DISSERTATIONES
MATHEMATICAE
UNIVERSITATIS
TARTUENSIS
127

KRISTI LÄLL

Risk scores and their predictive ability for common complex diseases

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
127

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
127

KRISTI LÄLL

Risk scores and their predictive ability for common complex diseases

UNIVERSITY OF TARTU
Press

Institute of Mathematics and Statistics, 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 mathematical statistics on 30th of April, 2019 by the Council of the Institute of Mathematics and Statistics, University of Tartu.

Supervisor

Prof. Krista Fischer
Institute of Mathematics and Statistics,
Faculty of Science and Technology, University of Tartu, Tartu, Estonia
Senior Research Fellow, Estonian Genome Center,
Institute of Genomics, University of Tartu, Tartu, Estonia

Opponents

Associate Prof. Juan R. González Ruiz
Barcelona Insitute for Global Health (ISGlobal),
Barcelona, Spain
Adjunct Prof. Department of Mathematics,
Autonomous University of Barcelona (UAB)

Associate Prof. Tanel Kaart
Chair of Animal Breeding and Biotechnology,
Institute of Veterinary Medicine and Animal Sciences,
Estonian University of Life Sciences, Tartu, Estonia

The public will take place on 12.06.2019 at 10:15 in J. Liivi 2.

The publication of this dissertation is granted by the Institute of Mathematics and Statistics, University of Tartu. This study was funded by EU H2020 grant 692145, Estonian Research Council Grant IUT20-60, IUT24-6, PUT1665, ETF9353 and European Union through the European Regional Development Fund Project No. 2014-2020.4.01.15-0012 GENTRANSMED end Project No. 2014-2020.4.01.16-0125, the Estonian Doctoral School of Mathematics and Statistics (NMTMM09577) (NLTMS16154) and TerVE programme grant PerMed I. This research was supported by national scholarship program Kristjan Jaak, which is funded and managed by Archimedes Foundation in collaboration with the Ministry of Education and Research. Data analyzes were carried out in part in the High-Performance Computing Center of University of Tartu.

ISSN 1024-4212
ISBN 978-9949-03-041-5 (print)
ISBN 978-9949-03-042-2 (PDF)

University of Tartu Press
www.tyk.ee

To my husband and colleagues from Estonian Genome Center for endless support and encouragement

CONTENTS

LIST OF ABBREVIATIONS

1000G:	1000 Genomes Project
AUC:	area under the curve
ACC/AHA:	American College of Cardiology/American Heart Association
ASCVD:	atherosclerotic cardiovascular disease
DNA:	deoxyribonucleic acid
EstBB:	Estonian Biobank
EGCUT:	Estonian Genome Center, University of Tartu
ESC:	European Society of Cardiology
FE:	fixed effect
GWAS:	genome-wide association study
GRS:	genetic risk score
dGRS:	doubly-weighted genetic risk score
LD:	linkage disequilibrium
MAF:	minor allele frequency
NICE:	UK National Institute for Health and Care Excellence
NRI:	net reclassification index
PCE:	Pooled Cohort Equation
QRISK2:	ASCVD Risk Estimator
SCORE:	Systematic COronary Risk Estimation
SIR:	standardized incidence ratio
SNP:	single nucleotide polymorphism
UKBB:	UK Biobank
WGS:	whole genome sequencing

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

This thesis is based on the following original publications, referred to in the text by Roman numerals (Ref. I to Ref. IV):

I Läll K, Mägi R, Morris A, Metspalu A, Fischer K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet Med.* 2017;19(3):322-329.

II Läll K, Lepamets M, Palover M, Esko T, Metspalu A, Tõnisson, N, Padrik P, Mägi R, Fischer K. Polygenic prediction of breast cancer: comparison of genetic predictors and implications for screening. *manuscript*.

III 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. Chaubey G, ed. *PLoS One.* 2017;12(7)

IV Saar A*, Läll K*, Alver M, Marandi T, Ainla T, Eha J, Metspalu A, Fischer K. (2018). Estimating the performance of three cardiovascular disease risk scores: the Estonian Biobank Cohort Study. *J Epidemiol Community Health*. 2019;73(3):272-277.

The publications listed above have been reprinted with the permission of the copyright owners.

My contributions to the listed publications were following:

Ref I: I was involved in planning the study and doing the majority of data management. I participated in analysing and visualising the data and writing of the manuscript.

Ref II: I participated in the study design, did majority of the data management, run all the analyses, prepared all the figures and drafted the first manuscript.

Ref III: I participated in the study design and data analysis and revised the manuscript.

Ref IV: I participated in the study design, performed all the data management, conducted the data analysis, prepared majority of the figures and co-wrote the first draft of the manuscript.

Publications not included in the thesis

V Elosua R, Lluís-Ganella C, Subirana I, Havulinna A, Läll K *et al.* Cardiovascular Risk Factors and Ischemic Heart Disease: Is the Confluence of Risk Factors Greater Than the Parts? A Genetic Approach. *Circ Cardiovasc Genet.* 2016;9(3):279-286.

VI Marioni RE, Ritchie SJ, Joshi PK, *et al.* Genetic variants linked to education predict longevity. *Proc Natl Acad Sci U S A*. 2016;113(47):13366-13371.

VII Mahajan A, Wessel J, Willems SM, *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat Genet.* 2018;50(4):559-571.

VIII Timmers PR, Mounier N, Läll K, *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife*. 2019;8.

IX Alver M, Palover M, Saar A, Läll K, *et al.* Recall by genotype and cascade screening for familial hypercholesterolemia in a population-based biobank from Estonia. *Genet Med*. 2018:1.

X Kals M, Nikopensius T, Läll K, Sikka T.T, Suvisaari J, Salomaa V, Ripatti S, Palotie A, Metspalu A, Palta P, Mägi R. (2018) Advantages of genotype imputation with ethnically-matched reference panel for rare variant association analyses. *Submitted*

INTRODUCTION

In the recent years, the cost for both genotyping and sequencing has been decreasing, making it possible to increase the number of individuals to be included in the genetics research. Large sample sizes together with detailed phenotypic and medical information have allowed researchers to address the role of genetic basis of several common complex diseases. Even though a lot is yet to be discovered about the genetic architecture of common complex diseases, scientists are already working on translating the current knowledge into advancements of everyday clinical setting.

One of the most studied (and also most common) source of genetic variation is single nucleotide polymorphisms (SNP). Even though SNPs explain only a fraction of the heritability of common complex diseases, their incremental value on top of classical risk factors has been shown to exist for many common complex diseases such as type 2 diabetes, coronary artery disease, breast cancer, etc. As each SNP usually has a small effect on a common complex disease, one needs to take into account the effects of many SNPs simultaneously to effectively estimate the genetic predisposition of a person. One option to do that is to use genetic risk scores (GRS, also referred to as "polygenic risk score").

GRS is essentially a sum of weighted effect allele counts of SNPs. However, their computation involves methodological challenges, as an optimal decision on the choice of SNPs and their weights needs to be made. The most popular choice used to be to include only a small number of SNPs, which association with a trait had been confirmed in several large studies. Later on more sophisticated methods have arisen since then.

It current thesis I first give an introduction to biological background related to genome-wide association studies (GWAS) and then describe in detail both the methodological and technical side of GWAS and meta-analysis. In the second part I introduce the basics of genetic risk scores and then focus on the new method which we have developed to improve genetic risk scores, called as the doubly-weighted GRS. I demonstrate its superiority compared to single-weighting with both simulations and with real data, using type 2 diabetes as an example (Ref I). In Ref II, I explain the idea of metaGRS by M. Inouye and colleagues and implement it to find the best predicting GRS for breast cancer. I also discuss the issue of non-uniqueness of the genetic risk score and its impact on genetic feedback. In Ref III, the effect of population admixture on genetic risk scores is assessed and the transferability of scores across populations is debated. Finally, in Ref IV, I focus on the aspects of validating and comparing non-genetic risk scores for cardiovascular diseases and cardiovascular death based on classical risk factors in Estonian Biobank. The goal was to assess whether internationally acknowledged scores are applicable in the Estonian population in the original form. I also compare the prevention guidelines in regards to statin recommendations to assess the overall cardiovascular disease risk levels in the Estonian Biobank.

1. BACKGROUND OF STATISTICAL ANALYSIS IN GENETICS

1.1. Introduction

The field of genetics has tight connections to medicine, as one of the current goals of research in genetics is to reveal the underlying biological mechanisms for diseases and traits. Screening for disorders with known genetic background is often already implemented in clinical practice. For example, prenatal genetic screening helps to identify babies with severe genetic disorders such as Down or Edwards syndrome and sickle cell anemia. There are inherited disorders such as familial hypercholesterolemia caused by mutations in certain genes. Screening these genes allows to identify disease causing mutations and use preventative measures for affected individuals[1]. However, exposing the underlying genetic structure for common complex diseases has remained a challenge.

In 2008, several leading scientists from the field of human genetics[2] stated that the ultimate goal – to fully describe the genetic architecture of common complex diseases and to translate the finding into clinical practice has remained unsolved despite of the efforts. Yet it was also said that the identification of the variants, genes and pathways which are involved in certain diseases provides routes to new therapies, advanced diagnosis and enchanted disease prevention[3]. Now, ten years later, various authors are starting to express hope that already known genetic information could be successfully implemented in prevention of at least some common complex diseases[4–6].

1.2. DNA

This chapter is written based on materials of National Human Genome Research Institute[7] and the book by prof. A. Heinaru[8].
DNA (deoxyribonucleic acid) is the hereditary material present mostly in the cell's nucleus. DNA could be seen as a code built with four letters (nitrogen bases): adenine (A), guanine (G), cytosine (C) and thymine (T). DNA bases pair up with each other (adenine pairs with thymine and cytosine pairs with guanine), forming units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. These three components (base, sugar and phosphate molecule) together are called a nucleotide. Nucleotides form a strand of DNA and two strands wound around each other form a spiral called a double helix. Human DNA consists of approximately 3 billion bases and that entire sequence is called a genome.

DNA molecule in the nucleus of each cell is packaged into thread-like structure called chromosomes. Normally, humans have 23 pairs of chromosomes, twenty-two pairs are called autosomes and the 23rd pair - the sex chromosomes- differs between males and females. Females have two copies of X chromosome and males have one X and one Y chromosome.

1.3. Types of genomic variation

Genetic variations can be present in many forms in human genome. In the broadest way, genetic variants can be divided into two different classes: single nucleotide variants and structural variants (see Figure 1). Single nucleotide variant is a variable position in human genome where the single nucleotide is substituted by another[9]. Most of the single nucleotide variants are di-allelic, meaning that only two alternative nucleotides (two alleles) can be detected in the specific position. The allele with smaller frequency in the population is referred to as a minor allele. Single nucleotide variants are classified to either as common or rare, depending on the frequency of minor allele (MAF) in the human population. Common single nucleotide variants have minor allele frequency of at least 1% (this is somewhat an arbitrary cut-off) in the population and they are often referred as single nucleotide polymorphisms (SNPs)[9–11]. Structural variants include insertions and deletions (indels), tandem repeats, copy number variations and other chromosomal rearrangements[9, 12].

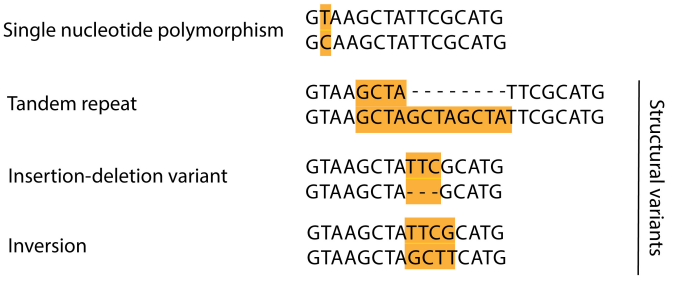

Figure 1. Different classes of human genetic variants. Insertion-deletion variant is occurring when a sequence of base pairs is present on some genome and missing in others. Inversion variant is present when an order of base pair sequence is reversed in a chromosome. Tandem repeat is a variant where a short sequence of base pairs is repeated. Adapted from[9].

According to the 1000 Genomes project, more than 88 million genetic variants were identified in the humans, approximately 84 million of them single nucleotide variants. Therefore, single nucleotide variants are the most frequent type of variation in the human genome. Most of them are rare, only 20 million are present with MAF >0.5% and 8 million with MAF 5% or more[13]. Due to abundance of single nucleotide variants and the fact that they are widespread across the entire genome makes them useful variables to study genomic alterations[10].

Single nucleotide variants are not always universal among major populations. Some are present in only one major population such as Europeans, but many of them are shared across all populations. Individuals from African ancestry population are currently believed to be most genetically diverse[13, 14].

The location of the single nucleotide variant might have several consequences at the phenotypic level. If the single nucleotide variant is in the coding area of the gene, it might alter the function of the encoded protein. However, most of the single nucleotide variants are located in the non-coding regions[15]. Recent research has been focusing on SNPs appearing in the non-coding, but regulatory regions (regions that do not code for proteins, but control the expression of coding regions) in hope for better understanding of the mechanism underlying complex traits[16, 17].

1.4. Haplotype and linkage disequilibrium

First, both terms - haplotype and linkage disequilibrium - have several definitions and only some of them are described here. Haplotype is a set of alleles in the same chromosome inherited together from a single parent[8]. They tend to be highly conserved areas which remain the same during many generations of reproduction[18]. When alleles are non-randomly inherited, then they are said to be in linkage disequilibrium (LD)[19]. The levels of LD tend to be higher when alleles are physically close to each other in the chromosome. There are several different measures to characterise LD, one of them being square of Pearson correlation.

When a set of SNPs are in high LD, then not all of them are usually needed to define haplotypes. The subset of SNPs representing a larger set of SNPs due to LD are called tag SNPs[20]. There are numerous methods to select tag SNPs, one of them is maximising the minimum correlation between non-tag and tag SNPs[20].

1.5. Genotyping, whole genome sequencing and imputation

A comprehensive method to analyse the genome is whole genome sequencing (WGS), which provides base-by-base view of the genome. This method (currently considered a "golden standard" for genetic testing[21]) allows capturing more complicated structural variants and *de novo* mutations which cannot be detected with genotyping array without beforehand knowledge of their existence and location.

Whole genome genotyping is a laboratory based approach, were SNP arrays allow the identification of hundreds of thousands of single nucleotide polymorphisms located all over the genome. There are several approaches how to genotype[15, 22]. The two major SNP genotyping array providers currently are Affymetrix[23] and Illumina[24].

The huge number of SNPs creates a problem while genotyping: it is not cost effective to genotype all of them[12]. Therefore, only a subset of tag SNPs are selected for genotyping and the ungenotyped SNPs in the genome are later imputed - their genotypes are predicted based on LD between tagged and untagged SNPs.

The idea of imputation is to predict the unobserved SNPs in a genotyped sample using a reference panel (usually whole-genome sequencing data from a similar population to the study sample). It requires that there is a sufficient overlap of genotyped SNPs in both study sample and the reference panel. Both reference data and genotype data must be phased (two independent haplotypes etc.)

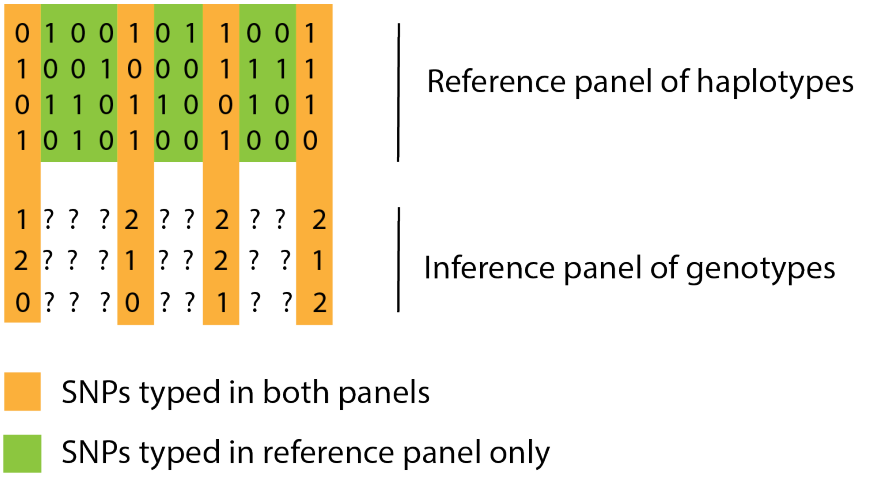

Figure 2. Explanatory drawing of imputation. Reference panel is usually constructed from a whole genome sequencing data with all positions known. Inference panel results from genotyping, where only tag SNPs are known, and the rest of them needs to be imputed. Adapted from[25].

Genotyping platforms often provide genotypic information, but no information about haplotypes. Even though haplotypes can be determined through molecular methods, it is often not done due to its cost and time consumption[26]. So haplotypes are determined from genotypic data using statistical methods. After constructing haplotypes and estimating their frequencies, imputation can be done, often based on hidden Markov models[25]. As a result of imputation, posterior probability of each possible genotype of a missing SNP for each individual is estimated, given the observed data.

A reference panel can be population specific (for example Estonian whole genome sequencing data) or mixture of different populations, such as publicly available 1000 Genome Project[13] or international HapMap resources[27]. The choice of reference panel is important as differences in LD patterns between reference study and imputed study reduces the accuracy of imputation[28]. Imputation is necessary to boost the power of genome-wide association studies and allow meta-analysing summary results from cohorts genotyped with different genotyping platforms[2, 28].

There are several different software solutions available for imputation, Estonian Biobank has been using IMPUTE2[25] and BEAGLE[29]. Both of those tools also provide measures to estimate imputation accuracy (as imputation is

done with uncertainty), which allows to filtering out and excluding low quality SNPs from the analysis. As most of my research uses imputed data filtered for imputation quality, the imputation quality (often called INFO score) is defined followingly[28].

It is assumed that for each SNP, three possible genotypes exist (for the sake of this example, lets say aa, Aa and AA). One allele is chosen to be counted ("coded allele", in this example allele "A") and the genotypes are presented as allele "A" counts, either 0, 1 or 2. It is usually assumed, that SNPs follow Hardy-Weinberg equilibrium, which states that in the absence of disturbing events the allele and genotype frequencies in a population will remain unchanged over time[30]. This means, that for one SNP with two possible alleles "A" with frequency p_A and "a" with frequency p_a, the genotype "aa" frequency can be calculated as p_a^2, genotype *Aa* frequency as $2p_ap_A$ and genotype "AA" frequency as p_A^2.

Let the genotype of the *i*th individual ($i = 1,...,N$) at the *j*th SNP be denoted as $G_{ij} \in \{0,1,2\}$. Let the set of haplotypes be denoted with H and the set of genotyped tag SNPs with G. Let the $p_{ijk} = P(G_{ij} = k|G,H)$ be the probability (obtained from imputation) of *i*th individual having the *k*th genotype for *j*th SNP. It is assumed that $\sum_{k=0}^{2} p_{ijk} = 1$. The expected coded allele dosage at the *i*th individual for *j*th SNP is defined as $e_{ij} = p_{ij1} + 2p_{ij2}$. We also define the expected squared allele dosage f_{ij} as $f_{ij} = p_{ij1} + 4p_{ij2}$. The unknown population coded allele frequency for the *j*th SNP is denoted with θ_j and estimated as $\hat{\theta}_j = \frac{\sum_{i=1}^{N} e_{ij}}{2N}$. The imputation INFO measure in IMPUTE for the *j*th SNP is based on a ratio of sample mean variance of the imputed genotypes and expected variance of *j*th SNP with frequency $\hat{\theta}_j$ under Hardy-Weinberg principle. It is estimated as

$$INFO_j = \begin{cases} 1 - \frac{\sum_{i=1}^{N}(f_{ij} - e_{ij}^2)}{2N\hat{\theta}_j(1-\hat{\theta}_j)} & \text{when } \hat{\theta}_j \in (0,1); \\ 1 & \text{when } \hat{\theta}_j \in \{0,1\}. \end{cases} \tag{1.1}$$

Common thresholds for INFO score vary from 0.4 for genome-wide association studies[31] to 0.7-0.8 for polygenic risk score construction studies[32].

The final choice of which array to use depends on many factors such as the research objectives, cost, array delivery schedules and available capacity of genotyping[2]. But even though the cost of sequencing has decreased dramatically over the past few years (being now 1000-2000 US dollars per genome[14]), majority of the genome-wide association studies for common complex traits/diseases (such as obesity, type 2 diabetes, breast cancer or coronary artery disease, etc) are still based on genotyped data together with imputation[4, 33–36].

1.6. Genome-wide association study

The idea of a genome-wide association study (GWAS) is to look for associations between a phenotype variable Y and each of the available SNPs across the entire genome. In the classical GWAS, it is assumed that individuals are unrelated. Due to the fact that the number of individuals n is often smaller than the number of SNPs, then each SNP is separately modelled to test for association in GWAS. Suppose there are m SNPs $X_1, \ldots, X_m$ we want to analyse. Let there be l covariates denoted as Z_j, $j = 1, ..., l$. The model for the ith SNP can be written in a form of generalized linear model with a link function $g(\cdot)$

$$g(E(Y|X_i, Z_1, ..., Z_l)) = \beta_0 + \beta_i X_i + \sum_{j=1}^{l} \gamma_j Z_j \tag{1.2}$$

Depending on how the SNP is coded, different genetic models can be investigated[37]. Using the arbitrary alleles "A" and "a" for a SNP like in the previous paragraph, genetic models can be additive - SNP genotype coded as "aa":0,"Aa":1,"AA":2; recessive - genotype coded as "AA":0,"Aa":0,"aa":1 or dominant - genotype coded as "aa":0,"Aa":1,"AA":1. Often additive models are applied in complex disease's studies.

If Y is a continuous random variable and linear relationship between Y and each SNP is assumed to hold, then identity link is appropriate. If Y is binary, then logit link function defined as

$$\text{logit}(p) = ln\left(\frac{p}{1-p}\right)$$

can be applied. Often - but not always - the covariates Z_j which are accounted for while modelling the SNP-phenotype association are age at recruitment, sex and a number of principal components to account for possible population stratification. The ith SNP is said to be genome-wide significant if the p-value from testing the null hypothesis that $\beta_i = 0$ is $\leq 5 \cdot 10^{-8}$ as it has been estimated that there are approximately 1 million independent common SNPs in the genome[38] and Bonferroni corrected significance threshold is $5 \cdot 10^{-8}$.

Several potential biases can occur in GWAS studies. There might be technical issues like different genotyping platforms for cases and controls[39]. Another major problem is the failure to properly account for population stratification or cryptic relatedness[2, 40]. Latter problems can be addressed with mixed linear models, however, their usage in practice until recent has been limited mainly due to computational issues[40, 41].

1.7. Meta-analysis of GWAS

Meta-analysis attempts to combine the analysis results from several individual studies to aggregate available information[42] and to provide pooled estimates.

It tackles the common concern of genetic data privacy, as no sharing of individual level data is required. It also avoids combining phenotype and genotype data together from several studies, which can be a harrowing job due to different disease coding systems and data formats. And finally, it allows for study-specific covariate adjustment[43].

The most common approach to perform meta-analysis with studies from similar ethnic background is fixed effect (FE) approach. FE approach assumes that there is no heterogeneity with respect to the true effect size of a SNP across all included studies[44].

To estimate the common underlying effect size by combining multiple observed effect sizes $\theta_1, \ldots, \theta_n$ together from n independent studies, weighted mean approach has been suggested:

$$\overline{\theta} = \frac{\sum_{i=1}^{n} \theta_i w_i}{\sum_{i=1}^{n} w_i}$$

Under the assumption, that θ_i are asymptotically normally distributed with the same mean, the maximum likelihood estimator will be obtained by choosing $w_i = 1/(Var(\theta_i))$[45]. This resulting estimator is called the inverse variance-weighted average effect size estimator. To test the null hypothesis that the common underlying effect is zero, z-statistic $Z_{IVW} = \dfrac{\overline{\theta}}{\sqrt{Var(\overline{\theta})}}$ can be calculated and p-value obtained as $p = 2\Phi(- \mid Z_{IVW} \mid)$.

In GWAS setting, θ_i are effect estimates such as linear regressions coefficients or log odds ratios denoted as $\hat{\beta}_1, ..., \hat{\beta}_n$ for a SNP from n independent studies and let $\hat{\sigma_1^2}, ..., \hat{\sigma_n^2}$ be their estimated variances, respectively. For simplicity, the estimated variances are often treated as known true variances[45, 46].

The inverse-variance-weighted (IVW) average effect size estimate for the common underlying effect then becomes

$$\hat{\beta}_{IVW} = \frac{\sum_{i=1}^{n} \hat{\beta}_i / \sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}$$

with variance

$$Var(\hat{\beta}_{IVW}) = \frac{1}{\sum_{i=1}^{n} 1/\sigma_i^2}$$

Normally, meta-analysis is done for millions of SNPs. The output is one huge file containing (but not limited to) information about SNP ID, its position in the chromosome, chromosome, reference allele, alternative allele, alternative allele frequency, sample size, inverse-variance weighted effect size estimator for an alternative allele and its variance and p-value from testing the null hypothesis that the effect size is zero. This information is essential of developing genetic risk scores.

Most of the GWA studies have been focusing on individuals with European ancestry. But in general, disease risk variants detected by GWASs seem to be shared across diverse populations[47, 48]. Transethnic meta-analyses are becoming increasingly popular to find population specific SNPs, increase the sample size of the studies as well as replicate previously found associations. It has been observed for type 2 diabetes for instance, that there is quite large directional consistency for allelic effects among different populations[49]. However, several challenges exist for transethnic studies. For instance, not all SNPs are polymorphic in all populations. Also, LD structures between SNPs as well as minor allele frequencies of SNPs might vary across populations[50]. It is also possible, that interactions between SNPs and environmental factors with different exposure levels between ethnic groups exist[51].

Two most popular programs to perform meta-analysis with studies from similar ethnic group are GWAMA[52] and METAL[43]. For trans-ethnic meta-analysis, various methods have been considered (for example, random effects meta-analysis, MANTRA or MR-MEGA)[47, 53].

2. GENETIC RISK SCORES

Over the years, genome-wide association studies have identified many SNPs associated with common complex diseases, however, most of them with relatively small effect on disease, deeming most of them alone worthless in clinical risk prediction. While searching for ways to make use of knowledge provided by GWAS, an idea of combining the effects of many SNPs into one variable called genetic risk score was presented. Genetic risk scores are nowadays one of the most researched topics in statistical/medical genetics field as they play an important role in explaining the genetic component of diseases' liabilities. In this paragraph, both the importance of genetic risk scores as well as different methods of composing genetic risk scores are introduced.

2.1. Common complex disease

By common complex disease we consider a frequently occurring chronic disease that has multifactorial aetiology, where both genetic susceptibility and environmental risk factors (including, but not limited to lifestyle and other health conditions) contribute[54]. Typical examples include coronary artery disease, type 2 diabetes, cancers, etc. Common complex diseases can affect the quality of life and as they also tend to have serious complications, it is desirable both for individual and health care system to prevent or postpone the onset of a disease as efficiently as possible. Estimating susceptibility of an individual to disease is a vital step in a clinical decision-making, especially as early disease detection and intervention is crucial to improve human health. Currently, most of the clinically used risk prediction tools are based on demographic and lifestyle variables such as age, sex, ethnicity, body mass index, alcohol and tobacco consumptions[55]. Clinical biomarkers and family history are also often incorporated[56, 57]. Even though one could argue that genetic predisposition is the earliest measurable component contributing to common complex disease, it is often missing from prediction tools. There are several reasons for that. First, it was studied that adding GRS into risk prediction tools already incorporating individual's family history and other risk factors would not improve the model[58], and therefore one could assume that family history at least partially already accounts for individual's genetic predisposition. Now it has been shown for several diseases that accounting for both individual's family history (if known) and genetic predisposition via genetic risk score results in the highest predictive ability[32, 59], proving them both as useful predictors. Second, the estimation of genetic risk was rather limited in the beginning and therefore showed little incremental value for clinical prediction tools[60, 61]. With more sophisticated statistical methods to compute genetic risk scores, its use in screening and prevention strategies has been encouraged[55, 62–64].

2.2. Heritability

Lets say there is a phenotype P which can be modelled in a simple case as the sum of environmental (E) and genetic (G) effects: $P = G + E$. Heritability is defined as the ratio of the genetic variation to the phenotypic variation, ie $H^2 = \frac{Var(G)}{Var(P)}$. Heritability estimates depend on several aspects, including the population they were derived in and the disease they were derived for[65]. Heritability is often estimated from twin studies with pairs of monozygotic twins (MZ) and dizygotic twins (DZ), where r_{MZ} and r_{DZ} are the correlations in monozygotic and in dizygotic twins for the same phenotype. Heritability is estimated as $H^2 = 2 * (r_{MZ} - r_{DZ})$[66]. Heritability estimates are useful because they help to determine the potential discriminative ability of predictors based on genetic variants. Assuming that the heritability estimates reflect the true parameter, phenotypic variance explained by linear predictor based on SNPs cannot be higher than heritability and the upper limit can be achieved only if the true causal SNPs together with their true effect sizes are known[67]. It has been estimated that the heritability of breast cancer ranges from 20% to 30%[68] and from 26%-69% for type 2 diabetes[69, 70], making both diseases eligible for genetic risk score research. In addition, for type 2 diabetes, it has been estimated earlier, that overall 550 independent SNPs are expected to be associated with T2D susceptibility[33], but only 140 have been found up to date[71]. Similar "expected true number of SNPs" calculation could not be found for breast cancer in the literature, but currently, 182 genome-wide significant SNPs have been found[72].

2.3. Genetic risk score

A **Genetic risk score (GRS)** for *i*th individual is defined as a weighted sum of coded allele dosages from k SNPs:

$$GRS_i = \sum_{j=1}^{k} w_j X_{ij}, \tag{2.1}$$

where X_{ij} denotes the dosage of coded alleles for jth SNP and ith individual and $w_j \in (-\infty, \infty)$ is the weight of the jth SNP. Dosage of ith individual's coded allele for jth SNP is $X_{ij} \in \{0,1,2\}$ if SNP is directly genotyped and $X_{ij} \in [0,2]$ if SNP is imputed and coded allele dosage is calculated based on genotype probabilities as described in paragraph 1.5.

It is unclear what is the optimal way of combining the effects of SNPs together to achieve the best possible predictive value. There are two main questions to be addressed in the process of computing the GRS: how to choose the set of SNPs to be included in the GRS and how to choose the weights w_j. In general, the following options are considered

1. Choice of SNPs:

(a) Use only uncorrelated genome-wide significant (a p-value threshold set to $p < 5 \cdot 10^{-8}$ in the GWAS or meta-analysis) SNPs
(b) Use less stringent p-value threshold to select SNPs
(c) Use all available (independent) SNPs

2. Choice of weights w_j:
   (a) All equal to 1, resulting in a sum of coded allele dosages
   (b) Estimated (logistic) regression parameters $\hat{w}_j$ from a discovery GWAS
   (c) Somehow modified $\hat{w}_j$ from discovery GWAS to take into account that SNPs might be correlated with each other

When estimated marginal effect sizes of SNPs from GWAS (choice of weights as mentioned in 2b) are chosen as weights, we call resulting GRSs **single-weighted genetic risk scores**.

2.4. Doubly-weighted polygenic risk score

2.4.1. Motivation

One of the main problems is that true effects of SNPs (w_i-s) are unknown. Despite of the fact that sample size of GWASs are large (reaching hundreds of thousands of individuals for many traits), the presence of sampling error is unavoidable. That also means that ordering of the markers and identification of the "top" markers in terms of Wald type statistic (or p-value) is a subject to uncertainty.

The power to detect association between SNP and phenotype in GWAS depends on sample size, MAF and effect size. SNPs with high minor allele frequency and with large effects are more likely to be picked up by GWAS[73]. But limiting the GRS to include only few number of highly significant SNPs would ignore the potential predictive ability of other SNPs. Even though including SNPs with weaker effect sizes in the GRS seems appealing, the accuracy of their effect size estimates tends to be low[67], causing the ordering of these SNPs according to their p-value to be unstable. Furthermore, when imposing a p-value threshold to select SNPs based on their p-value, one tends to systematically choose SNPs with effects overestimated by chance.

In other words, this introduces a problem called as *winners curse*, where by selecting SNPs according to their p-values we tend to choose more often SNPs which effect sizes are inflated compared to their corresponding true effects[74]. This is illustrated via simulation study in the Figure 3, where it can be seen that after setting p-value threshold to 0.00005, out of SNPs with true effect size one, only SNPs with estimated effect sizes larger than one would be included in GRS construction.

We propose a new method called **doubly-weighting**, where the SNP effect estimate from a GWAS study is additionally weighted by a coefficient that aims to correct at least partially for the possible overestimation of effect sizes due to

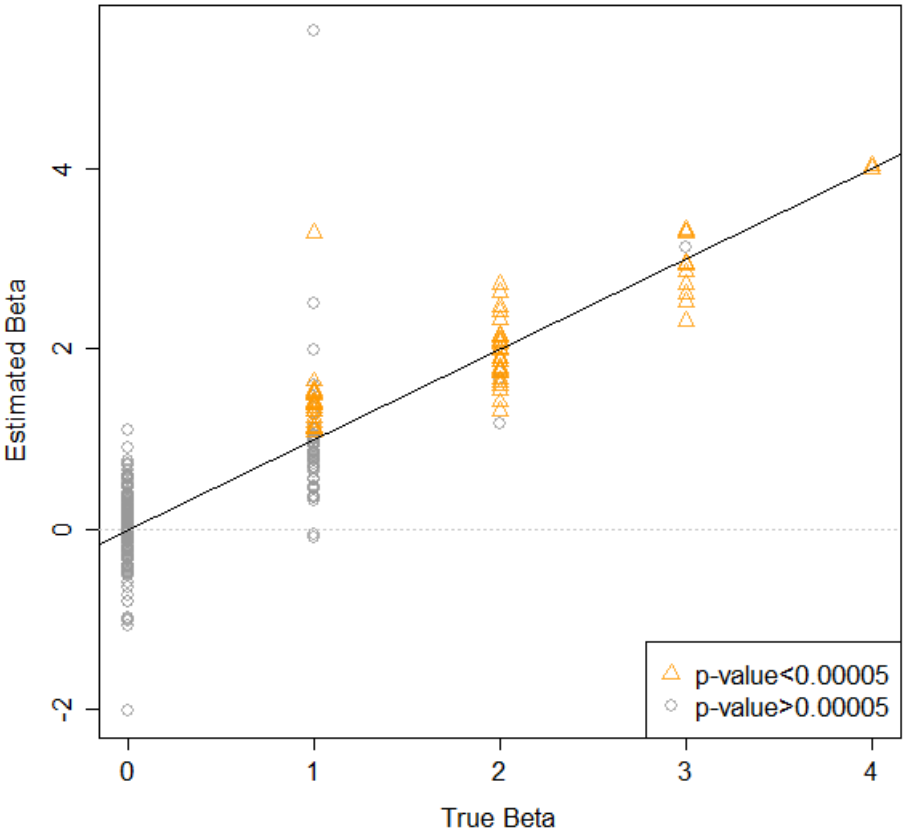

Figure 3. Comparison of true and estimated effect sizes for 300 SNPs in a simulated dataset of sample size 10000. The estimates with p-value below the threshold 0.00005 are shown as orange triangles.

winners curse. The idea of the method has been published in the Supplement of Ref I. We additionally show here with simulations that doubly-weighted GRSs tend to have a better predictive ability than the single-weighted GRSs.

2.4.2. Notation and methods

Let there be K independent SNPs, $X_1, \ldots, X_K$, tested for an association with a phenotype Y in a GWAS. For the simplicity, let us assume that SNPs have been standardized. If the phenotype Y is continuous, an additive linear regression model is assumed to hold for each SNP:

$$Y = \mu + \beta_i X_i + \varepsilon, \quad \text{with } \mathrm{E}(\varepsilon|X_i) = 0, \text{ for } i \in 1 \ldots K. \tag{2.2}$$

In case the the phenotype Y is binary, an additive logistic regression model is assumed to hold:

$$\text{logit}[\text{P}(Y=1|X_i)] = \mu + \beta_i X_i, \text{ for } i \in 1 \ldots K, \tag{2.3}$$

where the logit function is defined as $\text{logit}(p) = ln\left(\frac{p}{1-p}\right)$.

As mentioned, $\beta_1, \ldots, \beta_K$ are the true allelic effect sizes of K SNPs having fixed values. Let r_i be the rank of $|\beta_i|$ among $|\beta_1|, \ldots |\beta_K|$, so $r_i \in \{1, \ldots, K\}$. We define r_i as $r_i = \sum_j^K I(|\beta_i| - |\beta_j| \leqslant 0)$. The idea of doubly-weighting is to assume that there are at least k SNPs among K available SNPs having an additive genetic effect on the phenotype Y and we want to include the SNPs which have the k largest absolute effect. We do not explicitly assume that only k SNPs with largest absolute effects are associated with the phenotype. Still, we assume that the total effect of the SNPs not among the "top k" is not adding any significant contribution to the genetic risk potentially described by all K SNPs. We also assume the total effect of the k SNPs on the phenotype is additive and therefore define the polygenic risk score similarly to the formula 2.1

$$S_k = \sum_{i=1}^{K} f_{i(k)} w_i X_i, \tag{2.4}$$

where $f_{i(k)}$ is defined as

$$f_{i(k)} = f_{i(k)}(\beta_1, \ldots \beta_K) = \begin{cases} 1 & \text{if } r_i \leqslant k \\ 0 & \text{otherwise} \end{cases}$$

When Y is continuous, we assume that S_k is associated with the phenotype Y according to the following model:

$$Y = \mu + \alpha S_k + \varepsilon, \text{ with } \text{E}(\varepsilon|S_k) = 0 \tag{2.5}$$

When Y is binary, we assume

$$\text{logit}[\text{P}(Y=1|S_k)] = \mu + \alpha S_k, \text{ for } i \in 1 \ldots K, \tag{2.6}$$

For a continuous Y, $\alpha = 1$ in equation (2.5) if we select $w_i = \beta_i$ from equation (2.2) and the assumption, that X_i are independent, holds. For binary Y, due to non-collapsibility problem of odds ratios [75], selecting $w_i = \beta_i$ from (2.3) may not result in $\alpha = 1$.

When w_i-s and X_i-s are fixed and we have random variables $\hat{\beta}_1, \ldots \hat{\beta}_K$ instead of known effect sizes, then an estimate for S_k could be

$$\tilde{S}_k = \sum_{i=1}^{K} \hat{f}_{i(k)} w_i X_i \tag{2.7}$$

where $\hat{f}_{i(k)} = f_{i(k)}\left(\hat{\beta}_1, \ldots \hat{\beta}_K\right)$. The expected value of $\tilde{S}_k$ would be:

$$\mathrm{E}(\tilde{S}_k) = \sum_{i=1}^{K} E(\hat{f}_{i(k)}) w_i X_i \tag{2.8}$$

We propose an algorithm to estimate the $E(\hat{f}_{i(k)}) = P(\hat{f}_{i(k)} = 1)$ in 2.4.3 and the estimates will be denoted as $\pi_{i(k)}$. As w_i are also unknown in real life, we need to estimate them, and the simplest way would be to use $\hat{w}_i = \hat{\beta}_i$. Then the estimate becomes

$$\hat{S}_k = \sum_{i=1}^{K} \pi_{i(k)} \hat{\beta}_i X_i \tag{2.9}$$

2.4.3. Algorithm to identify optimal weights

In reality, it is unknown which SNPs among all available SNPs have the largest absolute effects for the trait, therefore we need to estimate $\mathrm{P}(\hat{f}_{i(k)} = 1)$ somehow.

We propose a following algorithm to obtain estimates $\pi_{i(k)}$ for given k and available set (from some GWA study) of $\hat{\beta}_1, \ldots \hat{\beta}_K$ and their corresponding standard errors $\hat{s}_1, \ldots \hat{s}_K$:

1. For each $i = 1, \ldots, K$, draw a random parameter value $\hat{\beta}_i^{(s)}$ from a normal distribution with mean $\hat{\beta}_i$ and standard deviation $\hat{s}_i$.
2. Order decreasingly the K estimated independent Wald-type statistics $|\hat{\beta}_i^{(s)}|/\hat{s}_i$, and according to the order, assign a rank $r_i^{(s)}$ for each $|\hat{\beta}_i^{(s)}|/\hat{s}_i$.
3. Repeat steps 1 and 2 M times, to obtain an empirical distribution of ranks $r_i^{(s)}$, $s = 1, \ldots, M$ for each i.
4. Estimate $\pi_{i(k)}$ for each i-th SNP is obtained as $\frac{1}{M}\sum I(r_i^{(s)} \leq k)$ (proportion of ranks assigned to the ith SNP that are not larger than k).

We could also choose any value for k, aiming to estimate the effect of k strongest SNPs with respect to their association with Y. One could try to vary k around the estimated true number of SNPs affecting the trait, if that estimate can be found from the literature.

2.5. MetaGRS

For some traits, there are several versions of GRSs already published, often based on different meta-analyses or computed using different methodologies. Alternative GRSs might be similar in regard to their predictive ability, however, they do not need to be highly correlated with each other. MetaGRS is a weighted average of several existing standardized genetic risk scores which should be based on different meta-analysis results. The idea was first proposed by M. Inouye and others[5]. The authors reason that all available genetic risk scores are imperfect

measures of true genetic risk due to many reasons, including incomplete coverage of genome, imputation uncertainty, limited variance explained by SNPs and errors in effect sizes of SNPs. They argue that it is desirable to improve precision of genetic risk estimation, as association between disease and risk factor (genetic risk in this case) measured with error can be attenuated. Combining several GRSs into one metaGRS could result in more precisely estimated genetic risk.

Mathematically, let $Z_{i1},...,Z_{ip}$ be p zero mean and unit-variance standardized genetic risk scores obtained with 2.1 for the ith individual and $\hat{\alpha}_1,...,\hat{\alpha}_p$ are effect size estimates for respective scores from the training set using model 2.5 or 2.6 depending on the type of trait Y. Pearson correlation ρ_{jk} is calculated between $Z_{\cdot j}$ and $Z_{\cdot k}$ scores in the training set. MetaGRS for the ith individual is defined as

$$MetaGRS_i = \frac{\sum_{j=1}^{p} \hat{\alpha}_j Z_{ij}}{\sqrt{\sum_{j=1}^{p} \hat{\alpha}_j^2 + 2\sum_{j=1}^{p}\sum_{k=j+1}^{p} \hat{\alpha}_j \hat{\alpha}_k \rho_{jk}}} \tag{2.10}$$

where the effect estimates of GRSs are treated as constants.

2.6. Simulation study I: Comparison of different genetic risk score's methods

2.6.1. Overview of simulation's workflow

To illustrate the benefit of doubly-weighting, the following simulation study is done.

Let the Y be a phenotype vector with $n \times 1$ dimensions. The genotype matrix is denoted as X with dimensions $n \times p$, where n is the number of unrelated individuals and p is the total number of SNPs. SNPs are coded additively, i.e they can take values 0, 1 or 2. We assume that SNPs are independent (i.e, not in LD) and that SNP values are drawn from binomial distribution, $X_{ij} \sim Bin(2, f_j)$, with f_j being the minor allele frequency for the jth SNP. Let Z be the standardized genotype matrix $Z_{ij} = (X_{ij} - 2f_j)/\sqrt{2f_j(1-f_j)}$, so that $E(Z_{ij}) = 0$ and $D(Z_{ij}) = 1$, where the X_{ij} is the number of minor alleles for the ith individuals and jth SNP and f_j is the frequency of the minor allele of jth SNP.

We define Y_i so that

$$Y_i = \sum_{j=1}^{m} \beta_j Z_{ij} + \varepsilon_i \tag{2.11}$$

so m SNPs are associated with Y and $p-m$ SNPs are not. We generate the effects for m SNPs from $\beta_j \sim N\left(0, \frac{\sigma_g^2}{m}\right)$, where σ_g^2 is the variance of total additive genetic effects. The noise is generated from $\varepsilon_i \sim N\left(0, 1-\sigma_g^2\right)$. The expectation of trait given genotypes is $\mathrm{E}(Y_i|Z_{i1},...,Z_{im}) = 0$ and the variance of the trait can

be partitioned to $Var(Y_i|Z_{i1},...,Z_{im}) = \sum_{j=1}^{m} \frac{\sigma_g^2}{m} Z_{ij}^2 + 1 - \sigma_g^2$. According to the total law of variance, $Var(Y_i) = Var(E(Y_i|Z_{i1},...,Z_{im})) + E(Var(Y_i|Z_{i1},...,Z_{im})) = 0 + \sum_{j=1}^{m} \frac{\sigma_g^2}{m} \cdot 1 + 1 - \sigma_g^2 = 1$.

So in this kind of setting, trait Y is normally distributed with mean zero and variance 1. Moreover, as the heritability is the ratio of variance explained by genetic factors divided by the total variance of the phenotype, then in this kind of simulation setting, the heritability $h^2 = \sigma_g^2$.

The simulations followed this pipeline:

1. Heritability $h^2 = \sigma_g^2$ was set to 0.2 for the first set of analyses and 0.5 for the second set of analyses.
2. Number of truly associated SNPs m was set to be $100, 450, 1000$ or 10000 for both heritability values.

This means that there were eight combinations of h^2 and m to generate Y and perform a GWA study. To do that, we generated dataset for each h^2 and m combination followingly:

1. Minor allele frequencies f_j were generated for $j = 1,...20000$ SNPs from Uniform $(0.01, 0.5)$ distribution.
2. The jth SNP for ith individual was generated from $X_{ij} \sim Bin(2, f_j)$ for $i = 1,...,15000$. Matrix X was later standardized.
3. Effects of m independent SNPs were drawn from $N\left(0, \frac{\sigma_g^2}{m}\right)$ and for the rest of the $20000 - m$ independent SNPs the effects were set to 0.
4. The outcome variable Y was generated as defined in 2.11, with error term ε_i generated from: $\varepsilon_i \sim N\left(0, (1 - \sigma_g^2)\right)$.

Finally, GWAS analyses were run using a linear model as defined in 1.2 without any additional covariates besides the single SNP to estimate the effect of each standardized SNP on Y under fixed h^2 and m values. For each SNP, regression parameter $\hat{\beta}_j$, its standard error and p-value were retrieved. In total, after the first part of simulation, 8 sets of GWAS summary statistics were obtained.

The purpose of the second part of the simulations was to construct GRSs based on GWAS results and estimate the association between different versions of GRSs and phenotype in independent datasets. For each combination of h^2 and m, 10 test datasets were generated followingly:

- The same f_j were used as in GWA study and SNPs were generated from $Bin(2, f_j)$, $j = 1,...,20000$ like before for 3000 unrelated individuals and then standardized.
- The outcome variable Y was generated as before in 2.11.

In each dataset, four different types of GRS were constructed by varying the value of k:

1. The true genetic risk score GRS_1, obtained according to 2.4 for a prespecified k using true effects sizes as weights (the beforehand generated β_i-s) and true ranks.
2. The genetic risk score GRS_2 calculated using the formula in 2.4 for a prespecified k, ie, true ranks r_i were used, but as weights, we used estimated regression coefficients $\hat{\beta}_i$-s from GWAS.
3. The single-weighted genetic risk score GRS_{SW} as defined in 2.1, including only k SNPs with smallest p-values from the GWAS and $\hat{w}_i$-s are taken from the same GWAS as p-values.
4. The doubly-weighted genetic risk score GRS_{DW} defined as 2.9 for a prespecified k, with estimated regression coefficients $\hat{\beta}_i$-s taken from GWAS and probabilities $\pi_{i(k)}$ estimated as previously proposed in 2.4.3.

For each score, its estimated effect size and standard error were calculated by regressing the outcome variable Y in the test dataset on the risk score. Also, the coefficient of determination - R^2- was obtained from each fitted model.

Finally, the entire simulation pipeline was repeated while increasing the GWAS sample size from 15000 to 30000 to study how GWAS sample size affects the predictive ability of GRSs.

2.6.2. Results of simulation

Simulations were done separately for two different heritability values: $h^2 = 0.5$ and $h^2 = 0.2$ and by varying the number of SNPs which actually have an effect on phenotype (denoted as m). GWASs for each h^2 and m combination were run twice, once with sample size 15000 individuals and secondly, with 30000 individuals. Number of associated SNPs m took values 100, 450, 1000 or 10000 out of 20000 simulated SNPs. Simulation results for different type of GRSs presented below are obtained by averaging the results of modelling Y and GRSs association over 10 test dataset simulations.

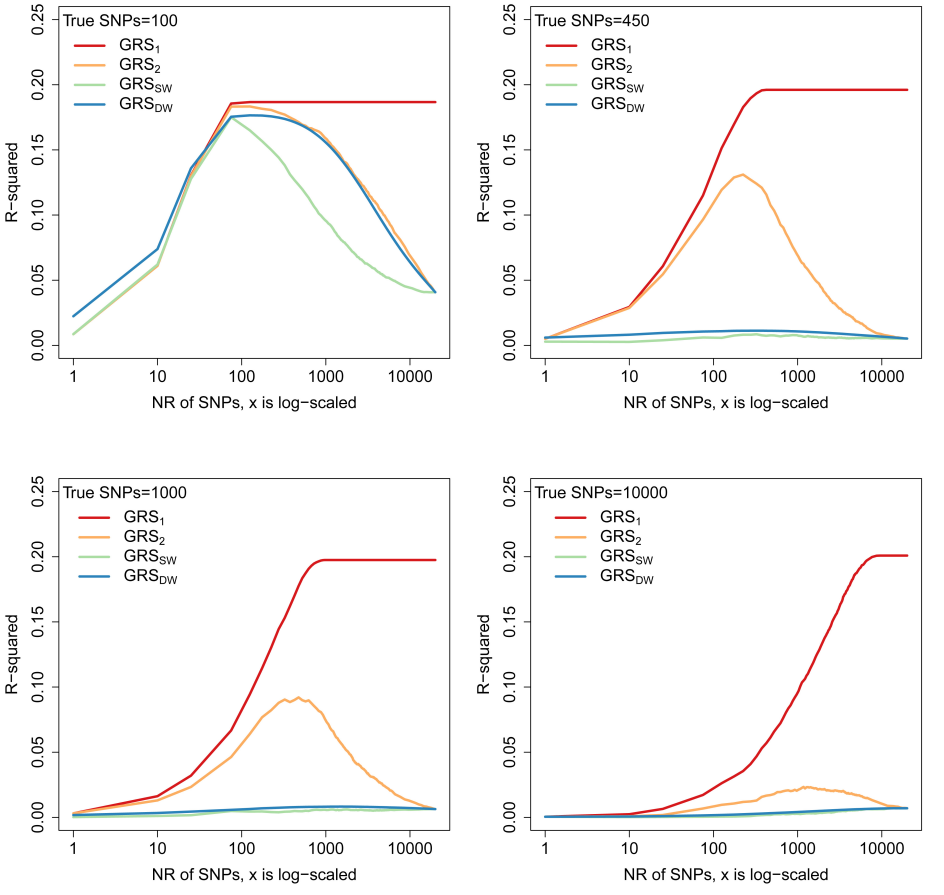

Figure 4. R^2 from the simple regression model with GRS as a single covariate in four different scenarios - the number of true associations with the trait is either 100, 450, 1000 or 10000 SNPs. In total, 20000 SNPs for 15000 individuals are generated for GWAS. Results are shown for four types of GRSs. On x axis, the meaning of the logarithmic value of number of SNPs (value of k) depends on the type of GRS. The heritability of the trait is set to be 0.2.

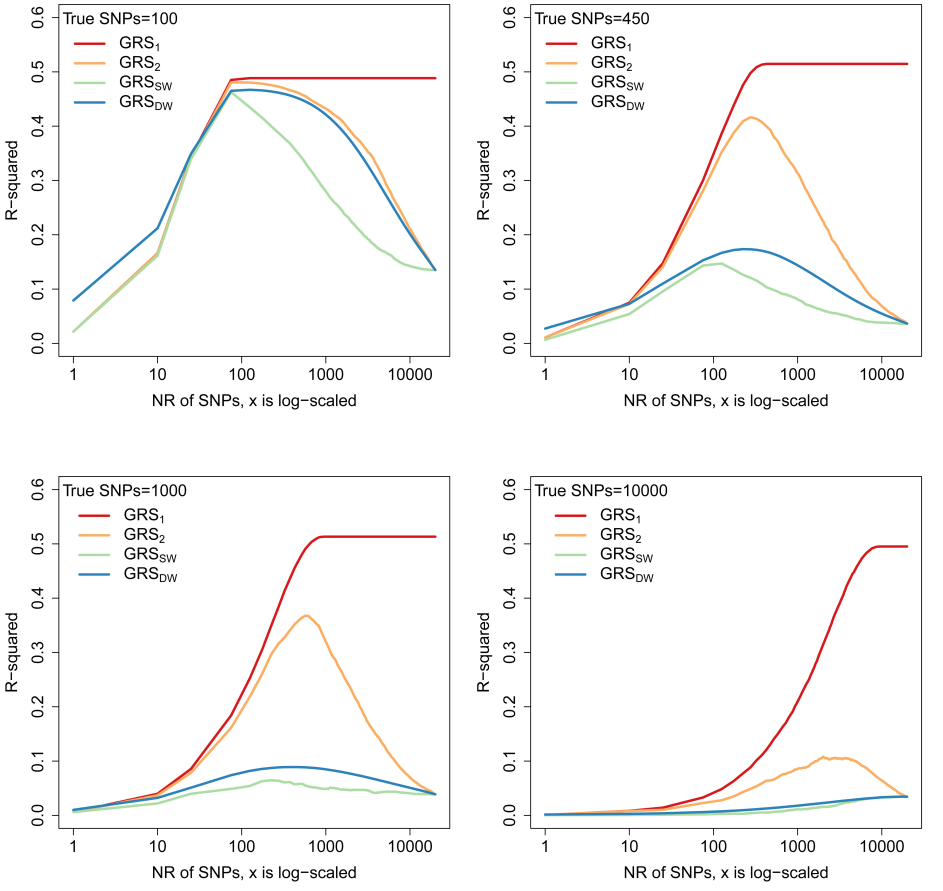

Figure 5. R^2 from the simple regression model with GRS as a single covariate in four different scenarios - the number of SNPs actually associated with the trait is either 100, 450, 1000 or 10000. In total, 20000 SNPs for 15000 individuals are generated for GWAS. The heritability of the trait is set to be 0.5.

First, we focus on using GWAS results to compose GRSs, when GWAS sample size was 15000. It can be seen from Figures 4 and 5, that doubly-weighted genetic risk scores (GRS_{DW}-s) tend to explain larger proportion of variance than single-weighted GRSs (GRS_{SW}-s). However, the benefit of GRS_{DW}-s depends on the number of causal SNPs and heritability of the trait. When the truly associated number of SNPs is 100, then with both heritability values, GRS_{DW}-s perform better than GRS_{SW}-s, but slightly worse than GRS_2-s. For all other scenarios of number of causal SNPs (450, 1000 and 10000), the benefit of doubly-weighting is more clearly present with scenarios of heritability value being 0.5.

For the scenario of 100 causal SNPs and $h^2 = 0.5$, the median difference in R^2 values (i.e median of $R^2_{DW} - R^2_{SW}$, taken over all tested k values, DW- doubly-weighted GRS, SW- single-weighted GRS) is 0.126 (25% percentile = 0.1 and 75% percentile = 0.142) while comparing doubly-weighted and single-weighted GRSs. In the case of 450 causal SNPs, the median difference in R^2 values is 0.041 (25% percentile = 0.034 and 75% percentile = 0.056); 0.025 (25% percentile = 0.02 and 75% percentile = 0.030) while there are 1000 causal SNPs and 0.005 (25% percentile = 0.003 and 75% percentile = 0.007) while number of causal SNPs is 10000. Results are similar in case of $h^2 = 0.2$ meaning that the benefit of doubly-weighting compared to single-weighting in regards of improvement in R^2 value is largest with small number of causal SNPs and the benefit decreases when number of causal SNPs increases. However, the medians of differences in R^2 values are smaller, varying between 0.0008-0.044.

When increasing the number on individuals in GWAS from 15000 to 30000 for the heritability 0.5, then the differences in R^2 values diminish very slightly for the majority of causal SNP scenarios, indicating, that the larger sample size improves the accuracy of GWAS estimated weights and mildly decreases the benefit of doubly-weighting. For the scenario of 100 causal SNPs, the median difference in R^2 values is 0.107 (25% percentile = 0.086 and 75% percentile = 0.114) while comparing GRS_{DW}-s and GRS_{SW}-s. In the case of 450 causal SNPs, the median difference in R^2 values is 0.040 (25% percentile = 0.031 and 75% percentile = 0.052); 0.024 (25% percentile = 0.019 and 75% percentile = 0.028) while number of causal SNPs is 1000 and 0.006 (25% percentile = 0.004 and 75% percentile = 0.008) while number of causal SNPs is 10000. R^2 values as well as regression coefficients for some scores are given in Table 1.

Table 1. Results from phenotype GRS modelling (formula 2.5), averaged over 10 simulations for $h^2 = 0.5$. Alphas are the estimated regression coefficients for GRS's.

k SNPs	Single-weighted			Doubly-weighted			$R^2_{DW} - R^2_{SW}$
	$\hat{\alpha}$	$SE(\hat{\alpha})$	R^2_{SW}	$\hat{\alpha}$	$SE(\hat{\alpha})$	R^2_{DW}	
GWAS n=15000 and causal SNP count =100							
10	0.974	0.041	0.161	1.301	0.046	0.212	0.05
125	0.905	0.019	0.435	0.989	0.019	0.467	0.032
275	0.801	0.018	0.388	0.968	0.019	0.462	0.074
475	0.717	0.018	0.349	0.941	0.019	0.451	0.102
725	0.641	0.017	0.310	0.909	0.019	0.437	0.126
1025	0.575	0.017	0.278	0.872	0.019	0.420	0.142
2025	0.462	0.016	0.224	0.767	0.018	0.371	0.147
5000	0.345	0.014	0.169	0.570	0.017	0.277	0.108
10667	0.287	0.013	0.141	0.399	0.015	0.195	0.054
GWAS n=30000 and causal SNP count =100							
10	0.950	0.040	0.160	1.114	0.043	0.186	0.026
125	0.952	0.019	0.464	0.983	0.019	0.474	0.01
275	0.887	0.019	0.432	0.974	0.019	0.472	0.04
475	0.826	0.018	0.404	0.960	0.019	0.467	0.063
725	0.771	0.018	0.377	0.942	0.019	0.458	0.082
1025	0.718	0.018	0.350	0.920	0.019	0.448	0.098
2025	0.614	0.017	0.299	0.853	0.018	0.415	0.116
5000	0.492	0.016	0.240	0.705	0.018	0.343	0.104
10667	0.429	0.015	0.210	0.549	0.017	0.268	0.058
GWAS n=15000 and causal SNP count =450							
10	0.775	0.059	0.054	1.659	0.108	0.073	0.019
125	0.515	0.023	0.147	1.095	0.045	0.167	0.02
275	0.357	0.018	0.121	0.854	0.034	0.173	0.052
475	0.272	0.015	0.101	0.689	0.028	0.167	0.066
725	0.225	0.013	0.090	0.567	0.024	0.155	0.065
1025	0.189	0.012	0.080	0.475	0.021	0.143	0.063
2025	0.135	0.010	0.061	0.323	0.016	0.115	0.054
5000	0.092	0.008	0.045	0.183	0.012	0.078	0.033
10667	0.075	0.007	0.038	0.112	0.009	0.053	0.015
GWAS n=30000 and causal SNP count =450							
10	0.826	0.061	0.057	2.658	0.144	0.102	0.045
125	0.478	0.023	0.126	1.165	0.047	0.170	0.044
275	0.360	0.017	0.124	0.879	0.035	0.171	0.048
475	0.269	0.015	0.099	0.695	0.029	0.162	0.063
725	0.213	0.013	0.082	0.566	0.025	0.149	0.067
1025	0.175	0.012	0.068	0.469	0.022	0.136	0.068
2025	0.128	0.010	0.055	0.313	0.017	0.107	0.052
5000	0.087	0.008	0.041	0.173	0.011	0.070	0.030
10667	0.069	0.007	0.033	0.104	0.009	0.047	0.014
GWAS n=15000 and causal SNP count =1000							
10	0.556	0.069 eal	0.022	1.400	0.141	0.032	0.012
125	0.336	0.026	0.054	1.000	0.061	0.082	0.028
275	0.273	0.019	0.064	0.744	0.044	0.088	0.025
475	0.215	0.016	0.058	0.586	0.034	0.089	0.031
725	0.175	0.014	0.051	0.480	0.028	0.087	0.036
1025	0.153	0.012	0.049	0.404	0.024	0.085	0.035
2025	0.120	0.010	0.046	0.281	0.018	0.076	0.030
5000	0.092	0.008	0.044	0.169	0.012	0.062	0.018
10667	0.078	0.007	0.040	0.110	0.009	0.050	0.01
GWAS n=30000 and causal SNP count =1000							
10	0.514	0.067	0.020	1.965	0.173	0.041	0.021
125	0.326	0.025	0.054	0.926	0.059	0.076	0.022
275	0.249	0.019	0.055	0.698	0.043	0.082	0.027
475	0.207	0.016	0.055	0.557	0.034	0.083	0.028
725	0.168	0.014	0.049	0.459	0.028	0.082	0.033
1025	0.151	0.012	0.049	0.387	0.024	0.079	0.030
2025	0.114	0.010	0.044	0.268	0.018	0.071	0.027
5000	0.085	0.008	0.039	0.158	0.012	0.057	0.019
10667	0.070	0.007	0.035	0.102	0.009	0.045	0.01

It was also investigated how doubly-weighting affects the winner's curse problem. Effect estimates $\hat{\beta}_i$-s were taken from GWASs with sample size 15000. Then, for each k value, the difference between true and estimated effects $(\hat{\beta}_i - \beta_i)$ was characterised. This means that for single-weights, $\hat{\beta}$ from GWAS for k SNPs with smallest p-values were taken and and their difference with true effect sizes calculated (i.e $(\hat{\beta}_i - \beta_i)$ for k SNPs). For doubly-weighted weights, the probabilities $\pi_{i(k)}$-s were estimated and $\hat{\beta}_i$-s modified by them. Then the difference between true effects and estimated effects were calculated (i.e $(\hat{\beta}_i \cdot \pi_i - \beta_i)$ for K SNPs). Results for some k values are in Table 2. "SD" stands for standard deviation.

Table 2. Characterising the differences between estimated and true effects for different GRS methods

k	Doubly-weighted		Single-weighted	
	$\frac{1}{K}\sum_{i=1}^{K}(\hat{\beta}_i \cdot \pi_{i(k)} - \beta_i)$	$\mathrm{SD}(\hat{\beta} \cdot \pi_{(k)} - \beta)$	$\frac{1}{k}\sum_{i=1}^{k}(\hat{\beta}_i - \beta_i)$	$\mathrm{SD}(\hat{\beta} - \beta)$
$h^2 = 0.2$ and causal SNP count =100				
10	$-3.26 \cdot 10^{-5}$	$2.35 \cdot 10^{-3}$	$1.89 \cdot 10^{-3}$	$9.02 \cdot 10^{-3}$
75	$-1.73 \cdot 10^{-6}$	$1.09 \cdot 10^{-3}$	$2.59 \cdot 10^{-3}$	$1.77 \cdot 10^{-2}$
125	$-1.00 \cdot 10^{-6}$	$1.09 \cdot 10^{-3}$	$4.23 \cdot 10^{-3}$	$2.03 \cdot 10^{-2}$
275	$4.24 \cdot 10^{-7}$	$1.22 \cdot 10^{-3}$	$1.87 \cdot 10^{-3}$	$2.10 \cdot 10^{-2}$
475	$1.85 \cdot 10^{-6}$	$1.44 \cdot 10^{-3}$	$5.75 \cdot 10^{-5}$	$2.04 \cdot 10^{-2}$
725	$3.36 \cdot 10^{-6}$	$1.71 \cdot 10^{-3}$	$-2.95 \cdot 10^{-4}$	$1.95 \cdot 10^{-2}$
1025	$4.91 \cdot 10^{-6}$	$2.00 \cdot 10^{-3}$	$-1.89 \cdot 10^{-4}$	$1.87 \cdot 10^{-2}$
2025	$8.90 \cdot 10^{-6}$	$2.78 \cdot 10^{-3}$	$7.98 \cdot 10^{-5}$	$1.68 \cdot 10^{-2}$
5000	$1.87 \cdot 10^{-5}$	$4.34 \cdot 10^{-3}$	$1.77 \cdot 10^{-4}$	$1.38 \cdot 10^{-2}$
10667	$3.19 \cdot 10^{-5}$	$6.13 \cdot 10^{-3}$	$9.17 \cdot 10^{-5}$	$1.80 \cdot 10^{-2}$
$h^2 = 0.2$ and causal SNP count =1000				
10	$6.18 \cdot 10^{-6}$	$3.27 \cdot 10^{-3}$	$2.84 \cdot 10^{-2}$	$6.30 \cdot 10^{-2}$
75	$1.43 \cdot 10^{-5}$	$3.32 \cdot 10^{-3}$	$3.53 \cdot 10^{-3}$	$5.68 \cdot 10^{-2}$
125	$1.85 \cdot 10^{-5}$	$3.39 \cdot 10^{-3}$	$3.68 \cdot 10^{-3}$	$5.38 \cdot 10^{-2}$
275	$2.64 \cdot 10^{-5}$	$3.65 \cdot 10^{-3}$	$4.65 \cdot 10^{-3}$	$4.95 \cdot 10^{-2}$
475	$3.39 \cdot 10^{-5}$	$4.03 \cdot 10^{-3}$	$3.47 \cdot 10^{-3}$	$4.65 \cdot 10^{-2}$
725	$4.02 \cdot 10^{-5}$	$4.50 \cdot 10^{-3}$	$1.10 \cdot 10^{-3}$	$4.40 \cdot 10^{-2}$
1025	$4.53 \cdot 10^{-5}$	$5.04 \cdot 10^{-3}$	$1.96 \cdot 10^{-3}$	$4.19 \cdot 10^{-2}$
2025	$5.59 \cdot 10^{-5}$	$6.59 \cdot 10^{-3}$	$1.05 \cdot 10^{-3}$	$3.74 \cdot 10^{-2}$
5000	$6.34 \cdot 10^{-5}$	$9.87 \cdot 10^{-3}$	$2.89 \cdot 10^{-4}$	$3.06 \cdot 10^{-2}$
10667	$6.21 \cdot 10^{-5}$	$1.38 \cdot 10^{-2}$	$1.10 \cdot 10^{-4}$	$2.40 \cdot 10^{-2}$

It can be seen from 2 that the winner's curse problem is reduced (but not completely corrected) while using doubly-weighting and the difference is more noticeable for smaller k values. Standard deviations of differences between estimated weights and true weights are also smaller for doubly-weighted weights. Results were similar for $h^2 = 0.5$ and are therefore not separately presented.

3. RESULTS AND DISCUSSION

3.1. General overview of datasets used in this thesis

During this PhD project, several datasets were used to develop genetic risk scores and investigate their features and associations with different variables. A broad overview of them is given in this paragraph and additional details are given in the specific sections of the corresponding references.

Estonian Biobank (EstBB or EGCUT) hosts a cohort of ∼52000 participants (34% are men and 66% women) aged 18 and older. Majority of the participants were recruited between 2002 and 2011. Cohort includes adults from all counties of Estonia and it accounts for approximately 5% of the Estonian adult population during the recruitment period[76]. Participants gave extensive information regarding their anthropometric, genealogical, lifestyle and educational characteristics as well as their medical history. A broad informed consent signed by participants allows the data to be used for various research purposes and it also enables follow-up of participants via linkage with national health-related databases and registries[77].

Genotyping has been done in several stages over time using different genotyping arrays. By the end of 2017, the genomes of most participants were either genotyped or sequenced, most of them with Illumina's Infinium Global Screening BeadChip. Analyses during this thesis were done using genetic data imputed with either 1000G or Estonian WGS data as a reference panel.

UK Biobank (UKBB) is a cohort study with ∼500 000 individuals aged 40-69 recruited during 2006-2010[78] in UK. Both deep phenotypic and genetic data has been collected, including biomarkers from blood and urine and images of brain and the body[79]. Participants are followed-up via linking to their health-related records. Majority of participants have reported to be of European origin[79]. Individuals are genotyped with two different genotyping arrays[79] and imputation used in this thesis had been done with Haplotype Reference Consortium (HRC) panel as reference. In this thesis, only women not included in the GWAS sample (∼46000 women) are used to assure independence of discovery and validation dataset.

The 1000 Genome project (1000G) includes in total 2504 individuals from 26 populations sampled from Africa, East-Asia, South-Asia, Europe and Americas[13]. All individuals were sequenced with whole-genome sequencing. The idea of the project was to provide characterization of the genetic variation in human genomes, both common and rare. The project also serves as a global reference for genotype imputation for many populations. However, phenotype information is not publicly available for this sample.

3.2. Polygenic risk scores for type 2 diabetes

3.2.1. Short description of materials and methods

The study planned to obtain the SNP effect sizes from meta-analysis for type 2 diabetes (T2D) by Morris *et al.*[33], but as subset of EstBB sample was included in this study, we first rerun the meta-analysis to exclude EstBB data as we intended to use all available genotypes in EstBB for GRSs development. The meta-analysis results without Estonian sample were then used in the following analysis. Independent set of SNPs ($r^2 \leq 0.05$) was obtained via clumping[80]. Clumping is a procedure that needs a GWAS summary statistics file and a reference file of WGS data or genotype dataset. Steps of clumping include taking a list of GWAS SNPs and their p-values and first sort them in an increasing order according to the p-value. Then it selects an index SNP with the smallest p-value from that list and calculates pairwise correlations between index SNP and all other SNPs in a list which are not physically further from each other than a predefined distance. All SNPs, which squared pairwise correlations with index SNP are larger than a user defined threshold, are excluded from the original GWAS list. Index SNP is kept. Next index SNP is then selected (second smallest p-value in the remaining GWAS list) and previous steps are repeated. In the end, clumping results in a list of uncorrelated (up to users definition) GWAS index SNPs. In addition, SNPs were also filtered by their imputation quality and minor allele frequency in our dataset, resulting in a set of 7502 SNPs for GRSs construction. More details about data management in [Ref I, Supplementary Materials].

A set of 10273 genotyped individuals (descriptive information in Ref I, Table 1) with 1181 prevalent T2D cases present and 386 incident T2D cases obtained via linking to National Health Insurance Fond and Causes of Death Registry was used in this article. The average follow-up time was 5.36 years. Baseline phenotype data including age, gender, body mass index (BMI), physical activity, etc as well as metabolic profiles for a subset of 6064 individuals were used in this analysis.

We constructed single-weighted GRSs (denoted as GRS_k) and doubly-weighted GRS (denoted as $dGRS_k$) by varying the number of k and compared them by modelling the prevalent status of T2D with logistic regression model adjusted for age, sex and genotyping platform and in the additional setting, also for BMI. Different version of GRSs were compared with Cox likelihood ratio test for non-nested models[81]. We investigated the effect of chosen GRS from the first step by modelling incident T2D, all-cause mortality and cardiovascular mortality using age as time scale while accounting for known classical risk factors of T2D among individuals aged 35-79.

Associations between classical risk factors and chosen GRS among diabetics and non-diabetics were investigated via linear regression analysis. Finally, incremental value of GRS was investigated with AUC and Harrell's c-statistics and via characterising properties of 5-year risk estimates for type 2 diabetes from models with and without GRS using reclassification indexes. Harrell's c- statistic aims

to estimate the probability of concordance between predicted and observed responses in survival setting[82]. Let $X_1, X_2...$ be the survival times of individuals in a given population. Let T be the time point after which all survival times are truncated, so individuals who do not develop an event by the time point T, have their survival time set to T. For a time point T, we will have two categories: events (who developed an outcome during follow-up) and non-events (who did not develop an outcome during follow-up). Let $Y_1, Y_2...$ be predicted probabilities of survival for the same individuals for any fixed time point. We take under consideration all pairs of subjects (i,j), assuming $i < j$, to avoid repetition. A pair is said to be concordant if $X_i < X_j$ and $Y_i < Y_j$ or $X_i > X_j$ and $Y_i > Y_j$. Only pairs of individuals, in which at least one is event, are usable. The c-statistic is defined as $c = P(Y_i < Y_j | X_i < X_j)$[83]. In a sample, it can be estimated followingly: Let there be a sample of n individuals, x_i-s their observed survival times and y_i-s their predicted probabilities of surviving until a given time point t, $i = 1, ...n$. Let us define c_{ij} which takes value 1 if $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$ and 0 otherwise. Let Q be the set of all usable pairs (i,j). Then

$$\hat{c} = \frac{1}{q} \sum_{(i,j) \in Q} c_{ij}$$

where q is the number of all usable pairs[83]. Value 0.5 shows that there is no predictive discriminative ability and value 1 shows perfect discrimination between individuals with different outcomes.

Net reclassification index (NRI) aims to quantify how well a new model correctly reclassifies subjects compared to old model[84]. Probabilities of having an event during fixed time of period from both old and new model are acquired for all individuals. Among events, NRI is defined as the difference in proportion of events whose probabilities are higher with new model compared to old model and the proportion of events whose probabilities are lower with the new model than with the old model. Among non-events, NRI is defined as the difference in proportion of non-events whose probabilities are lower with new model compared to old model and the proportion of non-events whose probabilities are higher with the new model than with the old model.

3.2.2. Associations of different GRSs and status of type 2 diabetes

Both BMI adjusted and unadjusted models were fitted to compare different versions of GRSs. No single-weighted score showed so high likelihood ratio test statistics [Ref I, Figure S1] as double-weighted scores with $300 < k < 2000$. The highest log-likelihood was reached with $dGRS_{1400}$ for BMI-unadjusted models and $dGRS_{800}$ for BMI-adjusted models [Ref I , Table 2]. As $dGRS_{1000}$ provided a fit that was not significantly different (Cox test $p > 0.05$) from the best-fitting GRSs either BMI adjusted or unadjusted models, it was chosen for all the following analyses. Up to our best knowledge, 65 SNPs included in the GRS by

Talmud and colleagues[85] was the most SNPs used to compose GRS for T2D before our study. Talmud *et al.* reported that OR corresponding to 1 standard deviation was 1.43 (95% CI 1.33-1.54) whereas in our study for $dGRS_{1000}$, OR per 1 SD was 1.56 (95% CI 1.45–1.68) in the BMI unadjusted model and 1.59 (95% CI: 1.46–1.72) in the BMI adjusted model. We also compared the GRS quintiles while analysing incident T2D, showing that the difference in hazards between lowest and highest quintile is more than threefold(HR = 3.45, 95% CI: 2.31–5.17) [Ref I, Table 3]. The $dGRS_{1000}$ showed also an association with all-cause and cardiovascular mortality.

We investigated associations between $dGRS_{1000}$ and T2D risk factors separately in prevalent T2D cases and controls. In the sample with individuals without prevalent T2D, $dGRS_{1000}$ showed positive association with triglycerides, plasma glucose levels and waist-hip ratio and negative association with high density lipoprotein levels [Ref I, Table S5].

3.2.3. Analysis of incremental value of GRS

Harrell's c-statistic increased by 1.2% (95% CI 0.004-0.023) after adjusting Cox proportional hazard model with BMI, waist circumference, waist-hip ratio, history of hypertension, history of high blood glucose, physical activity level, smoking, fruit and vegetable consumption additionally for $dGRS_{1000}$. Greater benefit of GRS was seen among individuals with BMI $\in (25, 35)$, were the increase of c-statistic was 2.1% (95% CI 0.006-0.039) [Ref I, Table S6]. Even though the added benefit of GRS is relatively low, majority of previous studies (given in the review by[60]) comparing discrimination of clinical risk models with and without including genetic risk scores for incident T2D had reported either no improvement in discrimination or it remained under 1%. One of the possible explanation for these findings could be that the risk factors also have genetic background (such as BMI for instance) and if the genetic risk score captures that, adjusting for both might not give additional information. The other reason could be that most of the scores mentioned in[60] included less than 20 SNPs and therefore might have had a limited predictive ability in the first place.

To study reclassification, status of incident T2D was determined after up to 5 years of follow-up. We investigated 5-year predictions from models with and without $dGRS_{1000}$. The net reclassification index for T2D events was 0.115 (95% CI 0.02-0.23) and for non-events, 0.209 (95% CI 0.182-0.23), showing that non-events gain more benefit in terms of prediction accuracy than events. Similar results were demonstrated in study by Talmud *et al.*[85] were they reported that adding *GRS* with 65 SNPs to a 10-year risk estimate composed of age, sex, parental history of T2D, BMI, blood pressure, HDL cholesterol, triglyceride, and fasting glucose level increased the event NRI by 13% and non-event NRI by 17%.

3.3. Polygenic risk scores for breast cancer

3.3.1. Description of materials and methods

Both Estonian ($n = 32557$) and UK Biobank ($n = 43827$) women (not included in UKBB GWAS data) are used during this analysis. Estonian Biobank data was divided into two parts: 1) derivation set with all prevalent breast cancer (BC) cases (317) and 2000 randomly selected controls and 2) validation set with 30240 women, including 308 incident BC cases. For UKBB, both status of BC is determined via combination of diagnosis linked from UK National Cancer Registries and self reported information, resulting in 3157 cases.

As meta-analysis summary results for BC were publicly unavailable until the end of 2017, genetic risk scores were limited to very few (less than 100) SNPs. We identified 2 previously published GRSs from the literature with originally 86 and 77 SNPs, but 70 and 75 SNPs of those were available respectively with high imputation accuracy in Estonian Biobank. Likewise to other complex diseases, we hypothesised that including more SNPs into genetic risk score would improve its predictive ability.

In 2017, Neale's Lab performed GWASs using $\sim$337000 unrelated individuals of British ancestry and made the results for more than 2000 phenotypes (including self reported BC) publicly available[86]. At the same time, Breast Cancer Association Consortium published their meta-analysis results performed with more than 100 000 cases and controls of European ancestry. Both summary statistics files were available for 11 million SNPs. We developed GRSs with both summary statistics files using PRSice[87] by first clumping the SNPs to obtain independent set of them and then varying p-value threshold for inclusion into GRS. As a next step, we tested prevalent BC status-GRS association in the case-control subset of Estonian Biobank with 318 BC cases and 2000 controls and chose the GRS from each study with the smallest p-value for further investigation and denoted them as GRS_{UK} and GRS_{ONCO}. Several versions of metaGRSs as described in section 2.5 were composed (see Figure 6). Association of all four GRSs and three metaGRSs were assessed in both UKBB with logistic regression model and EstBB's validation set with Cox proportional hazard model. Incremental value of GRSs were investigated by comparing Harrell's c-statistics from models adjusted for 10-year risk estimates obtained via National Cancer institute algorithm alone and together with different versions of GRSs. We also assessed the joint effect of family history status of breast cancer and genetic risk score in UKBB. Detailed overview of the workflow is given in Figure 6.

3.3.2. Comparison of predictive ability of GRSs

GRS_{75} and GRS_{ONCO} had the strongest effects on BC in both EstBB (OR per 1 SD 1.38, 95% CI1.22-1.57 and 1.44, 95% CI 1.27-1.64 , respectively) and in UKBB (OR per 1 SD 1.48, 95% CI 1.43-1.53 and 1.51, 95% CI 1.46-1.57) [REF II: Sup-

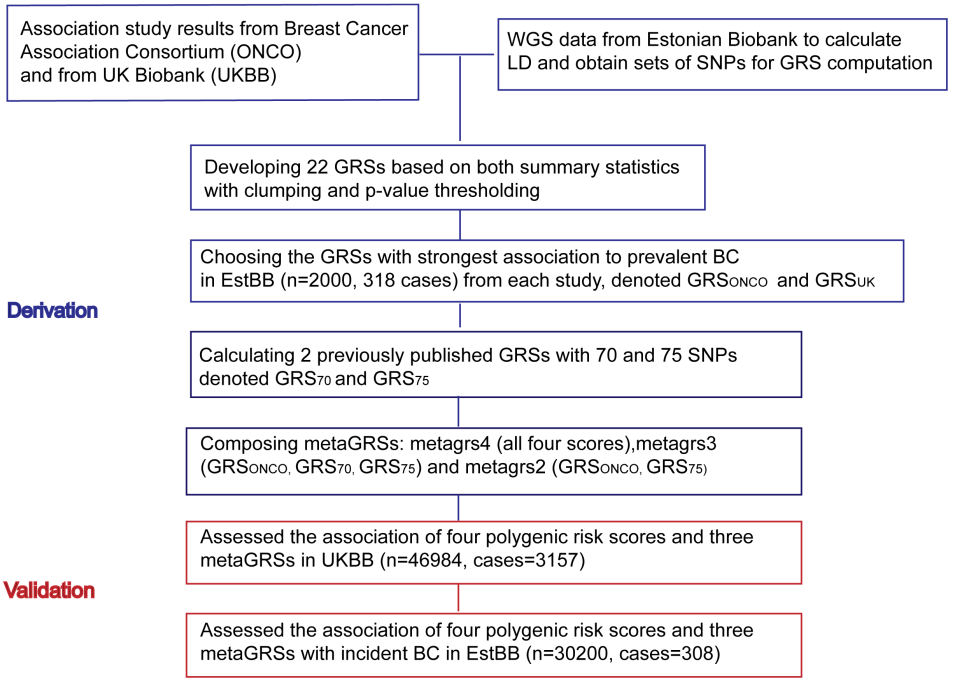

Figure 6. Overall workflow of breast cancer genetic risk score development and analyses.

plementary Table 2]. Combining these two score into $metaGRS_2$ resulted in the GRS with the strongest association in both cohorts [Ref II, Table 1, Supplementary Table 2]. Even more, adding $metaGRS_2$ to 10-year risk estimate taking into account other known risk factors (such as age, age of menarche, age of first childbirth, ethnicity, etc) for BC[88] increased the Harrell's c-statistic by 3.8% (from 0.677 to 0.715).

We looked into if and how the effect of GRS attenuates after adjusting for family history using UKBB data. Both family history of BC and any version of genetic risk score were statistically significant predictors for the status of prevalent BC, and including family history in the logistic regression model only marginally decreased the effect of any GRS [Ref II: Table S2]. The importance of both family history and genetic risk score is shown in Figure 7. Mavaddat *et al.*[89] also showed that GRS and family history modify BC risk together, adjusting for both attenuates the effect of family history by 12.6%.

3.3.3. Non-uniqueness of polygenic risk scores

Depending on a GWAS that is used to develop GRSs, different and not necessarily highly correlated GRSs can be produced for the same disease [Ref II: Supplementary Figure 2]. We hypothesize that GRSs might reflect the effects of different biological pathways or risk factors of the disease. For example, higher GRS_{UK} was associated with lower body mass index and waist circumference, more strongly among younger women whereas GRS_{ONCO} seemed to be weakly associated with

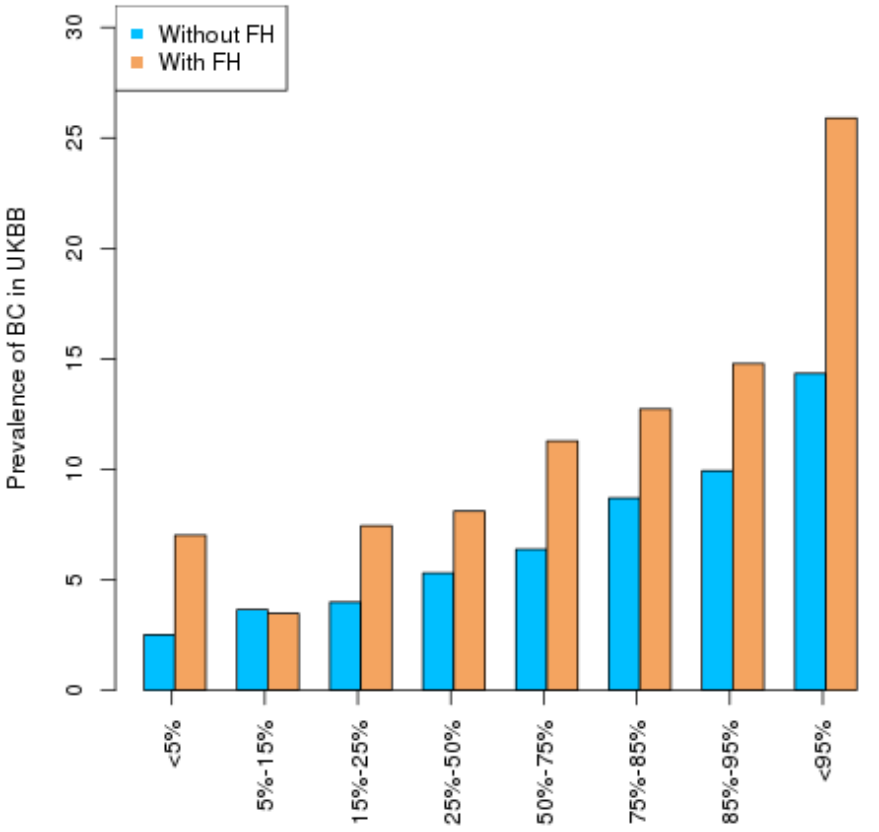

Figure 7. Prevalence of breast cancer in different $metaGRS_2$ categories depending on status of family history (FH) of breast cancer in UKBB data.

the status of smoking [Ref II: Supplementary Table 3-4]. The negative association between GRS_{UK} and BMI among younger women is biologically plausible, as low BMI has been shown to be a risk factor for breast cancer among pre-menopausal women[90, 91], even though the biological mechanism behind it is still unknown.

The fact that different GRSs for the same disease might not agree with each other has implications to genetic feedback and genetic risk based targeted screening. For example, individuals belonging to the top 5% with one GRS do not necessarily belong there with the other GRS [Ref II: Figure 2; Figure 8]. The overlap depends on the level of correlation between the two scores.

Currently, there are pilot programs going on in the world (like RITA in Estonia or WISDOM in US[92]), targeting women for earlier screening solely based on their value of GRS. This non-uniqueness of GRSs is an important issue for these type of projects and both feedback receivers and medical staff should be aware that GRSs at this state are still proxies of true genetic risk and are subjected to

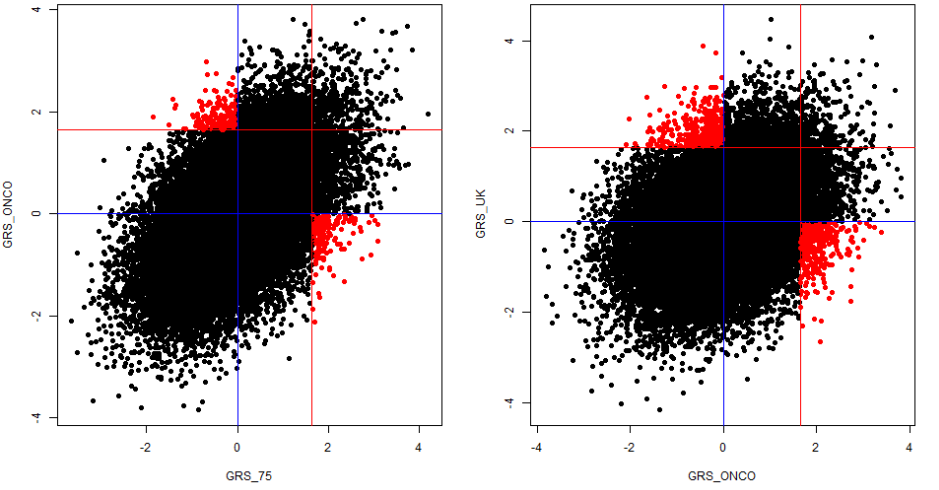

Figure 8. Scatter plots of different GRS. Red lines indicate the 95th percentile of the GRS and the blue line the median of GRS. Individuals belonging to top 5% category with one GRS and below median category with other are denoted with red dots.

change in the future. However, when GRS is not used alone, but is only one variable in the absolute risk model with similar effect size than other risk factors, its varying might not be an important problem.

Even though combing GRSs into one metascore resulted in a best predicting genetic estimator for BC, it remains unclear if using metaGRS instead of several different genetic risk scores is always the best practice – if biological mechanisms can be assigned to GRSs, more efficient prevention could be attributed. It seems also plausible, that optimal GRSs for different subtypes of diseases are not always the same. This has been recently demonstrated for breast cancer, where best predicting GRSs for BRCA1 and BRCA2 carriers were different and depended on whether the SNPs in GRS were associated with estrogen-receptor positive or negative breast cancer[93]. However, a lot more research needs to be done in subtype-specific GRS area before more conclusive inference can be drawn.

3.4. Polygenic risk score distributions in different ancestral populations

3.4.1. Description of materials and methods

2244 samples from the Estonian Biobank with whole-genome sequencing data available and 1000G Project data from phase 3 release with 2504 individuals from 5 populations (Europeans (EUR), East-Asia (EAS), South-Asia (SAS), Africa (AFR) and Americas (AMR)) were used in this analysis. We identified SNPs from both datasets for two published genetic risk scores - GRS_{T2D} and GRS_{CHD}. In to-

tal, for type 2 diabetes score, 7395 SNPs and for coronary heart disease, 45996 SNPs were used, respectively. Both GRSs had been developed based on GWA study mainly using European ancestry samples. Genetic risk scores were calculated with PLINK[80] and the distribution of both scores were plotted separately for all populations. Also, quintiles of GRSs depending on population were compared. Finally, principal components for both datasets were calculated to explore associations between them and genetic risk scores.

3.4.2. Characterization of distributions of polygenic risk scores in populations

Average GRS_{T2D} in Europeans is -0.73 (95% CI -0.69...-0.61) whereas average GRS_{T2D} among African population is 1.57 (95% CI 1.54-1.60)[Ref III: Table 1]. However, similar prevalences of type 2 diabetes[94] across different ancestral groups (age standardized prevalence 7.1% in African region vs 7.3% in European region) do not show the same disparity as genetic risk scores. Similar phenomena has been recently shown for schizophrenia[95, 96]. Bitarello and Mathieson also show that GRS for height developed based on UKBB GWAS effect estimates explains 1.7% of height variation in African Americans, compared to 5.5% in European Americans[97], raising an important question about the transferability of the GRSs between populations.

There are several possible explanations why higher GRS values in one population might not result in higher disease prevalence. One of them is that true effect sizes for the same SNP vary between populations. For example, it has been shown that there are SNPs which effects on total cholesterol depend on the ethnicity[98]. Furthermore, Brown *et al.*[99] compared SNP effect estimates obtain from GWASs based on individuals of European descent and from GWASs based on individuals of East-Asian descent. They showed that correlation between effect sizes for rheumatoid arthritis was 0.436 and for type 2 diabetes 0.606. However, heterogeneity of effect sizes for a SNP can also occur when the true effect sizes for a certain causal SNP are the same but the linkage disequilibrium structures differ between studies and therefore the level of correlation between tagging SNPs and causal SNPs differs as well, causing observed effects of tagging SNPs to vary[46, 95].

3.5. Predictive ability of non-genetic risk scores for atherosclerotic cardiovascular diseases and death in Estonian Biobank

3.5.1. Description of materials and methods

Even though my thesis otherwise focuses on genetic risk scores, non-genetic "risk scores" -risk assessment tools - also exist and research to study their predictive ability in different scenarios is ongoing. Risk assessment tools are often used to

support clinical decision making in primary prevention. There are several tools out there, all predicting partially overlapping endpoints. Three widely used risk scores were selected to assess their predictive ability in Estonian Biobank:

- The American Heart Association guideline (ACC/AHA) recommends a tool called PCE which estimates 10-year risk of developing hard atherosclerotic cardiovascular disease (ASCVD) (first fatal or non-fatal myocardial infarction or stroke and coronary heart disease death) on individuals aged 40–79.
- NICE guideline recommends a tool called QRISK2. QRISK2 was developed in UK and it estimates 10-year risk of developing fatal or non-fatal ASCVD (first diagnosis of coronary heart disease, stroke or transient ischaemic attack) on individuals aged 25–84 without type 1 diabetes.
- SCORE is a tool which estimates 10-year risk of cardiovascular mortality for individuals aged 40+ without chronic kidney disease or diabetes. It was developed based on data from 12 European countries and it is a tool recommended by the European Society of Cardiology (ESC).

Age, sex, systolic blood pressure, smoking status and total cholesterol are included in all risk score algorithms. Additionally, high-density lipoprotein, diabetes status, usage of antihypertensive medication and ethnicity are included in PCE and QRISK2. Finally, QRISK2 also includes family history of ASCVD, body mass index, social deprivation (using UK postcode as a proxy) and statuses of many diseases (such as rheumatoid arthritis). See also Figure 9.

Subsets of data corresponding to each score specific guidelines were defined as well as all three different endpoints of scores (Ref IV, Figure 1). Only individuals with measured metabolites available at recruitment were used in this study. Information to define endpoints and follow-up time were retrieved from National Health Insurance Fund and Estonian Causes of Death Registry (details in Ref IV, supplement).

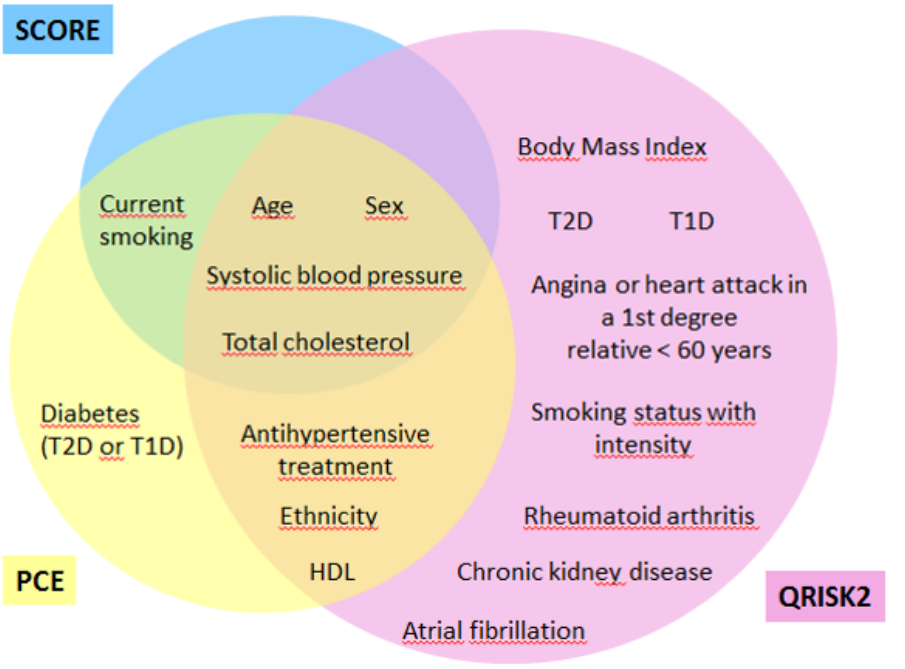

Figure 9. Venn diagram showing what risk factors are included in the risk scores and how risk factors are overlapping between scores

Predictive ability of all three scores was characterized with Harrell's c-statistic and standardized incidence ratios. Standardized incidence ratio (SIR) was defined as expected number of events divided by observed number of events. The expected number of events for each score specific outcome was calculated as the sum of 7-year risk estimates. To obtain 7-year risk estimates, original 10-year risks were modified using a constant hazard assumption. However, the modification depended on the status of the endpoint: for events, either 7-year risk or risk estimate corresponding to maximum follow-up length (if maximum possible follow up was less than 7 years from recruitment) was calculated. For non-events, the risk scores corresponded to the actual follow-up time. Viallon[100] has shown, that this method described above is unbiased compared to the modification method, where the risk scores are calculated for the actual length of follow-up regardless of the endpoint status. Surprisingly, the latter method is very popular in the literature, but it systematically lowers the expected number of events, bias depending on the proportion of events in the data (more details about methods in Ref IV, Supplement).

3.5.2. Predictive ability of risk scores in Estonian Biobank

Out of 8830 individuals, 4356, 7191 and 3987 individuals were eligible according to the guideline-specific criteria for the calculation of PCE, QRISK2 and

SCORE. During follow-up, which was censored after 7 years, 220 PCE-specific, 671 QRISK2-specific and 94 SCORE-specific outcomes occurred (Ref IV, Figure 1). Among three scores, SCORE showed the highest discriminative ability (Harrell's c = 0.865) and PCE the lowest (Harrell's c = 0.778) for score specific outcomes (Ref IV, Table 1). Regarding calibration, SCORE (SIR = 0.99, 95% CI 0.81 to 1.21) and PCE (SIR = 1.03, 95% CI 0.90 to 1.18) seemed to be well calibrated in Estonian Biobank, however, QRISK2 severely underestimated the the number of cases (SIR = 0.52, 95% CI 0.48 to 0.56) (Ref IV, Table 2).

While developing SCORE, the researchers did not use the status of diabetes for exclusions as suggested by the guidelines, as the information was not available for some cohorts. We also investigated, how inclusion of diabetics changes the discrimination and calibration of SCORE. When individuals with diabetes were included in the dataset, the overall SIR of SCORE decreased from 0.99 to 0.86 (95% CI 0.71-1.04). Among women, SIRs decreased the most - when including diabetics into the dataset, SCORE predicted 26% less cases than observed (48.1 expected vs 65 observed) compared to 14% less cases than observed without diabetics (38.8 expected vs 45 observed). Discrimination remained very similar among women and decreased among men (Ref IV, Supplement). These results support the recommendations by the European prevention guideline that diabetics should automatically categorize into a high or very high risk group, in whom the risk by SCORE is underestimated.

We also investigated how well for QRISK2 education level works as a proxy for level of social deprivation (estimated with UK postal code in QRISK2 calculator) in Estonia. The substitution was done as suggested by[101]. The predictive ability of QRISK2 was investigated with education level as a proxy and in a second scenario, leaving it blank (ie, using average value 0) as proposed by the developers of calculator in case UK postal code is unknown. The c-statistic's remained very similar. Even though estimated number of events increased marginally, QRISK2 still severely underestimated the risk of developing fatal or non-fatal ASCVD (Ref IV, Supplementary table 5), overall indicating that the level of education is not a very useful proxy for social deprivation in Estonia.

We also compared the treatment recommendations based on guideline-specific criteria. Depending on the combination of risk factor levels, each guideline gives suggestions regarding statin distribution: it can either be recommended, considered or deemed unnecessary. NICE guideline (QRISK2) was found to be most conservative and ESC (SCORE) most liberal while suggesting use of statins (Ref IV, Figure 3). However, statins for primary prevention were recommended to almost half of the men and quarter of women under investigation, illustrating high risk levels of ASCVD in Estonia.

4. CONCLUSION

The importance of prevention of common complex diseases has arisen as an important topic in personalized medicine. Interest in genetic predictors has increased, as they allow identification of individuals with higher predisposition for a trait already in an early age. One option to estimate individual's genetic predisposition for a common complex disease is via genetic risk scores.

At the beginning, genetic risk scores were usually a sum of risk alleles accounting for a few well established trait-associated SNPs. Pretty soon the idea arose to weight SNPs differently based on findings from GWASs, however, the scores were still limited to SNPs which had achieved genome-wide significance. In the beginning of this thesis, the largest GRS for type 2 diabetes included 65 SNPs and we set a goal to further improve the predictive ability of a GRS for type 2 diabetes.

Doubly-weighting is an alternative method to single-weighting while constructing genetic risk scores. Instead of setting a p-value threshold to include SNPs into the GRS as done with single-weighting, all uncorrelated SNPs can be included from GWAS. The benefit of doubly-weighting (as shown with simulations) depends on the heritability of the trait as well as number of causal SNPs. In the first article of this thesis, doubly-weighted GRS is a systematically stronger predictor for type 2 diabetes compared to single-weighted versions of GRSs in Estonian Biobank.

Over the years, many GWASs and meta-analyses are performed for the same trait and their summary statistics files are often publicly available. Several parameters of these studies vary: sample size, inclusion criteria of individuals, characteristics of samples as well as genotyping platforms used to generate the data or imputation reference used to impute the data. But this means that different GRSs can be constructed using different GWA studies to select both SNPs and weights. Therefore, the study that one uses as a base to construct GRSs will affect why and how GRS will be predicting a trait. Consequently, GRSs composed for the same trait based on diffent studies might not be highly correlated with each other. This phenomena is observed in the second article of this thesis for breast cancer, where four different base files are used to generate GRSs. To achieve the best predicting GRS, an idea of combining GRSs based on different studies is implemented and it results in a metaGRS combining two GRSs into one. However, the non-uniqueness of GRSs is an important issue especially for projects trying to stratify individuals for prevention solely based on GRS.

Another important aspect influencing the predictive ability of GRSs is the similarity of discovery and validation population. This is investigated in the third article, where it is showed that the distributions of GRSs for type 2 diabetes and coronary artery disease are different within different ancestral populations. Unfortunately, due to lack of phenotypes for the 1000G project data, we were unable to compare the predictive ability of the same GRS in different populations. How-

ever, other scientists have now shown that at least for schizophrenia and height, the GRSs developed using GWASs mainly based on European ancestry samples do not show the same predictive power for Europeans and African Americans, raising an important question about the overall transferability of GRSs between populations.

Finally, non-genetic risk scores for ASCVD are validated in the Estonian Biobank data. The idea was to investigate if and how the risk algorithms perform in their original form in Estonian Biobank. We found that the discriminative ability of all three scores were good. PCE and SCORE were well calibrated, but QRISK2 estimated almost twice as less cases than observed. We also compared the statin treatment recommendations based on guideline specific criteria. The most conservative out of three of them was NICE (QRISK2) and the most liberal was ESC (SCORE). However, statins for primary prevention were recommended to almost half of the men and quarter of women under investigation, illustrating high risk levels of ASCVD in Estonia.

BIBLIOGRAPHY

[1] R. Henderson, M. O'Kane, V. McGilligan et al. The genetics and screening of familial hypercholesterolaemia. *Journal of biomedical science*, 23:39, 2016. ISSN 1423-0127. doi: 10.1186/s12929-016-0256-1.

[2] M. I. McCarthy, G. R. Abecasis, L. R. Cardon et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008. ISSN 14710056. doi:10.1038/nrg2344.

[3] P. R. Burton, D. G. Clayton, L. R. Cardon et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007. ISSN 0028-0836. doi:10.1038/nature05911.

[4] K. Michailidou, S. Lindström, J. Dennis et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, 2017. ISSN 0028-0836. doi:10.1038/nature24284.

[5] M. Inouye, G. Abraham, C. P. Nelson et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. *Journal of the American College of Cardiology*, 72(16):1883–1893, 2018. ISSN 07351097. doi:10.1016/j.jacc.2018.07.079.

[6] J. L. Mega, N. O. Stitziel, J. G. Smith et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *The Lancet*, 385:2264–2271, 2015. ISSN 01406736. doi:10.1016/S0140-6736(14)61730-X.

[7] National Human Genome Research Institute. Deoxyribonucleic Acid (DNA) Fact Sheets. 2015. URL `https://www.genome.gov/25520880/`.

[8] A. Heinaru. *Geneetika*. Tartu Ülikooli Kirjastus, 2012.

[9] K. A. Frazer, S. S. Murray, N. J. Schork et al. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, 2009. ISSN 1471-0056. doi: 10.1038/nrg2554.

[10] I. Iacobucci, A. Lonetti, C. Papayannidis et al. Use of single nucleotide polymorphism array technology to improve the identification of chromosomal lesions in leukemia. *Current cancer drug targets*, 13(7):791–810, 2013. ISSN 1873-5576. doi:10.2174/15680096113139990089.

[11] L. J. Engle, C. L. Simpson and J. E. Landers. Using high-throughput SNP technologies to study cancer. *Oncogene*, 25(11):1594–1601, 2006. ISSN 0950-9232. doi:10.1038/sj.onc.1209368.

[12] C. S. Ku, E. Y. Loy, A. Salim et al. The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics*, 55(7):403–415, 2010. ISSN 1434-5161. doi:10.1038/jhg.2010.55.

[13] T. . G. P. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. ISSN 1476-4687. doi:10.1038/nature15393.

[14] A. Telenti, L. C. T. Pierce, W. H. Biggs et al. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(42):11901–11906, 2016. ISSN 1091-6490. doi:10.1073/pnas.1613365113.

[15] A.-C. Syvänen. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, 2(12):930–942, 2001. ISSN 1471-0056. doi:10.1038/35103535.

[16] A. C. Nica and E. T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620):20120362, 2013. ISSN 1471-2970. doi:10.1098/rstb.2012.0362.

[17] S. Kasela. *Genetic regulation of gene expression: detection of tissueand cell type-specific effects*. Ph.D. thesis, University of Tartu, 2017. URL `https://dspace.ut.ee/bitstream/handle/10062/57006/kasela{_}silva.pdf?sequence=1{&}isAllowed=y`.

[18] Nature Education. haplotype / haplotypes | Learn Science at Scitable. 2014. URL https://www.nature.com/scitable/definition/haplotype-haplotypes-142.

[19] K. G. Ardlie, L. Kruglyak and M. Seielstad. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309, 2002. ISSN 1471-0056. doi:10.1038/nrg777.

[20] D. O. Stram. Software for tag single nucleotide polymorphism selection. *Human genomics*, 2(2):144–51, 2005. ISSN 1479-7364. doi:10.1186/1479-7364-2-2-144.

[21] V. Kristensen, D. Kelefiotis, T. Kristensen et al. High-Throughput Methods for Detection of Genetic Variation. Technical report, Norwegian Radium Hospital, Oslo & Foundation for Research and Technology– Hella & University of Oslo, Oslo, 2001. URL www.res.ibm.com.

[22] M. Podder, J. Ruan, B. W. Tripp et al. Robust SNP genotyping by multiplex PCR and arrayed primer extension. *BMC medical genomics*, 1:5, 2008. ISSN 1755-8794. doi:10.1186/1755-8794-1-5.

[23] Thermo Fisher Scientific Inc. Affymetrix Arrays. 2017. URL https://www.affymetrix.com/products{_}services/arrays/index.affx{#}1{_}2.

[24] Illumina Inc. Illumina Microarray Kits. 2018. URL https://www.illumina.com/products/by-type/microarray-kits.html.

[25] B. N. Howie, P. Donnelly and J. Marchini. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, 5(6):e1000529, 2009. ISSN 1553-7404. doi:10.1371/journal.pgen.1000529.

[26] J. Marchini, D. Cutler, N. Patterson et al. A Comparison of Phasing Algorithms for Trios and Unrelated Individuals. *The American Journal of Human Genetics*, 78(3):437–450, 2006. ISSN 0002-9297. doi:10.1086/500808.

[27] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005. ISSN 0028-0836. doi:10.1038/nature04226.

[28] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010. ISSN 1471-0056. doi:10.1038/nrg2796.

[29] B. L. Browning and S. R. Browning. Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics*, 98(1):116–126, 2016. ISSN 0002-9297. doi:10.1016/J.AJHG.2015.11.020.

[30] Nature Education. Hardy-Weinberg equation | Learn Science at Scitable. 2014. URL https://www.nature.com/scitable/definition/hardy-weinberg-equation-299.

[31] D. Shungin, T. W. Winkler, D. C. Croteau-Chonka et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187–196, 2015. ISSN 1476-4687. doi:10.1038/nature14132.

[32] G. Abraham, A. S. Havulinna, O. G. Bhalala et al. Genomic prediction of coronary heart disease. *European heart journal*, 37(43):3267–3278, 2016. ISSN 1522-9645. doi:10.1093/eurheartj/ehw450.

[33] A. P. Morris, B. F. Voight, T. M. Teslovich et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9):981–990, 2012. ISSN 1061-4036. doi:10.1038/ng.2383.

[34] M. Nikpay, A. Goel, H.-H. Won et al. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10):1121–1130, 2015. ISSN 1061-4036. doi:10.1038/ng.3396.

[35] A. E. Locke, B. Kahali, S. I. Berndt et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015. ISSN 0028-0836. doi:10.1038/nature14177.

[36] F. R. Schumacher, A. A. Al Olama, S. I. Berndt et al. Association analyses of more than

140,000 men identify 63 new prostate cancer susceptibility loci. *Nature Genetics*, 50(7):928–936, 2018. ISSN 1061-4036. doi:10.1038/s41588-018-0142-8.

[37] F. Zhao, M. Song, Y. Wang et al. Genetic model. *Journal of cellular and molecular medicine*, 20(4):765, 2016. ISSN 1582-4934. doi:10.1111/jcmm.12751.

[38] J. Fadista, A. K. Manning, J. C. Florez et al. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European journal of human genetics : EJHG*, 24(8):1202–5, 2016. ISSN 1476-5438. doi:10.1038/ejhg.2015.269.

[39] J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005. ISSN 1471-0056. doi:10.1038/nrg1521.

[40] A. L. Price, N. A. Zaitlen, D. Reich et al. New approaches to population stratification in genome-wide association studies. *Nature reviews. Genetics*, 11(7):459–63, 2010. ISSN 1471-0064. doi:10.1038/nrg2813.

[41] P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, 2015. ISSN 1061-4036. doi:10.1038/ng.3190.

[42] A. B. Haidich. Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1):29–37, 2010. ISSN 1790-8019.

[43] C. J. Willer, Y. Li and G. R. Abecasis. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)*, 26(17):2190–1, 2010. ISSN 1367-4811. doi:10.1093/bioinformatics/btq340.

[44] J. Fleiss. Review papers : The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, 2(2):121–145, 1993. ISSN 0962-2802. doi:10.1177/096228029300200202.

[45] C. H. Lee, S. Cook, J. S. Lee et al. Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. *Genomics & informatics*, 14(4):173–180, 2016. ISSN 1598-866X. doi:10.5808/GI.2016.14.4.173.

[46] B. Han and E. Eskin. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American journal of human genetics*, 88(5):586–98, 2011. ISSN 1537-6605. doi:10.1016/j.ajhg.2011.04.014.

[47] R. Mägi, M. Horikoshi, T. Sofer et al. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Human molecular genetics*, 26(18):3639–3650, 2017. ISSN 1460-2083. doi:10.1093/hmg/ddx280.

[48] Y. R. Li and B. J. Keating. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome medicine*, 6(10):91, 2014. ISSN 1756-994X. doi:10.1186/s13073-014-0091-5.

[49] A. Mahajan, M. J. Go, W. Zhang et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3):234–244, 2014. ISSN 1061-4036. doi:10.1038/ng.2897.

[50] S. Myles, D. Davison, J. Barrett et al. Worldwide population differentiation at disease-associated SNPs. *BMC medical genomics*, 1:22, 2008. ISSN 1755-8794. doi:10.1186/1755-8794-1-22.

[51] A. P. Morris. Transethnic meta-analysis of genomewide association studies. *Genetic epidemiology*, 35(8):809–22, 2011. ISSN 1098-2272. doi:10.1002/gepi.20630.

[52] R. Mägi and A. P. Morris. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics*, 11(1):288, 2010. ISSN 1471-2105. doi:10.1186/1471-2105-11-288.

[53] J. P. Cook and A. P. Morris. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *European journal of human genetics : EJHG*, 24(8):1175–80, 2016. ISSN 1476-5438. doi:10.1038/ejhg.2016.17.

[54] N. Chatterjee, J. Shi and M. García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature reviews. Genetics*, 17(7):392–406, 2016. ISSN 1471-0064. doi:10.1038/nrg.2016.27.

[55] A. Torkamani, N. E. Wineinger and E. J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, 2018. ISSN 1471-0056. doi:10.1038/s41576-018-0018-x.

[56] J. Hippisley-Cox and C. Coupland. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ open*, 5(3):e007825, 2015. ISSN 2044-6055. doi:10.1136/bmjopen-2015-007825.

[57] J. Hippisley-Cox, C. Coupland, Y. Vinogradova et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ (Clinical research ed.)*, 336(7659):1475–82, 2008. ISSN 1756-1833. doi:10.1136/bmj.39609.449676.25.

[58] J. Wang, A. Stančáková, J. Kuusisto et al. Identification of Undiagnosed Type 2 Diabetic Individuals by the Finnish Diabetes Risk Score and Biochemical and Genetic Markers: A Population-Based Study of 7232 Finnish Men. *The Journal of Clinical Endocrinology & Metabolism*, 95(8):3858–3862, 2010. ISSN 0021-972X. doi:10.1210/jc.2010-0012.

[59] H. Li, B. Feng, A. Miron et al. Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the Breast Cancer Family Registry and kConFab. *Genetics in Medicine*, 19(1):30–35, 2017. ISSN 1098-3600. doi:10.1038/gim.2016.43.

[60] V. Lyssenko and M. Laakso. Genetic screening for the risk of type 2 diabetes: worthless or valuable? *Diabetes care*, 36 Suppl 2(Suppl 2):S120–6, 2013. ISSN 1935-5548. doi:10.2337/dcS13-2009.

[61] S. Ripatti, E. Tikkanen, M. Orho-Melander et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *The Lancet*, 376(9750):1393–1400, 2010. ISSN 0140-6736. doi:10.1016/S0140-6736(10)61267-6.

[62] Y. Shieh, D. Hu, L. Ma et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Research and Treatment*, 159(3):513–525, 2016. ISSN 0167-6806. doi:10.1007/s10549-016-3953-2.

[63] B. T. Helfand. A comparison of genetic risk score with family history for estimating prostate cancer risk. *Asian journal of andrology*, 18(4):515–9, 2016. ISSN 1745-7262. doi:10.4103/1008-682X.177122.

[64] A. V. Khera, M. Chaffin, K. G. Aragam et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, 2018. ISSN 1061-4036. doi:10.1038/s41588-018-0183-z.

[65] A. Tenesa and C. S. Haley. The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics*, 14(2):139–149, 2013. ISSN 1471-0056. doi:10.1038/nrg3377.

[66] A. J. Mayhew and D. Meyre. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. *Current genomics*, 18(4):332–340, 2017. ISSN 1389-2029. doi:10.2174/1389202918666170307161450.

[67] N. R. Wray, J. Yang, B. J. Hayes et al. Pitfalls of predicting complex traits from SNPs. *Nature reviews. Genetics*, 14(7):507–15, 2013. ISSN 1471-0064. doi:10.1038/nrg3457.

[68] S. Moller, L. A. Mucci, J. R. Harris et al. The Heritability of Breast Cancer among Women in the Nordic Twin Study of Cancer. *Cancer Epidemiology Biomarkers & Prevention*, 25(1):145–150, 2016. ISSN 1055-9965. doi:10.1158/1055-9965.EPI-15-0913.

[69] P. Poulsen, K. O. Kyvik, A. Vaag et al. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42(2):139–45, 1999. ISSN 0012-186X.

[70] P. Almgren, M. Lehtovirta, B. Isomaa et al. Heritability and familiality of type 2 diabetes

and related quantitative traits in the Botnia Study. *Diabetologia*, 54(11):2811–9, 2011. ISSN 1432-0428. doi:10.1007/s00125-011-2267-5.

[71] A. P. Morris. Progress in defining the genetic contribution to type 2 diabetes susceptibility. *Current Opinion in Genetics & Development*, 50:41–51, 2018. ISSN 0959-437X. doi:10.1016/J.GDE.2018.02.003.

[72] J. Lilyquist, K. J. Ruddy, C. M. Vachon et al. Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 27(4):380–394, 2018. ISSN 1538-7755. doi:10.1158/1055-9965.EPI-17-1144.

[73] P. M. Visscher, N. R. Wray, Q. Zhang et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017. ISSN 0002-9297. doi:10.1016/J.AJHG.2017.06.005.

[74] K. E. Lohmueller, C. L. Pearce, M. Pike et al. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, 33(2):177–182, 2003. ISSN 10614036. doi:10.1038/ng1071.

[75] C. Mood. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review*, 26(1), 2010. doi:10.1093/esr/jcp006.

[76] L. Leitsalu, T. Haller, T. Esko et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International journal of epidemiology*, 2014. ISSN 1464-3685. doi:10.1093/ije/dyt268.

[77] L. Leitsalu, H. Alavere, M.-L. Tammesoo et al. Linking a Population Biobank with National Health Registries—The Estonian Experience. *Journal of Personalized Medicine*, 5(2):96–106, 2015. ISSN 2075-4426. doi:10.3390/jpm5020096.

[78] C. Sudlow, J. Gallacher, N. Allen et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015. ISSN 1549-1676. doi:10.1371/journal.pmed.1001779.

[79] C. Bycroft, C. Freeman, D. Petkova et al. The UK Biobank resource with deep phenotyping and genomic data. 2018. doi:10.1038/s41586-018-0579-z.

[80] S. Purcell, B. Neale, K. Todd-Brown et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–75, 2007. ISSN 0002-9297. doi:10.1086/519795.

[81] D. R. Cox. Tests of Separate Families of Hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 105–123. The Regents of the University of California, 1961. ISSN 0097-0433. URL `http://projecteuclid.org/euclid.bsmsp/1200512162`.

[82] F. E. jr Harrell, K. L. Lee and D. B. Mark. Multivariable prognostic models: issues In developing models, evaluating Assumptions and adequacy, And measuring and reducing errors. *Statistics in medicine*, 15:361–387, 1996.

[83] M. J. Pencina and R. B. D'Agostino. OverallC as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, 2004. ISSN 0277-6715. doi:10.1002/sim.1802.

[84] M. J. Pencina, R. B. D'Agostino, E. W. Steyerberg et al. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine*, 30(1):11–21, 2011. ISSN 1097-0258. doi:10.1002/sim.4085.

[85] P. J. Talmud, J. A. Cooper, R. W. Morris et al. Sixty-Five Common Genetic Variants and Prediction of Type 2 Diabetes. *Diabetes*, 64(5):1830–1840, 2015. ISSN 0012-1797. doi:10.2337/db14-1504.

[86] Neale Lab. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank — Neale lab. 2017. URL `http://www.`

nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-/for-337000-samples-in-the-uk-biobank.

[87] J. Euesden, C. M. Lewis and P. F. O'Reilly. PRSice: Polygenic Risk Score software. *Bioinformatics*, 31(9):1466–1468, 2015. ISSN 1460-2059. doi:10.1093/bioinformatics/btu848.

[88] National Cancer Institute. Breast Cancer Risk Assessment Tool. 2011. URL https://www.cancer.gov/bcrisktool/Default.aspx.

[89] N. Mavaddat, P. D. P. Pharoah, K. Michailidou et al. Prediction of breast cancer risk based on profiling with common genetic variants. *Journal of the National Cancer Institute*, 107(5), 2015. ISSN 1460-2105. doi:10.1093/jnci/djv036.

[90] K. B. Michels, K. L. Terry and W. C. Willett. Longitudinal Study on the Role of Body Size in Premenopausal Breast Cancer. *Archives of Internal Medicine*, 166(21):2395, 2006. ISSN 0003-9926. doi:10.1001/archinte.166.21.2395.

[91] K. Liu, W. Zhang, Z. Dai et al. Association between body mass index and breast cancer risk: evidence based on a dose-response meta-analysis. *Cancer management and research*, 10:143–151, 2018. ISSN 1179-1322. doi:10.2147/CMAR.S144619.

[92] S. Rosenberg-Wohl, M. Eklund, J. Tice et al. Women informed to screen depending on measures of risk (WISDOM): A RCT of personalized vs. annual screening for breast cancer. *Journal of Clinical Oncology*, 34(15_suppl):TPS1594–TPS1594, 2016. ISSN 0732-183X. doi:10.1200/JCO.2016.34.15_suppl.TPS1594.

[93] K. B. Kuchenbaecker, L. McGuffog, D. Barrowdale et al. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *Journal of the National Cancer Institute*, 109(7), 2017. ISSN 1460-2105. doi:10.1093/jnci/djw302.

[94] WHO. GLOBAL REPORT ON DIABETES WHO Library Cataloguing-in-Publication Data Global report on diabetes. Technical report, 2016. URL http://www.who.int/about/licensing/copyright{_}form/index.html.

[95] A. R. Martin, C. R. Gignoux, R. K. Walters et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, 100(4):635–649, 2017. ISSN 15376605. doi:10.1016/j.ajhg.2017.03.004.

[96] T. R. de Candia, S. H. Lee, J. Yang et al. Additive Genetic Variation in Schizophrenia Risk Is Shared by Populations of African and European Descent. *The American Journal of Human Genetics*, 93(3):463–470, 2013. ISSN 0002-9297. doi:10.1016/J.AJHG.2013.07.007.

[97] B. Bitarello and I. Mathieson. Polygenic risk scores perform poorly across populations. 2018. URL https://eventpilot.us/web/page.php?page=IntHtml{&}project=ASHG18{&}id=180121278.

[98] M. M. Monir and J. Zhu. Comparing GWAS Results of Complex Traits Using Full Genetic Model and Additive Models for Revealing Genetic Architecture. *Scientific Reports*, 7(1):38600, 2017. ISSN 2045-2322. doi:10.1038/srep38600.

[99] B. C. Brown, C. J. Ye, A. L. Price et al. Transethnic Genetic-Correlation Estimates from Summary Statistics. *The American Journal of Human Genetics*, 99:76–88, 2016. doi:10.1016/j.ajhg.2016.05.001.

[100] V. Viallon, S. Ragusa, F. Clavel-Chapelon et al. How to evaluate the calibration of a disease risk prediction tool. *Statistics in Medicine*, 28(6):901–916, 2009. ISSN 02776715. doi:10.1002/sim.3517.

[101] A. P. Kengne, J. W. Beulens, L. M. Peelen et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. *The Lancet Diabetes & Endocrinology*, 2(1):19–29, 2014. ISSN 22138587. doi:10.1016/S2213-8587(13)70103-7.

[102] National Human Genome Research Institute. DNA Sequencing Costs: Data - National Human Genome Research Institute (NHGRI). 2018. URL https://www.genome.gov/sequencingcostsdata/.

ACKNOWLEDGEMENT

My profound gratitude belongs to my supervisor prof. Krista Fischer. She has taught me a lot about statistics and ways to communicate it to others. She has been very encouraging about me attending conferences, which were both useful and so much fun! On a personal level, I really admire her for treating her students as her equals and her strength to stand with and for her students even if it might put her in a tough spot. I'm also thankful for other opportunities she has sent in my way - because of that I discovered that I have both passion and knack for large-scale organizing.

I'm deeply grateful to Reedik Mägi. He has been such a delight to work with, teaching me so many things about genetics and life in academia. He took time to listen, discuss and debate my ideas and opinions, after what I often saw things in a completely new light. He also became a person who was not afraid to subtly push or guide me to evolve and work harder when he thought it was necessary. And in times when life got difficult, he was simply there to offer a sympathetic ear.

I'm indebted to prof. Andres Metspalu for believing in our research, sometimes even more, than we did ourselves. He made sure that I knew that what I was doing was important, and that motivated me a lot. I'm in awe of how a person who has been in research for so many years still manages to be inspired by developments in science and I hope I can take after him in that aspect.

I so grateful for my colleagues in Estonian Genome Center. They are so dedicated to the work we do, but at the same time, they don't forget to have fun along the way! I really appreciate the team spirit we have and hope that we can keep it up. There are too many of you to thank you all by name, but my life in EGV would not have been the same without the Merli (who is an a true inspiration when it comes to less whining and more living!), Annely, KreBsu, Elin, Maris, Mart, my beloved IT crowd and the lovely admin ladies on the 4th floor :).

I'm also thankful to the faculty of mathematical statistics, especially to Jüri Lember for all the help to improve the mathematical soundness of this work and to Meelis Käärik for all the practical tips and suggestions. Cheers to my awesome course mates, with whom we've shared the ups and downs in life for almost ten years now. I would not be where I am today without you.

I'm grateful to my family for teaching me the importance of family, appreciation of hard work and general concepts of being a decent human being. I'm so happy to have friends who make my life so adventurous. And finally, it is hard to express how much I owe to these two people – Silva and my husband Leonardo.

Silva taught me to pay attention to little annoying details, often being my moral compass and my main support person at work for so many years. Leo, you make my life much more spontaneous and fun! You have been my rock on occasions I have lost my nerve or needed cheering up and you take care that I stressed less and enjoyed more in life. Over the years, you two have shaped me to be a better version of myself. Thank you for being with me!

SISUKOKKUVÕTE

Riskiskoorid ja nende prognoosivõime komplekshaiguste jaoks

Võrreldes kümne aasta taguse ajaga on nii genotüpiseerimine kui ka sekveneerimine oluliselt odavamaks muutunud (10 miljonit dollarit 2007. a ühe genoomi sekveneerimiseks võrreldes $\sim$1000 dollariga 2017.a[102]). Selliste tehnoloogiate odavnemine on plahvatuslikult kasvatanud geneetiliste andmete valimimahtu, võimaldades nende ja olemasolevate meditsiiniliste ning fenotüübiliste andmete kombineerimisel paljude tunnuste ning haiguste geneetilist tausta põhjalikult uurida. Kuigi enamike tunnuste ja haiguste geneetiline arhitektuur on veel täielikult avastamata, töötatakse viimastel aastatel pidevalt selle kallal, et olemasolevaid geneetilisi teadmisi juba kuidagi inimeste tervise heaks rakendada.

Kõige uuritumad geneetilise varieeruvuse allikad on ühenukleotiidilised polümorfismid (SNPd). Enamasti on sagedased SNPid üsna väikese mõjuga ning seetõttu ühe SNPi kasutamine mingi tunnuse prognoosimiseks pole mõttekas. Kuid paljude SNPide efektide kombineerimisel saadud tunnus, mida nimetatakse geneetiliseks riskiskooriks, on aga mitmete komplekshaiguste nagu teist tüüpi diabeet, rinnavähk või südame isheemiatõbi geneetilise eelsoodumuse hindamiseks osutunud kasulikuks. Ka algoritmide, mida tavaliselt kasutatakse nimetatud haiguste 10a haigestumistõenäosuse hindamiseks, ennustusvõime on geneetilise riskiskoori lisamisel oluliselt paranenud.

Geneetiline riskiskoor on erinevate SNPide kaalutud alleelidooside summa.Antud töö eesmärgiks oli selgitada, milline võiks olla optimaalne SNPide ning nende kaalude valik sellise skoori moodustamisel. Kaua aega oli kõige populaarsem (ning ka arvutuslikult lihtsaim) viis väheste, ülegenoomselt oluliste SNPide kaasamine ülegenoomsetest assotsiatsiooniuuringutest või suurtest meta-analüüsidest, kas kaalumata alleelidooside summana või siis kasutades kaaludena ülegenoomsetes uuringutes hinnatud regressioonikordajaid. Käesoleva töö eesmärk oli uurida, kas geneetiliste riskiskooride ennustusvõimet saaks parandada, suurendades skoori kaasatud SNPide hulka ja korrigeerides nende kaalusid.

Töös tutvustatakse uut SNPide kaalumismeetodit - topeltkaalumist, kus SNPi kaalu modifitseeritakse vastavalt empiiriliselt hinnatud tõenäosusele, et antud SNP kuulub fikseeritud suurusega SNPide hulka, millel on uuritava tunnusega tegelik seos. Topeltkaalumise eeliseid näidatakse nii simulatsioonide abil kui ka pärisandmete peal, uurides Eesti Geenivaramu andmeteid.

Töös uuritakse ka erinevaid geneetilisi riskiskoore rinnavähile ning leitakse, et mitme erineva geneetilise riskiskoori kombineerimine üheks skooriks aitab geneetilist eelsoodumust rinnavähi jaoks kõige paremini hinnata. Samas tuleb esile, et erinevad geneetilised riskiskoorid, mis haigusega seotud, ei pruugi olla üksteisega korreleeritud ning seetõttu sõltub geneetilise eelsoodumuse hindamine tugevalt geneetilise riskiskoori valikust ega ole ühene hinnang.

Edasi uuritakse töös geneetiliste riskiskooride jaotust erinevates populatsioonides ning leitakse, et üksteisest geneetiliselt kaugel asuvate populatsioonide riskiskooride jaotused on erinevad. Selline teadmine näitab, et geneetilise eelsoodumuse määramisel geneetilise riskiskoori abil ei saa erinevaid populatsioone koos käsitleda. Samuti on hiljuti näidatud, et ühes populatsioonis välja töötatud riskiskooril ei pruugi teises populatsioonis prognoosivõimet olla[97].

Viimaks valideeritakse kolme erinevat mittegeneetilist riskiskoori TÜ Eesti Geenivaramu andmetes. Eestis on peamiseks surmapõhjuseks südameveresoonkonna haigused ning eriliselt murettekitav on see meeste hulgas, Eesmärgiks oli uurida, kas ja kuidas töötavad erinevad tuntud südameveresoonkonna haiguste jaoks mõeldud riskialgoritmid Eesti andmetel. Leidsime, et SCORE ja PCE algoritmid olid hästi kalibereeritud, kuid QRISK2 alahindas riski, prognoosides pea kaks korda vähem ASCVD juhte kui tegelikult tekkis. Erinevate riskiskooridega kaasas käivad ravijuhised erinevad oma statiinide määramise eeskirja pooles. Nende võrdlusel selgus, et NICE (QRISK2) on kõige konservatiivsem ning ESC(SCORE) ravijuhis kõige liberaalsem. Kahjuks soovitasid kõik kolm ravijuhist pea pooltele uuringus osalenud meestele ning veerandile uuringus osalenud naistele statiinide manustamist südameveresoonkonna haiguste riski vähendamiseks, mis näitab, et südameveresoonkonna riskitegurite tasemed on Eesti keskealistel inimestel väga kõrged.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name:	Kristi Läll
Date of birth:	29.01.1989
Nationality:	Estonian
E-mail:	kristi.lall@ut.ee

Education

2013-...	PhD, Mathematical Statistics, Institute of Mathematics and Statistics, University of Tartu, Estonia
2012 Autumn	Exchange student, Master of Biostatistics, Universiteit Hasselt, Belgium
2011-2013	Master's studies, Mathematical Statistics, Faculty of Mathematics and Computer Science, University of Tartu, Estonia
2008–2011	Bachelor's studies, Mathematical Statistics, Faculty of Mathematics and Computer Science, University of Tartu, Estonia
2005–2008	Nõo Secondary School of Science
1996-2005	Aravete Secondary School

Employment

2018-...	University of Tartu, Institute of Genomics, Specialist (0,50)
2015-2016	University of Tartu, Faculty of Medicine, course instructor for 'Epidemiology and biostatistics', biostatistics part
2012	University of Tartu, Faculty of Mathematics and Computer Science, Institute of Mathematical Statistics, course instructor of 'Data Analysis I'
2013	University of Tartu, Faculty of Mathematics and Computer Science, Institute of Mathematical Statistics, course instructor of 'Statistical analysis'
2013-2017	University of Tartu, Estonian Genome Center, data analyst (0,30)
2008–2011	University of Tartu, Faculty of Medicine, Faculty of Medicine Vivarium, lab worker (0.25)

Administrative work

2014-2017	Co-organizer of the course 'Statistical Practice in Epidemiology using R'
2014-2016	Ambassador of the science festival 'Researcher's night'
2014- ...	Member of the Nordic-Baltic Region of the International Biometric Society
2013- ...	Member of the Estonian Statistics Society
2017	Co-organizer of the international conference 'EMGM 2017'

Awards and stipends

2018	Scholarship of the Graduate School in Mathematics and Statistics, University of Tartu
2018	Kristjan Jaak Scholarship for foreign visits
2017	Erasmus + staff mobility (useR!2017)
2017	Kristjan Jaak scholarship for foreign visits
2016	Kristjan Jaak Scholarship for foreign visits
2015	Scholarship of Seitsmenda Samba Fund, University of Tartu
2014	Scholarship of the Graduate School in Mathematics and Statistics, University of Tartu
2012	ESF Dora 7 scholarship for studying in Belgium, Archimedes

Supervised dissertations

- Maia Arge, Master's Diploma, 2016, (sup) Kristi Läll; Pasi Korhonen; Krista Fischer, Generalized Estimating Equations: an overview and application in IndiMed study.
- Anett Tähiste, Bachelor's Diploma, 2014. Maris Alver, Kristi Läll, Andres Metspalu. Risk estimation of coronary artery disease based on genetic markers in Estonian Genome Center cohort.
- Maia Arge, Bachelor's Diploma, 2014. Reedik Mägi, Kristi Läll. Importance of independence of markers in genetic risk scores.
- Kaupo Koppel, Bachelor's Diploma, 2015. Kristi Läll, Silva Kasela. PheWAS in theory and in practise, based on data from Estonian Genome Center, University of Tartu.
- Sille Habakukk, Bachelor's Diploma, 2016. Krista Fischer, Kristi Läll. Estimation of disease risk using genetic markers or family history: a simulation study.
- Merli Mändul, Bachelor's Diploma, 2016. Krista Fischer, Kristi Läll. Effect of genetical feedback to treatment outcome.
- Kalder Maarand, Bachelor's Diploma, 2018. Mare Vähi, Silva Kasela, Kristi Läll. The demand for statistical competence in Estonian businesses.

ELULOOKIRJELDUS

Isikuandmed

Nimi:	Kristi Läll
Sünniaeg:	29.01.1989
Rahvus:	Eestlane
E-post:	kristi.lall@ut.ee

Haridus

2013-...	Doktoriõpe, matemaatiline statistika, loodus-ja täppisteaduste valdkond, Tartu Ülikool, Eesti
2012 Sügis	Vahetussemester magistriõppes, Universiteit Hasselt, Belgia
2011-2013	Magistriõpe, Matemaatiline statistika, Matemaatika-informaatikateaduskond, Tartu Ülikool, Eesti
2008–2011	Bakalaureuseõpe, matemaatiline statistika, Matemaatika-informaatikateaduskond, Tartu Ülikool, Eesti
2005–2008	Nõo Reaalgümnaasium (hõbemedal)
1996-2005	Aravete Keskkool

Teenistuskäik

2018-...	Tartu Ülikool, Tartu Ülikooli genoomika instituut, spetsialist (0.50)
2015-2016	Tartu Ülikool, Arstiteaduskond, Tervishoiu instituut, õppeülesannete täitja aine 'Epidemioloogia ja biostatistika' raames
2012	Tartu Ülikool, Matemaatika-informaatikateaduskond, Matemaatilise statistika instituut, õppeülesannete täitja aines 'Andmeanalüüs I'
2013	Tartu Ülikool, Matemaatika-informaatikateaduskond, Matemaatilise statistika instituut, õppeülesannete täitja aines 'Statistiline analüüs'
2013-2017	Tartu Ülikool, Tartu Ülikooli Eesti Geenivaramu, spetsialist (0,30)
2008–2011	Tartu Ülikool, Arstiteaduskond, Vivaarium, laborant (0.25)

Teadusorganisatsiooniline ja -administratiivne tegevus/kuuluvus

2014-2017	Rahvusvahelise kursuse "Statistical Practice in Epidemiology using R"korraldaja
2014-2016	Teadlaste öö festivali teadussaadik
2014- ...	Rahvusvahekuse Põhja-Balti regiooni biomeetriaühingu 'International Biometric Society' liige
2013- ...	Eesti Statistikaseltsi liige
2017	Rahvusvahelise konverentsi 'Euroopa Matemaatilise Geneetika konverents 2017' kaaskorraldaja

Stipendiumid

2018	Matemaatika ja statistika doktorikooli välisreisi stipendium
2018	Kristjan Jaagu välissõidu stipendium
2017	Erasmus + töötaja väliskoolituse stipendium (useR!2017)
2017	Kristjan Jaagu välissõidu stipendium
2016	Kristjan Jaagu välissõidu stipendium
2015	Tartu Ülikooli Seitsmenda Samba Fondi stipendium
2014	Matemaatika ja statistika doktorikooli välisreisi stipendium
2012	ESF Dora 7 stipendium välismaal õppimiseks, Sihtasutus Archimedes

Juhendatud väitekirjad

- Maia Arge, magistrikraad, 2016. Kristi Läll, Pasi Korhonen, Krista Fischer. Generalized Estimating Equations: an overview and application in IndiMed study (GEE: ülevaade ja rakendamine IndiMedi uuringus).
- Anett Tähiste, bakalaureusekraad, 2014. Maris Alver, Kristi Läll, Andres Metspalu. Geneetilistel markeritel põhinev südame isheemiatõve riski hindamine Eesti Geenivaramu kohordis.
- Maia Arge, bakalaureusekraad, 2014. Reedik Mägi, Kristi Läll. Markerite sõltumatuse olulisus geneetilistes riskiskoorides.
- Kaupo Koppel, bakalaureusekraad, 2015. Kristi Läll, Silva Kasela. PheWAS ja selle praktiline läbiviimine TÜ Eesti Geenivaramu andmete põhjal.
- Sille Habakukk, bakalaureusekraad, 2016. Krista Fischer, Kristi Läll. Geenimarkerite põhjal hinnatud haiguseriski võrdlus perekonnaajalooga.
- Merli Mändul, bakalaureusekraad, 2016. Krista Fischer, Kristi Läll. Geneetilise tagasiside mõju ravitulemusele.
- Kalder Maarand, bakalaureusekraad, 2018. Mare Vähi, Silva Kasela, Kristi Läll. Statistilise kompetentsi vajadus Eestis tegutsevates ettevõtetes.

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.
19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 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.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. Töö kaitsmata.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.
43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.

49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 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.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 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.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 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.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 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.
75. **Nadežda Bazunova.** Differential calculus $d^3 = 0$ on binary and ternary associative algebras. Tartu 2011, 99 p.
76. **Natalja Lepik.** Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Tartu 2011, 133 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.
80. **Marje Johanson.** $M(r, s)$-ideals of compact operators. Tartu 2012, 103 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
82. **Vitali Retšnoi.** Vector fields and Lie group representations. Tartu 2012, 108 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.
85. **Erge Ideon**. Rational spline collocation for boundary value problems. Tartu, 2013, 111 p.
86. **Esta Kägo.** Natural vibrations of elastic stepped plates with cracks. Tartu, 2013, 114 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
88. **Boriss Vlassov.** Optimization of stepped plates in the case of smooth yield surfaces. Tartu, 2013, 104 p.
89. **Elina Safiulina.** Parallel and semiparallel space-like submanifolds of low dimension in pseudo-Euclidean space. Tartu, 2013, 85 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.
93. **Kerli Orav-Puurand.** Central Part Interpolation Schemes for Weakly Singular Integral Equations. Tartu, 2014, 109 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.

95. **Kaido Lätt.** Singular fractional differential equations and cordial Volterra integral operators. Tartu, 2015, 93 p.
96. **Oleg Košik.** Categorical equivalence in algebra. Tartu, 2015, 84 p.
97. **Kati Ain.** Compactness and null sequences defined by ℓ_p spaces. Tartu, 2015, 90 p.
98. **Helle Hallik.** Rational spline histopolation. Tartu, 2015, 100 p.
99. **Johann Langemets.** Geometrical structure in diameter 2 Banach spaces. Tartu, 2015, 132 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.
105. **Md Raknuzzaman.** Noncommutative Galois Extension Approach to Ternary Grassmann Algebra and Graded q-Differential Algebra. Tartu, 2016, 110 p.
106. **Alexander Liyvapuu.** Natural vibrations of elastic stepped arches with cracks. Tartu, 2016, 110 p.
107. **Julia Polikarpus.** Elastic plastic analysis and optimization of axisymmetric plates. Tartu, 2016, 114 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.
115. **Tiina Kraav.** Stability of elastic stepped beams with cracks. Tartu, 2017, 126 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.

117. **Silja Veidenberg.** Lifting bounded approximation properties from Banach spaces to their dual spaces. Tartu, 2017, 112 p.
118. **Liivika Tee.** Stochastic Chain-Ladder Methods in Non-Life Insurance. Tartu, 2017, 110 p.
119. **Ülo Reimaa.** Non-unital Morita equivalence in a bicategorical setting. Tartu, 2017, 86 p.
120. **Rauni Lillemets.** Generating Systems of Sets and Sequences. Tartu, 2017, 181 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.
123. **Kaur Lumiste.** Improving accuracy of survey estimators by using auxiliary information in data collection and estimation stages. Tartu, 2018, 112 p.
124. **Paul Tammo.** Closed maximal regular one-sided ideals in topological algebras. Tartu, 2018, 112 p.
125. **Mart Kals.** Computational and statistical methods for DNA sequencing data analysis and applications in the Estonian Biobank cohort. Tartu, 2018, 174 p.
126. **Annika Krutto.** Empirical Cumulant Function Based Parameter Estimation in Stable Distributions. Tartu, 2019, 140 p.