The process of integrating both the developed pharmacogenomic diagnostic pipeline and genetic risk scores onto state level clinical practice is currently ongoing. To become an accepted part of routine clinical treatment, fitting these tools to clinical workflows and integrating them with existing (EHR) systems is critical. For instance, pharmacogenomic alerts should be shown only if the doctor is going to prescribe a medication which requires an adjusted dosage; polygenic risks of diseases should be presented only if a person is genetically a representative of the population that was used to build the risk model and if he/she has not already been diagnosed with the disease. This is why it is vital to understand the characteristics of the data that EHR systems and genetic databases contain – what is the quality of the data, how to analyze them and how to build cost-effective pipelines for making algorithm-based health-related recommendations. This thesis has increased our understanding in all of these aspects and helps us to step closer towards national-level implementation of personalized medicine.

There is strong evidence that the amount of genetic data is about to increase rapidly in the clinical setting during the upcoming years and decades. While being so far used mainly for scientific purposes, we are reaching a point where people are more and more expecting to use this information for individual clinical care. Therefore, bringing all pieces of information smoothly into clinical workflows requires additional work on semantic interoperability, and it can be foreseen that new health data standards will be developed. Several initiatives are already working on this topic, aiming to use health-data cross-border across European countries. This would not only make it easier for the patient to move around and get high-quality treatment anywhere but also allow using much larger datasets for clinical research.

Together with the growing amount of genetic data, our understanding of ethnic diversity is going to increase. Polygenic risk scores, currently suffering from ethnic bias, will be globally more balanced, containing more causal variants, leading to better disease risk estimation capabilities even in admixed populations.

As a result of increasing amounts of data, new associations between genetics and diseases could be detected, leading to better prevention, diagnosis and treatment. Therefore, a proper IT infrastructure, together with the skilled medical staff and data scientists, as well as legal and supporting environment for the multidisciplinary collaboration must be in place to take the full advantage of this new era. It is expected that several new nation-wide personalized medicine initiatives will be kicked off in the near future.

Though not discussed in this PhD thesis, it must not be overlooked that the field of using any individual-level data in general and using computer software in healthcare is getting more and more regulated. During this PhD study, the European Parliament approved General Data Protection Regulation (GDPR) in 2016, setting privacy issues into primary focus. Additionally, software used in the medical setting is now seen as medical devices, falling under appropriate regulations. In Europe, two new regulations – Medical Devices Regulation (2017/745/EU) and In-vitro Diagnostic Medical Devices Regulation (2017/746/EU), both published 2017 – apply in 2020 and 2022, and have a significant impact on medical software, possibly also on the results of this thesis. In order to improve clinical safety, they put more responsibility on computer software used in healthcare and on the manufacturer, requesting an external conformity assessment, post-market surveillance *etc.* While there is no doubt that stronger regulation improves the privacy, quality of the computerized assistance and safety of the patients, it will also require a lot of time and energy to bring new products to the clinical practice. Therefore, future studies have to put much more effort into satisfying various regulations.

Despite these limits and other challenges discussed in this thesis, the trend towards more personalized medicine is undeniable. The more information gets available, the more computers and advanced computer algorithms are needed to analyze the data, stratify patients and diseases into more granular subgroups and thereby helping clinicians to select the optimal therapy for each individual patient. However, there are no signs that the personal interaction between doctors and patients is about to disappear. We, as bioinformaticians, have to equip doctors with useful tools to support this interaction as much as possible, which, as a result, would lead to better health for everyone.

BIBLIOGRAPHY

1000Genomes. 1000 genomes project consortium and others. a global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

1000GenomesProject. 1000 genomes project consortium. a global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

G. Abraham, A. S. Havulinna, O. G. Bhalala, S. G. Byars, A. M. De Livera, L. Yetukuri, E. Tikkanen, M. Perola, H. Schunkert, E. J. Sijbrands, et al. Genomic prediction of coronary heart disease. *European heart journal*, 37(43): 3267–3278, 2016.

L. Aleksovska-Stojkovska and S. Loskovska. Clinical decision support systems: Medical knowledge acquisition and representation methods. In *2010 IEEE International Conference on Electro/Information Technology*, pages 1–6. IEEE, 2010.

O. Ali. Genetics of type 2 diabetes. *World journal of diabetes*, 4(4):114, 2013.

M. Aly, F. Wiklund, J. Xu, W. B. Isaacs, M. Eklund, M. D'Amato, J. Adolfsson, and H. Grönberg. Polygenic risk score improves prostate cancer risk prediction: results from the stockholm-1 cohort study. *European urology*, 60(1):21–28, 2011.

A. Alyass, M. Turcotte, and D. Meyre. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics*, 8(1): 33, 2015.

A. Antoniou, P. D. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper, N. Loman, H. Olsson, O. Johannsson, Å. Borg, et al. Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics*, 72(5):1117–1130, 2003.

E. A. Ashley. Towards precision medicine. *Nature Reviews Genetics*, 17(9):507, 2016.

D. W. Belsky and S. Israel. Integrating genetics and social science: Genetic risk scores. *Biodemography and social biology*, 60(2):137–155, 2014.

N. Blau, F. J. van Spronsen, and H. L. Levy. Phenylketonuria. *The Lancet*, 376 (9750):1417–1427, 2010.

T. F. Boat, M. J. Field, et al. *Rare diseases and orphan products: Accelerating research and development*. National Academies Press, 2011.

T. Botsis, G. Hartvigsen, F. Chen, and C. Weng. Secondary use of ehr: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010:1, 2010.

B. Braschi, P. Denny, K. Gray, T. Jones, R. Seal, S. Tweedie, B. Yates, and E. Bruford. Genenames. org: the hgnc and vgnc resources in 2019. *Nucleic acids research*, 47(D1):D786–D792, 2018.

J. Buchanan, S. Wordsworth, and A. Schuh. Issues surrounding the health economic evaluation of genomic technologies. *Pharmacogenomics*, 14(15):1833–1847, 2013.

C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018.

E. E. Calle, M. J. Thun, J. M. Petrelli, C. Rodriguez, and C. W. Heath Jr. Body-mass index and mortality in a prospective cohort of us adults. *New England Journal of Medicine*, 341(15):1097–1105, 1999.

C. S. Carlson, T. C. Matise, K. E. North, C. A. Haiman, M. D. Fesinmeyer, S. Buyske, F. R. Schumacher, U. Peters, N. Franceschini, M. D. Ritchie, et al. Generalization and dilution of association results from european gwas in populations of non-european ancestry: the page study. *PLoS biology*, 11(9): e1001661, 2013.

K. E. Caudle, H. M. Dunnenberger, R. R. Freimuth, J. F. Peterson, J. D. Burlison, M. Whirl-Carrillo, S. A. Scott, H. L. Rehm, M. S. Williams, T. E. Klein, et al. Standardizing terms for clinical pharmacogenetic test results: consensus terms from the clinical pharmacogenetics implementation consortium (cpic). *Genetics in Medicine*, 19(2):215, 2017.

C. A. Challener. Fda marks record year for new drug approvals. *Pharmaceutical Technology*, 43(1):30—-33, 2019.

D. Charles, M. Gabriel, and T. Searcy. Adoption of electronic health record systems among u.s. non-federal acute care hospitals: 2008-2014, 2015. `https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf`.

R. Chen and M. Snyder. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(1):73–82, 2013.

N. Chow, L. Gallo, and J. W. Busse. Evidence-based medicine and precision medicine: Complementary approaches to clinical decision-making. *Precision Clinical Medicine*, 1(2):60–64, 2018.

E. W. Chua and M. A. Kennedy. Current state and future prospects of direct-to-consumer pharmacogenetics. *Frontiers in pharmacology*, 3:152, 2012.

G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *BMC medicine*, 13(1):1, 2015.

I. H. G. S. Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

M. Curnutte. Regulatory controls for direct-to-consumer genetic tests: a case study on how the fda exercised its authority. *New Genetics and Society*, 36(3): 209–226, 2017.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

K. De Boeck, A. Zolin, H. Cuppens, H. V. Olesen, and L. Viviani. The relative frequency of cftr mutation classes in european patients with cystic fibrosis. *Journal of Cystic Fibrosis*, 13(4):403–409, 2014.

G. De Moor, M. Sundgren, D. Kalra, A. Schmidt, M. Dugas, B. Claerhout, T. Karakoyun, C. Ohmann, P.-Y. Lastic, N. Ammour, et al. Using electronic health records for clinical research: the case of the ehr4cr project. *Journal of biomedical informatics*, 53:162–173, 2015.

H. G. de Vries, J. M. Collée, H. E. de Walle, M. H. van Veldhuizen, C. T. S. Sibinga, H. Scheffer, and L. Ten Kate. Prevalence of δf508 cystic fibrosis carriers in the netherlands: logistic regression on sex, age, region of residence and number of offspring. *Human genetics*, 99(1):74–79, 1996.

J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

R. H. Dolin, L. Alschuler, C. Beebe, P. V. Biron, S. L. Boyer, D. Essin, E. Kimber, T. Lincoln, and J. E. Mattison. The hl7 clinical document architecture. *Journal of the American Medical Informatics Association*, 8(6):552–569, 2001.

F. Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348, 2013.

D. J. Duffy. Problems, challenges and promises: perspectives on precision medicine. *Briefings in bioinformatics*, 17(3):494–504, 2015.

H. M. Dunnenberger, K. R. Crews, J. M. Hoffman, K. E. Caudle, U. Broeckel, S. C. Howard, R. J. Hunkler, T. E. Klein, W. E. Evans, and M. V. Relling. Preemptive clinical pharmacogenetics implementation: current programs in five us medical centers. *Annual review of pharmacology and toxicology*, 55:89–106, 2015.

S. M. Edwards, Z. Kote-Jarai, J. Meitz, R. Hamoudi, Q. Hope, P. Osin, R. Jackson, C. Southgate, R. Singh, A. Falconer, et al. Two percent of men with early-onset prostate cancer harbor germline mutations in the brca2 gene. *The American Journal of Human Genetics*, 72(1):1–12, 2003.

J. Euesden, C. M. Lewis, and P. F. O'reilly. Prsice: polygenic risk score software. *Bioinformatics*, 31(9):1466–1468, 2014.

R. Evans. Electronic health records: then, now, and in the future. *Yearbook of medical informatics*, 25(S 01):S48–S61, 2016.

FDA. Table of pharmacogenomic biomarkers in drug labeling, 2019. Accessed on 27 March 2019:
`https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm`.

J. C. Feblowitz, A. Wright, H. Singh, L. Samal, and D. F. Sittig. Summarization of clinical information: a conceptual model. *Journal of biomedical informatics*, 44(4):688–699, 2011.

G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman. Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27(13): 1741–1748, 2011.

T. C. for the Study of Drug Development. Lack of clinically useful diagnostics hinder growth in personalized medicines. impact report, 2011.

H. Fröhlich, R. Balling, N. Beerenwinkel, O. Kohlbacher, S. Kumar, T. Lengauer, M. H. Maathuis, Y. Moreau, S. A. Murphy, T. M. Przytycka, et al. From hype to reality: data science enabling personalized medicine. *BMC medicine*, 16(1): 150, 2018.

C. D. Galloway, D. E. Albert, and S. B. Freedman. iphone ecg application for community screening to detect silent atrial fibrillation: a novel technology to prevent stroke. *International journal of cardiology*, 165:193–194, 2013.

J. Gantz and D. Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007(2012):1–16, 2012.

G. S. Ginsburg and J. J. McCarthy. Personalized medicine: revolutionizing drug discovery and patient care. *TRENDS in Biotechnology*, 19(12):491–496, 2001.

G. S. Ginsburg and K. A. Phillips. Precision medicine: from science to value. *Health Affairs*, 37(5):694–701, 2018.

B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, 2017.

O. Gottesman, H. Kuivaniemi, G. Tromp, W. A. Faucett, R. Li, T. A. Manolio, S. C. Sanderson, J. Kannry, R. Zinberg, M. A. Basford, et al. The electronic medical records and genomics (emerge) network: past, present, and future. *Genetics in Medicine*, 15(10):761, 2013.

HapMap. International hapmap consortium and others. the international hapmap project. *Nature*, 426(6968):789, 2003.

HapMap. International hapmap consortium and others. a haplotype map of the human genome. *Nature*, 437(7063):1299, 2005.

K. He, D. Ge, and M. He. Big data analytics for genomic medicine. *International journal of molecular sciences*, 18(2):412, 2017.

D. A. Hinds, L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and D. R. Cox. Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–1079, 2005.

G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2012.

G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574, 2015.

W. Hsu, R. K. Taira, S. El-Saden, H. Kangarloo, and A. A. Bui. Context-based electronic health record: toward patient specific healthcare. *IEEE Transactions on information technology in biomedicine*, 16(2):228–234, 2012.

G. James, S. Reisberg, K. Lepik, N. Galwey, P. Avillach, L. Kolberg, R. Mägi, T. Esko, M. Alexander, D. Waterworth, et al. An exploratory phenome wide association study linking asthma and liver disease genetic variants to electronic health records from the estonian biobank. *PloS one*, 14(4):e0215026, 2019.

J. V. Johansson, S. Langenskiöld, P. Segerdahl, M. G. Hansson, U. U. Hösterey, A. Gummesson, and J. Veldwijk. Research participants' preferences for receiving genetic risk information: a discrete choice experiment. *Genetics in Medicine*, page 1, 2019.

A. Katsnelson. Momentum grows to make'personalized'medicine more'precise', 2013.

D. Kazandjian, G. M. Blumenthal, W. Yuan, K. He, P. Keegan, and R. Pazdur. Fda approval of gefitinib for the treatment of patients with metastatic egfr mutation–positive non–small cell lung cancer. *Clinical Cancer Research*, 22(6):1307–1312, 2016.

T. J. Key, P. K. Verkasalo, and E. Banks. Epidemiology of breast cancer. *The lancet oncology*, 2(3):133–140, 2001.

R. Khan and D. Mittelman. Consumer genomics will change your life, whether you get tested or not. *Genome biology*, 19(1):120, 2018.

A. V. Khera, M. Chaffin, K. H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M. E. Haas, A. Bick, et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*, 177(3):587–596, 2019.

K. Koppel. Joller: aegunud tarkvara tõttu peab perearst patsiendi asemel ekraani vaatama. *Eesti Rahvusringhääling*, 2018. URL `https://www.err.ee/879036/joller-aegunud-tarkvara-tottu-peab-perearst-patsiendi-asemel-ekraani-vaatama`.

K. Läll, R. Mägi, A. Morris, A. Metspalu, and K. Fischer. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genetics in Medicine*, 19(3):322, 2017.

A. Laxmisan, A. B. McCoy, A. Wright, and D. F. Sittig. Clinical summarization capabilities of commercially-available and internally-developed electronic health records. *Applied clinical informatics*, 3(01):80–93, 2012.

L. Leitsalu, T. Haller, T. Esko, M.-L. Tammesoo, H. Alavere, H. Snieder, M. Perola, P. C. Ng, R. Mägi, L. Milani, et al. Cohort profile: Estonian biobank of the estonian genome center, university of tartu. *International journal of epidemiology*, 44(4):1137–1147, 2014.

H. Liivlaid, N. Eigo, and S. Reisberg. Eriarstiabi haigestumusstatistika võrdlus tervise arengu instituudi ja eesti haigekassa andmetel. *Eesti Arst*, 98:17–26, 2019.
`https://eestiarst.ee/eriarstiabi-haigestumusstatistika-vordlus-tervise-arengu-instituudi-ja-eesti-haigekassa-andmetel` (in Estonian).

J. Loscalzo and A.-L. Barabasi. Systems biology and the future of medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(6):619–627, 2011.

A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly. Current clinical use of polygenic scores will risk exacerbating health disparities. *BioRxiv*, 2019.

J. J. McCarthy, H. L. McLeod, and G. S. Ginsburg. Genomic medicine: a decade of successes, challenges, and opportunities. *Science translational medicine*, 5 (189):189sr4–189sr4, 2013.

S. A. McLean. *Critical interventions in the ethics of healthcare: Challenging the principle of autonomy in bioethics*. Ashgate Publishing, Ltd., 2013.

K. Mooses, M. Oja, S. Reisberg, J. Vilo, and M. Kull. Validating fitbit zip for monitoring physical activity of children in school: a cross-sectional study. *BMC public health*, 18(1):858, 2018.

S. Nawrocki. Molecular profiling of tumours for precision oncology–high hopes versus reality. *Contemporary Oncology*, 22(1A):3, 2018.

D. W. Nebert, L. Jorge-Nebert, and E. S. Vesell. Pharmacogenomics and "individualized drug therapy". *American Journal of Pharmacogenomics*, 3(6):361–370, 2003.

Z. Obermeyer and E. J. Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13): 1216, 2016.

K. Ozaki, Y. Ohnishi, A. Iida, A. Sekine, R. Yamada, T. Tsunoda, H. Sato, H. Sato, M. Hori, Y. Nakamura, et al. Functional snps in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, 32 (4):650, 2002.

N. Petrucelli, M. B. Daly, and T. Pal. Brca1-and brca2-associated hereditary breast and ovarian cancer. *GeneReviews*, 1998.

J. H. Phan, C. F. Quo, C. Cheng, and M. D. Wang. Multiscale integration of-omic, imaging, and clinical data in biomedical informatics. *IEEE reviews in biomedical engineering*, 5:74–87, 2012.

A. A. Philippakis, D. R. Azzariti, S. Beltran, A. J. Brookes, C. A. Brownstein, M. Brudno, H. G. Brunner, O. J. Buske, K. Carey, C. Doll, et al. The matchmaker exchange: a platform for rare disease gene discovery. *Human mutation*, 36(10):915–921, 2015.

A. M. Phillips. Only a click away—dtc genetics for ancestry, health, love… and more: A view of the business and regulatory landscape. *Applied & translational genomics*, 8:16–22, 2016.

K. A. Phillips, J. A. Sakowski, J. Trosman, M. P. Douglas, S.-Y. Liang, and P. Neumann. The economic value of personalized medicine tests: what we know and what we need to know. *Genetics in medicine*, 16(3):251, 2014.

R. Pivovarov and N. Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015.

PMC. The personalized medicine report 2017. opportunity, challenges, and the future. personalized medicine coalition., 2017. Accessed on 21 March 2019: `http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/The-Personalized-Medicine-Report1.pdf`.

A. Pokorska-Bocci, A. Stewart, G. S. Sagoo, A. Hall, M. Kroese, and H. Burton. 'personalized medicine': what's in a name? *Personalized Medicine*, 11(2): 197–210, 2014.

H. L. Rehm, S. J. Bale, P. Bayrak-Toydemir, J. S. Berg, K. K. Brown, J. L. Deignan, M. J. Friez, B. H. Funke, M. R. Hegde, and E. Lyon. Acmg clinical laboratory standards for next-generation sequencing. *Genetics in medicine*, 15(9):733, 2013.

K. Reinson. *Doctoral thesis: New diagnostic methods for early detection of inborn errors of metabolism in Estonia*. PhD thesis, University of Tartu, 2018.

S. Reisberg, H.-A. Talvik, K. Koppel, S. Laur, and J. Vilo. Description of the current status and future needs of the information architecture and data management solutions for the national personalised medicine pilot project, 2015.
`https://www.sm.ee/sites/default/files/content-editors/eesmargid_ja_tegevused/Personaalmeditsiin/description_of_the_current_status_and_future_needs_of_the_information_architecture_and_data_management_solutions_for_the_national_personalised_medicine_pilot_project.pdf`.

M. V. Relling and W. E. Evans. Pharmacogenomics in the clinic. *Nature*, 526 (7573):343, 2015.

J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, B. Shneiderman, et al. Interactive information visualization to explore and query electronic health records. *Foundations and Trends® in Human–Computer Interaction*, 5(3):207–298, 2013.

M. Rius and J. A. Darling. How important is intraspecific genetic admixture to the success of colonising populations? *Trends in ecology & evolution*, 29(4): 233–242, 2014.

J. Rüschoff, A. Lebeau, H. Kreipe, P. Sinn, C. D. Gerharz, W. Koch, S. Morris, J. Ammann, M. Untch, et al. Assessing her2 testing quality in breast cancer: variables that influence her2 positivity rate from a large, multicenter, observational study in germany. *Modern Pathology*, 30(2):217, 2017.

C. P. Schade, F. M. Sullivan, S. De Lusignan, and J. Madeley. e-prescribing, efficiency, quality: lessons from the computerization of uk family practice. *Journal of the American Medical Informatics Association*, 13(5):470–475, 2006.

E. S. Schellenberg, D. M. Dryden, B. Vandermeer, C. Ha, and C. Korownyk. Lifestyle interventions for patients with and at risk for type 2 diabetes: a systematic review and meta-analysis. *Annals of internal medicine*, 159(8):543–551, 2013.

M. Scholte, S. A. van Dulmen, C. W. Neeleman-Van der Steen, P. J. van der Wees, M. W. Nijhuis-van der Sanden, and J. Braspenning. Data extraction from electronic health records (ehrs) for quality measurement of the physical therapy process: comparison between ehr data and survey data. *BMC medical informatics and decision making*, 16(1):141, 2016.

F. Severin, P. Borry, M. C. Cornel, N. Daniels, F. Fellmann, S. V. Hodgson, H. C. Howard, J. John, H. Kääriäinen, H. Kayserili, et al. Points to consider for prioritizing clinical genetic testing services: a european consensus process oriented at accountability for reasonableness. *European Journal of Human Genetics*, 23 (6):729, 2015.

D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, and D. W. Bates. Grand challenges in clinical decision support. *Journal of biomedical informatics*, 41(2):387–392, 2008.

D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire. Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. *science*, 235(4785):177–182, 1987.

B. B. Spear, M. Heath-Chiozzi, and J. Huff. Clinical application of pharmacogenetics. *Trends in molecular medicine*, 7(5):201–204, 2001.

I. F. Tannock, J. A. Hickman, et al. Limits to personalized cancer medicine. *N Engl J Med*, 375(13):1289–1294, 2016.

B. M. Welch and K. Kawamoto. Clinical decision support for genetically guided personalized medicine: a systematic review. *Journal of the American Medical Informatics Association*, 20(2):388–400, 2012.

V. L. West, D. Borland, and W. E. Hammond. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, 22(2):330–339, 2014.

M. Whirl-Carrillo, E. M. McDonagh, J. Hebert, L. Gong, K. Sangkuhl, C. Thorn, R. B. Altman, and T. E. Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012.

G. R. Wilkinson. Drug metabolism and variability among patients in drug response. *New England Journal of Medicine*, 352(21):2211–2221, 2005.

O. Wolkenhauer, C. Auffray, R. Jaster, G. Steinhoff, and O. Dammann. The road from systems biology to systems medicine. *Pediatric research*, 73(4-2):502, 2013.

P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang. –omic and electronic health record big data analytics for precision medicine. *IEEE Transactions on Biomedical Engineering*, 64(2):263–273, 2017.

Y. Yang, D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *New England Journal of Medicine*, 369(16):1502–1511, 2013.

A. K. Yetisen, J. L. Martinez-Hurtado, B. Ünal, A. Khademhosseini, and H. Butt. Wearables in medicine. *Advanced Materials*, 30(33):1706910, 2018.

ACKNOWLEDGEMENTS

First, I would like to thank my employers – **University of Tartu**, **STACC** and **Quretec** – who have challenged me to lead the satisfying projects with giving enough responsibility together with the academic freedom, and provided their continuous support – both technical and financial – which altogether has allowed me to fully focus on this topic. It has also given a fantastic opportunity for me for building contacts with inspiring people in the healthcare field in Estonia. In addition, participating in the European level projects and developing international relationships with the top scientists in that field would not be possible without EMIF and EHDEN projects, funded jointly by European Parliament and European Federation of Pharmaceutical Industries and Associations. I'm very grateful to my supervisor professor **Jaak Vilo** who called me to these projects and gave me a great responsibility there.

I'm deeply grateful to all my co-authors, especially **Kristi Läll**, **Krista Fischer**, **Jaak Vilo**, **Kristi Krebs**, **Lili Milani**, **Volker M. Lauschke**, **Glen James**, **Katrina Loomis**, **Dawn Waterworth**, **Kaido Lepik**, **Liis Kolberg**, **Kerli Mooses**, **Tatjana Iljasenko**, **Hedi Liivlaid**, **Mart Kals**, **Reedik Mägi**, **Kristjan Metsalu**, **Maarja Lepamets**, **Tõnu Esko**, and **Raul Sirel** with the wonderful opportunity to do the research together. It has been a real pleasure for me. I hope there are several new projects ahead where we can collaborate in the future.

In science, nothing can be done alone. There is an inspiring health data team at STACC whose work and effort have allowed me to write large parts of this thesis. I'm deeply thankful to all of you, especially **Harry-Anton Talvik**, **Sven Laur**, **Dage Särg**, **Marek Oja**, **Mihkel Manguse**, **Kristjan Eljand** and **Evely Kukk**. Also, people from other institutions have helped me a lot, including **Kaur Alasoo**, **Elena Sügis**, **Tormi Reinson**, **Martin Kaljula**, **Ivar Koppel**, **Pille Banhard**, **Triin Tommingas**, **Marili Palover**, **Meeli Mets**, **Gea Pajula**, **Margus Jäger**.

Finally, I would also like to thank my **lovely family**. I'm deeply grateful for supporting me all that time.

SISUKOKKUVÕTE

Arvutuslikud meetodid personaalmeditsiini arendamiseks

Kuigi meditsiin on alati põhinenud patsiendi ja arsti vahelisel individuaalsel suhtlusel, on mõiste *personaalmeditsiin* tulnud laiemasse kasutusse alles viimastel aastakümnetel. Selle ajendiks on asjaolu, et senised raviskeemid toimivad küll paljudel patsientidel, kuid siiski mitte kõigil. Selle põhjuseks on patsientide individuaalsed, eriti geneetilised, omapärad. Personaalmeditsiini eristab traditsioonilisest meditsiinist püüe võtta maksimaalselt arvesse ka patsiendi individuaalset (geneetilist) tausta, mis võimaldaks senisest efektiisemat haiguste ennetust ja ravi.

Selline lähenemine on saanud võimalikuks tänu mitmele asjaolule. Esiteks on inimese geeniandmete tuvastamine bioloogilisest materjalist (verest) saanud laialt kättesaadavaks, mistõttu kasvab geeniandmete maht kogu maailmas kiiresti. Teiseks on tänapäeva arvutusvõimsus jõudnud geeniandmete analüüsimiseks ja töötlemiseks piisavale tasemele. Kolmandaks kasutab suurem osa arste oma töölaual elektroonilisi infosüsteeme, mis teeb võimalikuks kõigi kolme komponendi – geeniandmed, arvutusvõimsus, arsti töölaud – omavahelise ühendamise ja arstidele nn digitaalse otsustustoe pakkumise.

Mõnedes kliinilistes valdkondades on personaalmeditsiin juba jõudnud tavapraktikasse. Näiteks onkoloogias, kus täpne raviskeem valitakse vastavalt konkreetse patsiendi vähirakkude iseärasustele. Samuti teostatakse Eestis raseduseelset, rasedusaegset ja vastsündinute geneetilist testimist eesmärgiga varakult ära tunda võimalikke tõsiseid pärilikke haigusi. Kliinilisse praktikasse on jõudmas farmakogeneetiline testimine, mille eesmärgiks on tuvastada, kas konkreetsele patsiendile sobib konkreetse ravimi tavapärane doos või tuleks seda kohandada. Samuti on tänaseks välja töötatud mitmeid nn polügeenseid riskiskoore, mis aitavad hinnata keerukamate haiguste nagu 2. tüüpi diabeedi või südamehaiguste pärilikku riski ja on potentsiaalselt kasutatavad haiguse riskirühmade kindlakstegemiseks (nt sõeluuringutele kutsumiseks).

Personaalmeditsiini kontseptsiooni kliinilises praktikas rakendamine hõlmab mitmete eri osapoolte – arstide, bioinformaatikute, statistikute, bioloogide jne – koostööd, mistõttu on mitmetes riikides käivitatud selle toetamiseks riiklikul tasemel spetsiaalsed personaalmeditsiini programmid. Lisaks nõuab see paljude eri valdkondade küsimuste lahendamist, m.h regulatoorsete, tehniliste, eetiliste, hariduslike ja finantsiliste.

Ka Eestis on käivitatud riiklik personaalmeditsiini programm. Käesolev doktoritöö ongi seotud personaalmeditsiini rakendamisega Eesti tervishoiusüsteemis, kasutades Eesti Geenivaramu geeniandmeid ning käsitledes eelkõige selle arvutuslikke küsimusi.

Polügeensed riskiskoorid on matemaatilised arvutusmudelid, mis isiku geneetilise info põhjal ennustavad, kas isikul on madal, keskmine või kõrge pärilik risk teatud haiguse tekkimiseks. Need mudelid on tavaliselt välja töötatud teadustööks kättesaadavate geeniandmete puhul, mis on tänapäeval paraku valdavalt Euroopa päritolu. Seetõttu on ka välja töötatud riskiskoorimudelid “kallutatud” ega sobi riski hindamiseks mitte-eurooplastele. Käesoleva doktoritöö artikkel **Ref. I** oli üks esimesi artikleid, kus seda probleemi sõnastati ja ühtlasi esimene, kus näidati, et tuhandeid geenimutatsioone sisaldavate riskiskooride korral – näiteks 2. tüüpi diabeedi (7500 geenimutatsiooni) ja südamehaiguste riskihinnang (49 000 mutatsiooni) – võib teisest populatsoonist pärit isikule olla ennustatud riskihinnang risti vastupidine tema tegelikule riskile. Artiklis **Ref. II** on näidatud, et sama kehtib ka astma ja maksahaigustega seotud geenimutatsioonide kohta – eestlaste hulgas on nende mutatsioonide sagedused sarnased teistele eurooplastele, kuid erinevad muudest rahvastest. Seetõttu tuleb riskiskooride puhul vältida nende mudelite kasutamist sellistel isikutel/patsientidel, kelle päritolu oluliselt erineb mudeli väljatöötamise aluseks olevate isikute päritolust. Nimetatud artiklitel on ka olnud otsene mõju – Geenivaramu poolt oma doonoritele antavad personaalsel nõustamisel ei kasutata riskiskoore, kui geenidoonor pole oma rahvuseks märkinud eestlane või venelane (s.t on põhjust arvata, et tema genoom erineb oluliselt tüüpilise eestimaalase omast).

Lisaks geeniandmetele tuleb personaalmeditsiinis arvesse võtta ka muid erinevatest terviseinfosüsteemidest pärinevaid terviseandmeid. Paraku on kogutud terviseandmed ebaühtlase kvaliteediga ja nende ühtlustamise ning paremini kasutatavale kujule viimise osas on käesoleva doktoritöö autor viinud läbi mitmeid teadusuuringuid, mida on käsitletud peatükis 5. Samas on näidatud lahendusi, mida ühtlustatud andmeid kasutades on võimalik luua.

Valdavalt on senised geeniuuringud keskendunud konkreetse haiguse põhjuslikkuse uurimisele – otsides haigusega korreleeruvaid geenimutatsioone (nn genoomiülene assotsiatsiooniuuring). Seoses elektrooniliste terviseandmete laialdase levikuga, ning juhul kui neid on võimalik kokku viia geeniandmetega (nagu Eesti Geenivaramus), on sel kümnendil hakatud läbi viima ka vastupidiseid uuringuid – nn fenoomiüleseid assotsiatsiooniuuringuid. Nende sisuks on uurida, milliste haigustega mingi konkreetne geenimutatsioon on korreleerunud. See eeldab suuremahulist geeniandmete ja terviseandmete ühendamist, ning artiklis **Ref. III** vaadeldi, milliste muude haigustega on seotud mutatsioonid, mis varasemalt on seostatud astma ja maksahaigustega. Nimetatud artiklis kasutati ligi 27 000 geenidoonori geeniandmeid, kuid see valim osutus uute seoste leidmiseks liialt väikseks. Kuna aga Geenivaramu andmestik kasvab 2019. aasta lõpuks eeldatavasti 200 000 geenidoonorini, on võimalik sama uuringut lähitulevikus korrata. Lisaks näitasime artiklis, et vaadeldud andmestik sobib hästi varem leitud geenide ja haiguste vaheliste seoste kontrollimiseks.

Farmakogeneetikat on peetud üheks kõige potentsiaalikamaks personaalmeditsiini rakendusvaldkonnaks, sest paljude ravimite infolehtedel on juba praegu kirjas soovitused selle kohta, kuidas sõltuvalt geneetilisest taustast konkreetse ravimi annustamist tuleks kohandada. Samas ei olnud protsess, mis kirjeldaks täpselt, kuidas konkreetse isiku andmetest jõuda ravimisoovituseni, varem kirjeldatud, ning jättis palju tõlgendamisruumi. Käesoleva doktoritöö artiklis **Ref. IV** loodi tarkvara, mis selle töö 11 geeni puhul ära teeb ning mille abil koostasime farmakogeneetilise info raportid 44 tuhandele geenidoonorile. Selgus, et tervelt 99,8% geenidoonoritel esineb niisuguseid geenivariante, mis nõuaksid mõne ravimi puhul koguse kohandamist (iseasi, kas doonor seda ravimit kunagi tarbib). See number on suurem, kui varasemalt näidatud. Lisaks näitasime, et kuigi kõige parem on farmakogeneetiliseks testimiseks kasutada nn täissekveneerimise tulemusel saadud geeniandmeid, pakub ka kordades odavam alternatiiv – nn genotüpiseerimiskiipide kasutamine koos imputeerimisega – peaaegu samaväärseid tulemusi.

Selleks, et personaalmeditsiin jõuaks kliinilisse praktikasse, on oluline tuua selle IT-lahendused arstide igapäevastesse töövoogudesse, mis paljuski tähendab nende integreerimist olemasolevate terviseinfosüsteemidega. Näiteks polügeensete riskiskooride põhjal tehtud haigusriskide ennustusi on mõtet teha üksnes juhul, kui terviseandmete põhjal patsient seda haigust veel ei põe. Ka farmakogeneetilisi hoiatusi konkreetse ravimi kohta tasub kuvada vaid siis, kui arst kaalub selle ravimi väljakirjutamist. Seetõttu on olemasolevate terviseinfosüsteemide ja terviseandmete tundmine oluline – teada, milliseid andmeid on olemas, milline on nende kvaliteet ning kuidas luua arstidele digitaalseks otsustustoeks algoritmidel põhinevaid kuluefektiivseid IT-töövoogusid. Käesolev doktoritöö on kõigis neis aspektides meie teadmisi laiendanud ja aitab Eesti riigi tasemel personaalmeditsiini rakendamisele oluliselt kaasa.

Praeguseks on Eestis käivitatud mitmed personaalmeditsiini valdkonna alamprojektid, mis käsitlevad selle eri tahke. Üheks selliseks on riikliku IT-infrastruktuuri loomise projekt, et võtta **Ref. IV** kirjeldatud tarkvara ka kliinilises praktikas kasutusele. See infrastruktuur peaks võimaldama kasutada ka geneetilisi riskiskoore. Seega töö nimetatud teemadega jätkub ning loodetavasti saab personaalmeditsiinist varsti kasu juba suur osa Eesti inimestest.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name:	Sulev Reisberg
Date of birth:	3 Feb 1982
Nationality:	Estonian
E-mail:	sulev.reisberg@ut.ee

Education

2015–...	PhD studies, informatics, Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia
2005–2006	Master's studies, telecommunication, Tallinn University of Technology, Estonia
2000–2005	Bachelor's studies (cum laude), telecommunication, Tallinn University of Technology, Estonia
1997–2000	Tallinn Secondary School of Science, Tallinn, Eesti
1988–1997	Pelgulinna Gümnaasium, Tallinn, Eesti

Employment

2013–...	Programmer, Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia
2009–...	Project manager / researcher, STACC, Estonia
2009–...	Project manager / system analyst, Quretec, Estonia
2008–2009	Head of Software development department, Bigbank, Estonia
2008–2008	Project manager, Bigbank, Estonia
2006–2008	Programmer, Bigbank, Estonia
2005–2006	Assistant, Tallinn University of Technology, Estonia

Scientific work

Main fields of interest

- health informatics
- genetics
- personalized medicine

Research grants and scholarships

2017–2019	Smart specialisation scholarship for PhD student (co-funded by European Regional Development Fund)

ELULOOKIRJELDUS

Isikuandmed

Nimi:	Sulev Reisberg
Sünniaeg:	03.02.1982
Rahvus:	Eestlane
E-post:	sulev.reisberg@ut.ee

Haridus

2015–...	Doktoriõpe, informaatika, loodus-ja täppisteaduste valdkond, Tartu Ülikool, Eesti
2005–2006	Magistriõpe, telekommunikatsioon, Tallinna Tehnikaülikool, Eesti
2000–2005	Bakalaureuseõpe (cum laude), telekommunikatsioon, Tallinna Tehnikaülikool, Eesti
1997–2000	Tallinna Reaalkool, Tallinn, Eesti
1988–1997	Pelgulinna Gümnaasium, Tallinn, Eesti

Teenistuskäik

2013–...	Tartu Ülikool, arvutiteaduse instituut, programmeerija
2009–...	STACC, projektijuht/teadur
2009–...	Quretec, projektijuht-analüütik
2008–2009	Bigbank, tarkvaraarenduse osakonna juhataja
2008–2008	Bigbank, projektijuht
2006–2008	Bigbank, programmeerija
2005–2006	Tallinna Tehnikaülikool, assistent

Teadustegevus

Peamised uurimisvaldkonnad

- terviseinformaatika
- geneetika
- personaalmeditsiin

Saadud uurimistoetused ja stipendiumid

2017–2019	Nutika spetsialiseerumise doktorandistipendium (kaasrahastatud Euroopa Regionaalarengufondist)

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.