Transferability of Polygenic Risk Scores Across Ancestral Populations and Data Integration Methods for Improving Prediction on Small Sample Studies

by

Pedro Orozco del Pino

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Professor Sebastian Zöllner, Chair
Associate Professor Philip S. Boonstra
Assistant Professor Jean Morrison
Professor Bhramar Mukherjee
Associate Professor Jennifer Smith

Pedro Orozco del Pino
porozco@umich.edu
ORCID iD: 0000-0002-7373-3702

DEDICATION

To Mariana, thank you for being with me through all days. The good and the bad ones.

ACKNOWLEDGMENTS

It is hard to find the words to express my gratitude. My Ph.D. journey is filled with kind people, and it would be a whole chapter to mention them all. The most important one is, without a doubt, Mariana. She encouraged, supported, struggled, celebrated, and shared with me all the Ph.D. Without her, I would not have even started, let alone could have finished. I would not have done it without her. We have built a beautiful family, and this is just the beginning. Gracias Solecita.

During my Ph.D. I had the fortune to form a large and strong group of mentors who taught me how to be a statistician, a collaborator, a mentor, and a leader. But most importantly, they taught me that you could be all these things remaining a kind person. Thank you, Sebastian, Phil, Bhramar, Jean, Jen, Jeremy, Veera, Laura, Mike, Moussumi, Lu, Ananda, Zhenke, I have the highest respect and appreciation for all of them. Thank you, professors.

All of my family always showed unconditional support, which made me more resilient to the distance that separated us. I want to especially thank my parents, Maria Emilia and Alberto, and my mother-in-law Martha. I also want to thank the rest of my family back in Mexico. My siblings Beto, Emi, and Eduaro. My compadrito Alejandro and his spouse Lore. My Abu Maria Elena, who rests in peace, is my role model of kindness. My Abuelita Milin always made me feel I could do anything whenever she said "desde chiquitito eras bien abusado". My cousins, aunts, and uncles from my three families "Los Orozco", "Los del Pino", and "Los Cuevas". Their intentional and constant rooting made me never forget where I come from.

A special kind of family arises when you live far from your biological family. I want to thank Karl, Tracey, Gonzalo, and Paula for letting us be part of their lives. Our "asados" in the summer were nothing different from hanging out with family on weekends. We shared the feelings of being a graduate student with kids. I thank you for all you shared with me, especially the summer thirst.

My third family, the Mexicarbors. Our small but high-quality community of Mexicans in Ann Arbor. Thank you, Mar, Mau, Javi, y Mariana, for being our link between Mexico and UM.

My friends from the program were a fundamental part of keeping my chin up during (often shared) hard times. When preparing for Quals, studying for finals, preparing talks for ASHG, CSG, and MISSSISISSISSS (get it?). Thank you, Elizabeth, Fatema, Ford, Yichen, Kevin, Yuhua (best desk neighbor), Jionming, Alicia, Andy, Aubrey, Sarah, Emily, Madeline, Nate, Soumik, Irena, Nicky, Meg, Greg, Josh, Adam, and Holly. Friends outside the program helped me to not

take myself too seriously and take life lightly when needed. Thank you, Pedro, Carlo, Gary, Topo, Hector, David, Alets, and Paula. Finally, the staff from the Biostatistics department. Thank you, Nicole, Fatma, Davina, Tara, Sabrina, Mandi, Dan, Mike, Wendy, and Chrissy. Thank you for your outstanding work and for having so much patience with my constant need for guidance.

Thank you. Go Blue!

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AF Allele frequency

AFR African

AMR Admixed American

AUC area under the curve

CV cross validation

DELR data enriched linear regression

DEGLR data enriched generalized linear regression

EAS East Asian

EHR electronic health records

EUR European

GfG Genes for Good

GWAS genome wide association study

HRS Health and Retirement Study

LD linkage disequilibrium

MS most significant variant

MSE mean square error

PP Power Prior

PP-RSS power prior with regression with summary statistics

PRS polygenic risk scores

RCT randomized clinical trials

RSS regression with summary statistics

SAS South Asian

T2D type 2 diabetes

ABSTRACT

Polygenic risk scores (PRS) has been proven to help improve predictive models[66][65][34], build instrumental variables[76][77], study disease etiology [60], and contribute to risk assessment when combined with other diagnostic tools[38]. Thus their importance to research has increased over time. However, PRS has limited transferability across ancestral populations due to differences in linkage disequilibrium (LD), allele frequencies, and environmental exposure. In addition, most of the GWAS samples are of European ancestry, and diversity has not increased in recent years. As a result, a pre-mature inclusion of PRS into clinical practice could increase health disparities across ancestral groups. Currently, there are several projects exist that are very intentional in collecting samples from non-European populations [51][31][82][1]. However, increasing samples of under-represented groups must be done through community inclusion [28] and protecting against commodification[26]. Moreover, even after all the previously mentioned efforts, the over-representation of European samples has not changed significantly up to 2021[23]. Thus, there is a need for methods designed explicitly to improve the prediction and estimation of targeted populations to empower researchers from these communities to leverage the existing data efficiently.

Chapter 2 of this work quantifies the role of different LD structures on a PRS constructed using European genome-wide significant variants. We estimate the change across populations of a European-derived tag variant's predictive ability using extensive simulations with the 1000 Genome Project haplotypes. To isolate the effect of LD, we assume a genetic model with the same effect size of the true underlying risk variant across the five populations of 1000 genomes. Under this scenario, if the most significant variant is not the risk variant, then its predictive ability depends on the LD between the index and true causal variants. In our simulations across the genome, we found that even under an optimistic scenario, the index variant was not the risk variant around 60% of the time. If, most of the time, we are not finding the causal variant, then how much predictive ability do we expect to lose in different populations? Chapter 2 estimates that the reduction in the predictive ability of the most significant variant is modest in Admixed American, South Asian, and East Asian ancestral populations. However, in African populations, the loss of predictive ability can be substantial, reaching up to a 50% reduction in 22% of the time. Finally, in this chapter, we present evidence that suggests that LD score can be informative on the probability of tagging the true risk variant in a region.

We are interested in improving prediction measures such as mean square error (MSE) or area under the curve (AUC) by leveraging an external population with precise but biased information. Generally, when considering an external population as a valuable source of information, it is assumed that any inference that relies on that population is biased at the gain of a reduction in variance. In the context of the transferability of PRS in Chapter 2, we showed that different LD could cause significant bias in the prediction of PRS. Nevertheless, we also showed that this bias could be slight even in genetically distant populations. Chapter 3 proposes a method that dynamically adapts to LD and effect size differences across populations to increase the predictive ability in one of them, called the target population. The method works with GWAS summary statistics from two populations and returns an estimate of the multivariate effect size for a region for a target population. GWAS summary statistics are effect size estimates of the univariate regressions and thus are not comparable across populations with different LD structures. We use the Regression with Summary Statistics to infer the multivariate effect size in each population by using the joint likelihood of the marginal summary statistics. Nevertheless, the multivariate regression in a genomic region is still a miss specified model because it is impossible to include all the possible interactions. When this is the case, the multivariate effect size might differ between populations due to differences in allele frequencies and environmental exposures. To account for this possible scenario, we use a Power Prior to account for the heterogeneity across populations. We simulated GWAS data from European and African populations and showed that our method improves prediction in several measures when the genetic correlation is positive between populations. This method has promising results in increasing the predictive ability of PRS for populations where the sample size of the existing GWAS sample size is limited.

In non-genetics scenarios, there is extensive literature on leveraging existing studies to improve estimation or prediction in one study. Chapter 4 extends the Data Enriched Linear Regression[15] to generalized linear regression link functions. We show that the objective function of DELR is equivalent to the objective function of penalized regression, which means we can use existing software to obtain estimates. However, penalized regression does not differentiate between target and external data sources, and thus it requires a different algorithm to find the best penalty. We develop a Cross-Validation algorithm to find the penalty factor that would optimize prediction in the target population. Furthermore, we show through simulations that DEGLR improves prediction when bias is small and converges to ignoring the external study as bias increases. In a real data analysis, the Health and Retirement Study is our target data source, and the Genes for Good study is our external study. We use these data sources to explore the ability to increase the predictive ability of PRS as a covariate when the proportion of White participants is much higher in the external data source than in the target. We systematically split the HRS data into small training sets and increased the sample size gradually. From this analysis, we see that as the HRS's training

sample size increases, DEGLR adapts the weight of GfG to optimize the predictive ability. When the ratio of the synthetic GfG is large, the DEGLR uses the GfG data and increases the predictive ability of HRS. As this ratio decreases, the DEGLR method uses less GfG data and matches the predictive ability of HRS alone.

CHAPTER 1

Introduction

The world is more connected than ever, and the potential benefits to science are immense. Rapid technological advances produce large amounts of complex, personalized, and sensitive biomedical data. One example is the accelerated growth of genetic data. Since 2001 the cost of sequencing the human genome has gone from a hundred million to a thousand dollars [36]. Because of this, the amount of genetic studies has increased constantly. In particular, the number of genome wide association study (GWAS) and their sample sizes[10] has increased dramatically over the last two decades.

GWAS calculates the marginal association of an SNP with a trait across millions of SNPs. One application of GWAS is to build polygenic risk scores (PRS) with the resulting associations. To build PRS, we aggregate the highly associated variants from GWAS[17] to obtain a single number containing a trait's genetic information. Researchers use PRS to improve predictive models[66][65][38][34], build instrumental variables[76][77], and study disease etiology[60], and to contribute to risk assessment when combined with other diagnostic tools [39][24][38].

However, PRS has limited utility by the lack of diversity in the underlying GWAS data[22]. In addition, most of the samples are of European ancestry, and diversity has not increased in recent years [49][58][23]. A European over-representation in GWAS samples is a concern because many studies show that European-based PRS have a reduced predictive ability in non-European populations. For example, Curtis [19] showed a PRS build for schizophrenia based on a mainly European GWAS shown to be more powerful in predicting ancestry than schizophrenia. Martin et al.[49] found a lack of transferability of single-ancestry GWAS in eight well-studied phenotypes, such as height and type 2 diabetes. Mars et al. [48] found much less predictive ability of PRS in African populations for type 2 diabetes, coronary artery disease, and breast and prostate cancer when using six biobanks accounted for one million people altogether. As a result, a pre-mature inclusion of PRS into clinical practice could increase health disparities across ancestral groups.

The factors why PRS do not transfer across ancestral populations are: First, different LD structures will cause indexing SNPs to be less valuable surrogates for causal SNPs in different populations [11]. Second, true marginal effect sizes may differ substantially across ancestral populations

as different environments benefit from different interactions [82][57]. Third, differences in allele frequencies between populations may change the proportion of variance explained by a marker [37]. As a result of these factors, the predictive power of index variants decreases in populations genetically distant from the population of the GWAS. Understanding these factors allows the development of methods that can construct PRS that are more generalizable across ancestral populations.

Chapter 2 of this work quantifies the role of different LD structures on a PRS constructed using European genome-wide significant variants. To isolate the effect of LD, we assume a genetic model with the same effect size of the true underlying risk variant across the five populations of 1000 genomes [4]. For each variant with allele frequency greater than 0.1 in the European haplotypes of 1000 genomes [4], we sequentially simulated high-power European GWAS assuming that variant was the only risk variant. We select the variant with the lowest p-value as the index variant to use as a predictor in the four non-European populations from the 1000 genomes [4]. If the index variant is not the risk variant, then its predictive ability depends on the LD between the index and true causal variants. Since the initial (discovery) GWAS had exclusively European samples, we expect high LD in the European population. Thus, the LD will typically be lower in any non-European population, which means the predictive ability of the index variant will also be lower. In Chapter 2, we calculate the expected proportion of times for this event to happen across the genome and how much loss of predictive ability we should expect in Admixed American, South Asian, East Asian, and African populations.

Statistical power is a function of typically unknown quantities except for sample size. A simulation approach allows us to control all unknown parameters of statistical power and study them appropriately. In our simulations across the genome, we found that even under an optimistic scenario, the index variant was not the risk variant around 60% of the time. Even in the simulations where the power of a genome-wide association of the risk variant was above 0.99, the proportion of risk variants being index variants remained above 50%, which means that most of the GWAS discoveries are highly correlated variants with the causal variant. Thus, this first result highlights the difficulty of overcoming complex LD structures through increasing sample size.

If, most of the time, we are not finding the causal variant, then how much predictive ability do we expect to lose in different populations? When the index variant is not causal, its expected predictive ability is approximately the correlation with the causal variant times the effect size of the causal variant[59]. Chapter 2 estimates the correlation of the most significant variant with the true causal variant across the genome. We found a modest impact in Admixed American, South Asian, and East Asian ancestral populations. However, in African populations, the loss of predictive ability can be substantial, reaching up to a 50% reduction in 22% of the time. We compared the genome-wide distribution of the reduced predictive ability with the distribution using GWAS

catalog hits [10] from studies with only European samples and found no significant difference, suggesting our estimates transfer well into sites that are GWAS hits. Finally, we found that being in the lowest decile of the LD score of the risk variant significantly increases the probability of being the index variant. In practice, we do not know the causal variant, but we found that the index variant's LD score strongly correlates ($r^2 = 0.99$) with the LD score of the risk variant. This association gives a practical way to assess the probability of an index variant being causal. If the index variant is in the first decile of the LD score, it has a much higher probability of being the risk variant.

These conclusions are robust to assumptions about true effect size and sample size choices for simulations and that changing the power of the initial GWAS only has a modest effect on the probability of identifying the true risk variant and, thus, the transferability of the index variant. Chapter 2 shows that differences in LD across ancestral populations can largely explain the lack of transferability of PRS. Methods and frameworks that intend to increase the transferability of PRS should emphasize modeling LD differences carefully.

There are many attempts to improve the predictive ability of a PRS built with variants discovered with European GWAS in non-European populations. This research topic is currently very active, and most methods integrate large sample sizes of European ancestry with a similar or lower sample size of different ancestry. Marquez-Luna et al. [53] developed a method that leverages large sample size European studies to improve the prediction of much lower sample size studies of a target population by optimizing the linear combination of the single ancestry PRSs. Marnetto et al. [47] proposed a method that combines single ancestry PRS from different populations based on local ancestry inference, using the ancestrally closest PRS in each region. Coram et al. [18] proposed to declare index variants with trans-ethnic GWAS and estimate the effect with ethnic-specific GWAS. Grinde et al. [29] evaluated different approaches for selecting SNPs and estimating effect sizes using European and Latino samples. They found that selecting European variants works well for some traits but less for others. While estimating the effect sizes using the much smaller sample size Latino data set or a combination of both data sets was always better. These methods assume that underlying effect sizes are likely to be shared and use ancestral inference to differentiate populations. Nevertheless, do not model the differences in LD explicitly. These methods show that transferability improves when ancestral differences are acknowledged. However, all these methods consider ancestry as categorical information.

The statistical problem of using a different population to increase the efficiency of an estimator falls on the bias-variance trade-off. Generally, when considering an external population as a valuable source of information, it is assumed that any inference that relies on that population is biased at the gain of a reduction in variance. The objective overall is that reducing variance without a significant bias increase can decrease the estimand's means square error. In the context of the

transferability of PRS in Chapter 2, we showed that different LD could cause significant bias in the prediction of PRS. Nevertheless, we also showed that this bias could be small even in genetically distant populations.

Chapter 3 proposes a method that dynamically adapts to LD and effect size differences across populations to increase the predictive ability in one of them, called the target population. The method takes GWAS summary statistics from two populations and returns an estimate of the multivariate effect size for a region. GWAS summary statistics are effect size estimates of the univariate regressions and thus are not comparable across populations with different LD structures. We use the regression with summary statistics (RSS) to infer the multivariate effect size in each population. Nevertheless, the multivariate regression in a genomic region is still a miss specified model because it does not include all the possible interactions. When this is the case, if allele frequencies are different and or environmental exposures differ across populations, the multivariate effect size will differ between populations. To account for this possible scenario, we use a Power Prior (PP) to adjust for the heterogeneity across populations due to unadjusted interactions. We simulated GWAS data from European and African populations and showed that our method improves prediction in several measures when the genetic correlation is positive between populations. This method has promising results in increasing the predictive ability of PRS for populations where the sample size of the existing GWAS sample size is limited.

In non-genetics scenarios, there is extensive literature on leveraging existing studies to improve estimation or prediction in one study. Multi-task learning integrates several small tasks to increase prediction by leveraging common characteristics and preserving each task's unique characteristics [30][21]. Another example is transfer learning, which considers the setting of several data sources that are rich in sample size from which we can obtain information for a target source when they are "close"[70][73]. Several methods exist to leverage external summary statistics on a subset of the covariates of the target data set to improve estimation [7] [16] [69] [14]. However, the objective of these methods is not precisely the one we are looking at in this dissertation. Chen, Owen, and Shi [15] develop the data enriched linear regression (DELR) to leverage one potentially biased external source to improve the prediction of a target data source in the linear regression setting. However, most of the outcomes in biomedical data would can not be modelled with the linear regression and generalized linear models are instead preferred.

Chapter 4 develops data enriched generalized linear regression (DEGLR) to extend the DELR to several links. In this chapter, we show that the objective function of DELR is equivalent to the objective function of penalized regression, which means we can use existing software to obtain estimates. The *glmnet* R package is the state-of-the-art software to fit penalized linear regression. However, it does not differentiate between target and external data sources, and thus it requires a different algorithm to find the best penalty. We develop a cross validation (CV) algorithm to find

the penalty factor that would optimize prediction in the target population. Our effect size estimates match those using DELR (with a theoretical way to find the best penalty) and the DEGLR with the Gaussian link function. Furthermore, we show through simulations that DEGLR improves prediction when bias is small and converges to ignoring the external study as bias increases.

In our real data analysis, we have access to the Health and Retirement Study (HRS) as our target data source. This study is a longitudinal panel representative of the 50 years and older US population. For this study, we have access to genotyped data, so we construct a PRS for type 2 diabetes (T2D), and we want to evaluate the PRS's ability to predict the disease's status, adjusting for age, race, education, BMI, and sex. We use the Genes for Good (GfG) study as an external data source. The GfG participants are volunteers that answer questionnaires on a Facebook app. With access to genotype data, we construct a PRS for T2D and have access to the same covariates as for HRS. GfG participants are primarily White, while for HRS, White participants represent 64%. We assume then that there are two sources of bias in GfG, the selection bias, and the different ancestry distribution. However, the sample size of GfG is in the same order as HRS. The assumption of our method is to have a much larger sample size on the external data source. To match the assumptions of DEGLR, we create a large synthetic GfG data by sampling with replacement from GfG. We also systematically split the HRS data into training and testing data. From this analysis, we see that as the training sample size increases, DEGLR adapts the weight of GfG to optimize the predictive ability. When the ratio of the synthetic GfG is large, the DEGLR uses the GfG data and increases the predictive ability of HRS. As this ratio decreases, the DEGLR method uses less GfG data and matches the predictive ability of HRS alone.

In the final chapter, we discuss future work for the three chapters. We discuss the importance of utilizing external data sources to benefit a targeted population as a tool to aid underrepresented populations in exploiting existing studies for their benefit.

CHAPTER 2

The Role of Linkage Disequilibrium in the Change of Predictive Ability of Polygenic Risk Scores Across Populations

2.1 Introduction

Polygenic risk scores (PRS) quantify an individual's genetic risk for a given phenotype as a weighted sum of risk alleles associated with this phenotype [56]. The list of included variants and their weights are typically obtained from genome-wide association studies (GWAS)[17]. PRS are leveraged for improving prediction models [66][65][38][34], as instrumental variables for Mendelian randomization [76][77], and understanding the etiology of related traits [60], among others. Moreover, there is an ongoing debate about including PRS in clinical practice and management [40][5][46][72][43][42].

As the importance of PRS is increasing, their utility is limited by the lack of diversity in the underlying GWAS data[22]: Since 2014 only 20% of all GWAS studied non-European populations and only 4% study populations of African or Native American ancestry [58]; this proportion was constant at least up to 2019 [49]. Many studies show that such under-representation of non-European populations reduces the predictive power of PRS in non-European populations: Curtis [19] showed a PRS build for schizophrenia based on a mainly European GWAS shown to be more powerful to predict ancestry than schizophrenia in a sample of 38,131 cases and 114,674 controls, while Martin et al.[49] found a lack of transferability of single-ancestry GWAS from Europeans to Africans, South Asians, East Asians and Americans in eight well-studied phenotypes, such as height and type 2 diabetes. Moreover, Reisberg et al. [61] showed that the distribution of a European built PRS of coronary heart disease changes significantly across populations meaning that the cut off for risk assessment for European populations is not applicable in other populations, especially in African populations. Similarly, Mars et al. [48] found much less predictive ability of PRS in African populations for type 2 diabetes, coronary artery disease and breast and prostate

cancer when using six biobanks that accounted for one million people all together. As PRS may become a part of clinical care, this lack of transferability limits the benefit of this new approach to underrepresented populations thus aggravating health disparities [5]. It is critical to understand the causes of this lack of transferability and to identify strategies to ameliorate them.

Fundamentally, the lack of transferability reflects that variants identified by a European GWAS have a different effect on population risk in non-European populations than in Europeans. This difference can be explained by three biological factors: First, variants may have a different effect in different populations. Typically this is conceptualized as variants acting in different environments across populations. As GWAS effect sizes estimate the marginal effect size of a variant across all interactions within their environment, changing populations may change environments and thus change this marginal effect size and result in variants being more or less predictive for a phenotype [82]. Second, as allele frequencies differ between populations, the change in the genotype's variance will modify the proportion of variance explained by the variant. [37]. Third, LD differs between populations. As the variant causing the functional effect is unknown, PRS typically select the most significant variant in a region (index variant) as the predictor to be included in the statistic. If the index variant is not the functional variant, prediction accuracy depends on the LD between the index and the functional variant [11].

As a result of these factors the predictive power of index variants decreases in populations genetically distant to the population of the GWAS. Moreover, the distribution of the PRS differs between populations thus predictive models need to be adjusted accordingly [61]. While a range of statistical methods have been developed to overcome this lack of transferability [53][47][18][29][79][62], there is still a significant gap in the transferability of PRS.

Optimizing this transferability requires understanding the relative contribution of these three factors. Bitarello and Mathieson [6]have shown that changes in Allele frequency (AF) are unlikely to explain more than 8% of the changes in effect size. However, broad statements about similar effect sizes across all populations and all phenotypes are not possible as the change in marginal effect sizes likely depends on the unknown underlying biology of each trait that is not amendable to modelling[57]. Moreover, differences in estimated effect sizes may just reflect that the index variant differs from the true risk variant and changes in LD between populations cause changes in marginal effect size. On the other hand, the change in LD between populations is well understood, though the impact in transferability is complex and very challenging to model[78][64]. As LD structure differs across the genome and different parts of the genome are relevant to different traits, it is important to assess the whole genome in a systematic way to measure the change in predictive ability of PRS due to different LD structure. Here we use computer simulations to quantify the role of LD in the transferability of PRS based on resampling haplotypes from 1000 genomes [4]. We sequentially consider each variant with minor allele frequency (MAF)> 0.1 as risk variant and

simulate a high-powered European GWAS to identify this variant. From this GWAS, we use the most significantly associated variant (index variant) as predictor for the trait and estimate the effect size of this predictor in four non-European populations. If the initial index variant is not the true risk variant, this effect size depends on its LD with the causal variant. To effectively isolate the effect of LD, we assume that the effect size of the true underlying risk variant is the same across populations. Under this model, the predictive ability of a risk variant can only differ between populations, if the initial index variant is not the true risk variant.

In our simulations across the genome, we found that in 61.1% of simulations, the risk variant is not the index variant. Thus, changes in LD result in an average reduction of 25% in the effect size of the index variant in African populations. For South Asian, American, and East Asian populations, the average reduction is 7.3%, 7.9%, and 9%, respectively. Consistent with empirical observations [50][48][25], loss of information is limited for East Asian, South Asian, and Admixed American populations and substantial for African populations. We compared the genome wide distribution of this statistic with the distribution of European GWAS hits from the GWAS catalog and predict that 22% of the European GWAS hits have a reduced predictive ability of at least 50% in African ancestry. These conclusions are robust to assumptions about true effect size and sample size choices for simulations. We varied true effect size and sample size in repeated simulation in chromosome 22 and observed consistent results with our genome-wide simulations with one true effect size. Furthermore, with this set of simulations in chromosome 22 observed that changing the power of the initial GWAS only has a modest effect on the probability of identifying the true risk variant and thus the transferability of the index variant. Thus, our work shows that much of the observed lack of transferability can be explained by the differences in LD between populations. Strategies to overcome this lack of transferability thus should account for uncertainty in identifying the true underlying causal variant.

2.2 Methods

We assess the transferability of polygenic risk scores by simulating genome-wide association studies in a European population and then estimating the effect of the most significant variant (MS) in other populations. We sequentially consider every autosomal variant with European allele frequency (AF) between 0.1 and 0.5 as causal variant affecting a binary trait, simulate a GWAS to identify this variant, assess the most significant (MS) variant in that GWAS and estimate the effect size of that MS variant in other populations assuming that the underlying risk variant has the same effect in all populations. Thus we estimate the transferability of the causal variant.

2.2.1 Genetic model

We assume a binary phenotype Y where 1 indicates affected status and 0 indicates unaffected status. We assume a single causal variant G_c, let g denote the genotype of the risk allele of that causal variant. We assume the following data generating mechanism for a population p.

$$\text{logit}(P(Y = 1|G_c = g)) = \beta_{0p} + \beta g, \tag{2.1}$$

where $\text{logit}(x) = log(x)/log(1 - x)$. We chose β_{0p} based on the allele frequency of G_c in population p to ensure the same prevalence in all populations. The prevalence and the value of β is always the same for all populations.

2.2.2 Simulation Algorithm

For each autosomal variant G_c with European AF between 0.1 and 0.5, we simulate a dataset of cases and controls assuming that G_c is a functional variant. We then test all variants within 40kb for association with the trait. To generate cases and controls, we sample with replacement haplotypes from the Phase 3 release of 1000 Genome project [4]. Samples thus have LD structures that are very similar to the reference sample. To ensure a preset number of cases and controls, we sample haplotypes conditional on their affection status. The probability of being sampled for each haplotype, conditional on disease status is provided by Equation 2.3.

$$P(D_i|H_i = h) = P(D_i|H = h)\frac{P(H_i = h)}{P(D_i)} \tag{2.2}$$

$$= P(D_i|G_{ci} = g)\frac{P(H_i = h)}{P(D_i)}$$

$$= \left(\frac{e^{\beta_o+g\beta}}{1+e^{\beta_0+g\beta}}\frac{P(H_i = h)}{P(D_i)}\right)^{D_i}\left(\frac{1}{1+e^{\beta_0+g\beta}}\frac{P(H_i = h)}{P(D_i)}\right)^{1-D_i}, \tag{2.3}$$

where h is a vector with dosages of a haplotype, G_{ih} is the dosage of the risk variant in haplotype h of individual i, and θ is a vector that is zero for all non risk variants and β for the risk variant. The conditional probability of the disease status given the haplotype only depends on the genotype of the causal variant $P(D_i|H_i = h) = P(D_i|G_{ci} = g)$.

We simulate cases by setting $Y_i = 1$, and sampling with replacement n_A diploid cases. We simulate controls by setting $Y_i = 0$ and sampling with replacement n_U diploid controls. After generating the case control data set we test for association at all polymorphic variants within 100kb of the risk variant.

2.2.3 Simulation procedure

For each variant with an allele frequency between 0.1 and 0.5 in the EUR population, we performed the following simulation pipeline.

1. Simulate n_A =30,000 cases and n_U =30,000 controls with European haplotypes in a window of 200kb centered in the causal variant.

2. Test each polymorphic SNP in the haplotype region for significant association with Z-tests.

* Stop if there is no genome-wide significant variant (p-value $<$ 5e-8) and move to the next variant.

3. Identify the variant with the lowest p-value as the MS variant. If more than one variant are tied with the lowest p-value because $r^2 = 1$ between them, choose one at random.

4. Estimate the effect size of the MS variant in the simulated European by running a logistic regression.

5. Simulate 30,000 cases and 30,000 controls for each of the super populations, including a second sample of EUR, assuming the same causal variant as in step 1.

6. Estimate selected statistics with each of the newly simulated data sets, including the new European data set.

We run the simulation procedure one time per variant in all autosomal chromosomes except chromosome 22. In the case of chromosome 22 we run the simulation procedure 20 times per variant.

2.2.4 GWAS catalog analysis

We download the files *all associations v1.0.2* and *all ancestry data v1.0* from the GWAS catalog on August 26, 2022. The first file contains all associations in the GWAS catalog and the second contains the ancestry of the samples for each study. From this files we kept associations that came from studies that only had European ancestry samples according to the field *BROAD.ANCESTRAL.CATEGORY* from *all ancestry data v1.0*, a p value less than 5e-8 according to the field *P.VALUE* from *all associations v1.0.2*, an allele frequency of the risk variant between 0.1 and 0.9 according to the field *RISK.ALLELE.FREQUENCY* from *all associations v1.0.2*, and that did not have *NA* in the field *OR.or.BETA* from *all associations v1.0.2*. We then matched by the field *SNPS* from *all associations v1.0.2* with our simulations from chromosome 22 to obtain an estimate of ρ_* for each population.

2.2.5 Ldscore and probability of the MS variant being causal

We obtained the ldscore of all variants from chromosome 22 in the European population. We took the ldscore of a variant as the sum of pairwise correlations with that variant that were equal or greater than 0.1. We used *plink v1.90b6.2* to obtain a table with all r^2 of the variants and used *R* to aggregate the results. We then used the repeated simulations from chromosome 22 to obtain a point estimate of the probability of the MS variant to be the risk variant for each variant with and AF between 0.1 and 0.5 in the European population. We then obtained the ldscore for the MS variant in each of the repeated simulations and aggregate it across all 20 repetitions for each variant.

2.2.6 Measures of transferability

The expected predictive ability of a variant G_j's GWAS effect size estimate is approximately $\beta\rho_j$ [59], where $\rho_j = \text{cor}(G_c, G_j)$. As we assume the same β for all populations, the difference in predictive ability of a variant G_j across populations is determined by the difference in ρ_j across populations. Let G_* be the most significant variant in a European GWAS, then $\rho_* = \text{cor}(G_c, G_*)$. Through extensive simulations we estimate ρ_*, assuming model 2.1 holds for each of the five super populations of the 1000 Genome project [4]: African (African (AFR)), American (Admixed American (AMR)), East Asian (East Asian (EAS)), European (European (EUR)), and South Asian (South Asian (SAS)). Thus, we estimate ρ_*^{AFR}, ρ_*^{AMR}, ρ_*^{EAS}, ρ_*^{EUR}, and ρ_*^{SAS}. We indicate with

the population's abbreviation as a sub index to indicate the population used to estimate statistics in the non-discovery data. The statistics that we estimate with the simulations are:

a) Using the European discovery data set we calculate the probability that the MS variant is the causal variant.

b) For Population $j \in \{\text{AFR,AMR,EAS,EUR,SAS}\}$ using simulated data not used for testing: $\hat{\beta}_{jc}$ effect size estimates of the causal SNP, $\hat{\beta}_{j*}$ effect size estimates of the index SNP, $\hat{\rho}_{j*} = cor_j(\hat{\beta}_{jc}, \hat{\beta}_{j*})$ correlation between the most significant variant and the causal variant.

c) For Population $j \in \{\text{AFR,AMR,EAS,EUR,SAS}\}$ using simulated data not used for testing for association we calculate the relative estimated effect size of the index variant $\hat{\beta}_{j*}/\beta$.

d) A column to indicate one of four possible outcomes: the causal variant was the most significant, a non causal variant was the most significant, the causal variant was the only variant genome wide significant, and at least one non causal variant that was perfectly correlated with the causal variant was the most significant. When we encounter a set of multiple variants with pairwise LD $r^2 = 1$ being the most significant, we choose the MS variant at random to be the index variant.

From the previous data, we will calculate two statistics, which we will use to measure the transferability of PRS. The correlation coefficient between the MS variant and the causal variant $\hat{\rho}_{j*}$, and the probability of the MS variant being the causal variant α. The first statistic estimates the change in the predictive ability of an index variant. The second statistic estimates the proportion of times the index SNP has the same predictive ability in all populations.

2.3 Results

We simulated discovery GWAS studies using European haplotypes from the 1000 Genome Project, modeling each variant with minor AF >0.1 as a risk variant (5,482,942 variants total) with a 1.1 odds ratio. A total of 4,929,437 (90%) simulations had at least one genome-wide significant variant. We then asses the ability of this GWAS's most significant (MS) variant in the vicinity of that risk variant to predict the phenotype as part of a PRS across European, South Asian, Admixed American East Asian, and African, populations, assuming that the marginal effect of the true risk variant is unchanged. Here, we have to consider two data configurations: (1) either the MS variant is the underlying risk variant. Then it is equally predictive in all populations. (2) The MS variant is not the risk variant, but in LD with it. Then the ability of the MS variant to predict disease risk changes between populations as this LD changes.

2.3.1 Identifying the true risk variant

Across the genome, we observed that in 25.7% of simulations, the MS variant was the causal variant while in 61.1% of simulations, the MS is not the causal variant. In the remaining 13.2% of simulations, a set of multiple variants with perfect pairwise LD ($r^2 = 1$) in 1000 Genomes, including the true risk variant were equally most significant. For the following analyses, we choose the predictive variant at random from this set of perfectly correlated variants. Thus, in 67.7% of all simulations, the predictive variant is a nearby tag SNP whose effect size depends on the LD between the MS SNP and the risk variant.

2.3.2 Correlation between MS variant and risk variant

The loss of informativeness of the MS variant is determined by the correlation between the MS variant and the risk variant ρ_*[59]. This correlation averages $\rho_* = 1$ for the 32% of cases where the MS variant is the risk variant and the LD (r^2) between the MS variant and the risk variant for the 68% of cases where they differ. In Europeans, the mean $\hat{\rho}_*$ is 0.96 and 95% of MS variants have high LD ($\hat{\rho}_* > 0.79$) with the causal variant Figure 2.1. In the the South Asian, Admixed American, East Asian and African population, the mean (median) $\hat{\rho}_*$ is 0.93 (0.99), 0.92 (0.98), 0.91 (0.99) and 0.75 (0.93). However, this distribution of $\hat{\rho}_*$ has a long tail, in South Asian, Admixed American, East Asian and African population, 95% of MS variants have a $\hat{\rho}_* > 0.61$, 0.63, 0.45, and 0.09.

If we consider $\hat{\rho}_* < 0.9$ as a meaningful loss of information, we see that for 48% of variants, $\hat{\rho}_*$ is below this threshold in the African population, a considerably larger proportion than in Europeans (13%), South Asian (21%), Admixed Americans (25%), and East Asians (22%) (Table

2.1). Variants with $\hat{\rho}_* < 0.5$ lose at least half their predictive power when transferred between populations. Almost one quarter of variants (24%) showed this pattern in the African population while it was uncommon ($\leq 6\%$) in the other populations.

To assess if known GWAS hits have different patterns of transferability than the genome average, we calculate the distribution of $\hat{\rho}_*$ from our genome wide simulations for 109,843 GWAS hits across 5,761 traits from the GWAS catalog (see Methods for details). We also evaluated this distribution individually for 9 diseases with > 90 significant findings. Across all GWAS hits, the distribution of $\hat{\rho}_*$ was very similar to the distribution of the entire genome(Table 2.1) with a mean $\hat{\rho}_*$ of 0.97 for European, 0.93 for South Asian, 0.93 for Admixed American, 0.91 for East Asian, and 0.77 for Africans. The same pattern was observed across all individual traits with mean $\hat{\rho}_*$ ranging from 0.95 to 0.97 for Europeans, 0.92 to 0.94 for South Asians, 0.90 to 0.94 for Admixed Americans, 0.89 to 0.93 for East Asians, and 0.74 to 0.80 for Africans (Figure2.4).

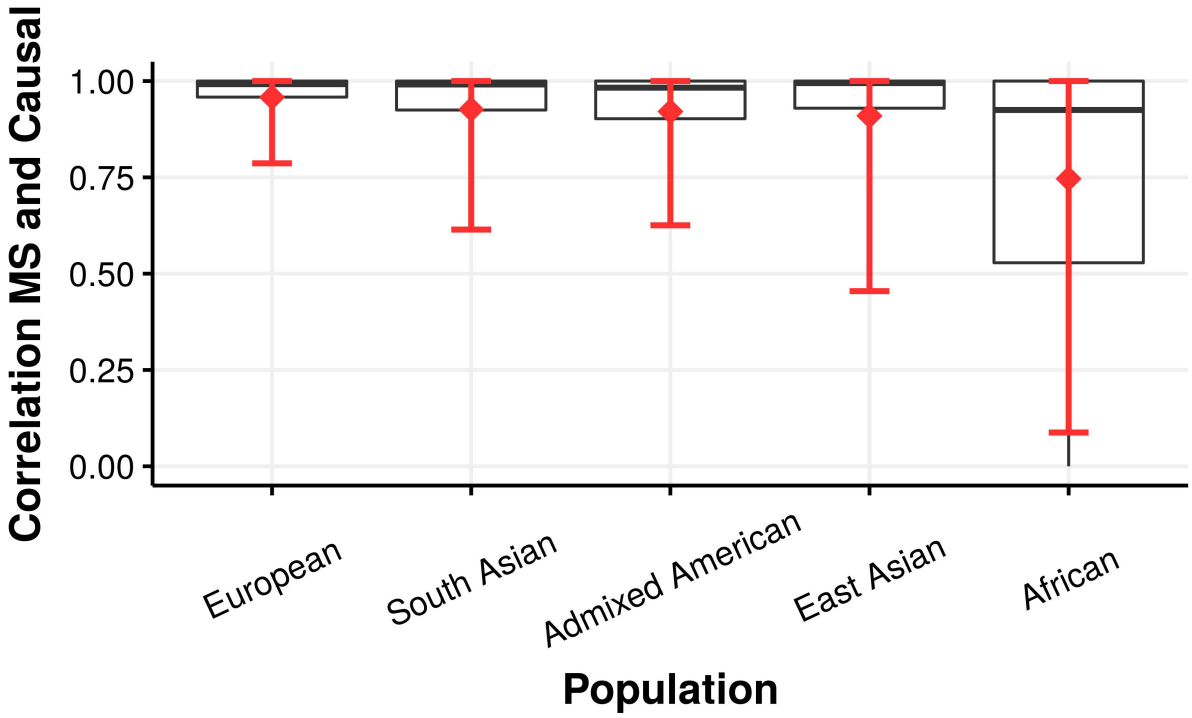

Figure 2.1: Boxplot of the squared correlation between the most significant and the causal variant for each population. The red interval represents the 0.05 and 0.95 quantiles and the red diamond is the mean.

Genome wide results					
Population	$\hat{\rho}_* < 0.99$	$\hat{\rho}_* < 0.95$	$\hat{\rho}_* < 0.90$	$\hat{\rho}_* < 0.75$	$\hat{\rho}_* < 0.50$
European	0.46	0.22	0.13	0.04	0.005
South Asian	0.49	0.30	0.21	0.10	0.03
Admixed American	0.55	0.36	0.25	0.10	0.03
East Asian	0.45	0.29	0.22	0.12	0.06
African	0.62	0.53	0.48	0.37	0.24
GWAS hits results					
Population	$\hat{\rho}_* < 0.99$	$\hat{\rho}_* < 0.95$	$\hat{\rho}_* < 0.90$	$\hat{\rho}_* < 0.75$	$\hat{\rho}_* < 0.50$
European	0.41	0.21	0.12	0.04	0.006
South Asian	0.44	0.29	0.20	0.10	0.03
Admixed American	0.47	0.32	0.22	0.10	0.03
East Asian	0.41	0.27	0.20	0.12	0.06
African	0.52	0.45	0.41	0.32	0.22

Table 2.1: Proportion of times the correlation of the most significant variant with the causal variant is below the threshold indicated in the columns. In the top table we present results using simulations across all predictive variants. In the table below we present the simulation results for variants that are GWAS hits from the GWAS catalog.

2.3.3 Distribution of estimated effect sizes

The differences in $\hat{\rho}_*$ translate directly into differences of informativeness for estimates of effect sizes of the MS variant: Aggregate across all variants, the mean (median) relative estimated effect size for Europeans was 0.97 (0.97) , for South Asians 0.96 (0.97), for Admixed Americans 0.95 (0.96), for East Asians 0.94 (0.96), and for Africans and 0.83 (0.88) (Table2.2). To understand the change of the estimated effect size across populations, we stratified simulations on whether the MS was the causal variant, the MS was not the causal variant, and the MS was among multiple markers in perfect LD with the causal variant (Figure 2.2). As expected under our model, there is no difference in estimated effect size if the MS variant is also the functional variant. For variants where MS was not the causal variant, the mean (median) relative estimated effect size was 0.96 (0.96) for Europeans, 0.93 (0.94) for South Asians, 0.93 (0.94) for Admixed Americans, 0.91 (0.94) for East Asians, and 0.75 (0.80) for Africans (Table 2.2). For a few variants, the effect size of MS changes signs between populations, this occurred for 0.09% of variants in the South Asians, for 0.03% in the Americans, for 1.9% in the East Asians, and for 2.6% in the African sample. If multiple variants are in perfect LD with the causal variant and one of these is MS, the mean (median) relative estimated effect size was 1.00 (1.00) for Europeans, 0.99 (0.99) for South Asians, Admixed Americans, East Asians, and 0.93 (0.95) for Africans.

To quantify the impact of these differences on a PRS composed of many risk variants, we constructed a set of PRS that had 1, 10, 20, and 30 causal variants with allele frequency between

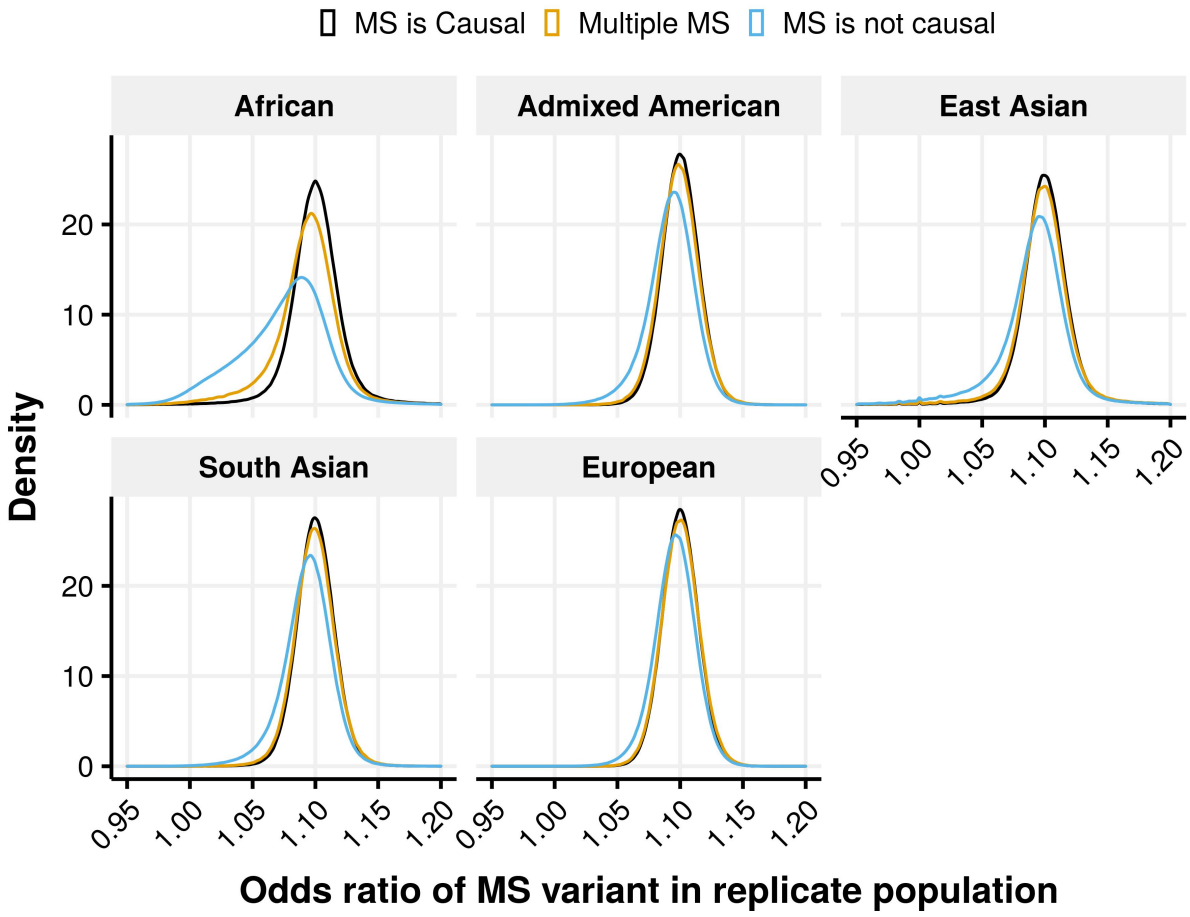

Figure 2.2: Distribution of estimated odds ratio of the most significant variant in the replicate population stratified by the cases in which the most significant variant is not the causal variant (blue), most significant variant is the causal variant (red) and the causal variant has $r^2 = 1$ with another variant and both are the most significant (green). Vertical lines show the median of each distribution.

All simulations					
	European	**South Asian**	**Admixed American**	**East Asian**	**African**
Mean	0.97	0.96	0.95	0.94	0.83
Median	0.97	0.97	0.96	0.96	0.88
Simulations were MS is not causal					
	European	**South Asian**	**Admixed American**	**East Asian**	**African**
Mean	0.96	0.93	0.93	0.91	0.75
Median	0.96	0.94	0.94	0.94	0.80

Table 2.2: Mean and median of the estimated effect size divided by the true effect size across the five populations for all predictive variants. The top table is for all simulations, the bottom table is for simulations in which the MS variant was not the causal variant.

0.1 and 0.5 in the European population by sampling this number of variants from our simulation and calculating PRS based on the effect sizes of the MS variants. We evaluated these PRS by calculating the difference in prevalence between the individuals with a PRS in the bottom 1% and individuals with a PRS in the top 99%. In this model based on 10 variants, we calculate a mean relative risk (RR) of 1.81 in the European sample, 1.76 in the South Asian and Admixed American sample, 1.70 in the East Asian Sample and 1.57 in the African sample. Increasing the number of predictive variants increased the relative risks. When normalizing each population's RR by the RR in Europeans, we see that the relative difference between populations is similar across the number of risk variants simulated A.1. American and South Asian populations have normalized RR of 0.97, East Asian samples have normalized RR of 0.92 and a African samples have normalized RR of 0.83 relative to the European RR.

2.3.4 Local LD patterns and study power modify transferability

To assess if local LD patterns allow predicting whether the MS variant is the true risk variant, we calculated the LD score in the European population for every variant and correlated it with the probability of tagging the true variant, observing a modest correlation ($r^2 = 0.17$) (Figure A.1). In practice, the true risk variant is unknown and we can only assess the LD-score of the MS variant. However, the LD-score of the true risk variant and the LD-score of the MS variant are highly correlated ($r^2 = 0.99$) and thus the LD-score of the MS variant and the probability of MS being the true risk variant is also $r^2 = 0.17$ (Figure A.2). Thus the LD patterns of the MS variant provide provide some evidence for our ability to identify the functional variant. For variants in the lowest decile of LD-scores (LD score< 16) the mean (median) probability of the MS variant being the functional variant is 0.33 (0.21). For all other deciles, the mean (median) probability

is below 0.14 (0.05). In the highest decile of the LD-score distribution (LD score$>$ 193), the probability of tagging the right variant is 0.04 (0.0). To evaluate the impact of allele frequency of the causal variant, we correlated $\hat{\rho}_*$ with minor allele frequency across markers and observe a positive correlation with MAF (Figure 2.3). In the lowest frequency quantile, the mean $\hat{\rho}_*$ ranges from 0.937 in the European population to 0.669 in the African population, while in the highest frequency quantile, the mean $\hat{\rho}_*$ ranges from 0.966 in the European population to 0.786 in the African population. However, MAF and power of the discovery study are confounded in our initial study design. Variants in the lowest allele frequency bin had a mean power of 0.75, while variants in the highest allele frequency bin had a mean power of 0.99996.

We performed extra simulations varying allele frequency and expected power (See Methods) to untangle this relationship and assess which one of the two factors was primarily driving the results above. In Figure A.4 (see Appendix) we show that power increases $\hat{\rho}_*$ at every bin for all populations. In the case of African populations the impact of power is more notable as it increases the mean $\hat{\rho}_*$ from 0.65 for a power between 0.2 and 0.4 to a mean $\hat{\rho}_*$ of 0.75 for a power above 0.9. On the other hand, the mean $\hat{\rho}_*$ has almost no change for different intervals of allele frequency of the causal variant. Except for African populations in which we can observe a mild change in mean $\hat{\rho}_*$ when the allele frequency of the causal variant is between 0.4 and 0.5.

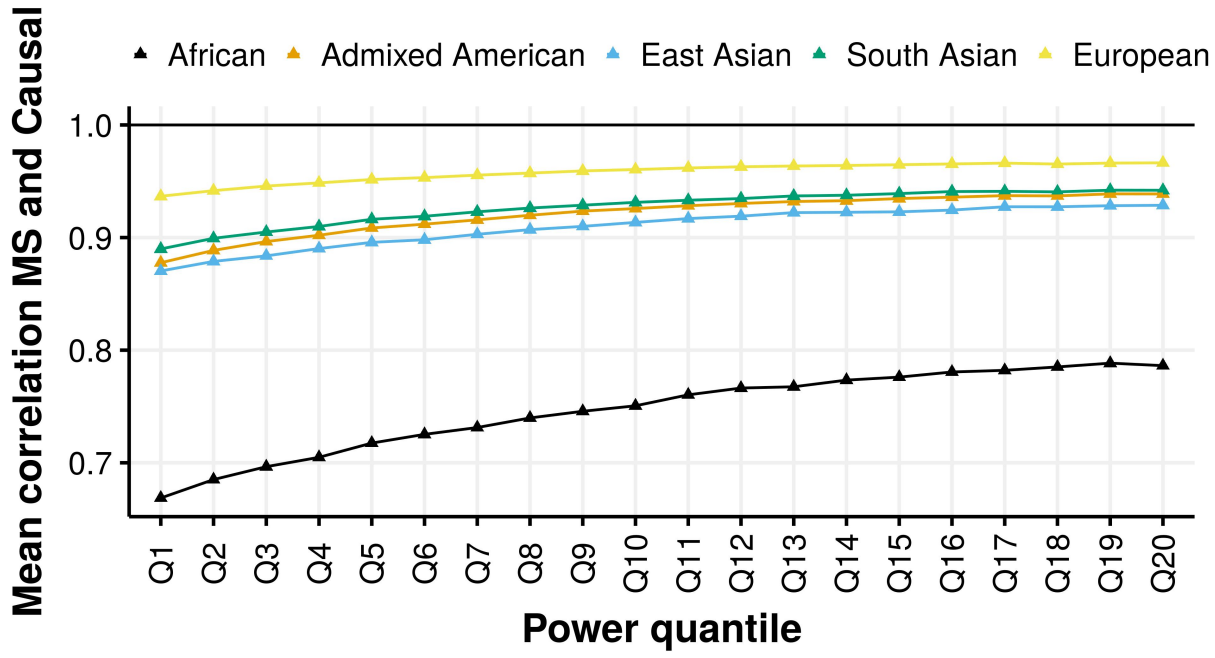

Figure 2.3: Relationship between power and the correlation between the most significant and the causal variant. Each bin in the x-axis has 5% of all simulations and the y-axis has the average correlation between the MS and the causal variant for each population.

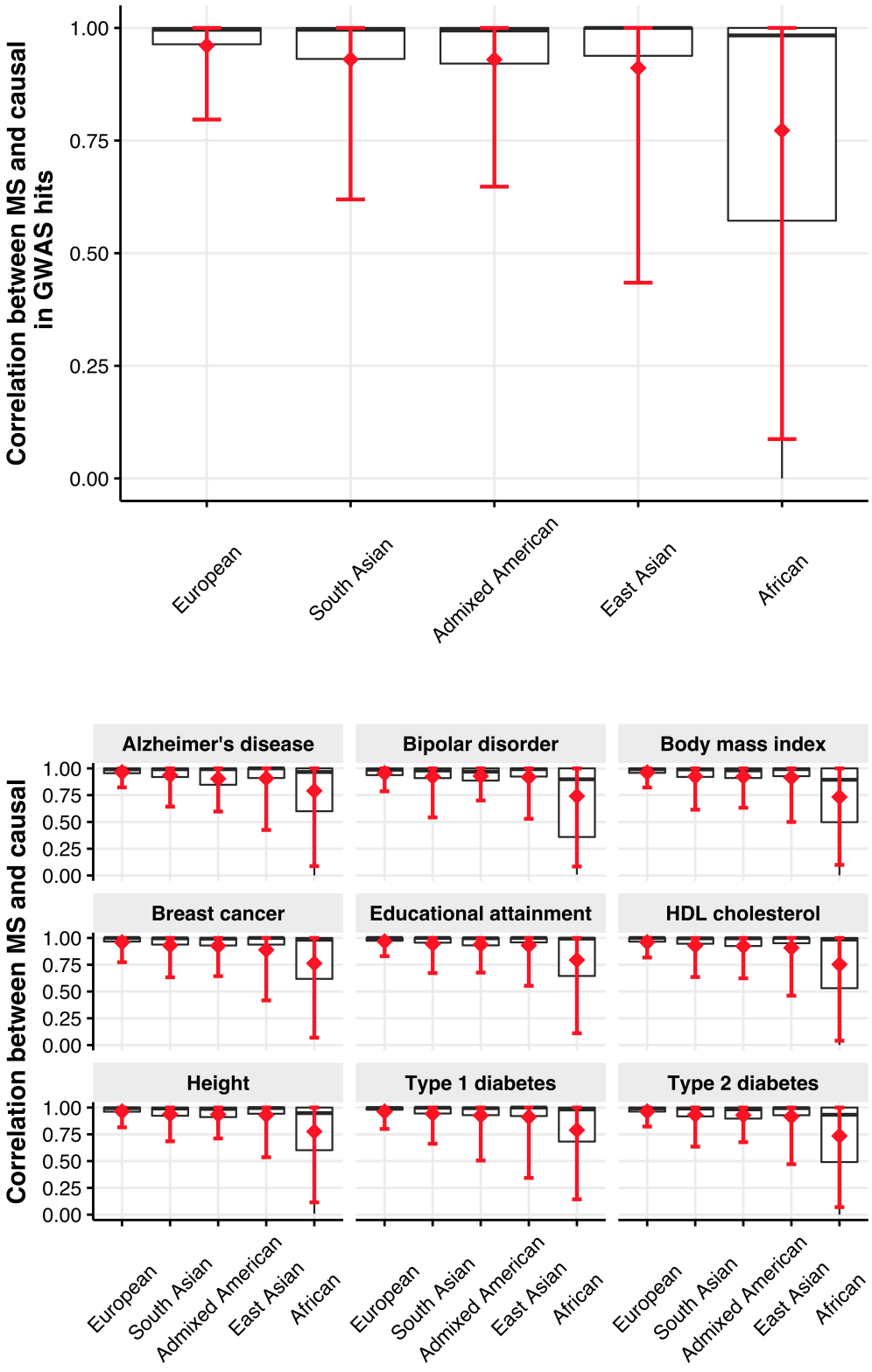

Figure 2.4: Distribution of the correlation of the most significant variant with the causal variant among GWAS hits from studies with samples of European ancestry. In the top plot we present all GWAS hits , in the bottom plot we present the distribution for some selected traits.

2.4 Discussion

While PRS increase in importance in biomedical research, it is now well established that these statistics are less predictive in non-European populations. Here, we assessed how differences between populations patterns of LD affect transferability of PRS by simulating large discovery GWAS in Europeans, modeling each variant with minor AF > 0.1 as a potential risk variant. We identify the most significant (MS) variant in the vicinity of that risk variant and considered this MS variant as the best estimate for the underlying risk variant. We then assess the predictive ability of the MS variant in African, East Asian, South Asian, and Admixed American populations assuming the same risk variants acts with the same effect size in all populations. Under this scenario, the predictive ability of a GWAS finding only differs between populations when the MS variant is not the true risk variant. The magnitude of the difference then depends on the difference in LD patterns between populations. Across all common variants in the genome, the MS variant was the true risk variant in only 32.3% of simulations, even though our discovery GWAS had power > 0.75. As a result, the MS variant lost predictive power in populations other than the European population. Across all simulations, the mean loss of predictive power was around 8%, in East Asians, Admixed Americans and South Asians populations, which have LD patterns similar to the European discovery population [4]. The mean predictive power of the MS variant is reduced by 25% in the African population, consistent with African LD patterns being less similar to European LD patterns with overall lower levels of LD in Africa [71]. To assess if GWAS hits were different than random common markers in the genome, we focused on 109,843 variants that had been identified as genome-wide significant in the GWAS catalogue [10] and observed no difference in the simulated mean change in effect size across populations between the set of all common variants and the set of GWAS hits. We also assessed, if patterns of LD in the discovery population provide insight in the change of effect size and saw that the LD score of the most significant variant is correlated with its probability of being causal ($r^2 = 0.17$) Interestingly, it is sufficient to calculate the LD score of the MS variant as it is very highly correlated with the LD score of the true risk variant. To evaluate how these single-variant results would affect a PRS constructed out of many variants, we constructed PRS from 10, 20 or 30 simulated variants and calculated the relative risk between PRS deciles. The informativeness of these PRS not only depended on changes in LD, but also changes in allele frequency. For these PRS, we observed that the relative risk between the lowest decile and the highest decile of the PRS decreases by a factor of 0.97 in South Asians and Admixed Americans, a factor of 0.92 in Each Asians, and a factor of 0.83 African samples, regardless of the number of variants included in the PRS. These the pattern of reduction in effect size very comparable to the patterns observed in comparative studies of PRS informativeness; Martin et al. [50] showed predictive accuracy across 17 traits to be 1.6 fold lower in American and South

Asian samples, 2.0 fold lower in East Asian samples, and 4.5 fold lower in African samples. Veturi et al. [74] estimated the correlation of effect sizes for the MS variants for height to be between 0.5 and 0.75 between European Americans and African Americans. Under a model where these differences in effect size are driven by LD differences, this result is consistent with our result that the mean correlation between the mean correlation between the MS variant and the functional variant is 0.75 in an African population.

Thus it seems plausible that differences in LD between populations are the mayor driver for the lack of transferability of PRS. This hypothesis is supported by several empirical studies Shi et al. [67], stratified variants across nine traits by posterior probability after fine mapping and showed that variants with high likelihood of correctly identifying the functional variant, had highly correlated effect sizes between European and East Asian populations. Cavazos [12] showed that using Africans as a discovery population can increase transferability of a PRS.

While some of these effects could also be driven by changes in allele frequency, Bitarello and Mathieson [6] estimate that differences in the site frequency spectrum explain at most, 8% of the decrease in partial r^2 in non-European samples. Hou et all [32] recently showed that in admixed individuals, effect sizes are similar across ancestries. Patel et al [57] showed that genetic interactions are the drive a large portion of the heterogeneity of causal effect sizes.

If we consider that changes in LD are the major driver for differences in transferability, how can we use this model to improve transferability? Our results indicate that the transferability of a marker modestly increases with the power of the initial discovery study, as increasing the power increase the probability of the true risk variant being the MS variant. However, the impact of this approach is limited: even for studies with power > 0.99, risk variants still had their effect size reduced by 30% in African samples. This is not surprising, in many regions of the genome have many markers in very tight LD in Europeans and each of those variants has about the same probability of being the MS variant. Thus the increasing sample sizes of European GWAS will not suffice to overcome the problem of reduced transferability[35]. Functional annotation may improve fine-mapping, but so far, many attempts to annotate GWAS hits, which are typically regulatory, has been challenging [52]. Diverse samples have also been leveraged to improve fine-mapping [82]. Here we have two approaches: (1) combining samples and generating effect size estimates from the combined sample or (2) using the combined sample to fine-map the risk variant. Note that if the goal is to identify the true functional variant, we don't necessarily need a sample from the population we want to calculate the PRS in, including diverse samples with different patterns of LD will improve the ability of identifying the true functional variant.

Without specifically focussing on fine-mapping, combining samples across diverse populations has been shown to improve PRS across many traits [12][62][53]. Not only can such studies provide more precise estimates of effect sizes, they can also help identify variants that are of low frequency

in Europeans. However, how to optimally combine data across multiple populations when some population samples are much bigger than others without drowning out the smaller populations is an open question.

We obtained these conclusions using a relatively simplistic PRS consisting of variants that are genome-wide significant in the discovery GWAS. Other methods aim to accumulate evidence across variants with less evidence for association while carefully modelling the LD in the discovery population[75][62].However, considering complex methods that jointly analyze the genome is not possible in the systematic fashion employed here where equal weight is given to every variant. In practice, PRS derived from such methods is not substantially different from a PRS based on GWAS : Bench marking several methods, Kulm et al. [41] found little to no improvement in power and transferability when using complex methods over more straightforward methods. Cecile et al. [13] found that using genome-wide significant variants in PRS has more clinical utility and the same predictive ability than using millions of variants as in LDpred. Moreover, it seems unlikely that the role of LD differences between populations is smaller in methods that leverage detailed models of LD to derive the best predictor.

In summary, our simulations of common variants a cross the genome estimate the transferability of European-derived PRS under the assumption that the genetic architecture of the trait is the same in all populations, thus providing an upper bound for this transferability. Our results indicate that differences in LD between population is likely a major driver for the performance of PRS. We discuss several strategies for improving transferability under this scenario and show that even if the genetic architecture is identical across populations, we can expect substantive loss of information when applying European-derived instruments to non-European populations, which can only be avoided by including substantial non-European samples in future analyses.

CHAPTER 3

Increasing the Prediction Ability of Polygenic Risk Scores in a Target Low Sample GWAS by an Adaptive Integration of Large Sample GWAS of an External Population

3.1 Introduction

PRS assess an individual's genetic risk for a phenotype calculated as a weighted sum of risk variants obtained from GWAS[17]. They are leveraged for improving prediction models [66][65][38][34], as instrumental variables for Mendelian randomization [76][77], and understanding the etiology of related traits [60], among others. For example, studies use PRS as a surrogate variable of the genetic contribution of a phenotype. In clinical research, PRS can be used to investigate the increase in the risk of early onset of a disease under different environmental factors. Aas et al. [2] investigated the genetic load of Schizophrenia and cannabis use before onset by comparing Schizophrenia PRS between people that use cannabis and people who do not. Thus, in combination with other sources of information, PRS will contribute to the risk assessment of complex diseases as it becomes better at capturing the genetic contribution.

One advantage of PRS is that researchers can build them using publicly available GWAS summary statistics[17]. Thus allowing several studies to leverage existing GWAS at the same time without compromising privacy of participants. The down side is that PRSs are limited because most samples used in GWAS are of European ancestry[22]. In 2016 only 20% of all GWAS analyzed non-European populations, with 16% being of Asian ancestry [58]. This proportion was constant since 2014 and at least up to 2019 [49]. Even multi-ethnic GWAS have much larger European sample sizes than any other population [48][44][45]. Such under-representation of non-European populations has been shown in many studies to reduce the predictive power of PRS in non-European populations. Curtis [19] showed a PRS build for Schizophrenia based on a mainly

European GWAS shown to be more powerful in predicting ancestry than Schizophrenia. Martin et al.[49] found a lack of transferability of single-ancestry GWAS in eight well-studied phenotypes, such as height and type 2 diabetes. The lack of predictive ability of European GWAS in other populations means that research about populations genetically distant from the European population are much less benefited by the massive amount of publicly available GWAS.

Some studies have shown that GWAS associations from large sample populations can be used in other populations to increase their predictive ability. Marquez-Luna et al. [53] developed a method that linearly combines PRS from two populations and shows that increased the over all predictive ability in both populations. Marnetto et al. [47] propose a method that uses PRS from different populations based on local ancestry inference. Grindle et al. [29] found that the optimal way to use meta-analysis to pick the best tag SNP between a European GWAS and Hispanic Latinos from the HCHS/SOL was the trait and sample size of the discovery GWAS dependent. Coram et al. [18][22] propose a method that declares a tag SNP in trans-ethnic GWAS and estimates its effect with ethnic-specific GWAS. Weissbrod et al. [80] developed a method that leverages functional data and GWAS from multiple populations to improve the generalizability of PRS. Ruan et al. [63] developed a method that uses population-specific parameters of a continuous shrinkage prior to increasing the predictive of a PRS in multiple populations. Most of these methods rely upon access to data beyond publicly available summary statistics.

Multi-ethnic PRS methods focus on increasing the predictive ability of PRS in all populations involved in the calculation of the PRS. While this is the end goal of a multi-ethnic PRS, the objective of our method is to focus on increasing the predictive ability of the PRS in a target population with a small or moderate sample size. The proposed method leverages existing huge sample GWAS from a different population, generally of European ancestry, that we call the external GWAS. We rely on the external GWAS's large sample to detect the small effect sizes that small samples GWAS would not have the power to detect. However, different environmental exposures and gene-gene interactions across populations will bias the effect size estimates. Cavazos and Witte [12] showed that the bias of European estimates in effect size estimates increases as genetic divergence increases. We are willing to accept biased estimates of the effect sizes in return for decreasing the estimates' variance and thus increasing the predictive ability of the resulting PRS.

Some parts of the genome might be more similar than others across populations, meaning the bias's magnitude and direction might differ across regions in the genome. Weighting regions differently is crucial for leveraging the external GWAS without diluting the target GWAS population-specific information. We propose to use a Power Prior [33] model that integrates an external source of information by modeling the effect size heterogeneity across populations. The Power Prior is a Bayesian model that gives more weight to the external GWAS when data suggests that populations are similar and gives less weight otherwise. This approach uses the European data as part

of the prior and regulates the strength of the European summary statistics in the analysis based on heterogeneity through a parameter called *power parameter*.

The original Power Prior is used with complete data, which we assume is unavailable. Thus we use the Regression with Summary Statistics likelihood, which enables us to infer the multivariate effect sizes from marginal summary statistics and allows us to assume the same likelihood for the target and external population. We show that the Power Prior PRS increases the concordance correlation with the phenotype, increases the heritability explained by the PRS, and reduces the mean squared error of the causal variants in comparison to using only the target GWAS, only the external GWAS, or using external GWAS as prior without weighting. We show that under a different effect size scenario, the optimal weight is strictly below one and above zero, which supports the idea of dynamic weight across regions. The Power Prior with Regression with Summary Statistics (PP-RSS) is a robust and powerful method that considers heterogeneity and different LD structures to construct PRS based on multivariate effect size estimates. PP-RSS integrates GWAS summary statistics from two populations with significantly different sample sizes and genetic architecture that emphasizes the low sample population to avoid diluting population-specific regions.

3.2 Methods

In our method we assume that we only have access to GWAS summary statistics from both population. Because of LD structure the marginal associations from GWAS can be directly integrated when the external and the target population are have different LD structure. We infer the joint effect size from the marginal association using the regression with summary statistics (RSS) likelihood. However, the joint effect size might be different due to different environment exposures. We use the power prior that can adaptively weight the external likelihood to account for differences in the joint effect size. This section is organized in the following way. In subsection 3.2.1 we describe the RSS likelihood for one population. Subsection 3.2.2 describes how to construct the power prior when we have access to data from two populations that assume a similar likelihood. Subsection 3.2.3 puts together the idea of using the RSS likelihood and power prior. Finally, in subsection 3.2.4 we describe how to choose the tuning power parameter to optimize the prediction on the target population.

3.2.1 The Regression with Summary Statistics (RSS) likelihood

Suppose that we have summary statistics from a Genome Wide Association Study, this is effect sizes and standard deviation of the single variant regressions. We assume the GWAS had individuals from a single ancestral population; hence we use data source and population interchangeably. In the rest of the document we parametrize Gaussian distributions with mean and precision; thus $N(x; \mu, \Omega)$ is the density of a Gaussian distribution with mean vector μ, precision matrix Ω evaluated at x.

Consider the following multivariate regression model.

$$Y = X\beta + \epsilon, \tag{3.1}$$

where $\epsilon \sim N(0, \phi \mathrm{I})$ is a $n \times 1$ vector of independent errors and ϕ is a scalar precision parameter, $Y \in \mathbb{R}^n$ is the standardized phenotype, $X \in \mathbb{R}^{n \times m}$ is a centered genotype matrix information at a region with m variants.

Assuming known ϕ, the likelihood for β is $L(\beta|Y, X, \phi) = N(Y; X\beta, \phi\mathbb{I})$. However, we do not have access to individual-level data. Instead, we have access to summary statistics from the m univariate regressions $Y = \alpha_j X_j$, $j = 1, ..., m$ from a GWAS. Let $(\hat{\alpha}_j, \hat{\sigma}_j^2)$ be the OLS estimator of α_j and the OLS estimator o the variance of $\hat{\alpha}_j$ respectively. Let $s_j = sd(\hat{\beta}_j)$ be the standard deviation of the j element of the estimated multivariate effect size. Then $S = diag(s)$ can be estimated with $\hat{S} = diag(\hat{s}_1, ..., \hat{s}_m)$ with $\hat{s}_j^2 = \hat{\sigma}_j^2 + \hat{\alpha}_j^2/n$. Finally let R be the LD matrix of the genotype matrix of the GWAS X and $\hat{R}$ be the LD or correlation matrix estimated with a reference

panel. The Regression with Summary Statistics (RSS) likelihood proposed by Zhu & Stephens [85] is a likelihood of the multivariate effect sizes β conditional on marginal summary statistics $\hat{\alpha}, \hat{S}, \hat{R}$.

The RSS likelihood of β is

$$L(\beta|\hat{\alpha}, \hat{S}, \hat{R}) = N(\hat{\alpha}; \hat{S}\hat{R}\hat{S}^{-1}\beta, (\hat{S}\hat{R}\hat{S})^{-1}), \tag{3.2}$$

where $\hat{S}\hat{R}\hat{S}^{-1}$ is a matrix that maps the joint effect sizes to the expected effect sizes of marginal models and $(\hat{S}\hat{R}\hat{S})^{-1}$ is the precision matrix of the model. Zhu & Stephens proved that the RSS likelihood is proportional to the multivariate likelihood from model 3.1 we would obtain with individual level data, this inference about β depends only on summary data $\hat{\alpha}, \hat{S}, \hat{R}$, in other words:

$$L(\beta|Y, X, \phi) \propto L(\beta|\hat{\alpha}, \hat{S}, \hat{R}). \tag{3.3}$$

3.2.2 Power Prior

Let D_t and D_e be the data for a target and external population respectively. We assume that the external and target populations have the same underlying model that is indexed by a parameter β. Let likelihoods $L(\beta|D_t)$ and $L(\beta|D_t)$ be the corresponding likelihood for the parameter β. An important assumption is that the sample size of D_e is much larger than that of D_t.

Given data D_e from the external source and a parameter of interest β, the power prior is defined as follows.

$$P(\beta|a_0, D_e, b) \propto L(\beta|D_e)^{a_0}\pi_0(\beta|b) \tag{3.4}$$

where $a_0 \in [0, 1]$, π_0 is the initial prior of β, b is a set of hyper parameters for π_0.

The power prior is a family of priors to perform Bayesian inference when an external data source is available. Its unique characteristic is the inclusion of the *power parameter* a_0 which weights the influence of the external population. When $a_0 = 1$, the power prior is equivalent to a Bayesian update with the full use of the external data source information. In the case of a_0, the power prior is equivalent to ignoring the external data source. The power prior's strength resides in the flexibility to regulate the influence that the external population will have in the inference on β through a sensible choice of a_0. In essence the power parameter is an scale parameter that regulates the tails of the external likelihood to increase or decrease the importance of the external data.

Equation 3.4 assumes a fixed value of a_0. Such value has to correspond with the heterogeneity of effect sizes among the external data D_e and the target data D_t. A way to find a_0 is to define a

function $f(a_0)$ that has a monotonic behaviour with the predictive ability in the target population. A candidate for a_0 would be such that optimizes $f(a_0)$.

Another option is to assume a prior distribution on a_0. In that case, Equation 3.4 is no longer a proper prior[54]. In order to have an adequate prior we need to divide by the appropriate constant $c(a_0) = 1/\int L(\beta|D_e)^{a_0}\pi_0(\beta|b)d\beta$. Thus the power prior with $\pi(a_0|\gamma)$ as a prior for a_0 is:

$$P(\beta, a_0|D_e, b, \gamma) \propto L(\beta|D_e)^{a_0}c(a_0)\pi(a_0|\gamma)\pi_0(\beta|b), \tag{3.5}$$

where γ are the hyper parameters for $\pi(a_0|\gamma)$.

Finally, the power prior from Equation 3.4 or Equation 3.5 is used to construct the posterior distribution of β conditional on the target data, external data and the hyper parameters. Thus the posterior distribution of β is

$$P(\beta, a_0|D_t, D_e, b, \gamma) \propto L(\beta|D_t)L(\beta|D_e)^{a_0}c(a_0)\pi(a_0|\gamma)\pi_0(\beta|b). \tag{3.6}$$

3.2.3 Using RSS likelihood and the power prior to integrate GWAS from different populations

Let $D_t = (Y_t, X_t)$ and $D_e = (Y_e, X_e)$ be the target and external data sources when we have access to individual level data. We assume multivariate model 3.1 for the target and external population. Thus, a natural way to integrate both studies is through a Bayesian approach in which $L(\beta|D_t)$ and $L(\beta|D_e)$ is the target and external contribution to the posterior of β. However, in genetic data model 3.1 is most likely to be miss-specified due to unaccounted gene gene and gene environment interactions. This interactions will impact the expected value of $\hat{\beta}$ in each population creating heterogeneity across the populations that depends on allele frequencies and population average environment exposures. The we could use posterior 3.6 with likelihoods $L(\beta|D_t) = N(Y_t; X_t\beta, \phi_t\mathbb{I})$ and $L(\beta|D_e) = N(Y_e; X_e\beta, \phi_e\mathbb{I})$ to model this heterogeneity through a_0.

However, we do not have access to these likelihoods because we do not have access to individual-level data in either of the two populations. In addition to the bias introduced through the miss specification of model 3.1 we know that our GWAS summary statistics will differ in expectation because of different LD structures. The expected value of the GWAS estimated effect sizes depend on the pairwise correlation with all the variants times their multivariate effect size, this is $E[\hat{\alpha}_j] = s_j \sum_{i=1}^{m} r_{ij}\beta_i/s_i$. Where r_{ij} is the correlation between variant i and j and s_i is the variance of the variant. Thus, if we want to use external data from a different population than the target, we have to model LD to account for different s_j's and r_{ij}'s across populations.

Let $\tilde{D}_t = (\hat{\sigma}_t, \hat{\alpha}_t, \hat{S}_t, R_t)$, $\tilde{D}_e = (\hat{\sigma}_e, \hat{\alpha}_e, \hat{S}_e, R_e)$ be the target and external summary statistics

respectively. Instead of the multivariate likelihoods from model 3.1 we use the RSS likelihood 3.2. Thus, the posterior distribution of β is conditional on the target and external summary statistics, and the power parameter is:

$$P(\beta_t|a_0, \tilde{D}_t, \tilde{D}_e) = N(\hat{\alpha}_t; \hat{\theta}_t\beta, \hat{\eta}_e)N(\hat{\alpha}_e; \hat{\theta}_e\beta, \hat{\eta}_e)^{a_0}N(\beta; 0, \tau\mathbb{I}). \tag{3.7}$$

where $\hat{\eta}_t = (\hat{S}_t R_t \hat{S}_t)^{-1}$, $\theta_t = \hat{S}_t R_t \hat{S}_t^{-1}$, $\hat{\eta}_e = (\hat{S}_e R_e \hat{S}_e)^{-1}$, and $\theta_e = \hat{S}_e R_e \hat{S}_e^{-1}$ with their respective population suffixes. We chose a ridge prior for β for two reasons. The first reason is that we want to evaluate the method with the least influence of a prior possible, this is easily attainable with a ridge prior with small τ. The second reason is that a ridge prior makes computation very efficient. Here we denote R_e, and R_t as known, which means they are calculated with the same data as the GWAS. In practice, we can estimate R for each population with reference panels. Conditional on a_0 as in 3.7 we obtain an analytical expression for the posterior distribution, this is

$$\begin{aligned} P(\beta|a_0, \tilde{D}_t, \tilde{D}_e) &\propto exp\left(-\frac{1}{2}\left((\hat{\alpha}_t - \theta_t\beta)^\top \eta_t(\hat{\alpha}_t - \theta_t\beta) + a_0(\hat{\alpha}_e - \theta_e\beta)^\top \eta_e(\hat{\alpha}_e - \theta_e\beta) + \tau\beta^\top\beta\right)\right) \\ &\propto exp\left(-\frac{1}{2}\left(\beta^\top(\theta_t\eta_t\theta_t + a_0\theta_e\eta_e\theta_e + \tau\mathbb{I})\beta - 2\beta^\top(\theta_t\eta_t\hat{\alpha}_t + a_0\theta_e\eta_e\hat{\alpha}_e)\right)\right) \\ &\propto exp\left(-\frac{1}{2}\left((\hat{\beta}_{te} - \beta)^\top \eta_{te}(\hat{\beta}_{te} - \beta)\right)\right), \end{aligned} \tag{3.8}$$

where $\nu_{te} = \theta_t\eta_t\theta_t + a_0\theta_e\eta_e\theta_e + \tau\mathbb{I}$ and $\hat{\beta}_{te} = \nu_{te}^{-1}(\theta_t\eta_t\hat{\alpha}_t + a_0\theta_e\eta_e\hat{\alpha}_e)$ the posterior mean (see Appendix for an expression in terms of $\hat{S}$ and R). Finally, we construct a PRS as the standardized predicted phenotype for the genetic information X_t^*, i.e. $PRS_i = X_i^*\hat{\beta}_{te}$ (see section 3.2.5).

3.2.4 Choice for the power parameter a_0

The objective of the power parameter is to optimize the weight of the external information for the prediction in the target population. From posterior 3.8 we obtained the posterior mean $\hat{\beta}_{te}$, as a function of a_0. The predicted value of the phenotype in the target population $\hat{Y}_t(a_0) = X_t\hat{\beta}_{te}$ is also a function of a_0. Thus, we define $f(a_0)$ (mentioned above) as a metric of the prediction ability of $\hat{Y}_t(a_0)$ to predict Y_t. The final objective of this project is to propose $f(a_0)$ that do not depends on having any individual level training data. Consider the following four candidates for $f(a_0)$.

1. Pearson correlation between phenotype and predicted phenotype.

$$f(a_0; D_t^+) = cor\left(Y_t, \hat{Y}_t(a_0)\right) \tag{3.9}$$

2. Sums of squares of the phenotype and predicted phenotype.

$$f(a_0; D_t^+) = ||Y_t - \hat{Y}_t(a_0)||_2^2 \tag{3.10}$$

3. Cross validation of mean square error with K folds

$$f(a_0; D_t^+) = CV_K(||Y_t - \hat{Y}_t(a_0)||_2^2) \tag{3.11}$$

4. Pseudo correlation

$$f(a_0) = \frac{\hat{\beta}_{te}^\top \text{diag}(X_t^{*\top} X_t^*)\hat{\alpha}_t}{\hat{\beta}_{te}^\top X_t^{*\top} X_t^* \hat{\beta}_{te}} \tag{3.12}$$

Where D_t^+ is an individual level training data, and X_t^* is the reference panel of the target population used to obtain $\hat{\beta}_{te}$. Note that $f(a_0)$ 1. and 2. will not provide the same estimate because a_0 appears in the mean and the variance of the posterior distribution.

In this project want to compare the performance of optimizing the pseudo correlation function against the performance of optimizing any of the other three functions that depend on training data. To measure performance we will use several metrics that are all related to prediction accuracy in different ways. We use a testing data set $D^{test} = (Y_t^{test}, X_t^{test})$, to evaluate the performance. The first metric is the relative concordance correlation between $\hat{Y}_t^{test}(a_0)$ and Y_t^{test}. Concordance correlation of two vectors measures the linear agreement between them, this is a useful metric of predictive ability for our simulations because the linear data generating mechanism. We divide the concordance correlation of the prediction by the concordance correlation of the true genetic contribution $X_t^{test}\beta$.

We describe in more detail the expression for the Pseudo correlation objective function $f(a_0)$. The intuition of the function is to write the correlation between the phenotype Y_t and the predicted phenotype $\hat{Y}_t(a_0)$ to find an expression that does not depend on individual-level data. First note that since Y_t is centered then

$$cor(\hat{Y}_t(a_0), Y_t) = \frac{\hat{\beta}_{te}^\top X_t^\top Y_t}{\sqrt{\hat{\beta}_{te}^\top X_t^\top X_t \hat{\beta}_{te} Y^\top Y_t}}. \tag{3.13}$$

$$\tag{3.14}$$

Then note that $diag(X_t^\top X_t)\hat{\alpha}_t = X_t^\top Y_t$, which means that $n_t/n diag(X_t^{*\top} X_t^*)\hat{\alpha}_t = X_t^\top Y_t$, where n is the sample size of the reference panel (when the target GWAS data is not used as reference panel) and n_t is the sample size of X_t. Thus, optimizing $f(a_0)$ is equivalent to optimizing 3.13 because they are proportional in terms of a_0.

$$cor(\hat{Y}_t(a_0), Y_t) \propto \frac{\hat{\beta}_{te}^{\top} \text{diag}(X_t^{*\top} X_t^*) \hat{\alpha}_t}{\hat{\beta}_{te}^{\top} X_t^{*\top} X_t^* \hat{\beta}_{te}} \tag{3.15}$$

3.2.5 Polygenic risk score with multivariate effect size estimates

Let $D^* = (Y^*, X^*)$, be a testing data set of sample size n, and let $\hat{\beta}$ be an estimate of the multivariate effect size of model 3.1. We use the standardized predicted value of the multivariate model to construct the polygenic risk score (PRS). Thus, the PRS for the i^{th} individual is

$$PRS_i = \frac{X_i^* \hat{\beta} - \hat{E}[X^* \hat{\beta}]}{\sqrt{\hat{V}ar(X^* \hat{\beta})}}. \tag{3.16}$$

We will compare four methods to construct PRS such as in Equation 3.16. Each method relies on having an estimate of β that is plugged in Equation 3.16. Our proposed method is to use the Power Prior estimate $\hat{\beta}_{te}$ in Equation 3.16. The External method is to use the posterior mean obtained with the RSS likelihood using exclusively the External data. The Target method is to use the posterior mean obtain with the RSS likelihood using exclusively the target data, this coincides with the Power Prior method at $a_0 = 0$. The Full method is to fully combine the RSS likelihoods from the external and target data without any weighting, which is equivalent to the Power Prior at $a_0 = 1$.

3.2.6 Simulations

We first simulate the target population in which we draw effect sizes from $\beta_t \sim N(0, m/h^2)$, where h^2 is the heritability, and m is the number of causal SNPs. Let ρ be the correlation of the true effect sizes between the target and the external population. Thus, given β_j the external population's effect sizes will be simulated from $\beta_{ej}|\beta_{tj} \sim N(\rho m/h^2 \beta_{tj}, (h^2/m - \rho^2 m/h^2)^{-1})$. We then select different values of ρ to represent the heterogeneity of effect sizes across populations. If $\rho = 1$, the effect sizes are the same in both populations. If $0 < \rho < 1$, the effect sizes have the same sign but different magnitude, this represents cases in which interactions with genes or environment are different but in the same direction across populations. If $0 > \rho \geq -1$ the effect sizes have different signs and magnitude, this represents cases in which interactions with genes or environment are different in direction and magnitude across populations. We first describe how we simulate the target population, and then we will describe how we simulate the external population in the three scenarios.

3.2.6.1 Target population

We select m causal variants from a set of variants with at least 0.005 MAF in both populations. We use 1k Genome Project [4] haplotypes. For each causal variant j draw $\psi_{jt} \sim N(0, m/h^2)$, where h^2 is the heritability of the phenotype. Define the genetic contribution with an additive model $Z_t = \sum_{j=1}^{m} z_{jt}\psi_{jt}$, where z_{jt} is the dosage of causal variant j. Then we standardize it as $X_t = \frac{(Z_t - \bar{Z}_t)h}{sd(Z_t)}$ to ensure to have variance equal to h^2. Finally, simulate the environment contribution as $E_t \sim N(0, 1/(1-h^2))$ and define the phenotype $Y_t = X_t + E_t$. True effect sizes are then $\beta_t = \psi_t \sqrt{\sum_{j=1}^{m} var(z_{jt})}$.

3.2.6.2 External population

Simulate $\beta_{ej}|\beta_{tj} \sim N(\rho m/h^2 \beta_{tj}, h^2/m - \rho^2 m/h^2)$ with the effect sizes from the target population β_j. Calculate the genetic contribution with $Z_e = \sum_{j=1}^{m} z_{je}\beta_{je}$ and center it $X_e = Z_e - \bar{Z}_e$. To ensure the same effect sizes when $\rho = 1$, we do not standardize the genetic contribution Z_e as we did in the target population. Thus, the heritability in the external population will differ due to differences in allele frequencies. However, that difference will cancel out on average when we perform the simulations in several regions. The variance of the genetic component of the external population is $h^2/m \sum_{j=1}^{m} v_{je}$, where v_{je} is the variance of variant j in the external population.

To evaluate if the power prior has the potential to capture the heterogeneity of effect sizes, we perform a simulation analysis in a 500-kilo base pair region from chromosome 22. We fix the heritability at 0.001 for all simulations. We simulate a testing data set from the target population, denoting $D^* = (Y_t^*, X_t^*)$. Let β_t be the effect sizes that generated D^*, then we simulate an external data set (sample size 100,000) using haplotypes of European ancestry as described in 3.2.6.1, we denote β_e the true effect sizes that generate the external data $D_e = (Y_e, X_e)$. We then simulate a target data set $D_t = (Y_t, X_t)$ with sample size n_t using β_t. We choose the correlation between the effect sizes as $\rho = \{1, 0.75, 0.25, 0.0, -0.25\}$, and the target sample size as $n_t = \{1k, 5k, 10k\}$.

We set the ancestry of the target population as European to asses how well the method captures the heterogeneity of true effect sizes. We then set the ancestry of the target population to African to evaluate the ability of the method to capture differences in LD in combination with heterogeneity. The metric to evaluate methods are the concordance correlation between the phenotype Y_t^* and the standardize predicted phenotype of the multivariate model $X_t^* \hat{\beta}$.

3.3 Results

3.3.1 Power parameter identifies best way to combine information

In Figure B.1 we show the heritability explained by the risk models built with the Power prior for different values of the power parameter compared to risk models with no external data (Target), full use of external data (Full), and ignoring target data (External). Different panels present different configurations of target sample size (columns) and correlation of effect sizes between populations (rows). The external population had a sample size of 100,000 in all cases. Power prior and the Full method present the highest concordance correlation when the effect sizes in both populations are the same (top row of panels). In this scenario, the improvement in concordance correlation by joining two data sources is considerable when the ancestry of the target and the external data is different compared to using any of the single data sources approaches. When ancestry is the same in both populations, the improvement is negligible compared to using only the external data source. The power prior method shows a plateau in the concordance correlation as values of the power parameter increase and agrees in its performance with the full use of the external information. For intermediate non-negative correlation values (second and third rows of panels), the Power prior obtains a higher concordance correlation than all methods and ties with the Full method when the correlation is 0.75. When correlation is 0.25 (3rd row), the approach of using only the target data source outperforms the full use of the data source, though the Power prior is always better. When the correlation with effect sizes is 0 or negative (fourth and fifth rows), the best approach is to ignore the external data source. In that case, the Power prior collapses in that same approach as the clear optimum is to choose the power parameter of 0.

In the Appendix B, we present the same plot for the explained heritability. The characteristic curves are more challenging to interpret because the explained heritability is always positive. For the positive correlations, the interpretation is the similar as for the concordance correlations. However, for the non-positive correlations, the best method depends on which data source explains heritability most. However, the Power prior is always tied with whichever is the best method. For example, when the target data has the same ancestry as the external data (top plot) and the target sample size is 1,000, the External method is best, so the optimal Power prior shifts to close to 1, with a plateau in high values. When the target data have different ancestry than the external data (bottom plot), and the target sample size is 5,000, the External method is better, so the optimal Power prior shifts to close to 1, with a plateau in high values.

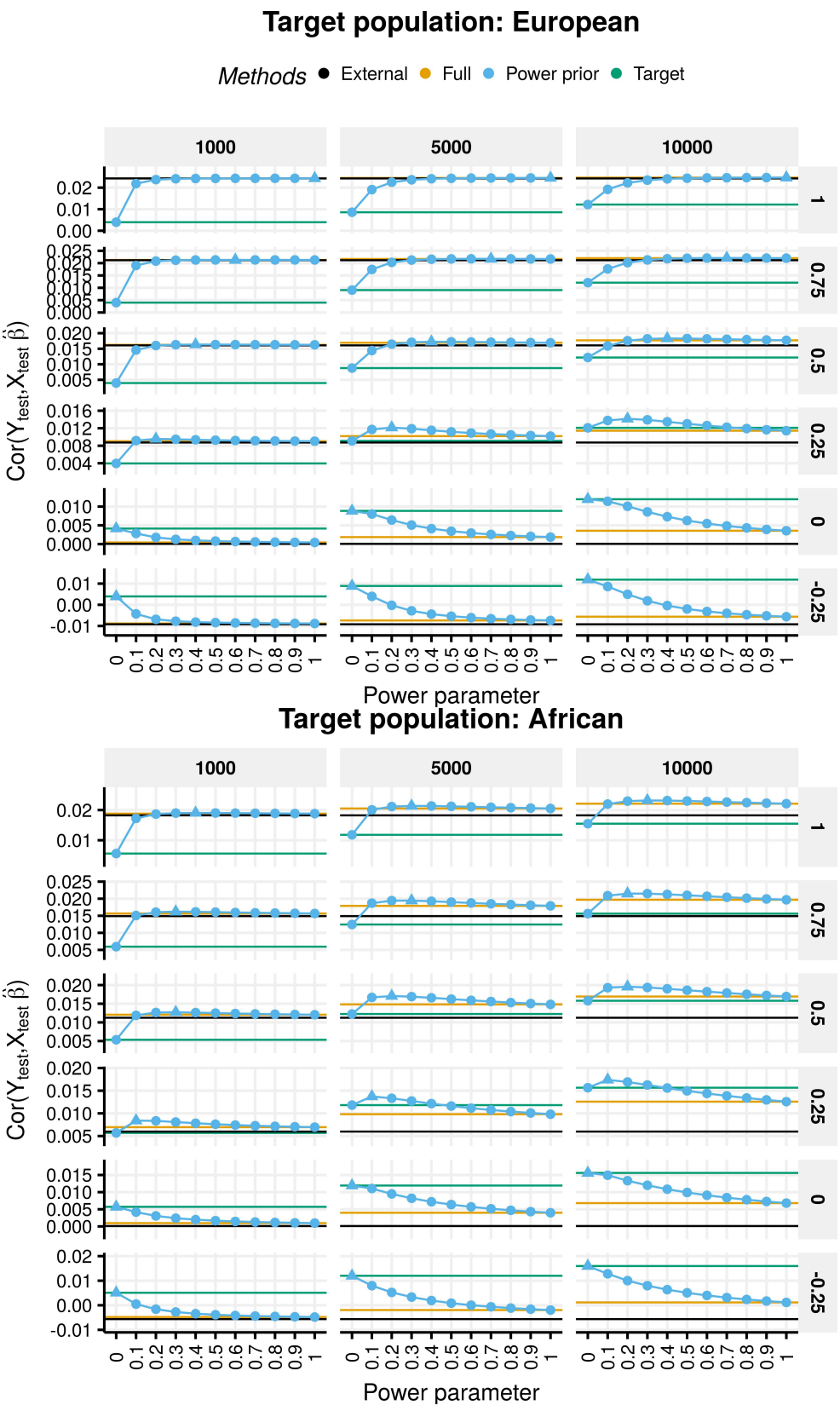

Figure 3.1: We show the concordance correlation with different choices of the power parameter. Color represents the four different methods. Both plots have an external population with European ancestry. Panel columns indicate sample size of the target population and panel rows indicate correlation of true effect sizes across populations.

3.3.2 Comparison of the methods to construct risk models

We assess the different methods of using external information by taking the mean of the proportion of heritability explained across 703 non-overlapping regions of 50 kilo-base pairs in chromosome 22, shown in Figure 3.2. The 703 different regions showed a large variability in the mean proportion of heritability explained, ranging from 0.1 to 0.8 (see Appendix Figure B.2). The Power prior with Regression with Summary Statistics (PP-RSS) has the highest proportion of heritability explained for all sample sizes when the correlation of the effect sizes is positive (see Appendix for results on selecting the objective function). When the correlation is zero or negative, the Target method is the highest, with the PP-RSS method being very close. The External method has the same proportion of heritability explained for low and null correlations, even for the negative correlation. Nevertheless, when the correlation is negative, the proportion of heritability explained by the Full method is notably less than the Target and PP-RSS. When the sample size of the target data source is small, the PP-RSS has minimal improvement over the best among the remaining methods; however, when the sample size is 5,000 or 10,000, the improvement can be very significant depending on the correlation of the effect sizes.

Another metric for the predictive ability of the risk model is the relative concordance correlation. We define the relative concordance correlation as the concordance correlation of the PRS based on the estimated multivariate effect size divided by the concordance correlation of the PRS based on the true multivariate effect size. In Figure 3.3 we present the mean concordance correlation of the four methods. The PP-RSS, in all cases the best or the second best method. Only when the correlation is negative, and the sample size of the target population does the PP-RSS present a negative concordance correlation. The PP-RSS improves the concordance correlation at least twofold for high and moderate correlations compared to the Target method. When the correlation is zero, the PP-RSS and the Target method have the same concordance correlation. As expected, the External method concordance correlation declines rapidly as the correlation between effect sizes decreases. Notably, the effect of the target sample size is different across correlations for the Full method, in which the concordance correlation becomes lower than the External method when the correlation is negative, and the target sample size is 1,000. The PP-RSS's concordance correlation is always parallel to the Target method regardless of the correlation, which implies that the effect of sample size is weaker than the effect of heterogeneity for the PP-RSS.

We compare the four methods in their ability to estimate the effect size of the causal variants. In Figure 3.4 we show the mean squared error (MSE) for the causal variants. As expected, when the correlation is one, the Full method is far from best when the target population is European. In the case of an African target population, the Target data source has a very similar MSE for a sample size of 5,000 and perfect correlation. As the correlation between true effect sizes decreases, the MSE of the External method increases rapidly up to two orders of magnitude. PP-RSS decreases

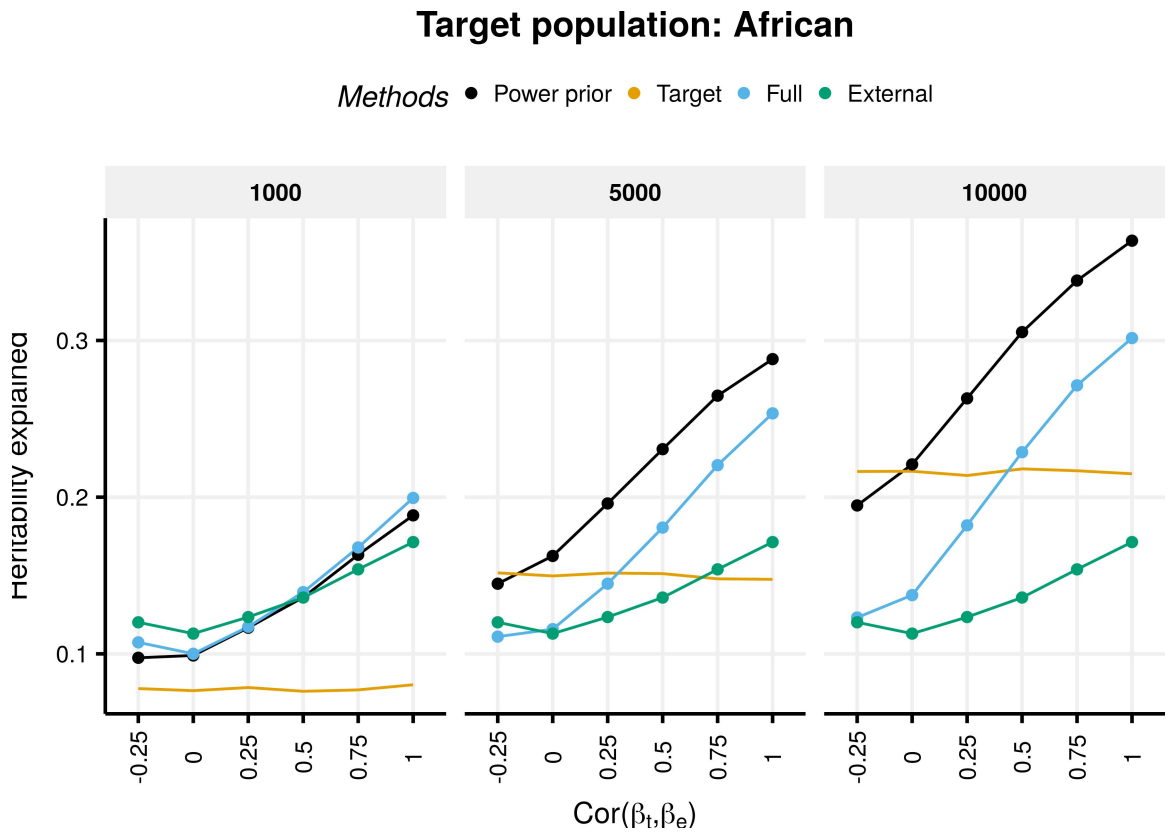

Figure 3.2: We show the proportion of heritability explained by the four methods. Color represents the four different methods: External is using an only external data source, Target is using an only target data source, Full is combining external and target data sources with the full weight of the external, and Power prior is the proposed method to weight differently the external data source. Panels represent the different correlations between the effect sizes of the target and the external population. The external population had European ancestry with a sample size of 100,000. Finally, we compute the Power prior approach with the pseudo correlation objective function. Panel columns indicate sample size of the target population.

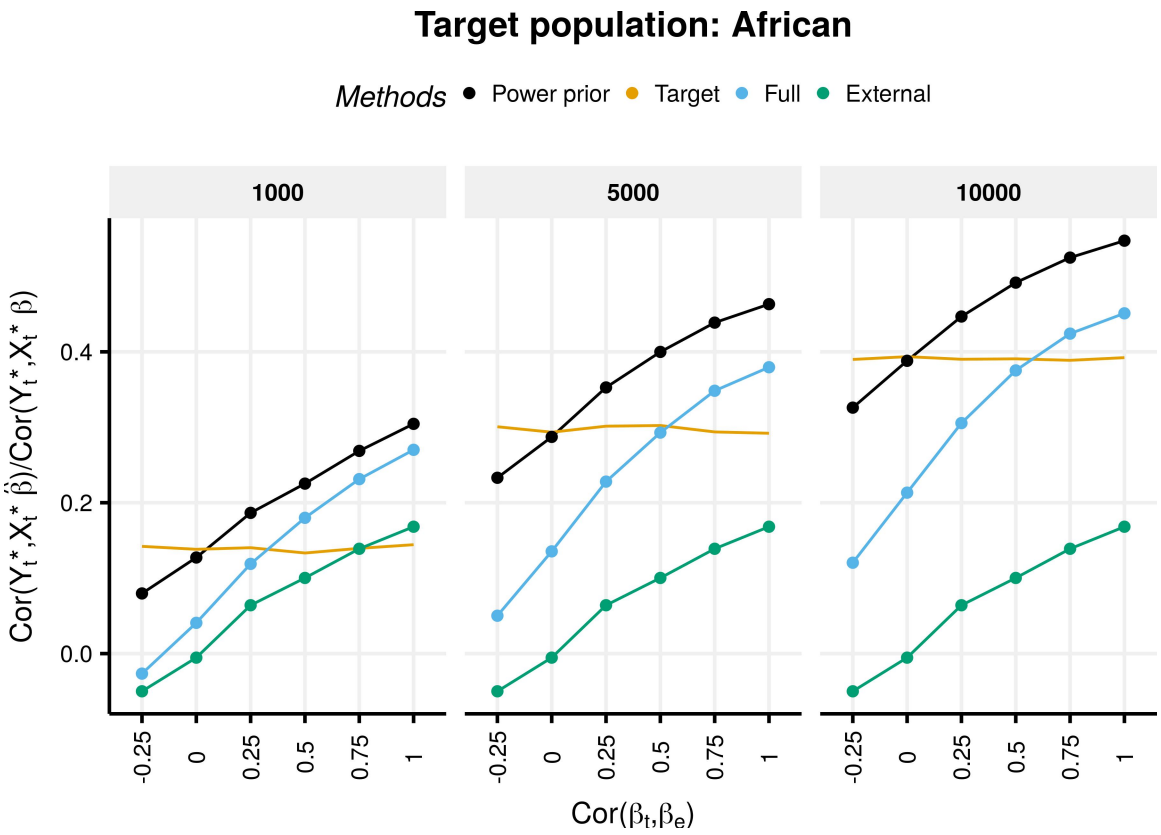

Figure 3.3: Mean relative concordance correlation for all four methods. Color represents the four different methods: External is using an only external data source, Target is using an only target data source, Full is combining external and target data sources with the full weight of the external, and Power prior is the proposed method to weight differently the external data source. Panels in columns represent the different correlations of effect sizes between the target and the external population. The external population had European ancestry with a sample size of 100,000. We computed the Power prior with the pseudo correlation objective function. Panel columns indicate sample size of the target population.

the MSE of the causal variants by one order of magnitude when the effect sizes' correlation is one compared to the Target method. When the number of causal variants is 1, the Target method is best for low correlations and sample sizes of 5,000 and 10,000. The PP-RSS method is always better when the number of causal variants is 2 or 3. The PP-RSS has the lowest MSE only when the correlation is negative, and the number of causal variants per region is greater than 1. However, PP-RSS is always close to the best method. The number of causal variants has a considerable influence on the MSE. The difference in MSE when the correlation is 1 is up to 2 orders of magnitude when there is one causal variant. However, this difference is less than one order of magnitude when the number of causal variants is 3.

3.4 Calculate power parameter with only summary statistics gives similar results than training data

We evaluate four methods to find a power parameter (see Methods for details). For the cross validation method we used $K = 5$. The sample size of the training data for the correlation and sums of squares methods is the same as the sample size of the target data source use to estimate the risk model. Simulations are across 703 regions of 50 kilo base pairs in chromosome 22. In Figure 3.5 we show the relative correlation concordance for different sample sizes of the target study and correlation of true effect sizes. The concordance correlation increases as target sample size or correlation between true effect sizes increases for all four methods. All four methods give very similar results in all cases.

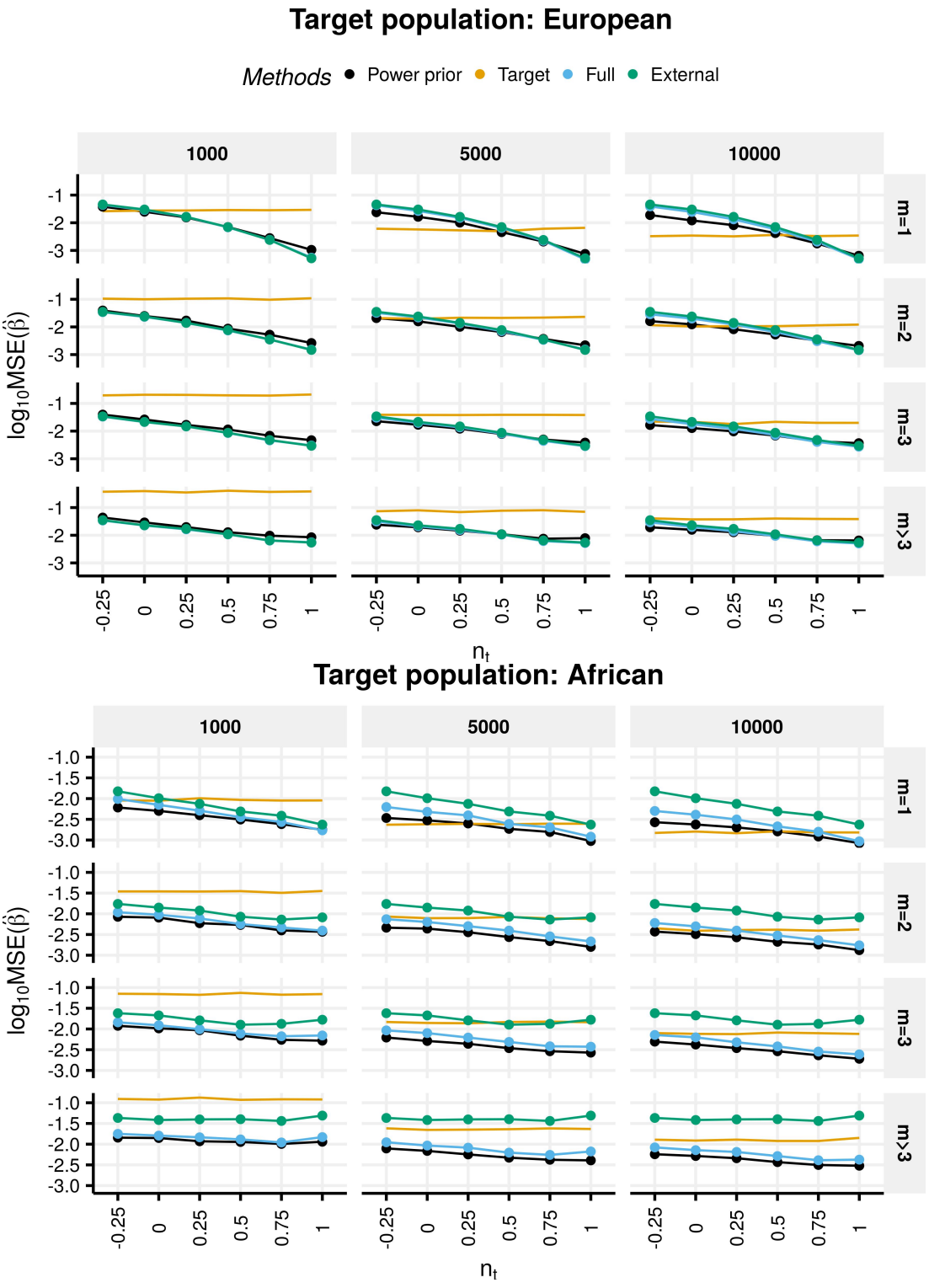

Figure 3.4: Mean squared error in the effect size of causal variants by the four methods. Color represents the four different methods. Panel columns indicate sample size of the target population. Panels in rows are the number of causal variants per analyzed region. We compute the Power prior with the pseudo correlation objective function.

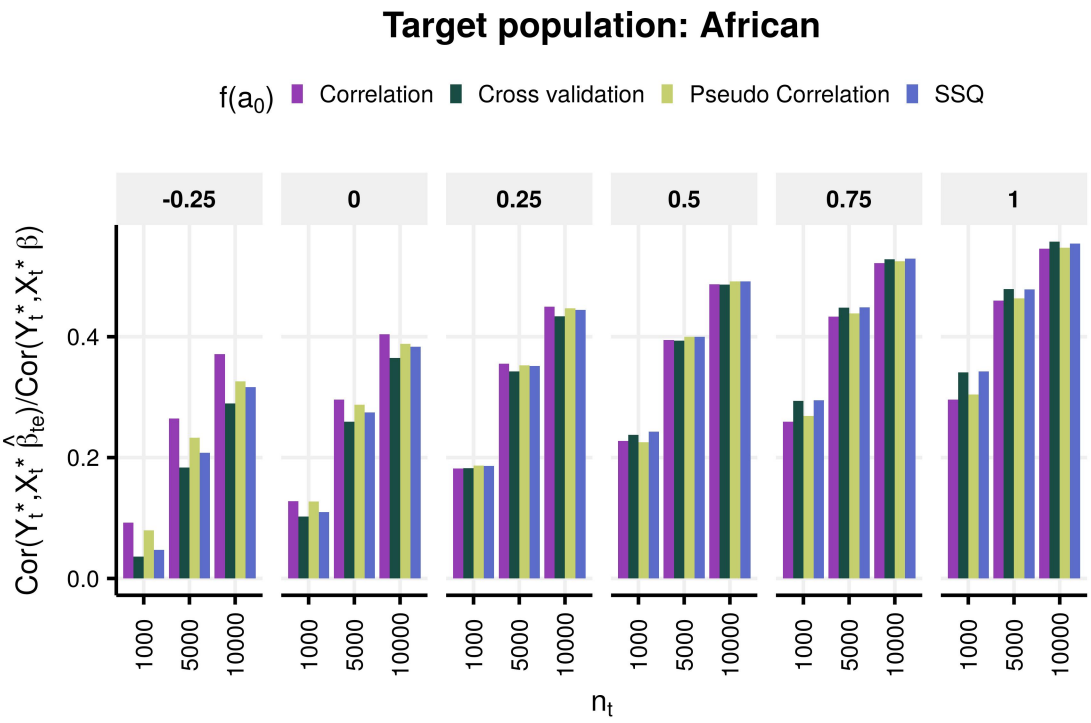

Figure 3.5: Relative concordance correlation is the correlation of phenotype and predicted phenotype divided by the correlation of the phenotype and the true genetic risk model. Color represents the three different methods: correlation between phenotype and predicted phenotype, sums of squares of difference of phenotype and predicted phenotype, and pseudo correlation of phenotype. Panels indicate the correlation between the true effect sizes of target and external population.

3.5 Discussion

We present a robust method called Power Prior with Regression Summary Statistics (PP-RSS) to increase the predictive ability of a PRS in a target population by adaptively weighting GWAS summary statistics from an external population. The method has two components. The first component assumes a common likelihood of the multivariate effect sizes conditional on univariate summary statistics, called a Regression with Summary Statistics likelihood [85]. The underlying assumption of the common likelihood is that the multivariate effect size is the same in both populations, which is a restrictive assumption. In response to the common likelihood assumption, the second component adaptively weights the likelihood of the external population to account for the heterogeneity of the true effect sizes between the populations. We perform adaptive weighting with a power prior[33] with a fixed power parameter on the likelihood of the external population. We propose to optimize a function of the power parameter that calculates the expected correlation of the predicted phenotype and the true phenotype using only summary statistics to calculate the optimal weight. We present simulations that show that our method increases a target population's PRS explained heritability, concordance correlation with the true phenotype, and reduces the MSE of causal variants compared to combining both populations without accounting for heterogeneity.

In our simulations, we presented results for the case where the target population has different LD than the external population and different levels of heterogeneity. However, different LD is not the only possible aspect between two populations that may differ. Social determinants have considerable impacts on health outcomes[68][81][20], Thus, even in genetically homogeneous populations, we have to consider that those different environments will have a different impact on the transferability of PRS. Our method assumes no specific reason for different effect sizes and includes the same LD case. We show that when the LD structure is the same and effect sizes are equal, the PP-RSS converges to fully use the external information and thus optimally uses the external GWAS. However, whenever the effect size is different, the PP-RSS method improves predictive ability and estimation compared to the full use of data in every scenario or even ignoring the external data source. Thus, we can use the PP-RSS to leverage any external population to improve the prediction and estimation of a target population.

We show that an adaptive weight is better than using all the external data or none in almost all scenarios. In the first project, we showed that LD could have a meaningful impact on the transferability of PRS. We model LD to infer multivariate effect sizes from summary statistics; thus, the inference is the same across different LD structures (assuming a shared multivariate model). We show that in the case of different LD structure, our method identifies an optimal weight that differs from using all the external data even when the multivariate effect sizes is the same. This result shows the robustness of PP-RSS to jointly account for model miss specification and different cor-

relation structures. Only when the direction of the effect sizes is opposite in each population using an external population is detrimental. In this case, the PP-RSS converges to ignore the external data in these cases, considerably reducing the detriment of using non-beneficial external data. Our method generalizes cases when all the external data is valuable and when it is preferable to ignore it.

We show the strength of modeling whole regions and adaptively weight the contribution of the external data considering different LD. In the first project, we show that regions differ in how much LD impacts the transferability of PRS. PP-RSS models regions independently to compute a different power parameter for each analyzed region. In addition, complex diseases usually have sets of causal variants in LD[3], which can make it particularly difficult to transfer PRS from populations with different LD structures[22] When a region had one causal variant, the PP-RSS method showed an inferior or similar performance in estimating causal variants, with either the Full method or the Target being the better approach depending on the heterogeneity of the effect sizes. However, when having multiple causal variants per region, our method is always the best in reducing MSE, with a more clear reduction when LD differs.

The presented model has a fixed power parameter and a ridge prior that assumes all variants in a region have a non-null effect size. Nevertheless, we can extend the proposed method in several ways. The first possible extension is assuming a prior distribution for the power parameter. We presented in the Appendix the conditional distribution of the power prior when the prior of the effect size is the ridge prior. A prior distribution for the power parameter could be helpful when the target sample size is minimal, which is the case in which our presented method was the least robust. The second extension is to use a different prior for the multivariate effect size. The ridge-prior distribution we assumed in this work implicitly assumes that all variants have a small true effect size, which is an implausible assumption. We can use an extensive collection of priors to analyze genetic data (see Zhou et al. [84]). For example, Zhu and Stephens [85] propose to use a mixture of Normal distributions in their original paper on the RSS likelihood. The PP-RSS has the flexibility to incorporate any prior, which can be adapted to suit several genetic models. The third extension is to include several external populations. A simple way to achieve this is to include a power parameter for each external population and repeat the analysis assuming all external populations are independent analyses. Furthermore, we can acknowledge the correlation across different external populations and model the power parameters with a prior distribution in the power parameter with shared parameters.

We present a powerful and flexible family of models that can focus on improving PRS in a target population. Most current multi-ethnic methods focus on creating a PRS with similar predictive ability in all populations. Our approach is best suitable for studies that are interested in a specific population but are limited in sample size. Increasing the predictive ability of PRS in such studies

also improve equity in genetic research as it increases the utility of existing European GWAS in non-European populations. This work presents an opportunity to develop methods using high sample GWAS without diluting population-specific results.

CHAPTER 4

Data Enriched Generalized Linear Regression

4.1 Introduction

In several biomedical settings, we are interested in predicting an outcome in a target population. High-quality data (randomized clinical trials, probability sampling, etc.) has always been expensive and complicated, nonetheless the gold standard in statistical analysis. Researchers have more access to external biomedical data from observational sources like Electronic Health Records or Bio Banks. These data sources are massive in sample size but usually do not represent the target population of interest well. However, there is an argument that we can borrow information from external data to increase prediction in the target population if we adequately model the differences in both data sources.

There is a wide range of methods that account for differences in data sources. The multi-task learning setting target the scenario in which several small tasks are integrated to infer the structure of each of them[30][21]. Another example is transfer learning, which considers the setting of several data sources that are rich in sample size from which we can obtain information for a target source when they are "close"[70][73]. There are several methods in the context of leveraging external summary statistics on a subset of the covariates of the target data set to improve prediction and estimation. [7][16][69][14].

In this chapter, we discuss the extension of the Data Enriched Linear Regression (DELR), proposed by Chen, Owen, and Shi [15] to study the case of one external source to improve the prediction of a target data source in the linear regression setting. Motivated by the DELR method, Cheng[83] described an alternative approach for DELR in logistic regression based on finding an optimal weight matrix that combines the single data source estimators. In this project, we extend DELR to generalized linear models. Thus we called the extension Data Enriched Generalized Linear Regression (DEGLR). We use the package *glmnet* with a tailored cross-validation method to fit the model. In principle, our approach works with all the link functions in the *glmnet* package, which are: gaussian, binomial, poisson, multinomial, cox, and multi-response Gaussian.

In the first section, we review the DELR proposed by Owen and Shi [83] and present some valuable results for our extension. The second section of this work presents the extension to generalized linear regression in which we show that this problem is equivalent to the penalized generalized linear regression from *glmnet* in R with some considerations. The third section presents two simulations. The first simulation evaluates the equivalence of DEGLR with DELR in the continuous outcome case. We show that we obtain the exact estimates as DELR with a mean concordance correlation of 0.99 across multiple scenarios. In the second simulation, we show that DEGLR predictive ability is always greater than or equal to pooling or ignoring the external data. We showed these scenarios where covariates have a high correlation and several predictors.

The fourth and final section presents a real data example in which we consider as target data the Health and Retirement Study (https://hrs.isr.umich.edu/) as a target data source. This HRS is a longitudinal panel study representing the American population over 50 in which participants are surveyed on health-related issues. Due to the nature of the study, we consider this data a high-quality data set that could potentially be used for prediction. We consider the Genes for Good (GfG) study as the external study. This study consists of volunteers 18 years and older that answer health-related questions through a Facebook app. Since this data is from volunteers that optionally answer questionnaires, we consider this data a lower quality data. Both studies have genotype data, which constructs the Polygenic Risk Score on Type 2 Diabetes (T2D). Due to selection bias and non-transferability of PRS across ancestral populations, the GfG study is very likely to be biased. Selection bias will arise because GfG represents a different population than HRS. The transferability of PRS is essential in our analysis because GfG has 95% of non-Hispanic White participants while HRS has 67% of non-Hispanic White participants. The data sources have a similar sample size, which is not the assumption of DEGLR. We compared the predictive ability of pooling GfG and DEGLR with a reduced sample size training HRS data source. We varied the training sample size to evaluate how DEGLR adapts as the relative importance of GfG increases. Our analysis shows that GfG has a limited predictive ability even for the case when the training sample size is 13.3 times smaller than GfG. As the training sample size increases, the predictive ability of DEGLR improves consistently, showing that the method adapts correctly to improve prediction ability. In addition to our simulations, this analysis shows that our extension of the DELR works as expected.

4.2 Methods

Let Y_e, Y_t be outcomes vectors for the external population and the target population respectively, let n_e and n_t denote sample size, X_e, X_t are $p+1$ are matrices that has p covariates and a column of 1 for the intercept. We assume outcomes and covariates to be standardized. Finally, we assume that $n_e > n_t$.

4.2.1 Data Enriched Linear Regression

Consider the following linear regression model for the target population.

$$Y_t = X_t\beta + \epsilon \tag{4.1}$$

where $\beta \in \mathbb{R}^{p+1}$ and ϵ are independent errors with $E[\epsilon] = 0$ and $var(\epsilon) = \sigma_t^2$. Now consider the following for the external population

$$Y_e = X_e\alpha + \epsilon \tag{4.2}$$

where $\alpha = \beta + \gamma$, $\gamma \in \mathbb{R}^{p+1}$ is a vector of bias and ϵ are independent errors with $E[\epsilon] = 0$ and $var(\epsilon) = \sigma_e^2$.

The data enrichment approach is to fit a model that includes both data sources and a quadratic penalty $\lambda \in \mathbb{R}^+$ on the prediction bias $||X_t\gamma||_2^2$. Thus, the DELR estimate is such that it minimizes the following objective function.

$$f(\beta, \gamma; \lambda) = ||Y_t - X_t\beta||_2^2 + ||Y_e - X_e(\beta + \gamma)||_2^2 + \lambda||X_t\gamma||_2^2 \tag{4.3}$$

Given a value of λ the function f is minimized by

$$\hat{\beta}_\lambda = W_\lambda\hat{\beta} + (\mathbf{I} - W_\lambda)\hat{\alpha}, \tag{4.4}$$

where $W_\lambda = ((\lambda+1)V_e + \lambda V_t)^{-1}(V_e + \lambda V_t)$ (see Appendix C.1 for details on normal equations and simplification of W_λ). From this expression it is easy to see that when $\lambda = 0$ then $W_\lambda = \mathbf{I}$ and thus $\hat{\beta}_\lambda = \hat{\beta}$. On the other hand when $\lambda \to \infty$ then $W_\lambda = (V_e + V_t)^{-1}V_t$ and $\mathbf{I} - W_\lambda = (V_e + V_t)^{-1}V_e$ which means that $\hat{\beta}_\lambda \xrightarrow{\lambda\to\infty}$ goes to the estimator we would get by pooling the data from both sources (details in the Appendix C.2). Thus the DELR estimate fluctuates between ignoring the external data and using it fully depending on the amount of penalty we impose on the prediction bias.

Choosing a value for λ

We are interested in improving the prediction accuracy in the target population. Thus the DELR uses the AICc of the target population to choose λ. To compute the AICc, we need to calculate the model's degrees of freedom. We define the degrees of freedom as the normalized covariance between the predicted outcome and the outcome in the target population.

$$df(\lambda) = \frac{1}{\sigma_t^2} tr(cov(\hat{Y}, Y_t)). \tag{4.5}$$

In the DELR case, $\hat{Y} = X_t\hat{\beta}_\lambda$, thus obtaining a closed form of the degrees of freedom.

$$\begin{aligned} df(\lambda) &= \frac{1}{\sigma_t^2} tr(cov(\hat{Y}, Y_t)) = \frac{1}{\sigma_t^2} tr(cov(X_t(W_\lambda\hat{\beta} + (\mathbf{I} - W_\lambda)\hat{\alpha}), Y_t)) \\ &= \frac{1}{\sigma_t^2} tr\left(cov(X_t W_\lambda (X_t^\top X_t)^{-1} X_t^\top Y_t, Y_t)\right) = \frac{1}{\sigma_t^2} tr(X_t W_\lambda (X_t^\top X_t)^{-1} X_t^\top)\sigma_t^2 \\ &= tr(W_\lambda). \end{aligned}$$

With an expression for the degrees of freedom, we can obtain the AICc as.

$$AIC(\lambda) = n_t log(\hat{\sigma}_t^2(\lambda)) + n_t \left(1 + \frac{df(\lambda)}{n_t}\right) / \left(1 - \frac{df(\lambda) + 2}{n_t}\right), \tag{4.6}$$

where $\hat{\sigma}_t^2(\lambda) = \frac{1}{n_t - p}||Y_t - X_t^\top\hat{\beta}_\lambda||_2^2$. This is the Hurvich and Tsai AIC definition, which penalizes models that are not parsimonious relative to the sample size. With this conservative definition of AIC we will tend to choose larger values of λ.

Algorithm to find λ

Note that $df(0) = p$ because $\hat{\beta}_\lambda = \hat{\beta}$ and from the alternative expression of $df(\lambda)$ in the Appendix C.3 we can see that $d(\infty) = \sum_{j=1}^{p} \frac{\nu_j}{1+\nu_j}$. We then exploit this monotonicity of $df(\lambda)$ to define a grid search of λ. Take a grid of df in the interval $[df(\infty), p]$ and find the correspondent λ for each. Use λ such that minimizes 4.6 to construct the DELR estimate as in 4.4.

4.2.2 Extension to generalize linear models

Most outcomes in biomedical problems are not continuous, which limits DELR. A natural way to extend the DELR to generalized linear models is to change the sums of squares from the objective function 4.3 for negative log-likelihoods. The two unique components of DELR are kept the same in our generalization; the first, we model the bias of the external population as a bias vector that

is additive to the parameter of interest; the second, we add a penalty on the bias prediction in the linear scale.

Consider the following model for the target data set.

$$\eta(E[Y_t|X_t]) = X_t\beta, \tag{4.7}$$

and the following for the external data set

$$\eta(E[Y_e|X_e]) = X_e\alpha, \tag{4.8}$$

where $\alpha = \beta+\gamma$, and both models have the same link function η. Models 4.7 and 4.8 induce two likelihoods, for model 4.7 and the target data we denote with $l_t(\beta; X_t, Y_t)$ the log likelihood of β. Similarly, for model 4.8 and the external data $l_e(\alpha; X_e, Y_e)$ denotes the log likelihood of $\alpha = \beta+\gamma$. Thus the Data Enriched Generalized Linear Regression estimates minimize the following objective function.

$$g(\beta, \gamma; \lambda) = -l_t(\beta; X_t, Y_t) - l_e(\beta + \gamma; X_e, Y_e) + \lambda||X_t\gamma||_2^2. \tag{4.9}$$

This objective function is very similar to the objective function of the penalized regression. In Appendix C.4 we show the connection with the penalize regression implementation of the *glmnet* R package. Which means that for a fixed λ we can use standard software to obtain estimates of β.

4.2.3 Assumptions of the model

The DEGLR assumes that model 4.7 and model 4.8 are the same up to a difference in the effect sizes in the covariates included in the model. This means that if we do not include covariates correlated with the covariates included in the model, it will not impact the ability to distinguish when the external data is valid. This is because this difference will be captured by γ, which is the only parameter that impacts the penalty.

A vital characteristic of DEGLR is that it has assumptions about the distribution of X_t and X_e. Both populations can have different distributions without impacting the method, which is the desired assumption in our context. We assume we are unsure if the external information should be used or to what extent we can trust it. DEGLR identifies the best way to use external data to benefit the target data. Which is a difference from the data integration methods, for which similar distribution of covariates is always a necessary assumption.

DEGLR assumes we have access to the same covariates in both populations. We can relax this assumption by modifying the extended data used in the penalized regression. This is beneficial if some inclusion/exclusion criteria for the target population are too restrictive for the external

population. Then we assume that the conditional associations help increase prediction in the target data. For example, the target data excludes young people but including them in the external data might increase the sample size dramatically. We can include younger people from the external population by adding a covariate that identifies them. By doing this, we assume that the conditional association of the remaining covariates is still useful. We will not be able to distinguish between β and γ for the covariate of young people, but we also do not care about it since is people we excluded from the target population.

4.2.4 Finding the optimal λ^* in non linear cases

The objective of the DEGLR method is to find an optimal way to combine external data with target data to increase the prediction of the target data. Standard penalized regression methods use cross-validation approaches that treat equally all observations in the training data would be inappropriate. Thus, we have to tailor the cross-validation method to obtain an optimal λ^* because it will calculate the prediction error in the combined data set. However, an advantage of using *glmnet* R package is that it calculates the whole path of $\hat{\beta}_\lambda$ efficiently for a grid of values of λ. We use K fold cross-validation to estimate the prediction error in the path and select the λ^* that gives the lowest prediction error in the target data. Since the target prediction error does not depend on the external data, we do not partition the external data for cross-validation. In other words, the estimate of the prediction error we obtain remains unbiased if we do not partition the external data.

In theory, we know that $\hat{\beta}_\lambda$ converges to the target estimator when $\lambda \to 0$ and converges to the estimator using target and external data pooled when $\lambda \to \infty$. But in practice there exist a range for $(\lambda_{min}, \lambda_{max})$ such that $\hat{\beta}_{\lambda_{min}} \approx \hat{\beta}_t$ and $\hat{\beta}_{\lambda_{max}} \approx \hat{\beta}_{pool}$. Thus, we will search for the best λ in a grid on the log scale in such a grid. To find λ_{max}, we take advantage of *glmnet* with the augmented data searching for that same λ in their algorithm. Then we first use *glmnet* with the augmented data and the largest λ from *glmnet* is λ_{max}. We specify the minimum value of the grid of λ by specifying a ratio between the maximum and the minimum prior, and we check for convergence internally. We describe the algorithm in detail below.

Cross validation algorithm to find λ

1. Fit DEGLR using the *glmnet*. The result includes a grid of λ in which the maximum value is λ_{max}, i.e.$\hat{\beta}_{\lambda_{max}} \approx \hat{\beta}_{pool}$.

2. Set a grid of λ that is equally distant in the log scale setting $\lambda_{min}/\lambda_{max} = r$, with a predefined r.

3. With the complete target data compute L such that $(L^{-1})^\top$ is the Cholesky decomposition of V_t.

4. Split the target data into K roughly equal parts.

5. Randomly select a subset of the external data of size $n_e(K-1)/K$ so that the ratio

6. For each of the k^{th} fit *glmnet* as in C.5 using the pre-computed L matrix from step 3, all the external data, the other $k-1$ parts of the target data, and the grid of λ from step 2.

7. Calculate $CV(\lambda) = \frac{1}{n_t}\sum_{k=1}^{K}\sum_{j=1}^{n_t^{(k)}}(Y_{jt}^{(k)} - \hat{f}^{(-k)}(X_{jt}))^2$, where $\hat{f}^{(-k)}(s) = s^\top \hat{\beta}_\lambda^{(-k)}$ is the prediction for s using model from fitted as in step 5 without the k data.

8. Find $\lambda_{optim} = \underset{\lambda}{} CV(\lambda)$.

Two data exclusions stand out from the cross-validation algorithm proposed above. First, we compute L with all the target data instead of computing $L^{(-k)}$ for each data split. This is because to compute L we need V_t to be a positive definite matrix, which means that if $n_t \leq p$ we cannot do it. Thus, the DEGLR can only work in the $n_t < p$ scenario, which is more restrictive if we compute $L^{(-k)}$ since we would need $n(K-1)/K < p$. Furthermore, computing L would not create bias in our estimate of the mean square error because L does not contain any information about Y_t. The second exclusion is that we do not partition Y_e. Similarly, there is no information about Y_t in the external data; thus, this exclusion will not create bias in our estimate. An alternative would be stratified cross-validation, in which we maintain the ratio between external and data sources in each data split. However, this will increase the variance of our mean square error estimate without any benefit in reducing bias.

4.3 Simulations

We present two simulation studies. The first simulation aims to validate that DEGLR, which finds the optimal λ through cross-validation, obtains the same $\hat{\beta}_\lambda$ as the DELR that has a theoretical way to define a search space of λ. The second simulation evaluates the DEGLR in a logistic regression case.

4.3.1 Validation of relationship between DELR and *glmnet* with continuous outcome

We verify Equation C.6 by comparing $\hat{\beta}_\lambda$ from DELR and from *glmnet* in the linear case. To that end, we simulate data from a target population using model 4.1 with sample size $n_t = 500$ and data from an external population using model 4.2 with sample size $n_e = 20,000$. We assumed X_t and X_e had a multivariate Gaussian distribution with mean 0_p and correlation matrix C with a uniform structure in the covariates with $\rho = 0, 0.05, 0.15$, for the dimension of the covariates let $p = 3, 10, 50$, target sample size $n_t = 500$, and external sample size $n_e = 20,000$. Then we simulated β from a uniform hypersphere of dimension p. We fixed the variance explained $r^2 = 0.2$ and calculated $\sigma_t^2 = \beta^\top C\beta(1-r^2)/r^2$. We simulated Y_e from a Gaussian distribution with mean 0 and variance σ_t^2. For the external population we first sample γ^* from a uniform hyper sphere of dimension p and then scaled it have a relative bias, thus $\gamma = \sqrt{\kappa/n_t}\sigma_t^2$, with $\kappa = 0, 1, 2, 3, 4, 5, 6, 7$. Then we calculated $\sigma_e^2 = (\beta+\gamma)^\top C(\beta+\gamma)(1-r^2)/r^2$ to simulate Y_e from a Gaussian distribution with mean 0 and variance σ_e^2. With this simulation mechanism, we imply that σ_e^2 is larger than σ_t^2, representing a lesser quality data source in the external population. In addition, the variance explained by the covariates is the same within each population but the variance explained by the external population covariates in the target population decreases as the bias increases.

We compared the median λ^* from DELR with the scale $2\lambda/(n_t+n_e)$ from DEGLR (see Methods for details on this relationship) to observe the agreement on the penalty factor. From Figure 4.1, when there is no bias, the DEGLR has a larger median penalty than the DELR, which means, in this case, it is giving larger weight to the pooled estimate. However, the DELR has a larger penalty for all the levels of bias, which means that DEGLR is more conservative in the presence of bias.

However, we found very little difference in the final estimates, which means that the difference in penalty has very little influence on the final estimates. The median concordance correlation across all the twelve scenarios was above 0.999 see Figure 4.2. When the dimension of the covariates is 3, the distribution of the concordance correlation has a notably smaller variance. Apart from this case, the distribution of the concordance correlation looks very similar in all the cases with no

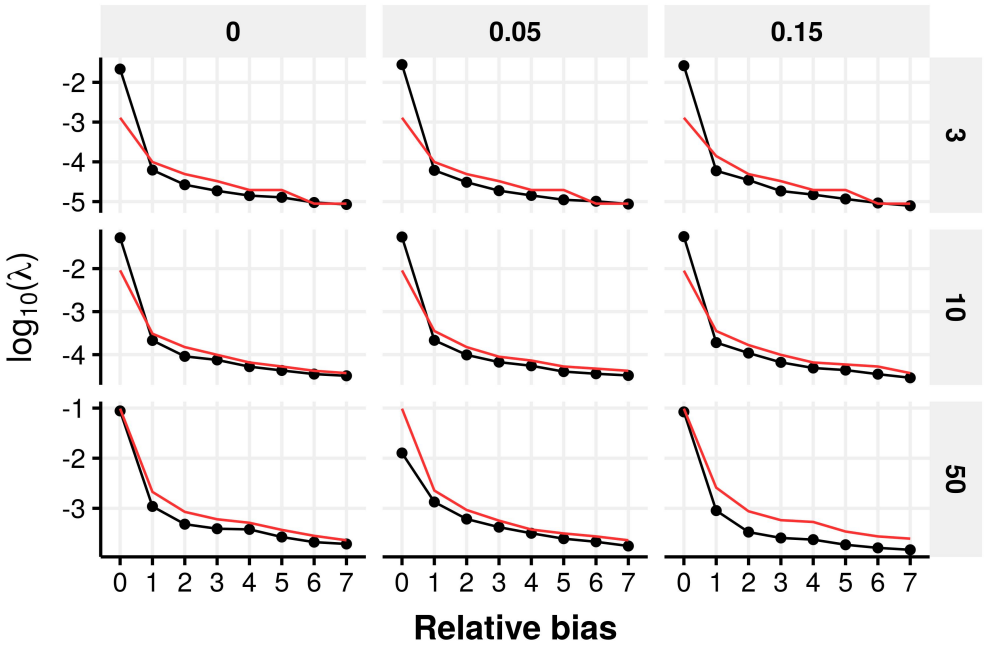

Figure 4.1: We present in black the median λ^* in the log scale, and in red, we have the median $2\lambda/(n_e + n_t)$ in the log scale. We present this for twelve different scenarios of correlation in the covariates, dimension of the covariates, and magnitude of bias of the external population.

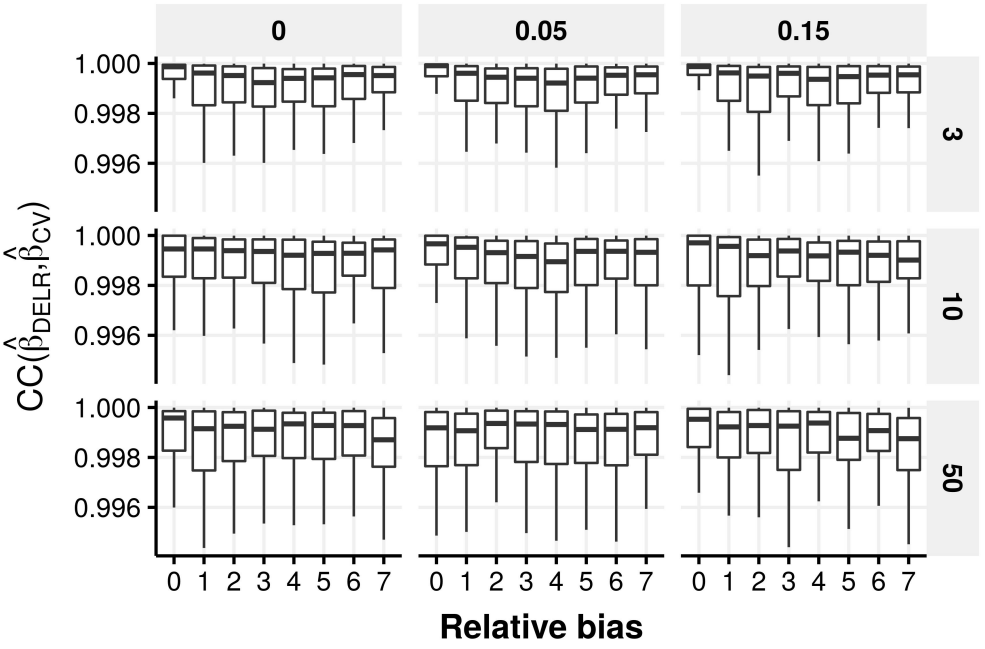

Figure 4.2: We present the concordance correlation between $\hat{\beta}_\lambda$ obtained with DELR and with DEGLR with the gaussian link function.

clear trend.

4.3.2 Simulations under the logistic case for the DEGLR

The second simulations assume the same data generating mechanism for the covariates as in the linear case simulations; the sample size in the target is $n_t = 500$ and the sample size in the external data is $n_e = 5000$. Similarly, in the linear case, we wanted to choose β and γ to imply that the covariates had the same predictive ability of the outcome within data sources. In the linear case, we have the variance parameter in the outcome that allowed us to regulate the predictive ability of the covariates once we fixed β and γ. However, the mean and variance in a binary outcome are not separable as in the linear case. In order to have the same covariate predictive ability within populations we choose β and γ such that $||\beta||_2^2 = ||\beta+\gamma||_2^2 = 1$. We first simulate β from a uniform hypersphere of dimension p and then generate γ in a two-step procedure. Let Z be a matrix that is 0 in the diagonal and $\frac{c}{p-1}$ and then define $\beta^* = \beta - Z\beta$. Note that $\beta_j^* = \beta_j - \frac{c}{p-1}\sum_{i\neq j}\beta_i$, which means that$\beta_j^*$ is shrinks β_j changed by a factor c and the mean effect size ignoring β_j. This means that large (relative to the rest) effect sizes change less than small effect sizes and that we have c to increase the overall change in effect sizes. The second step is to normalize β^* to have norm 1 and

let $\gamma = \beta^*/||\beta^*||_2^2 - \beta$. Thus $||\beta+\gamma||_2^2 = ||\beta+\beta^*/||\beta^*||_2^2 - \beta||_2^2 = 1$. In our simulations we choose $c = (0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3)p$, we multiply by p to have the similar magnitude change in the effect size across different p. To obtain λ for the DEGLR we use a $K = 5$ fold cross-validation approach as described in Section 4.2.4.

In Figure 4.3, we see that as the relative bias increases then, the Brier score and the AUC of the External method show a consistent loss of predictive ability. When there is little or no bias, the External method has better predictive ability than the target data set. This is because the External data set has a tenfold larger sample size for the no-bias case. As the relative bias increases, the reduction of predictive ability is a trade-off between bias and variance. The same behavior is observed for the pooling method, with better predictive ability than the External population in all cases because of the influence of the Target data set. On the other hand, the DEGLR method matches the predictive ability of the Pool/External methods for no bias. Similarly to Pool and External methods, the predictive ability for the DEGLR method decreases as the relative bias increases. The key difference is that the predictive ability of the DEGLR method is bounded by the predictive ability to use only the Target data set. This means that the DEGLR is down-weighting the external data information as the bias increases. For no bias, DEGLR is almost equivalent to Pooling both data sources, and for large bias, DEGLR is equivalent to ignoring the external data. Most importantly, in intermediate bias cases, DEGLR is the best method. Even for some of the non-zero relative bias, the DEGLR improves predictive ability on the target population while Pooling the data reduces the predictive ability. These cases can be seen for $p = 50$ and relative bias larger than 0.2 with a correlation of 0.0.

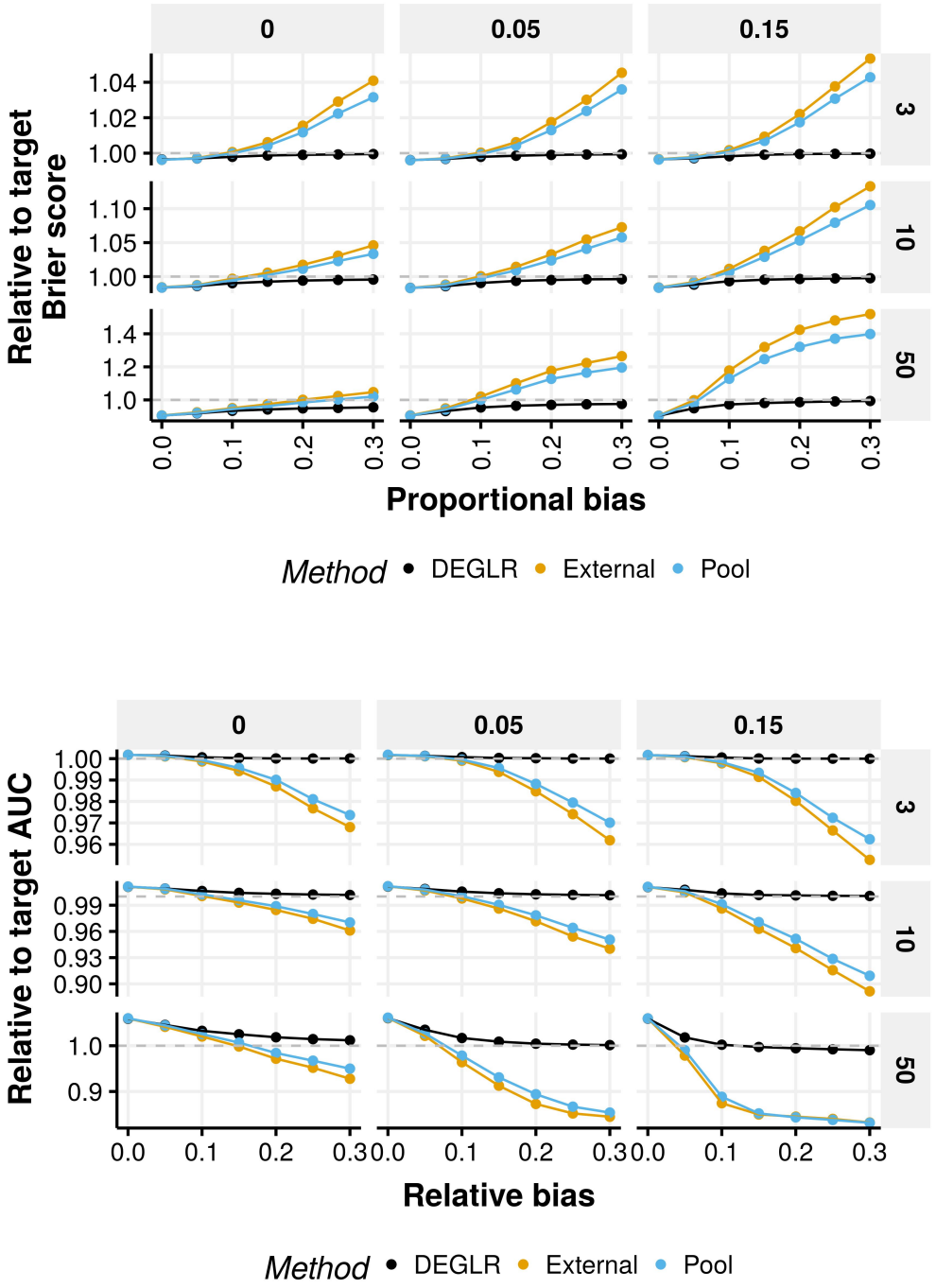

Figure 4.3: Relative Brier score and mean relative AUC of each method. Both metrics are relative to analysis with only the target data. We show in black the proposed DEGLR method, in yellow the estimate obtained with the external data, and in blue using the estimate obtained with the pooled data.

4.4 Real data example

As our target data, we obtained a sub-sample from the Health and Retirement Study (HRS) (https://hrs.isr.umich.edu). The HRS data is a longitudinal panel representing people over 50 years old who will retire or are already retired in the United States of America. This study's participants answer rigorous surveys on aging, health, cognition, and financial well-being. Because of this, we consider the HRS as high quality. We consider the Genes for Good (GfG) study for the target data source. GfG collects data from volunteers 18 years and older through a Facebook app in which participants answer health-related questions. Therefore, we consider only participants over 50 years old. Since the GfG's participants are volunteers and the questionnaires are answered optionally, we consider this a lower-quality data source than HRS.

The data consist of 2284 participants from the HRS and 1522 from the GfG for which we have affected status for diabetes, race (non-Hispanic White, non-Hispanic Black), Gender, Born in USA, BMI, Education (degree less than Bachelor), age (categorized in 5 groups). We also included a Polygenic Risk Score for Type 2 Diabetes calculated with about 223,000 overlapping SNPs. Table 4.1 presents the distribution across all covariates except PRS of both studies. The mean age, education, gender, and race are different in both data sources. Only BMI and the proportion of US-born participants are similar in both studies. We do not show the mean of the PRS because it was mean-centered before we obtained the data. However, as previously discussed in this dissertation, PRS have different predictive abilities across populations. It is mainly well studied that transferability from European to African populations is limited, which is essential in this analysis because GfG has 95% of non-Hispanic White participants while HRS has 67% of non-Hispanic White participants. Since the PRS are constructed with a different ancestral distribution of the samples, we expect that the PRS will not be as predictive in HRS as in GfG. Thus, we consider two potential sources of bias: the selection bias introduced through the non-probability sampling of GfG and the bias in the PRS variable in the different studies. This presents an excellent opportunity for our method to adjust for bias that could impact prediction ability if we pool both studies.

The DEGLR setting assumes that the external data is larger in sample size, which is not the case for HRS and GfG. Because of this, if we apply DEGLR in this scenario, it would not be surprising that there is very little to gain, and probably analyzing HRS by itself is the best route to follow. However, we can still assess the performance of DEGLR by splitting the HRS data in training and testing, keeping the testing sample size fixed, and systematically increasing the sample size of the training data. For this, we randomly set apart 20% of the samples from the HRS data for testing. Then we randomly select 20% samples for training to evaluate AUC with DEGLR, using only HRS, GfG, and pooling HRS and GfG. Then we sample an additional 5% of samples from HRS to the same training data (25% in total) and repeat. We continue adding groups of 5% samples

	Mean in HRS	Mean in GfG
Diabetes	0.21 (0.19,0.23)	0.16 (0.14,0.18)
Born between '24 - '30	0.04 (0.029,0.05)	0.005 (0.002,0.009)
Born between '31 - '41	0.13 (0.11,0.14)	0.04 (0.03,0.05)
Born between '42 - '47	0.07 (0.06,0.08)	0.14 (0.12,0.16)
Born between '48 - '53	0.16 (0.14,0.17)	0.35 (0.32,0.37)
Born between '54 - '59	0.59 (0.57,0.61)	0.46 (0.43,0.49)
Less than Bachelor	0.33 (0.31,0.35)	0.51 (0.48,0.53)
Masculine	0.45 (0.43,0.47)	0.32 (0.29,0.34)
White	0.67 (0.65,0.69)	0.95 (0.94,0.96)
US-born	0.95 (0.94,0.96)	0.96 (0.95,0.97)
BMI	29.24 (28.97,29.51)	29.57 (29.22,29,92)

Table 4.1: Means of variables in the HRS and GfG. Confidence intervals are calculated with normal approximation.

until we reach 80% of the sample size and thus use the complete data (80% for training and 20% for testing). We repeat this analysis 1000 times to smooth results. Taking 10% of HRS's sample size, GfG is 6.7 times larger in sample size, which might not be large enough. We sample with replacement from GfG to generate a synthetic larger version of GfG and use that larger version as the external data. We use a $K = 10$ fold cross-validation as in Section 4.2.4 to obtain the λ for the DEGLR. In Table 4.2 we observe that when analyzing the testing HRS data with the synthetic GfG we obtain an AUC of 0.682. The AUC of the synthetic GfG is the same regardless of the training sample size since it ignores the HRS data. The difference in AUC between GfG and HRS is only 0.003 when the GfG is 20.0 folds larger than the training sample size of HRS, which speaks to the reduced predictive ability of GfG in comparison to HRS. When the ratio of GfG and HRS testing sample size is larger than 8.00 folds, the DEGLR has the largest AUC, followed by Pool. This shows that DEGLR is able to optimize prediction even when Pooling the GfG data would increase prediction compared to only GfG. However, when GfG is ten times larger, Pool does not do better than only HRS, this is also true for all smaller ratios. This speaks to the risks of pooling a larger but biased data set. As the sample size of the training HRS increases, the AUC of all methods (except GfG alone) increases as well, which is expected since all of them benefit from a larger training sample size of HRS. The DEGLR is either best in all cases or extremely similar to the best method, which is the desired behavior of the method.

In simulations, we can regulate bias and see how the DEGLR's penalty changes. We showed in simulations that this relationship is inverse, the larger the bias the smaller the penalty. In our real data application is impossible to know or control the bias. Thus, in this analysis, we modify the ratio between the target sample size (training HRS) and external (GfG) to assess the change in the penalty. Increasing the training sample size should also have an inverse relationship with

n_{GfG}/n_{train}	n_{train}/n_{HRS}	AUC Pool	AUC HRS	AUC DEGLR	AUC GfG	$\|\|\hat{\gamma}\|\|_2$
20.0	0.10	0.690	0.677	0.692	0.681	1.16
13.3	0.15	0.694	0.691	0.697	0.681	1.22
10.0	0.20	0.696	0.697	0.699	0.681	1.38
8.00	0.25	0.697	0.701	0.702	0.681	1.55
6.66	0.30	0.699	0.704	0.704	0.681	1.64
5.71	0.35	0.700	0.707	0.706	0.681	1.79
5.00	0.40	0.702	0.708	0.708	0.681	1.93
4.44	0.45	0.703	0.709	0.709	0.681	2.07
4.00	0.50	0.704	0.710	0.710	0.681	2.18
3.63	0.55	0.704	0.711	0.711	0.681	2.28
3.33	0.60	0.705	0.712	0.712	0.681	2.33
3.08	0.65	0.706	0.713	0.712	0.681	2.43
2.86	0.70	0.707	0.713	0.713	0.681	2.47
2.67	0.75	0.707	0.714	0.713	0.681	2.55
2.50	0.80	0.708	0.714	0.714	0.681	2.57

Table 4.2: The first column is the ratio between the sample size of GfG and the training HRS. The second column indicates the ratio between the sample size of the partition of HRS used to fit the models and the complete sample size of the HRS data reserved for testing. The third, fourth, fifth, and sixth columns have the AUC obtained with each of the four methods. The pooling method uses HRS and GfG staked as one data source, HRS uses only HRS, DEGLR corresponds to the Data Enriched Generalized Linear Regression, and GfG uses only the GfG data. The last column has the norm of the bias γ.

the penalty. An interesting effect of observing is if the penalty decreased as the training sample size increased. However, the magnitude of λ is not comparable across the training sample size, making a direct comparison of the penalty impossible. On the other hand, we can conceptualize a true value of γ. When $\lambda \to \infty$ we expect that $\hat{\beta}_\lambda$ is the same as the estimator from pooling both data sources. For this case γ is unidentifiable and those the DEGLR estimates it as zero, thus $\lim_{\lambda\to\infty} E[\hat{\gamma}_\lambda] = 0$. Following the same logic when $\lambda \to 0$ the DEGLR converges to ignoring the external data, which means the estimate for γ should be unbiased, thus $\lim_{\lambda\to 0} E[\hat{\gamma}_\lambda] = \gamma$. In Table 4.2 we see the norm of γ increases as the sample size ratio increases, which is what we expect. Finally, in Figure C.1 in the appendix we show that the cross validation approach selects a λ that attains a value of the AUC that is very close to the maximum possible in a testing data set.

The larger gamma the less we are using the external data for that covariate, because of this we define a measure of relative weight of the external data as the norm of beta divided by the sum of the norm of beta and gamma. In Figure

We assess if the maximum AUC is attained by the λ from the cross-validation from 4.2.4. In Figure 4.4 we show the AUC for three of the ratio sample sizes of the analysis. The AUC of

the predictor with the seleted λ is close to the maximum but is not the maximum in any of the cases. This could be because the current algorithm does not takes into account that the sample size ratio between external and target data is different in the complete training data that in the stratified training data. In the panels the ratio of GfG over traing HRS is 20, 13.3, and 10. We used a cross-validation with $K = 10$ folds which means that in each fold the sample size ratio is 22.2, 14.78, and 11.1, which is around 10% higher external weight. A higher value of K so that the sample size ratios are more similar, or a modified version of the CV algorithm could improve the attainability of the optimal value. However, it is notable that even under the unequal sample size ratio the algorithm approximates the maximum well.

The larger the value of $\hat{\gamma}$ the more we are ignoring the external information, this is shown in Table 4.2 for this example. This poses the question the interpretation of the values of $\hat{\gamma}$ for each covariate. In Figure 4.5 we can see that the estimates of γ differ across covariates with some of them being close to zero and others having sizable values. It is very interesting that all of the values, except for Born between 1924 and 1930, are non decreasing. This is expected since we observed that the norm of $\hat{\gamma}$ was increasing. Nevertheless, there are two interesting cases for the values of $\hat{\gamma}$. The first is when the estimates increases considerably, which we interpret as covariates that initially were useful for increasing the prediction but as the sample size of the target data increased the emphasis on using the external information for those covariates decreased. The second case is less intuitive, it is the case for covariates for which the value of $\hat{\gamma}$ remained practically constant. This could have two opposite interpretations. Either the external information of the covariate was not useful to predict the target or that even with very similar sample sizes the information from the external data is useful for the target data. To be able to choose which of these interpretations is appropriate we can think about the magnitude of $\hat{\gamma}_j$ for each covariate. Similar to deciding when an effect size is clinically relevant, this will depend on the scale of the covariate. In Figure 4.5 all covariates are standardize to the target population and so the interpretation of the effect sizes is the change from the mean by one standard deviation. From these we can see that most of the covariates have small and constant values for the bias effect size with only Born '42-'47, Born '48-'53, and Born '54-'59 have a large effect size and they increase. These suggest that most of the change in prediction comes from ignoring the external data from these covariates.

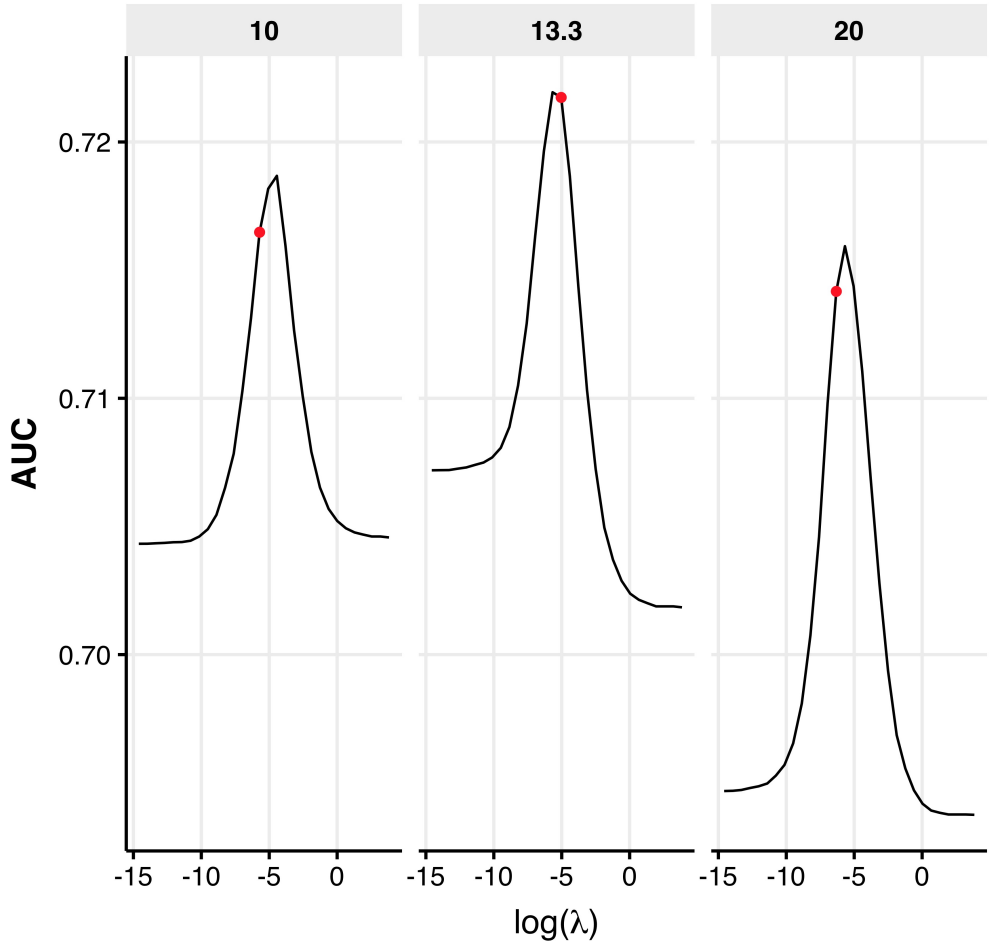

Figure 4.4: In the x-axis we present the value of the penalty λ in the log scale. In the y-axis is the AUC of the HRS data using $\hat{Y}_{HRS} = X_{HRS}\hat{\beta}_{DEGLR}(\lambda)$ as a predictor. With a red dot we mark the corresponding AUC of the predictor that uses the penalty from the cross validation.

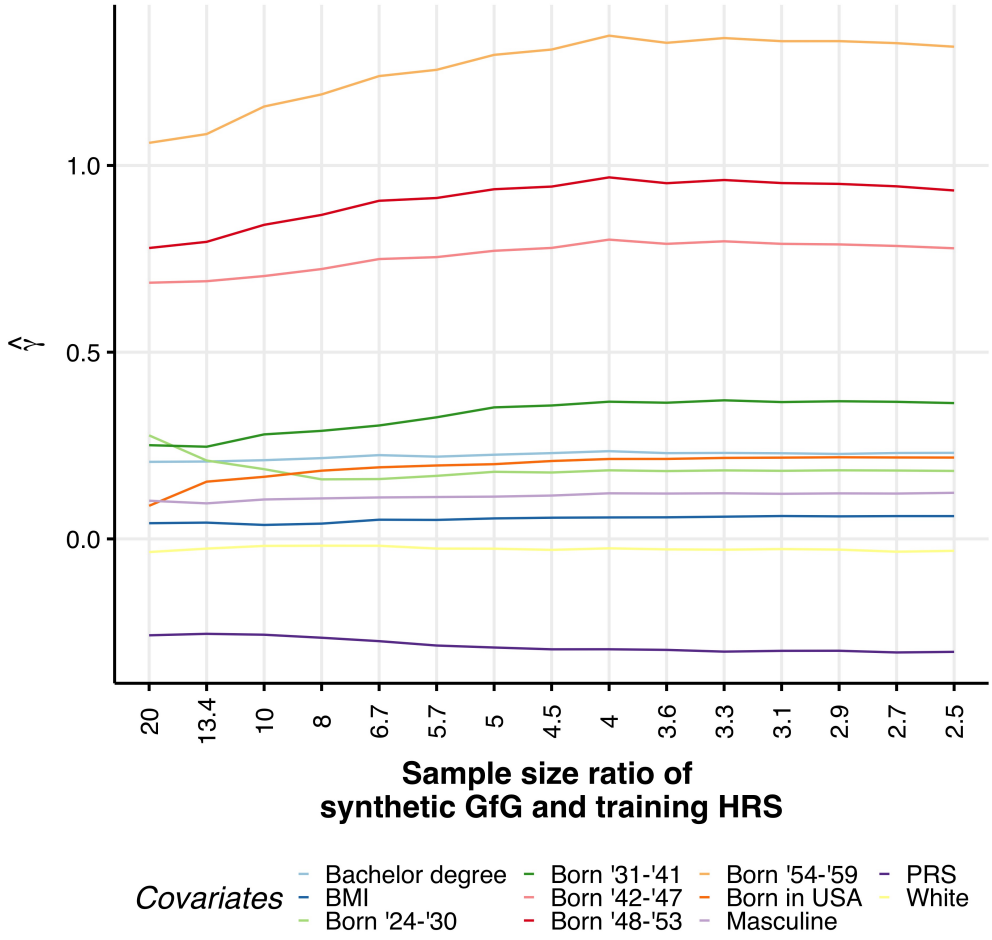

Figure 4.5: In the y-axis we see the $\hat{\gamma}$ for each of the covariates in the real data analysis. In the x-axis we see the sample size ratio of the synthetic GfG and the training HRS. For these analysis all covariates are standardize so the scale of the effect size is change for one standard deviation.

4.5 Discussion

We develop the Data Enriched Generalized Linear Regression (DEGLR) method in this work. We show that DEGLR is equivalent to a penalized linear regression problem when we augment the data to include a penalization on the target linear prediction bias with an appropriated cross-validation method to select the penalty λ. Our equivalence results do not depend on the functional form of the likelihood; thus, DEGLR can be fit through penalized linear regression for any model with likelihood. The objective of DEGLR is to leverage external data to improve prediction in target data. We use *glmnet* exclusively to obtain estimates of the DEGLR, and we propose a cross-validation algorithm that finds the λ that regulates the amount of information to borrow from the external data source. This is because penalized linear regression does not differentiate between target and external information, so it has no mechanism to ignore the external data if we choose the penalty λ as penalized regression would do. DEGLR chooses λ by cross-validation in the target data, which means that the loss function in the cross-validation only depends on the target data. Through simulations, we obtain the exact estimates as DELR across 12 different scenarios under the linear model by appropriately scaling the penalty (median concordance correlation of 0.999). Thus, DEGLR extends the Data Enriched Linear Regression (DELR) method from Chen, Owen, and Shi [15] for binomial, poisson, multinomial, cox, and multi-response Gaussian links, and we call it Data Enriched Generalized Linear Regression (DEGLR).

The DEGLR shines when pooling both data sources could harm more than benefit. How much the external information can improve or decrease prediction ability ultimately depends on a bias-variance trade-off. The DEGLR assumes a larger sample size in the external data source. If both data sources have the same sample size, the external data source could be helpful only under a very small bias. However, in this work, we focus on the case of having access to large studies with lower-quality data. A method to solve a similar problem is presented by Boonstra [8]. However, the difference is that this method assumes we have access to low-quality observations on all samples for high-quality observations only on a subset of the samples. This would be equivalent to our scenario if the target data set were a subset of the external data set.

In our real data analysis, we explored the utility of applying this method when one of the co-variates is a PRS, and the ancestral distribution of the data is different. This is related to chapters 2 and 3 of this dissertation. As Chapter 2 shows the impact of the predictive ability of PRS in non-European populations due to differences in LD, we expect this variable to have different predictive abilities in both data sources. This work relates to Chapter 3 because it attempts to address the problem of transferability after the PRS has all predictive variants aggregated, and Chapter 3 focuses on the problem before summing the risk variants. In our real data analysis, we see that the GfG has a lower AUC than HRS even when the sample size of GfG is 13.3 larger than the training

sample of HRS. This is not surprising when considering the results from Chapters 2 and 3.

Chen, Owen, and Shi presented Stein-type results on the inadmissibility of using only the target sample for a model with more than five predictors and more than 10 degrees of freedom. We did not explore if such a type of result extends to other link functions. However, we believe this could be interesting future work. Another interesting future work is to explore the method in the other link functions that are available in the *glmnet* package. We provide code to implement DEGLR that can include these other functions by only including the loss function associated with their likelihoods. Thus, this work increases the reach of Data Enriched Linear Regression to several link functions and provides a useful tool to improve prediction in a small sample with high-quality data sets when there is a large data set with overlapping covariates.x

CHAPTER 5

Discussion and Future Work

The driving topic of this dissertation is to study how different studies can be transferable or integrated. In Chapter 2, we studied the effect of LD on the transferability of European GWAS to non-European populations through extensive simulations. In our simulations, we isolated the effect of LD by assuming the same effect size in all populations. Then we conducted a GWAS in a European simulated data set and used the most significant variant (MS) as a predictor for the trait. Under our scenario, the predictive ability of the MS variant is the same for all populations if it is the true risk variant. The predictive ability of the MS variant when it is not causal is the effect size of the causal variant time their correlation. Since LD structures are different across populations, then the predictive ability of the MS variant will change across populations when it is not the causal variant. We estimated the probability of this scenario and found it to be significantly high even when the expected power is high. Thus our results suggest that increasing sample sizes of European GWAS will not suffice to overcome the problem of reduced transferability, which agrees with [35]. Thus, we estimated the correlation of the most significant variant with the true underlying risk variant in high-powered GWAS as a transferability measure. By this measure, we found a significant reduction in the predictive ability of a European MS variant in the African population and a mild reduction in Admixed American, East Asian, and South Asian populations due exclusively to the differences in LD.

Other factors impact the transferability of PRS. GWAS fits a regression model to each variant, adjusting for covariates. However, we do not account for all interactions of the variant with other variants or relevant environmental factors. Thus, the marginal associations contain the mean effect of the ignored interactions. As mean environmental exposure and allele frequencies are different across populations, then the marginal associations are going to be different across populations[82]. An additional effect of different allele frequencies manifests through different proportions of variance explained by the same variant in different populations[37]. This is a much more complex problem because environmental interactions ideally should be studied longitudinally, they are highly complex as even within families members can vary significantly, and the impact of environmental exposures varies significantly across traits. Some methods estimate the genome-wide

correlation of causal variant effect sizes at shared SNPs that account for different LD structures [9][27]. However, none of these methods can quantify the effect of genetic and environmental interactions separately. Patel et al. [57] developed a method to study how genetic interactions drive the heterogeneity of causal effect sizes. Measuring the environment interaction by itself has proved to be more challenging. Because of this, a better understanding of the impact of interactions on the transferability of PRS is a potential area for future work that will help develop methods.

The genetic component of traits is exceptionally diverse. Some traits have a substantial genetic component with few causal variants. Others have thousands of variants with small effects, which could explain a large portion of the trait's variance. In addition, genetics very often acts through the environment. Some environments might be beneficial for one trait but detrimental for another. With all the different possibilities for genetics to manifest in human traits, it is naive to overcome transferability through one universal solution. Thus, there is a wide range of needs in method development that has to be studied.

Currently, there are several projects exist that are very intentional in collecting samples from non-European populations [51][31][82][1]. However, increasing samples of underrepresented groups must be done through community inclusion [28] and protecting against commodification[26]. Moreover, even after all the previously mentioned efforts, the overrepresentation of European samples has not changed significantly up to 2021[23]. Thus, there is a need for methods designed explicitly to improve the prediction and estimation of targeted populations to empower researchers from these communities to leverage the existing data efficiently.

There are several methods that improve transferability of PRS across populations.[53] [47] [29] [18] [80][63] However, these methods focus on constructing a PRS that could have similar prediction ability across populations. This type of PRS might hinder population-specific discoveries in underrepresented populations. Chapter 3 presents a Power Prior approach to model regions using summary statistics. We show that it is robust to an unbalanced sample size and unequal correlation. Also, the method works with only summary statistics, making it possible to use it with publicly available data. However, the method only considers one external population, which is a limitation. A first extension to the Power Prior from Chapter 3 would include more than one external population by adding their likelihoods with a power parameter. The concern of that approach is that this assumes that all external populations are independent and have no overlap information. This assumption can overestimate the external information about specific parameters making the optimization of the power parameters inefficient and inaccurate. Thus, extending Chapter 3 to several external populations is not a trivial problem and poses an exciting area for future work.

Leveraging research across studies is not limited to genetic studies. In recent years access to observational data has increased dramatically. For example, electronic health records (EHR) collect people's medical history every time they attend a hospital. Researchers are interested in

utilizing this massive information source to investigate health-related questions. This area of research is highly active, with a strong focus on causal inference. One example is that due to inclusion/exclusion criteria for RCT, often, their results are valid only for a subset of a population. On the other hand, extensive observational studies represent a more significant portion of the population but often have incomplete information on confounding variables. Thus, it is often of interest to increase the transferability of randomized clinical trials (RCT) to broader populations by integrating RCT with observational studies, borrowing sample size strength from the observational study and unbiased estimators from RCT. Another example is the multi-task learning scenario which integrates several small studies. The objective is that each task benefits from the shared structure and, at the same time, retains information that is task-specific[30][21]. Also, there is an extensive literature on leveraging historical studies to increase estimation efficacy on current studies with a larger set of covariates [7] [16] [69] [14]. Similarly, as we mentioned in the genetic setting, none of these methods accounts for the case in which the interest resides in uplifting one population.

The setting for the DELR is to increase prediction in a target data set by leveraging a large external study [15]. The downside of DELR is that it only applies to continuous outcomes with a Gaussian error. Most biomedical outcomes do not follow this assumption, which significantly limits its DELR. Chapter 4 extends DELR to DEGLR. This extension leverages the connection between the objective function of DEGLR and penalized linear regressions for using the existing software *glmnet*. We implemented a tailor CV algorithm to find the optimal penalty to increase prediction in the target population. With simulations and a real data example, we show that DEGLR can increase prediction ability when an external data set has a low bias but continuously shifts the weight of the external information as bias increases to eventually ignore it. Even though DEGLR focuses on prediction, the equivalence with penalized regression makes an easy transition to use the method for estimation by changing the penalty to the bias vector γ. The CV algorithm can remain the same, only modifications on the loss function if desired.

An essential gain of using DEGLR for estimation is removing a limiting assumption. In the case of prediction, we use the Cholesky decomposition of the matrix of sums of squares of the target data. Cholesky decomposition applies to a positive definite matrix, implying that the target sample size has to be less than the number of variables. An assumption that penalized regression not only does not make but is one of the motivations for its development. Thus, DEGLR for estimation also constitutes interesting future work as extensive literature on penalized regression is already available.

There is always going to work to be done to help reduce disparities. Statistics, Data Science, Computer Science, and other sciences involved in using data to develop prediction models have unintentionally increased inequalities[55]. This dissertation develops two methods to intentionally uplift small data sets by safely utilizing large sample studies. The problem presents itself as

something very similar to existing problems, but in this dissertation, we learned that there are some specific challenges to this setting. A critical challenge of this setting is that the bias variance trade-off plays an important role. When the external studies have a much larger relative sample size than the target study, the gain in variance reduction can overwhelm bias. In principle, this is not a problem when prediction is the interest. However, in the case of estimation, this can make the methods very unstable. For example, in the case of the Power Prior, the optimal power parameter would most likely be close to the boundaries of the space parameter for this case. Further research on this specific statistical setting is needed as it can help researchers from underrepresented populations adequately use the massive amount of research done in better-studied populations.

APPENDIX A

Supplemental material for Chapter 2

Simulating Case-Control Data Sets To maintain the LD structure around a causal variant, we sample with replacement haplotypes from the 1000 Genome Project from the four super populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). We will restrict our set of causal variants to those that have AF between 0.1 and 0.5 in the European population to ensure high power in the European GWAS. To simulate each data set, we first simulate the cases by calculating the probability of being sampled for each haplotype. Let D_i be a binary variable that takes the value 1 when the individual is a case and 0 when is a control, G_i be the dosage of the causal variant, β is the effect size in log odds ratio scale, and β_0 is the intercept that depends on the prevalence of the disease. Thus, Equation (1) is the probability of sampling a haplotype with a dosage of G_i of the causal variant, conditional on disease status.

$$P(H_i = h|D_i) = P(D_i|H_i = h)\frac{P(H_i = h)}{P(D_i)} \tag{A.1}$$

$$P(H_i = h|D_i) = \left(\frac{e^{\beta_o + h^\top\theta}}{1 + e^{\beta_0 + h^\top\theta}}\frac{P(H_i = h)}{P(D_i)}\right)^{D_i}\left(\frac{1}{1 + e^{\beta_0 + h^\top\theta}}\frac{P(H_i = h)}{P(D_i)}\right)^{1-D_i} \tag{A.2}$$

$$= \left(\frac{e^{\beta_o + G_{ih}\beta}}{1 + e^{\beta_0 + G_{ih}\beta}}\frac{P(H_i = h)}{P(D_i)}\right)^{D_i}\left(\frac{1}{1 + e^{\beta_0 + G_{ih}\beta}}\frac{P(H_i = h)}{P(D_i)}\right)^{1-D_i}, \tag{A.3}$$

where G_{ih} is the dosage of the risk variant in haplotype h of individual i, and θ is a vector that is zero for all non risk variants and β for the risk variant.

We first simulate the cases by fixing $D_i = 1$, and then sample with replacement until we reach the desired sample size. Then repeat the same for controls by fixing $D_i = 0$ to get the probability of being sampled for each haplotype. The code for the sampling by replacement procedure was done with R and can be found here https://github.com/orozcodelpinopedro/MS.variant.simulations. We will run the simulation pipeline one time per variant across all variants with an allele frequency between 0.1 and 0.5 in the European population.

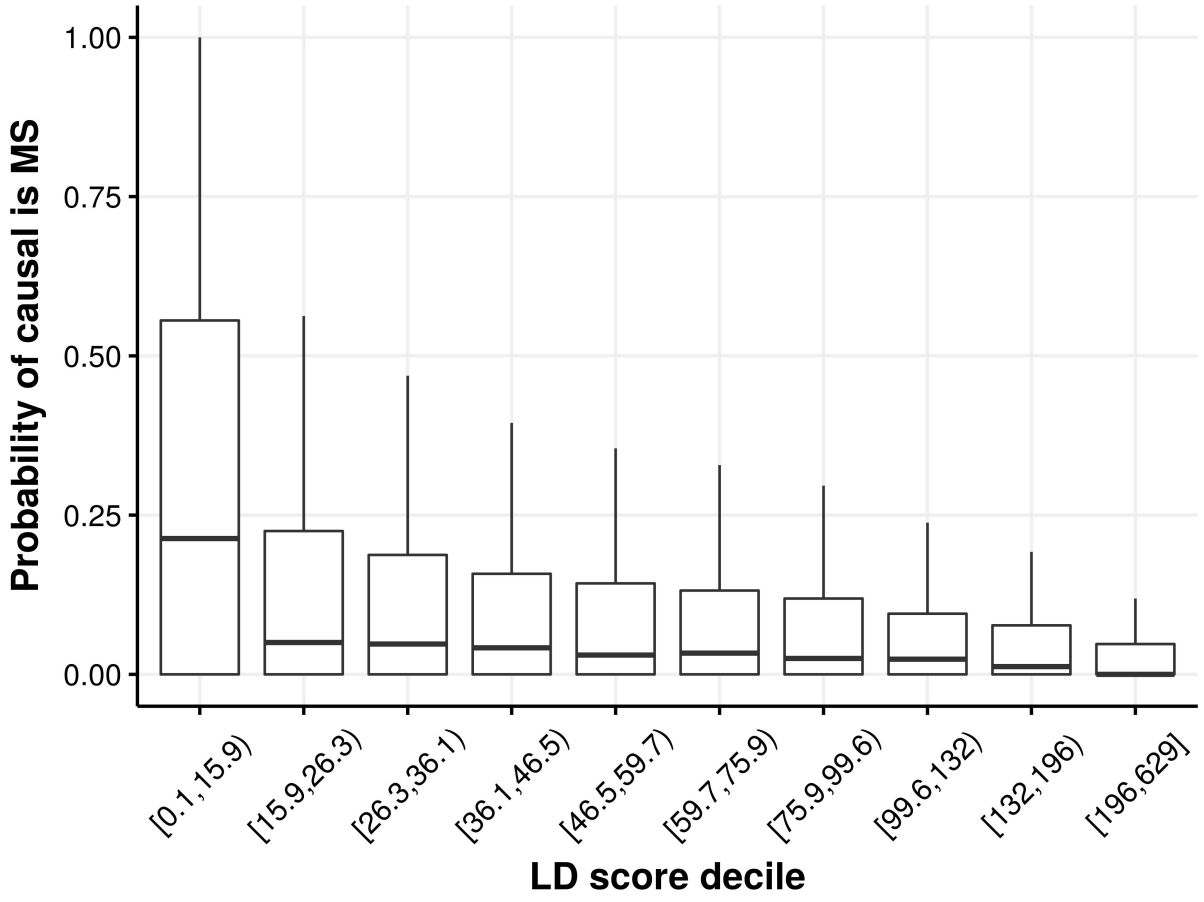

Figure A.1: We show the distribution of the probability of the causal variant being most significant for deciles of LD-score.

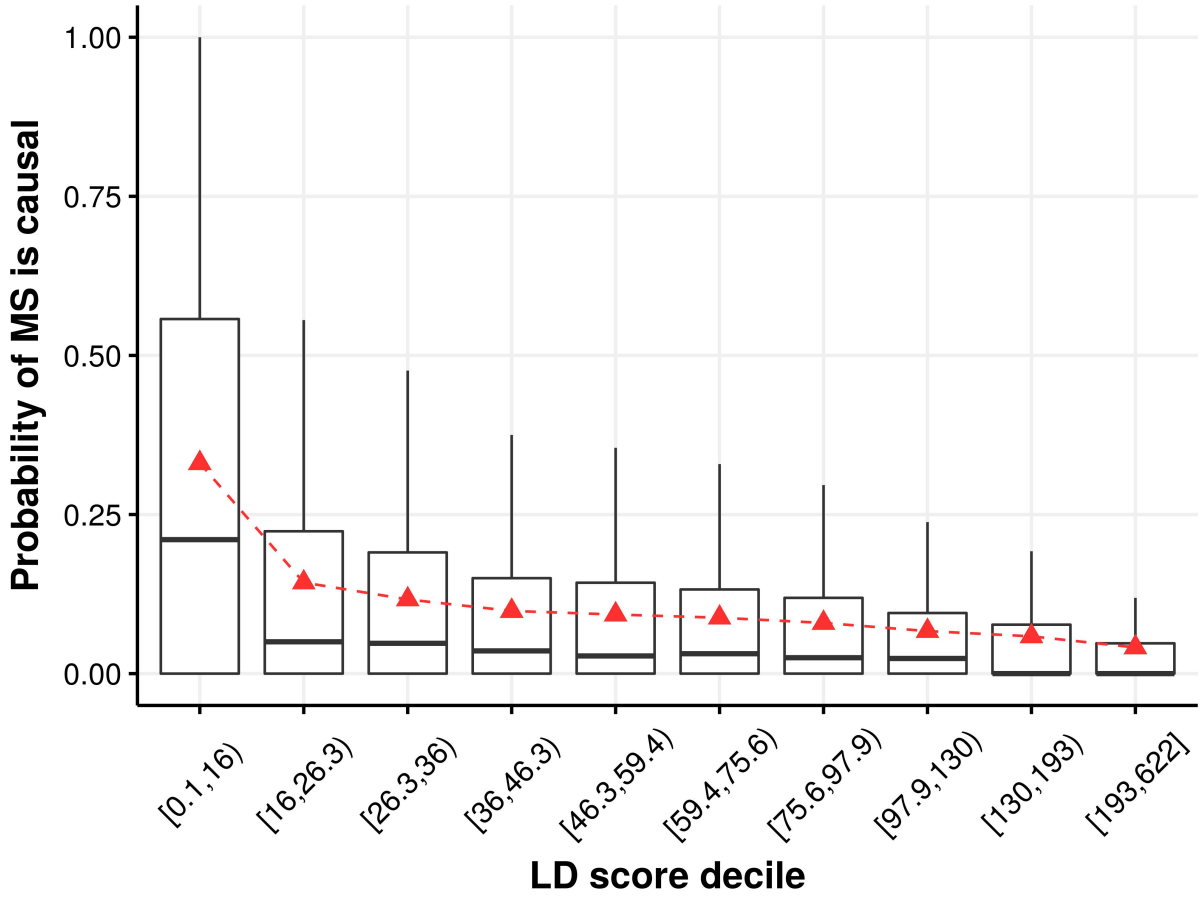

Figure A.2: We show the distribution of the probability of the MS variant being causal for each decile of LD-score for all simulated variants in chromosome 22. In red we present the mean probability of the MS variant being the causal variant.

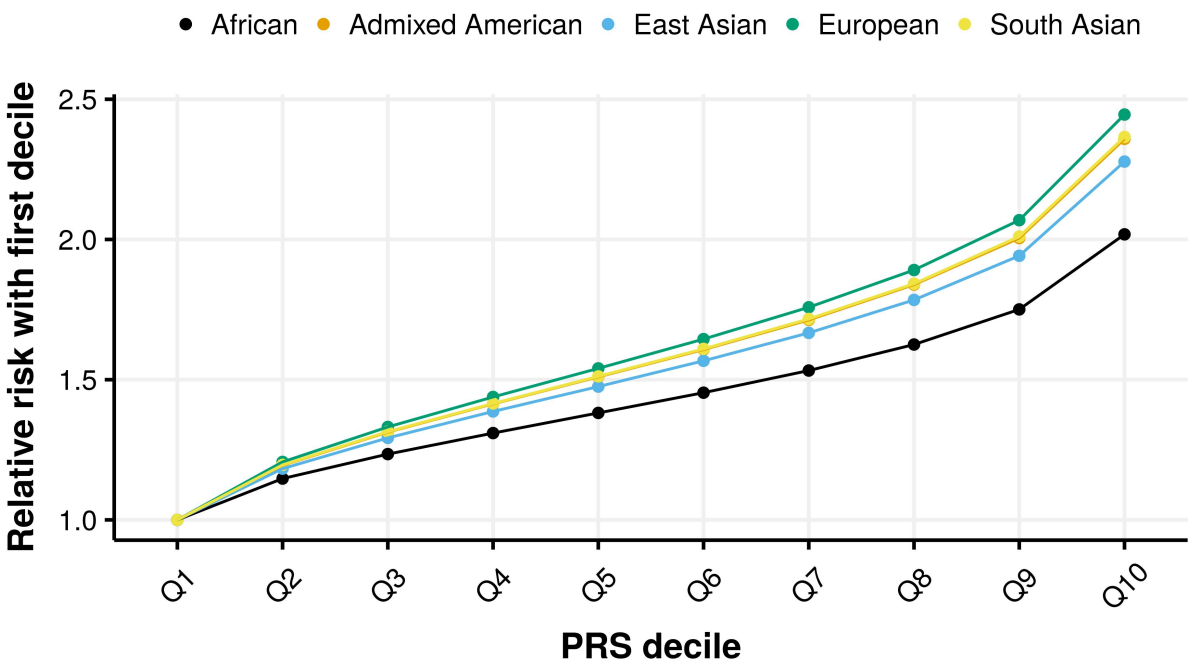

Figure A.3: We sample from the distribution of $\hat{\rho}_*$ to generate a PRS of 30 risk variants with same effect size for all 30 variants and same true effect size across populations. We then calculated the prevalence for different quantiles of the PRS distribution for all five populations.

Number of risk variants in the PRS	Population	Relative risk	Normalized relative risk
5*10	European	1.81	1.00
	South Asian	1.76	0.97
	Admixed American	1.76	0.97
	East Asian	1.70	0.94
	African	1.57	0.87
5*20	European	2.20	1.00
	South Asian	2.13	0.97
	Admixed American	2.13	0.97
	East Asian	2.04	0.92
	African	1.87	0.83
5*30	European	2.45	1.00
	South Asian	2.37	0.97
	Admixed American	2.36	0.96
	East Asian	2.28	0.93
	African	2.02	0.83

Table A.1: Estimated reduction in predictive ability of European based MS variants for different number of risk variants in the PRS. The relative risk is calculated as the ratio of the prevalence of the trait between individuals in the bottom 10% of the distribution of PRS and individuals in the top 10% of the distribution of PRS. Normalized relative risk is the relative risk divided by the European relative risk.

Impact of under estimating odds ratio on the predictive ability of a PRS

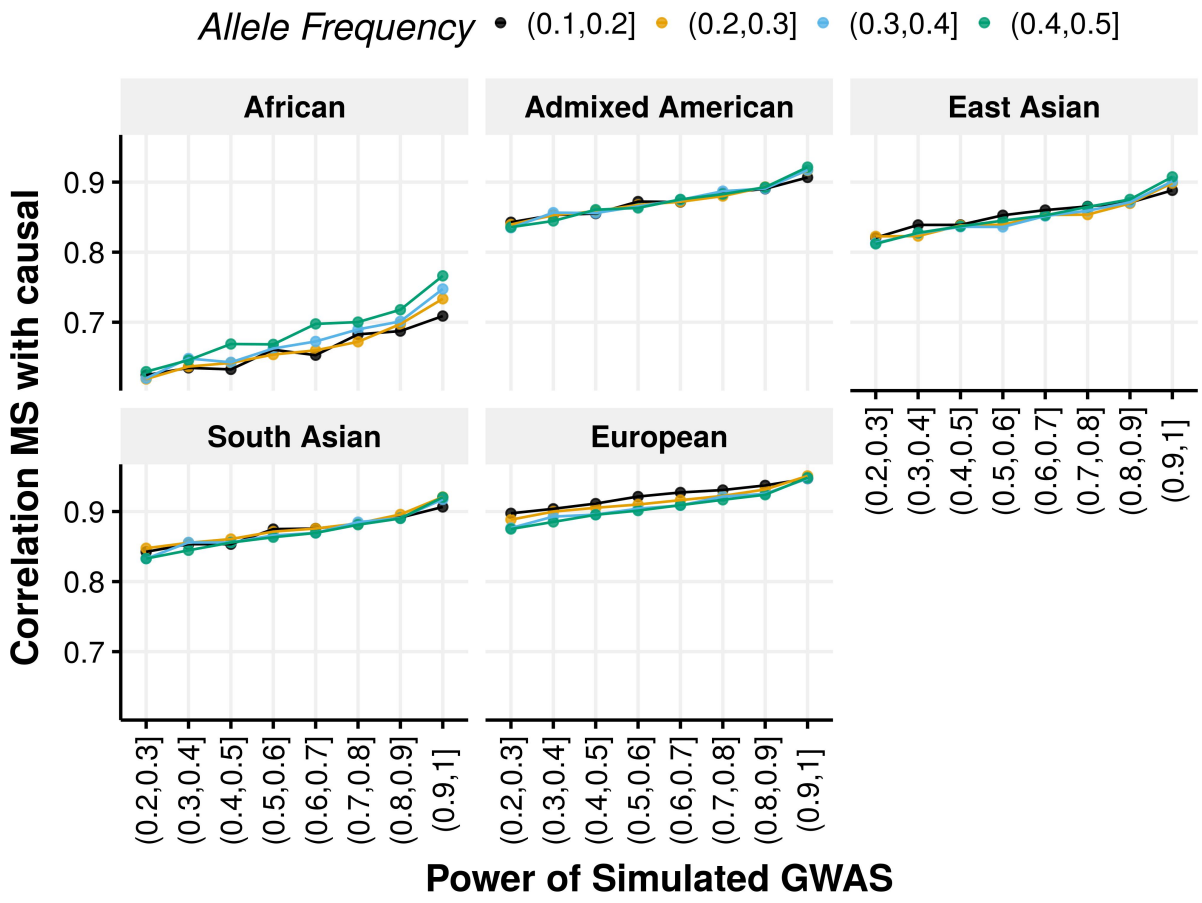

Figure A.4: In the horizontal axis we plot the power of the European discovery GWAS. The vertical axis has the mean correlation coefficient between the most significant variant and the causal variant. The color is the allele frequency of the causal variant.

APPENDIX B

Supplemental material for Chapter 3

B.1 Power prior model for a multivariate model

In this section we describe a model to apply at a locus level. We parametrize Gaussian distributions with mean and precision, thus $N(x; \mu, \tau\mathbb{I})$ is the density of a Gaussian distribution with mean vector μ, precision matrix $\tau\mathbb{I}$ evaluated at x. Consider genetic data in which data source refers to genetic data sampled from individuals from the same ancestral population, thus we use data source and population interchangeably. Let $Y_e \in \mathbb{R}^{n_e}$ be the standardized phenotype from an external population, $X_e \in \mathbb{R}^{n_e \times m}$ a centered genotype matrix information from an external population at a region with m variants. Consider the following multivariate regression model.

$$Y_e = X_e\beta + \epsilon, \tag{B.1}$$

where $\epsilon \sim N(0, 1/\tau_e)$ is a $n_e \times 1$ vector of independent errors and τ_e is a precision parameter. Assuming known τ_e, the likelihood for β is $L(\beta|Y_e, X_e, \tau_e) = N(Y_e; X_e\beta, \tau_e^{-1}\mathbb{I})$. In addition with the likelihood and a ridge prior $\pi_0(\beta|\tau) = N(\beta; 0, \tau\mathbb{I})$ the power prior conditional on the external population data, a_0, and the ridge parameter τ is $P(\beta|Y_e, X_e, \tau_e, a_0, \tau) = N(Y_e; \mu_e, \nu_e)$, where $\nu_e = a_0\tau_e X_e^\top X_e + \tau\mathbb{I}$, and $\mu_e = \nu_e^{-1} a_0\tau_e X_e^\top Y_e$. For $a_0 = 0$ the power prior coincides with the ridge prior. When $a_0 = 1$ the power prior coincides with using the posterior distribution of the external population with a ridge prior as a prior. A convenient result of having Gaussian priors is that the power prior conditional on a_0 remains as a Gaussian distribution for any prior $\pi(a_0|\gamma)$ (See Appendix for details).

Let $Y_t \in \mathbb{R}^{n_t}$ be the standardized phenotype from a target population, $X_t \in \mathbb{R}^{n_t \times m}$ a centered genotype matrix information from a target population at the same region as the external population with m variants. Assume the same model as 4.1 but with a precision parameter τ_t. The power prior posterior distribution conditional on a_0, the target data, and the external data for a multivariate model is

$$
\begin{aligned}
P(\beta|X_t, Y_t, \tau_t, X_e, Y_e, \tau_e, \tau, a_0) &= N(Y_t; X_t\beta, \tau_t\mathbb{I})N(\mu_e; \beta, \nu_e) \\
&\propto exp\left(-\frac{1}{2}\left(\tau_t(Y_t - X_t\beta)^\top(Y_t - X_t\beta) + (\mu_e - \beta)^\top\nu_e(\mu_e - \beta)\right)\right) \\
&\propto exp\left(-\frac{1}{2}\left(\beta^\top(\tau_t X_t^\top X_t + \nu_e)\beta - 2\beta^\top(\tau_t X_t^\top Y_t + \underbrace{\nu_e\mu_e}_{a_0\tau_e X_e^\top Y_e})\right)\right) \\
&= exp\left(-\frac{1}{2}\left(\beta^\top\nu_{te}\beta - 2\beta^\top\nu_{te}\nu_{te}^{-1}(\tau_t X_t^\top Y_t + a_0\tau_e X_e^\top Y_e)\right)\right) \\
&\propto exp\left(-\frac{1}{2}\left(\mu_{te} - \beta\right)^\top\nu_{te}(\mu_{te} - \beta)\right),
\end{aligned}
$$

where $\nu_{te} = \tau_t X_t^\top X_t + a_0\tau_e X_e^\top X_e + \tau\mathbb{I}$, and $\mu_e = \nu_{te}^{-1}(\tau_t X_t^\top Y_t + a_0\tau_e X_e^\top Y_e)$. Thus the posterior distribution of β conditional on the target data, the external data, and the power parameter is $N(\mu_{te}; \beta, \nu_{te})$.

B.2 Power prior conditional on a_0 is Gaussian

We calculate the normalizing constant $c(a_0)^{-1} = \int L(\beta|Y_e, X_e, \tau_e)^{a_0}\pi_0(\beta|\tau)d\beta$

$$
\begin{aligned}
c(a_0)^{-1} &= \int L(\beta|Y_e, X_e, \tau_e)^{a_0}\pi_0(\beta|\tau)d\beta \\
&= \sqrt{\frac{\tau_e^m\tau}{(2\pi)^{(m+1)}}}\int exp\left(-\frac{1}{2}\left(a_0\tau_e(Y_e - X_e\beta)^\top(Y_e - X_e\beta) + \tau\beta^\top\beta\right)\right)d\beta \\
&= \sqrt{\frac{\tau_e^m\tau}{(2\pi)^{(m+1)}}}\int exp\left(-\frac{1}{2}\left(a_0\tau_e(Y_e^\top Y_e - 2\beta^\top X_e^\top Y_e + \beta^\top X_e^\top X_e\beta) + \tau\beta^\top\beta\right)\right)d\beta \\
&= \sqrt{\frac{\tau_e^m\tau}{(2\pi)^{(m+1)}}}\int exp\left(-\frac{1}{2}\left(\beta^\top(a_0\tau_e X_e^\top X_e + \tau\mathbf{I})\beta - 2a_0\tau_e\beta^\top X_e^\top Y_e + a_0\tau_e Y_e^\top Y_e\right)\right)d\beta \\
&= \sqrt{\frac{\tau_e^m\tau}{(2\pi)^{(m+1)}}}exp\left(-\frac{1}{2}\left(a_0\tau_e Y_e^\top Y_e - \mu^\top\nu\mu\right)\right)\underbrace{\int exp\left(-\frac{1}{2}\left((\mu - \beta)^\top\nu(\mu - \beta)\right)\right)d\beta}_{\text{Kernel of } N(\mu;\beta,\nu)} \\
&= \sqrt{\frac{\tau_e^m\tau|\nu^{-1}|}{2\pi}}exp\left(-\frac{1}{2}(a_0\tau_e Y_e^\top Y_e - \mu^\top\nu\mu)\right).
\end{aligned}
$$

Thus, we can now see the conditional distribution of β given a_0, τ, the external information (X_e, Y_e), and τ_e.

$$
\begin{aligned}
\pi(\beta|a_0, X_e, Y_e, \tau_e, \tau) =& L(\beta|Y_e, X_e, \tau_e)^{a_0}\pi_0(\beta|\tau)c(a_0) \\
=& \frac{\sqrt{\frac{\tau_e^m\tau}{(2\pi)^{(m+1)}}}exp\left(-\frac{1}{2}\left(a_0\tau_e(Y_e - X_e\beta)^\top(Y_e - X_e\beta) + \tau\beta^\top\beta\right)\right)}{\sqrt{\frac{\tau_e^m\tau|\nu^{-1}|}{2\pi}}exp\left(-\frac{1}{2}(a_0\tau_e Y_e^\top Y_e - \mu^\top\nu\mu)\right)} \\
=& \frac{1}{\sqrt{(2\pi)^m|\nu^{-1}|}}\frac{exp\left(-\frac{1}{2}a_0\tau_e Y_e^\top Y_e\right)exp\left(-\frac{1}{2}\left(\beta^\top\nu\beta - 2\beta^\top\nu\mu\right)\right)}{exp\left(-\frac{1}{2}(a_0\tau_e Y_e^\top Y_e\right)exp\left(\frac{1}{2}\mu^\top\nu\mu\right))} \\
=& \frac{1}{\sqrt{(2\pi)^m|\nu^{-1}|}}exp\left(-\frac{1}{2}\left(\mu - \beta\right)^\top\nu(\mu - \beta)\right) \\
=& N(\mu; \beta, \nu).
\end{aligned}
$$

Note that the resulting Gaussian is the same as the one when a_0 was considered fixed.

B.3 Power prior with regression summary statistics posterior mean in terms of S and R

From Equation 3.8 we obtain an analytic solution to the posterior mean of the PP-RSS method. The expression depends on the matrix product $\theta\nu\theta$, where $\theta = SRS^{-1}$ and $\nu = (SRS)^{-1}$. Here are the calculations of $\theta\nu\theta$ in terms of S and R.

$$
\begin{aligned}
\theta\nu\theta &= SRS^{-1}(SRS)^{-1}SRS^{-1} \\
&= SRS^{-1}S^{-1}R^{-1}S^{-1}SRS^{-1} \\
&= SRS^{-3}.
\end{aligned}
$$

The other expression in the posterior mean is the product $\theta\nu$ which in terms of S and R is $\theta\nu = SRS^{-2}(RS)^{-1}$. Thus the posterior mean from Equation 3.8 in terms of S and R is.

$$
\hat{\beta}_{te} = (S_tR_tS_t^{-3} + a_0S_eR_eS_e^{-3} + \tau\mathbb{I})^{-1}(S_tR_tS_t^{-2}(R_tS_t)^{-1}\tilde{\beta}_t + a_0S_eR_eS_e^{-2}(R_eS_e)^{-1}\tilde{\beta}_e). \quad \textbf{(B.2)}
$$

B.4 Explained heritability with different power parameters

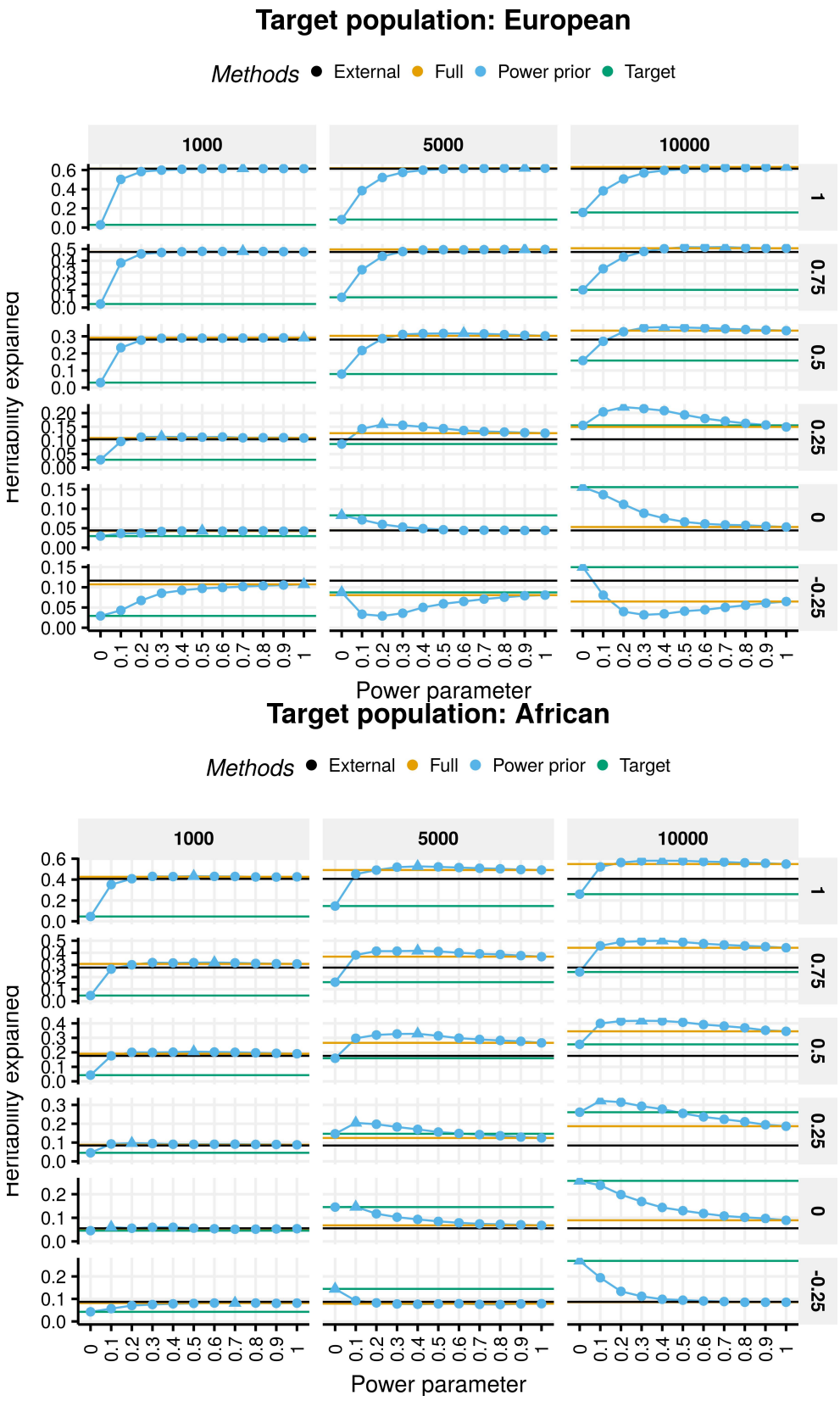

Figure B.1: Explained heritability obtained with different choices of the power parameter. Color represents the four different methods. Both plots have an external population with European ancestry.

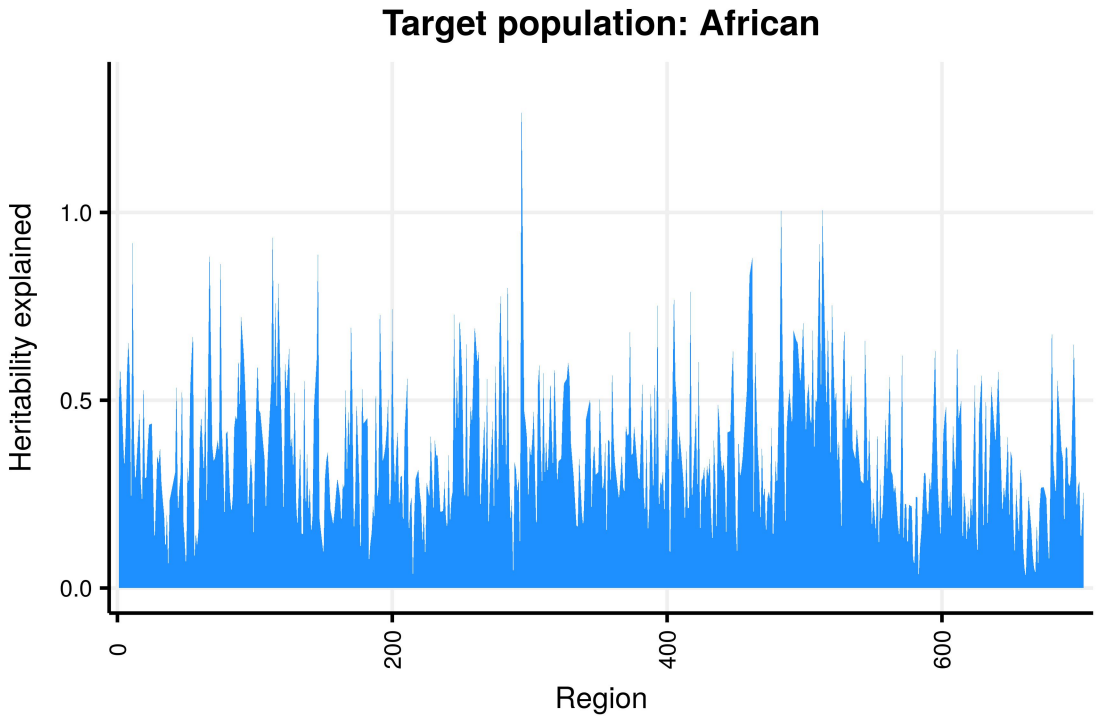

Figure B.2: Heritability explained by the Power Prior with pseudo correlation method in 100 simulations for the 703 regions (See Methods for details).

APPENDIX C

Supplemental material for Chapter 4

C.1 Details on the DELR estimate

The DELR estimate is such that minimizes Equation 4.3. By deriving f with respect to β and γ we can obtain the following normal equations.

$$(V_t + V_e)\hat{\beta}_\lambda = V_t\hat{\beta} + V_e\hat{\alpha} - V_e\hat{\gamma}_\lambda \tag{C.1}$$

$$(\lambda V_t + V_e)\hat{\gamma}_\lambda = V_e\hat{\alpha} - V_e\hat{\beta}_\lambda \tag{C.2}$$

We solve for $\hat{\gamma}_\lambda$ in C.2 and plug it in C.1.

$$(V_t + V_e - V_e(\lambda V_t + V_e)^{-1}V_e)\hat{\beta}_\lambda = V_t\hat{\beta} + V_e\hat{\alpha} - V_e(\lambda V_t + V_e)^{-1}V_e\hat{\alpha} \tag{C.3}$$

Let $\hat{W}_\lambda = (V_t + V_e - V_e(\lambda V_t + V_e)^{-1}V_e)^{-1}V_t$, then

$$\begin{aligned}
\hat{\beta}_\lambda &= \hat{W}_\lambda\hat{\beta} + \hat{W}_\lambda V_t^{-1}(V_e\hat{\alpha} - V_e(\lambda V_t + V_e)^{-1}V_e\hat{\alpha}) \\
&= \hat{W}_\lambda\hat{\beta} + \hat{W}_\lambda V_t^{-1}(\hat{V}_t\hat{\alpha} + V_e\hat{\alpha} - V_e(\lambda V_t + V_e)^{-1}V_e\hat{\alpha} - \hat{V}_t\hat{\alpha}) \\
&= \hat{W}_\lambda\hat{\beta} + \hat{W}_\lambda V_t^{-1}(\hat{V}_t\hat{\alpha} + V_e\hat{\alpha} - V_e(\lambda V_t + V_e)^{-1}V_e\hat{\alpha}) - \hat{W}_\lambda\hat{\alpha} \\
&= \hat{W}_\lambda\hat{\beta} + \hat{W}_\lambda V_t^{-1}(\hat{V}_t + V_e - V_e(\lambda V_t + V_e)^{-1}V_e)\hat{\alpha} - \hat{W}_\lambda\hat{\alpha} \\
&= \hat{W}_\lambda\hat{\beta} + \hat{\alpha} - \hat{W}_\lambda\hat{\alpha} \\
&= \hat{W}_\lambda\hat{\beta} + (\mathbf{I} - \hat{W}_\lambda)\hat{\alpha}
\end{aligned}$$

We now want to show that W_λ simplifies to $[(\lambda + 1)V_e + \lambda V_t]^{-1}(V_e + \lambda V_t)$.

$$
\begin{aligned}
\hat{W}_\lambda &= (V_t + V_e - V_e(\lambda V_t + V_e)^{-1}V_e)^{-1}V_t \\
&= \left([(V_t + V_e)V_e^{-1}(\lambda V_t + V_e) - V_e](\lambda V_t + V_e)^{-1}V_e\right)^{-1} V_t \\
&= [(V_t V_e^{-1} + \mathbf{I})(\lambda V_t + V_e) - V_e]^{-1}(\lambda V_t + V_e)V_e^{-1}V_t \\
&= [\lambda V_t V_e^{-1}V_t + \lambda V_t + V_t + V_e - V_e]^{-1}(\lambda V_t + V_e)V_e^{-1}V_t \\
&= [\lambda V_t V_e^{-1} + (\lambda + 1)\mathbf{I}]^{-1}V_t^{-1}(\lambda V_t + V_e)V_e^{-1}V_t \\
&= [\lambda V_t V_e^{-1} + (\lambda + 1)\mathbf{I}]^{-1}(\lambda V_e^{-1} + V_t^{-1})V_t \\
&= [\lambda V_t V_e^{-1} + (\lambda + 1)\mathbf{I}]^{-1}(\lambda V_e^{-1}V_t + \mathbf{I}) \\
&= [\lambda V_t V_e^{-1} + (\lambda + 1)\mathbf{I}]^{-1}V_e^{-1}(\lambda V_t + V_e) \\
&= [\lambda V_t + (\lambda + 1)V_e]^{-1}(\lambda V_t + V_e)
\end{aligned}
$$

C.2 Limit of W_λ

To find the limits of W_λ when $\lambda \to \infty$ it is best to express as the sum of two components $W_\lambda = [\lambda V_t + (\lambda + 1)V_e]^{-1}V_e + [\lambda V_t + (\lambda + 1)V_e]^{-1}\lambda V_t$. It is trivial that $\lambda \approx \lambda + 1$ for any big λ. Thus $[\lambda V_t + \lambda V_e]^{-1}V_e + [\lambda V_t + \lambda V_e]^{-1}\lambda V_t$ has the same limit as W_λ. Thus

$$
\begin{aligned}
\lim_{\lambda\to\infty} W_\lambda &= \lim_{\lambda\to\infty} \left([\lambda V_t + \lambda V_e]^{-1}V_e + [\lambda V_t + \lambda V_e]^{-1}\lambda V_t\right) \\
&= \lim_{\lambda\to\infty} \left(\frac{1}{\lambda}[V_t + V_e]^{-1}V_e + [V_t + V_e]^{-1}V_t\right) \\
&= [V_t + V_e]^{-1}V_t
\end{aligned}
$$

C.3 Degrees of freedom of DELR

Here we give a different expression of the degrees of freedom of Model 4.3 that is useful to understand the monotonicity of the degrees of freedom ad a function of λ.

Let $M = V_t^{1/2}V_e^{-1}V_t^{1/2}$, now let $M = UDU^\top$ be the eigen value decomposition of M. Then we can write W_λ in terms of M.

$$
\begin{aligned}
W_\lambda &= [V_e + \lambda V_t + \lambda V_e]^{-1}(V_e + \lambda V_t) \\
&= [V_e + \lambda V_t + \lambda V_e]^{-1}(V_e^{-1}V_t)^{-1}(V_e^{-1}V_t)(V_e + \lambda V_t) \\
&= [V_t + \lambda V_t V_e^{-1} V_t + \lambda V_t]^{-1}(V_t + \lambda V_t V_e^{-1} V_t) \\
&= [V_t^{1/2}(\mathbf{I} + \lambda V_t^{1/2} V_e^{-1} V_t^{1/2} + \lambda \mathbf{I}) V_t^{1/2}]^{-1} V_t^{1/2}(\mathbf{I} + \lambda V_t^{1/2} V_e^{-1} V_t^{1/2}) V_t^{1/2} \\
&= [\mathbf{I} + M + \lambda \mathbf{I}]^{-1}(\mathbf{I} + \lambda M)
\end{aligned}
$$

Then $tr\,(W_\lambda) = tr([\mathbf{I} + M + \lambda \mathbf{I}]^{-1}(\mathbf{I} + \lambda M))$, and thus another expression for the degrees of freedom is

$$
df(\lambda) = \sum_{j=1}^{p} \frac{1 + \lambda \nu_j}{1 + \lambda + \lambda \nu_j}, \tag{C.4}
$$

where $\nu = diag(D)$.

C.4 Relationship between DEGLR and the penalized regression implemented by *glmnet*

The objective function from Equation 4.9 is very similar to the objective function of the penalized regression implementation from the package *glmnet*. For outcomes that are not continuous with log likelihood $l(\theta; X, Y)$ for θ, and data (X, Y), where Y is the outcome, X are the covariates, and N is the sample size the penalized regression objective function can be written as follows.

$$
f(\theta; \lambda^*) = -l(\theta; X, Y)/N + \lambda^*/2||\theta||_2^2. \tag{C.5}
$$

Let $Y^\top = (Y_t^\top, Y_e^\top)$, $\theta = (\beta, \gamma^*)$, $N = n_t + n_e$, and $X = \begin{pmatrix} X_t & 0_{p \times p} \\ X_e & X_e L \end{pmatrix}$, where L will be specified later. Then we can show that function C.5 is equivalent to function 4.9 when we put no penalty on β and adjusting λ by a factor of $n_e + n_t$. Since we will do differential shrinkage so that we do not impose any shrinkage on β, then from now on, we include only γ^* in the penalty instead of θ. Thus,

$$
\begin{aligned}
f(\theta;\lambda^*) &= -l(\theta;X,Y)/N + \lambda^*/2||\gamma^*||_2^2 \\
&= -1/N\sum_{i=1}^{n_t} l(\theta;X_i^\top,Y_i) - 1/N\sum_{i=n_t+1}^{n_e+n_t} l(\theta;X_i^\top,Y_i) + \lambda^*/2||\gamma^*||_2^2 \\
&= -1/N\sum_{i=1}^{n_t} l(\theta;(X_{ti}^\top,0_{1\times p}),Y_{ti}) - 1/N\sum_{i=1}^{n_e} l(\theta;(X_{ei}^\top,X_{ei}^\top L),Y_{ei}) + \lambda^*/2||\gamma^*||_2^2 \\
&= -1/N\sum_{i=1}^{n_t} l_t(\beta;X_{ti},Y_{ti}) - 1/N\sum_{i=1}^{n_e} l_e(\beta+L\gamma^*;X_{ei},Y_{ei}) + \lambda^*/2(||\gamma^*||_2^2) \\
&= -l_t(\beta;X_t,Y_t) - l_e(\beta+L\gamma^*;X_e,Y_e) + \lambda^* N/2{\gamma^*}^\top\gamma^*
\end{aligned}
$$

Let $\gamma = L\gamma^*$ then $L^{-1}\gamma = \gamma^*$ which gives the following equivalence

$$
f(\theta;\lambda^*) = -l_t(\beta;X_t,Y_t) - l_e(\beta+\gamma;X_e,Y_e) + \lambda^* N/2\gamma^\top(L^{-1})^\top L\gamma
$$

If we choose L such that $(L^{-1})^\top$ is the Cholesky decomposition of V_t, (i.e. $(L^{-1})^\top L^{-1} = V_t$)then $||L^{-1}\gamma||_2^2 = \gamma^\top V_t\gamma = ||X_t\gamma||_2^2$ and thus

$$
\begin{aligned}
f(\theta;\lambda^*) &= -l_t(\beta;X_t,Y_t) - l_e(\beta+\gamma;X_e,Y_e) + \lambda^* N/2||X_t\gamma||_2^2 \\
&= g(\beta,\gamma;\lambda^* N/2).
\end{aligned} \tag{C.6}
$$

Because of this relationship with penalized regression and DEGLR we can use the *glmnet* Package to fit DEGLR for gaussian, binomial, poisson, multinomial, cox, and multi-response Gaussian.

C.5 Penalization obtained through cross validation

We assess if the selected λ with the cross validation approach from Section 4.2.4 attains the optimum value in the real data analysis. In Figure C.1 we show that the cross validation method does not obtains the optimal value in the testing data. However, it does attains a value of AUC that is close to the maximum.

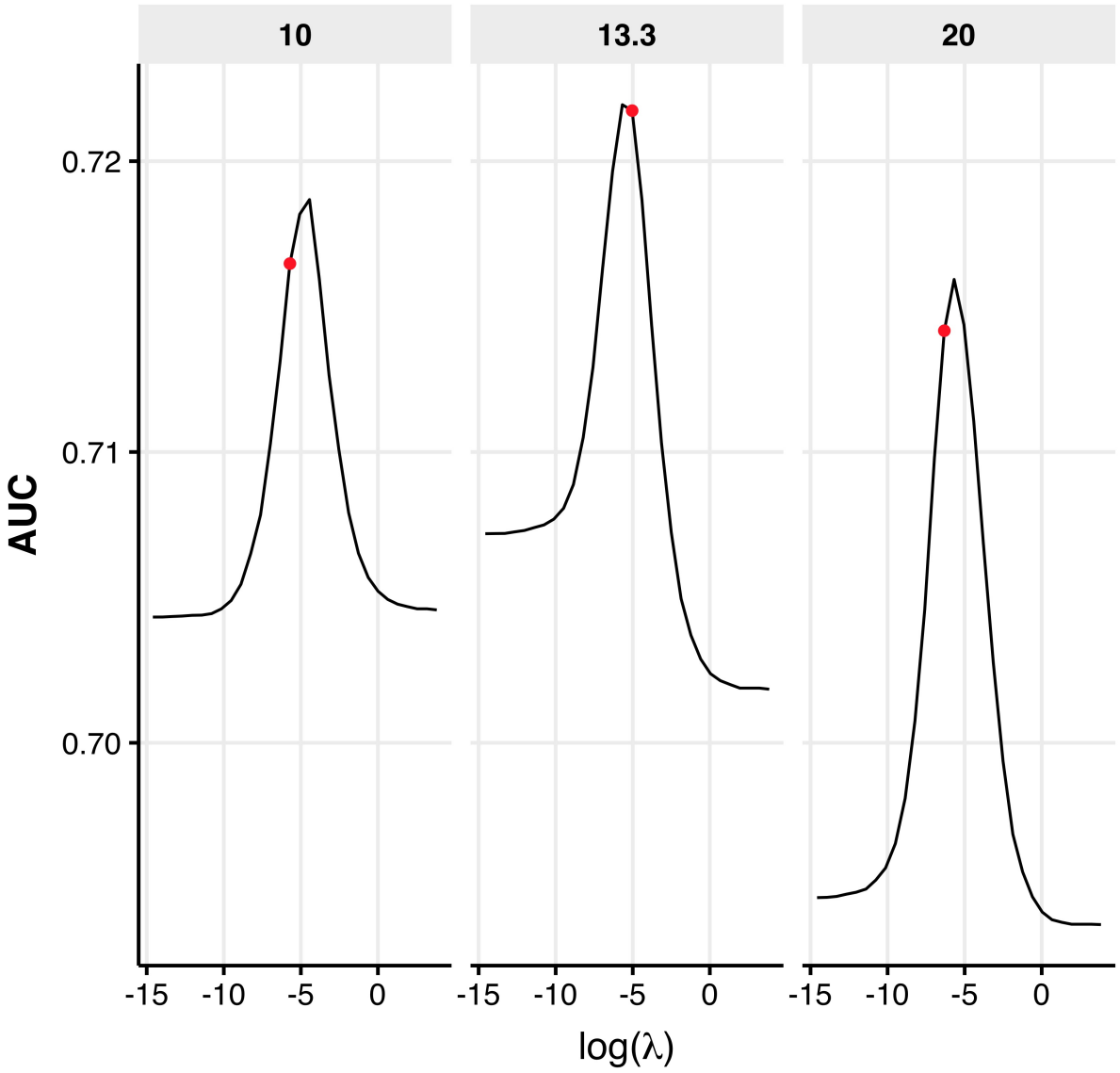

Figure C.1: In the x axis is the log value of the penalty factor λ. With a red dot we indicate the value of λ that the cross validation from section 4.2.4 selected as optimal using the training data of HRS. In the y-axis we show the AUC in the test data of HRS for different values of λ. The panels of the plot are three different sample size ratios between the GfG sample size and the training sample size of HRS.

BIBLIOGRAPHY

[1] The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019. PMID: 31412182.

[2] M. Aas, I. Melle, F. Bettella, S. Djurovic, S. Le Hellard, T. Bjella, P. A. Ringen, T. V. Lagerberg, O. B. Smeland, I. Agartz, O. A. Andreassen, and M. Tesli. Psychotic patients who used cannabis frequently before illness onset have higher genetic predisposition to schizophrenia than those who did not. *Psychological Medicine*, 48:43–49, 1 2018.

[3] Nathan S. Abell, Marianne K. DeGorter, Michael J. Gloudemans, Emily Greenwald, Kevin S. Smith, Zihuai He, and Stephen B. Montgomery. Multiple causal variants underlie genetic associations in humans. *Science*, 375:1247–1254, 3 2022.

[4] Adam Auton, Gonçalo R Abecasis, David M Altshuler, Richard M Durbin, Gonçalo R Abecasis, David R Bentley, Aravinda Chakravarti, Andrew G Clark, Peter Donnelly, Evan E Eichler, Paul Flicek, Stacey B Gabriel, Richard A Gibbs, Eric D Green, Matthew E Hurles, Bartha M Knoppers, Jan O Korbel, Eric S Lander, Charles Lee, Hans Lehrach, Elaine R Mardis, Gabor T Marth, Gil A McVean, Deborah A Nickerson, Jeanette P Schmidt, Stephen T Sherry, Jun Wang, Richard K Wilson, Richard A Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G Reid, Yiming Zhu, Jun Wang, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S Lander, David M Altshuler, Stacey B Gabriel, Namrata Gupta, Neda Gharani, Lorraine H Toji, Norman P Gerry, Alissa M Resch, Paul Flicek, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach, Ralf Sudbrak, Marcus W Albrecht, Vyacheslav S Amstislavskiy, Tatiana A Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R Mardis, Richard K Wilson, Lucinda Fulton, Robert Fulton, Stephen T Sherry, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A McVean,

Richard M Durbin, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P Schmidt, Christopher J Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Adam Auton, Christopher L Campbell, Yu Kong, Anthony Marcketta, Richard A Gibbs, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Donna Muzny, Aniko Sabo, Zhuoyi Huang, Jun Wang, Lachlan J M Coin, Lin Fang, Xiaosen Guo, Xin Jin, Guoqing Li, Qibin Li, Yingrui Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T Marth, Erik P Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N Ward, Jiantao Wu, Mengyao Zhang, Mark J Daly, Mark A DePristo, Robert E Handsaker, David M Altshuler, Eric Banks, Gaurav Bhatia, Guillermo del Angel, Stacey B Gabriel, Giulio Genovese, Namrata Gupta, Heng Li, Seva Kashin, Eric S Lander, Steven A McCarroll, James C Nemesh, Ryan E Poplin, Seungtai C Yoon, Jayon Lihm, Vladimir Makarov, Andrew G Clark, Srikanth Gottipati, Alon Keinan, Juan L Rodriguez-Flores, Jan O Korbel, Tobias Rausch, Markus H Fritz, Adrian M Stütz, Paul Flicek, Kathryn Beal, Laura Clarke, Avik Datta, Javier Herrero, William M McLaren, Graham R S Ritchie, Richard E Smith, Daniel Zerbino, Xiangqun Zheng-Bradley, Pardis C Sabeti, Ilya Shlyakhter, Stephen F Schaffner, Joseph Vitti, David N Cooper, Edward V Ball, Peter D Stenson, David R Bentley, Bret Barnes, Markus Bauer, R Keira Cheetham, Anthony Cox, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E Kenny, Mark A Batzer, Miriam K Konkel, Jerilyn A Walker, Daniel G MacArthur, Monkol Lek, Ralf Sudbrak, Vyacheslav S Amstislavskiy, Ralf Herwig, Elaine R Mardis, Li Ding, Daniel C Koboldt, David Larson, Kai Ye, Simon Gravel, The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of M I T, Harvard, Coriell Institute for Medical Research, European Bioinformatics Institute European Molecular Biology Laboratory, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, U S National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, and N I H National Eye Institute. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

[5] Amy R Bentley, Shawneequa Callier, and Charles N Rotimi. Diversity and inclusion in genomic research: why the uneven progress? *Journal of Community Genetics*, 8:255–266, 2017.

[6] Bárbara D Bitarello and Iain Mathieson. Polygenic scores for height in admixed populations. *G3 Genes—Genomes—Genetics*, 10:4027–4036, 11 2020.

[7] Philip S Boonstra and Ryan P Barbaro. Incorporating historical models with adaptive Bayesian updates. *Biostatistics*, 21(2):e47–e64, 2018.

[8] Philip S. Boonstra, Bhramar Mukherjee, and Jeremy M. G. Taylor. Bayesian shrinkage

methods for partially observed data with many predictors. *The Annals of Applied Statistics*, 7(4):2272–2292, 2013.

[9] Brielin C. Brown, Chun Jimmie Ye, Alkes L. Price, and Noah Zaitlen. Transethnic genetic-correlation estimates from summary statistics. *The American Journal of Human Genetics*, 99(1):76–88, 2016.

[10] Cerezo M Harris LW Hayhurst J Malangone C McMahon A Morales J Mountjoy E Sollis E Suveges D Vrousgou O Whetzel PL Amode R Guillen JA Riat HS Trevanion SJ Hall P Junkins H Flicek P Burdett T Hindorff LA Cunningham F Buniello A, MacArthur JAL and Parkinson H. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47 (Database issue): D1005-D1012, 2019.

[11] Christopher S Carlson, Tara C Matise, Kari E North, Christopher A Haiman, Megan D Fesinmeyer, Steven Buyske, Fredrick R Schumacher, Ulrike Peters, Nora Franceschini, Marylyn D Ritchie, David J Duggan, Kylee L Spencer, Logan Dumitrescu, Charles B Eaton, Fridtjof Thomas, Alicia Young, Cara Carty, Gerardo Heiss, Loic Le Marchand, Dana C Crawford, Lucia A Hindorff, Charles L Kooperberg, and for the PAGE Consortium. Generalization and dilution of association results from european gwas in populations of non-european ancestry: The page study. *PLOS Biology*, 11:e1001661–, 9 2013.

[12] Taylor B Cavazos and John S Witte. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Human Genetics and Genomics Advances*, 2:100017, 2021.

[13] A Cecile, J W Janssens, and Michael J Joyner. Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: Is more, better? *Clinical Chemistry*, 65:609–611, 5 2019.

[14] Nilanjan Chatterjee, Yi-Hau Chen, Paige Maas, and Raymond J Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016.

[15] Aiyou Chen, Art B. Owen, and Minghui Shi. Data enriched linear regression. *Electronic Journal of Statistics*, 9, 1 2015.

[16] Wenting Cheng, Jeremy MG Taylor, Pantel S Vokonas, Sung Kyun Park, and Bhramar Mukherjee. Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statistics in medicine*, 37(9):1515–1530, 2018.

[17] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15:2759–2772, 2020.

[18] Marc A Coram, Huaying Fang, Sophie I Candille, Themistocles L Assimes, and Hua Tang. Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *The American Journal of Human Genetics*, 101:218–226, 2017.

[19] David Curtis. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatric Genetics*, 28, 2018.

[20] Ozan Dikilitas, Daniel J. Schaid, Matthew L. Kosel, Robert J. Carroll, Christopher G. Chute, Joshua C. Denny, Alex Fedotov, QiPing Feng, Hakon Hakonarson, Gail P. Jarvik, Ming Ta Michael Lee, Jennifer A. Pacheco, Robb Rowley, Patrick M. Sleiman, C. Michael Stein, Amy C. Sturm, Wei-Qi Wei, Georgia L. Wiesner, Marc S. Williams, Yanfei Zhang, Teri A. Manolio, and Iftikhar J. Kullo. Predictive utility of polygenic risk scores for coronary heart disease in three major racial and ethnic groups. *The American Journal of Human Genetics*, 106:707–716, 5 2020.

[21] Yaqi Duan and Kaizheng Wang. Adaptive and robust multi-task learning, 2022.

[22] L Duncan, H Shen, B Gelaye, J Meijsen, K Ressler, M Feldman, R Peterson, and B Domingue. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, 10:3328, 2019.

[23] Emma Fabian. Little progress in the fight against health research inequality: Action needed. https://www.demographicscience.ox.ac.uk/post/little-progress-in-the-fight-against-health-research-inequality-action- 2021.

[24] Akl C. Fahed, Krishna G. Aragam, George Hindy, Yii-Der Ida Chen, Kumardeep Chaudhary, Amanda Dobbyn, Harlan M. Krumholz, Wayne H.H. Sheu, Stephen S. Rich, Jerome I. Rotter, Rajiv Chowdhury, Judy Cho, Ron Do, Patrick T. Ellinor, Sekar Kathiresan, and Amit V. Khera. Transethnic transferability of a genome-wide polygenic score for coronary artery disease. *Circulation: Genomic and Precision Medicine*, 12 2020.

[25] Akl C. Fahed, Krishna G. Aragam, George Hindy, Yii-Der Ida Chen, Kumardeep Chaudhary, Amanda Dobbyn, Harlan M. Krumholz, Wayne H.H. Sheu, Stephen S. Rich, Jerome I. Rotter, Rajiv Chowdhury, Judy Cho, Ron Do, Patrick T. Ellinor, Sekar Kathiresan, and Amit V. Khera. Transethnic transferability of a genome-wide polygenic score for coronary artery disease. *Circulation: Genomic and Precision Medicine*, 14, 2 2021.

[26] Keolu Fox. The illusion of inclusion — the "all of us" research program and indigenous peoples' dna. *New England Journal of Medicine*, 383(5):411–413, 2020. PMID: 32726527.

[27] Kevin J. Galinsky, Yakir A. Reshef, Hilary K. Finucane, Po-Ru Loh, Noah Zaitlen, Nick J. Patterson, Brielin C. Brown, and Alkes L. Price. Estimating cross-population genetic correlations of causal effect sizes. *Genetic Epidemiology*, 43:180–188, 3 2019.

[28] Nanibaa'A Garrison, Māui Hudson, Leah L Ballantyne, Ibrahim Garba, Andrew Martinez, Maile Taualii, Laura Arbour, Nadine R Caron, and Stephanie Carroll Rainie. Genomic research through an indigenous lens: understanding the expectations. *Annual Review of Genomics and Human Genetics*, 20:495–517, 2019.

[29] Kelsey E Grinde, Qibin Qi, Timothy A Thornton, Simin Liu, Aladdin H Shadyab, Kei Hang K Chan, Alexander P Reiner, and Tamar Sofer. Generalizing polygenic risk

scores from europeans to hispanics/latinos. *Genetic Epidemiology*, 43:50–62, 2 2019. https://doi.org/10.1002/gepi.22166.

[30] Samuel M. Gross and Robert Tibshirani. Data shared lasso: A novel tool to discover uplift. *Computational Statistics Data Analysis*, 101:226–235, 9 2016.

[31] Deepti Gurdasani, Tommy Carstensen, Fasil Tekola-Ayele, Luca Pagani, Ioanna Tachmazidou, Konstantinos Hatzikotoulas, Savita Karthikeyan, Louise Iles, Martin O. Pollard, Ananyo Choudhury, Graham R. S. Ritchie, Yali Xue, Jennifer Asimit, Rebecca N. Nsubuga, Elizabeth H. Young, Cristina Pomilla, Katja Kivinen, Kirk Rockett, Anatoli Kamali, Ayo P. Doumatey, Gershim Asiki, Janet Seeley, Fatoumatta Sisay-Joof, Muminatou Jallow, Stephen Tollman, Ephrem Mekonnen, Rosemary Ekong, Tamiru Oljira, Neil Bradman, Kalifa Bojang, Michele Ramsay, Adebowale Adeyemo, Endashaw Bekele, Ayesha Motala, Shane A. Norris, Fraser Pirie, Pontiano Kaleebu, Dominic Kwiatkowski, Chris Tyler-Smith, Charles Rotimi, Eleftheria Zeggini, and Manjinder S. Sandhu. The african genome variation project shapes medical genetics in africa. *Nature*, 517:327–332, 1 2015.

[32] Kangcheng Hou, Yi Ding, Ziqi Xu, Yue Wu, Arjun Bhattacharya, Rachel Mester, Gillian Belbin, David Conti, Burcu F Darst, Myriam Fornage, Chris Gignoux, Xiuqing Guo, Christopher Haiman, Eimear Kenny, Michelle Kim, Charles Kooperberg, Leslie Lange, Ani Manichaikul, Kari E North, Natalie Nudelman, Ulrike Peters, Laura J Rasmussen-Torvik, Stephen S Rich, Jerome I Rotter, Heather E Wheeler, Ying Zhou, Sriram Sankararaman, and Bogdan Pasaniuc. Causal effects on complex traits are similar across segments of different continental ancestries within admixed individuals. *medRxiv*, page 2022.08.16.22278868, 1 2022.

[33] Joseph G Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749, 2015.

[34] Douville Nicholas J, Surakka Ida, Leis Aleda, Douville Christopher B, Hornsby Whitney E, Brummett Chad M, Kheterpal Sachin, Willer Cristen J, Engoren Milo, and Mathis Michael R. Use of a polygenic risk score improves prediction of myocardial injury after noncardiac surgery. *Circulation: Genomic and Precision Medicine*, 13:e002817, 8 2020. doi: 10.1161/CIRCGEN.119.002817.

[35] Ying Ji, Jirong Long, Sun-Seog Kweon, Daehee Kang, Michiaki Kubo, Boyoung Park, Xiao-Ou Shu, Wei Zheng, Ran Tao, and Bingshan Li. Incorporating european gwas findings improve polygenic risk prediction accuracy of breast cancer among east asians. *Genetic Epidemiology*, n/a, 3 2021. https://doi.org/10.1002/gepi.22382.

[36] Wetterstrand KA. Dna sequencing costs: Data from the nhgri genome sequencing program (gsp). www.genome.gov/sequencingcostsdata, 2022.

[37] Eimear E Kenny, Nicholas J Timpson, Martin Sikora, Muh-Ching Yee, Andrés Moreno-Estrada, Celeste Eng, Scott Huntsman, Esteban González Burchard, Mark Stoneking, Carlos D Bustamante, and Sean Myles. Melanesian blond hair is caused by an amino acid change in tyrp1. *Science*, 336:554, 5 2012.

[38] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50:1219–1224, 2018.

[39] Amit V. Khera, Mark Chaffin, Kaitlin H. Wade, Sohail Zahid, Joseph Brancale, Rui Xia, Marina Distefano, Ozlem Senol-Cosar, Mary E. Haas, Alexander Bick, Krishna G. Aragam, Eric S. Lander, George Davey Smith, Heather Mason-Suares, Myriam Fornage, Matthew Lebo, Nicholas J. Timpson, Lee M. Kaplan, and Sekar Kathiresan. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*, 177, 4 2019.

[40] Joshua W Knowles and Euan A Ashley. Cardiovascular disease: The rise of the genetic risk score. *PLOS Medicine*, 15:e1002546–, 3 2018.

[41] Scott Kulm, Jason Mezey, and Olivier Elemento. Benchmarking the accuracy of polygenic risk scores and their generative methods. *medRxiv*, page 2020.04.06.20055574, 1 2020.

[42] Samuel A Lambert, Gad Abraham, and Michael Inouye. Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*, 28:R133–R142, 11 2019.

[43] Anna C F Lewis and Robert C Green. Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Medicine*, 13:14, 2021.

[44] Anubha Mahajan, Cassandra N. Spracklen, Weihua Zhang, Maggie C. Y. Ng, Lauren E. Petty, Hidetoshi Kitajima, Grace Z. Yu, Sina Rüeger, Leo Speidel, Young Jin Kim, Momoko Horikoshi, Josep M. Mercader, Daniel Taliun, Sanghoon Moon, Soo-Heon Kwak, Neil R. Robertson, Nigel W. Rayner, Marie Loh, Bong-Jo Kim, Joshua Chiou, Irene Miguel-Escalada, Pietro della Briotta Parolo, Kuang Lin, Fiona Bragg, Michael H. Preuss, Fumihiko Takeuchi, Jana Nano, Xiuqing Guo, Amel Lamri, Masahiro Nakatochi, Robert A. Scott, Jung-Jin Lee, Alicia Huerta-Chagoya, Mariaelisa Graff, Jin-Fang Chai, Esteban J. Parra, Jie Yao, Lawrence F. Bielak, Yasuharu Tabara, Yang Hai, Valgerdur Steinthorsdottir, James P. Cook, Mart Kals, Niels Grarup, Ellen M. Schmidt, Ian Pan, Tamar Sofer, Matthias Wuttke, Chloe Sarnowski, Christian Gieger, Darryl Nousome, Stella Trompet, Jirong Long, Meng Sun, Lin Tong, Wei-Min Chen, Meraj Ahmad, Raymond Noordam, Victor J. Y. Lim, Claudia H. T. Tam, Yoonjung Yoonie Joo, Chien-Hsiun Chen, Laura M. Raffield, Cécile Lecoeur, Bram Peter Prins, Aude Nicolas, Lisa R. Yanek, Guanjie Chen, Richard A. Jensen, Salman Tajuddin, Edmond K. Kabagambe, Ping An, Anny H. Xiang, Hyeok Sun Choi, Brian E. Cade, Jingyi Tan, Jack Flanagan, Fernando Abaitua, Linda S. Adair, Adebowale Adeyemo, Carlos A. Aguilar-Salinas, Masato Akiyama, Sonia S. Anand, Alain Bertoni, Zheng Bian, Jette Bork-Jensen, Ivan Brandslund, Jennifer A. Brody, Chad M. Brummett, Thomas A. Buchanan, Mickaël Canouil, Juliana C. N. Chan, Li-Ching Chang, Miao-Li Chee, Ji Chen, Shyh-Huei Chen, Yuan-Tsong Chen, Zhengming Chen, Lee-Ming Chuang, Mary Cushman, Swapan K. Das, H. Janaka de Silva, George Dedoussis, Latchezar Dimitrov, Ayo P. Doumatey, Shufa Du, Qing Duan, Kai-Uwe Eckardt, Leslie S. Emery, Daniel S. Evans, Michele K. Evans, Krista Fischer, James S. Floyd, Ian Ford, Myriam Fornage, Oscar H. Franco, Timothy M. Frayling, Barry I. Freedman, Christian Fuchsberger, Pauline Genter, Hertzel C. Gerstein, Vilmantas Giedraitis, Clicerio González-Villalpando, Maria Elena González-Villalpando,

Mark O. Goodarzi, Penny Gordon-Larsen, David Gorkin, Myron Gross, Yu Guo, Sophie Hackinger, Sohee Han, Andrew T. Hattersley, Christian Herder, Annie-Green Howard, Willa Hsueh, Mengna Huang, Wei Huang, Yi-Jen Hung, Mi Yeong Hwang, Chii-Min Hwu, Sahoko Ichihara, Mohammad Arfan Ikram, Martin Ingelsson, Md Tariqul Islam, Masato Isono, Hye-Mi Jang, Farzana Jasmine, Guozhi Jiang, Jost B. Jonas, Marit E. Jørgensen, Torben Jørgensen, Yoichiro Kamatani, Fouad R. Kandeel, Anuradhani Kasturiratne, Tomohiro Katsuya, Varinderpal Kaur, Takahisa Kawaguchi, Jacob M. Keaton, Abel N. Kho, Chiea-Chuen Khor, Muhammad G. Kibriya, Duk-Hwan Kim, Katsuhiko Kohara, Jennifer Kriebel, Florian Kronenberg, Johanna Kuusisto, Kristi Läll, Leslie A. Lange, Myung-Shik Lee, Nanette R. Lee, Aaron Leong, Liming Li, Yun Li, Ruifang Li-Gao, Symen Ligthart, Cecilia M. Lindgren, Allan Linneberg, Ching-Ti Liu, Jianjun Liu, Adam E. Locke, Tin Louie, Jian'an Luan, Andrea O. Luk, Xi Luo, Jun Lv, Valeriya Lyssenko, Vasiliki Mamakou, K. Radha Mani, Thomas Meitinger, Andres Metspalu, Andrew D. Morris, Girish N. Nadkarni, Jerry L. Nadler, Michael A. Nalls, Uma Nayak, Suraj S. Nongmaithem, Ioanna Ntalla, Yukinori Okada, Lorena Orozco, Sanjay R. Patel, Mark A. Pereira, Annette Peters, Fraser J. Pirie, Bianca Porneala, Gauri Prasad, Sebastian Preissl, Laura J. Rasmussen-Torvik, Alexander P. Reiner, Michael Roden, Rebecca Rohde, Kathryn Roll, Charumathi Sabanayagam, Maike Sander, Kevin Sandow, Naveed Sattar, Sebastian Schönherr, Claudia Schurmann, Mohammad Shahriar, Jinxiu Shi, Dong Mun Shin, Daniel Shriner, Jennifer A. Smith, Wing Yee So, Alena Stančáková, Adrienne M. Stilp, Konstantin Strauch, Ken Suzuki, Atsushi Takahashi, Kent D. Taylor, Barbara Thorand, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Brian Tomlinson, Jason M. Torres, Fuu-Jen Tsai, Jaakko Tuomilehto, Teresa Tusie-Luna, Miriam S. Udler, Adan Valladares-Salgado, Rob M. van Dam, Jan B. van Klinken, Rohit Varma, Marijana Vujkovic, Niels Wacher-Rodarte, Eleanor Wheeler, Eric A. Whitsel, Ananda R. Wickremasinghe, Ko Willems van Dijk, Daniel R. Witte, Chittaranjan S. Yajnik, Ken Yamamoto, Toshimasa Yamauchi, Loïc Yengo, Kyungheon Yoon, Canqing Yu, Jian-Min Yuan, Salim Yusuf, Liang Zhang, Wei Zheng, Sina Rüeger, Pietro della Briotta Parolo, Yoonjung Yoonie Joo, M. Geoffrey Hayes, Leslie J. Raffel, Michiya Igase, Eli Ipp, Susan Redline, Yoon Shin Cho, Lars Lind, Michael A. Province, Craig L. Hanis, Patricia A. Peyser, Erik Ingelsson, Alan B. Zonderman, Bruce M. Psaty, Ya-Xing Wang, Charles N. Rotimi, Diane M. Becker, Fumihiko Matsuda, Yongmei Liu, Eleftheria Zeggini, Mitsuhiro Yokota, Stephen S. Rich, Charles Kooperberg, James S. Pankow, James C. Engert, Yii-Der Ida Chen, Philippe Froguel, James G. Wilson, Wayne H. H. Sheu, Sharon L. R. Kardia, Jer-Yuarn Wu, M. Geoffrey Hayes, Ronald C. W. Ma, Tien-Yin Wong, Leif Groop, Dennis O. Mook-Kanamori, Giriraj R. Chandak, Francis S. Collins, Dwaipayan Bharadwaj, Guillaume Paré, Michèle M. Sale, Habibul Ahsan, Ayesha A. Motala, Xiao-Ou Shu, Kyong-Soo Park, J. Wouter Jukema, Miguel Cruz, Roberta McKean-Cowdin, Harald Grallert, Ching-Yu Cheng, Erwin P. Bottinger, Abbas Dehghan, E-Shyong Tai, Josée Dupuis, Norihiro Kato, Markku Laakso, Anna Köttgen, Woon-Puay Koh, Colin N. A. Palmer, Simin Liu, Goncalo Abecasis, Jaspal S. Kooner, Ruth J. F. Loos, Kari E. North, Christopher A. Haiman, Jose C. Florez, Danish Saleheen, Torben Hansen, Oluf Pedersen, Reedik Mägi, Claudia Langenberg, Nicholas J. Wareham, Shiro Maeda, Takashi Kadowaki, Juyoung Lee, Iona Y. Millwood, Robin G. Walters, Kari Stefansson, Simon R. Myers, Jorge Ferrer, Kyle J. Gaulton, James B. Meigs, Karen L. Mohlke, Anna L. Gloyn, Donald W. Bowden, Jennifer E. Below, John C. Chambers, Xueling Sim, Michael Boehnke, Jerome I. Rotter, Mark I. McCarthy, and Andrew P. Morris. Multi-

ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nature Genetics*, 54:560–572, 5 2022.

[45] Rainer Malik, Ganesh Chauhan, Matthew Traylor, Muralidharan Sargurupremraj, Yukinori Okada, Aniket Mishra, Loes Rutten-Jacobs, Anne-Katrin Giese, Sander W. van der Laan, Solveig Gretarsdottir, Christopher D. Anderson, Michael Chong, Hieab H. H. Adams, Tetsuro Ago, Peter Almgren, Philippe Amouyel, Hakan Ay, Traci M. Bartz, Oscar R. Benavente, Steve Bevan, Giorgio B. Boncoraglio, Robert D. Brown, Adam S. Butterworth, Caty Carrera, Cara L. Carty, Daniel I. Chasman, Wei-Min Chen, John W. Cole, Adolfo Correa, Ioana Cotlarciuc, Carlos Cruchaga, John Danesh, Paul I. W. de Bakker, Anita L. DeStefano, Marcel den Hoed, Qing Duan, Stefan T. Engelter, Guido J. Falcone, Rebecca F. Gottesman, Raji P. Grewal, Vilmundur Gudnason, Stefan Gustafsson, Jeffrey Haessler, Tamara B. Harris, Ahamad Hassan, Aki S. Havulinna, Susan R. Heckbert, Elizabeth G. Holliday, George Howard, Fang-Chi Hsu, Hyacinth I. Hyacinth, M. Arfan Ikram, Erik Ingelsson, Marguerite R. Irvin, Xueqiu Jian, Jordi Jiménez-Conde, Julie A. Johnson, J. Wouter Jukema, Masahiro Kanai, Keith L. Keene, Brett M. Kissela, Dawn O. Kleindorfer, Charles Kooperberg, Michiaki Kubo, Leslie A. Lange, Carl D. Langefeld, Claudia Langenberg, Lenore J. Launer, Jin-Moo Lee, Robin Lemmens, Didier Leys, Cathryn M. Lewis, Wei-Yu Lin, Arne G. Lindgren, Erik Lorentzen, Patrik K. Magnusson, Jane Maguire, Ani Manichaikul, Patrick F. McArdle, James F. Meschia, Braxton D. Mitchell, Thomas H. Mosley, Michael A. Nalls, Toshiharu Ninomiya, Martin J. O'Donnell, Bruce M. Psaty, Sara L. Pulit, Kristiina Rannikmäe, Alexander P. Reiner, Kathryn M. Rexrode, Kenneth Rice, Stephen S. Rich, Paul M. Ridker, Natalia S. Rost, Peter M. Rothwell, Jerome I. Rotter, Tatjana Rundek, Ralph L. Sacco, Saori Sakaue, Michele M. Sale, Veikko Salomaa, Bishwa R. Sapkota, Reinhold Schmidt, Carsten O. Schmidt, Ulf Schminke, Pankaj Sharma, Agnieszka Slowik, Cathie L. M. Sudlow, Christian Tanislav, Turgut Tatlisumak, Kent D. Taylor, Vincent N. S. Thijs, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Steffen Tiedt, Stella Trompet, Christophe Tzourio, Cornelia M. van Duijn, Matthew Walters, Nicholas J. Wareham, Sylvia Wassertheil-Smoller, James G. Wilson, Kerri L. Wiggins, Qiong Yang, Salim Yusuf, Joshua C. Bis, Tomi Pastinen, Arno Ruusalepp, Eric E. Schadt, Simon Koplev, Johan L. M. Björkegren, Veronica Codoni, Mete Civelek, Nicholas L. Smith, David A. Trégouët, Ingrid E. Christophersen, Carolina Roselli, Steven A. Lubitz, Patrick T. Ellinor, E. Shyong Tai, Jaspal S. Kooner, Norihiro Kato, Jiang He, Pim van der Harst, Paul Elliott, John C. Chambers, Fumihiko Takeuchi, Andrew D. Johnson, Rainer Malik, Ganesh Chauhan, Matthew Traylor, Muralidharan Sargurupremraj, Yukinori Okada, Aniket Mishra, Loes Rutten-Jacobs, Anne-Katrin Giese, Sander W. van der Laan, Solveig Gretarsdottir, Christopher D. Anderson, Michael Chong, Hieab H. H. Adams, Tetsuro Ago, Peter Almgren, Philippe Amouyel, Hakan Ay, Traci M. Bartz, Oscar R. Benavente, Steve Bevan, Giorgio B. Boncoraglio, Robert D. Brown, Adam S. Butterworth, Caty Carrera, Cara L. Carty, Daniel I. Chasman, Wei-Min Chen, John W. Cole, Adolfo Correa, Ioana Cotlarciuc, Carlos Cruchaga, John Danesh, Paul I. W. de Bakker, Anita L. DeStefano, Marcel den Hoed, Qing Duan, Stefan T. Engelter, Guido J. Falcone, Rebecca F. Gottesman, Raji P. Grewal, Vilmundur Gudnason, Stefan Gustafsson, Jeffrey Haessler, Tamara B. Harris, Ahamad Hassan, Aki S. Havulinna, Susan R. Heckbert, Elizabeth G. Holliday, George Howard, Fang-Chi Hsu, Hyacinth I. Hyacinth, M. Arfan Ikram, Erik Ingelsson, Marguerite R. Irvin, Xueqiu Jian, Jordi Jiménez-Conde, Julie A. Johnson, J. Wouter Jukema, Masahiro

Kanai, Keith L. Keene, Brett M. Kissela, Dawn O. Kleindorfer, Charles Kooperberg, Michiaki Kubo, Leslie A. Lange, Carl D. Langefeld, Claudia Langenberg, Lenore J. Launer, Jin-Moo Lee, Robin Lemmens, Didier Leys, Cathryn M. Lewis, Wei-Yu Lin, Arne G. Lindgren, Erik Lorentzen, Patrik K. Magnusson, Jane Maguire, Ani Manichaikul, Patrick F. McArdle, James F. Meschia, Braxton D. Mitchell, Thomas H. Mosley, Michael A. Nalls, Toshiharu Ninomiya, Martin J. O'Donnell, Bruce M. Psaty, Sara L. Pulit, Kristiina Rannikmäe, Alexander P. Reiner, Kathryn M. Rexrode, Kenneth Rice, Stephen S. Rich, Paul M. Ridker, Natalia S. Rost, Peter M. Rothwell, Jerome I. Rotter, Tatjana Rundek, Ralph L. Sacco, Saori Sakaue, Michele M. Sale, Veikko Salomaa, Bishwa R. Sapkota, Reinhold Schmidt, Carsten O. Schmidt, Ulf Schminke, Pankaj Sharma, Agnieszka Slowik, Cathie L. M. Sudlow, Christian Tanislav, Turgut Tatlisumak, Kent D. Taylor, Vincent N. S. Thijs, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Steffen Tiedt, Stella Trompet, Christophe Tzourio, Cornelia M. van Duijn, Matthew Walters, Nicholas J. Wareham, Sylvia Wassertheil-Smoller, James G. Wilson, Kerri L. Wiggins, Qiong Yang, Salim Yusuf, Najaf Amin, Hugo S. Aparicio, Donna K. Arnett, John Attia, Alexa S. Beiser, Claudine Berr, Julie E. Buring, Mariana Bustamante, Valeria Caso, Yu-Ching Cheng, Seung Hoan Choi, Ayesha Chowhan, Natalia Cullell, Jean-François Dartigues, Hossein Delavaran, Pilar Delgado, Marcus Dörr, Gunnar Engström, Ian Ford, Wander S. Gurpreet, Anders Hamsten, Laura Heitsch, Atsushi Hozawa, Laura Ibanez, Andreea Ilinca, Martin Ingelsson, Motoki Iwasaki, Rebecca D. Jackson, Katarina Jood, Pekka Jousilahti, Sara Kaffashian, Lalit Kalra, Masahiro Kamouchi, Takanari Kitazono, Olafur Kjartansson, Manja Kloss, Peter J. Koudstaal, Jerzy Krupinski, Daniel L. Labovitz, Cathy C. Laurie, Christopher R. Levi, Linxin Li, Lars Lind, Cecilia M. Lindgren, Vasileios Lioutas, Yong Mei Liu, Oscar L. Lopez, Hirata Makoto, Nicolas Martinez-Majander, Koichi Matsuda, Naoko Minegishi, Joan Montaner, Andrew P. Morris, Elena Muiño, Martina Müller-Nurasyid, Bo Norrving, Soichi Ogishima, Eugenio A. Parati, Leema Reddy Peddareddygari, Nancy L. Pedersen, Joanna Pera, Markus Perola, Alessandro Pezzini, Silvana Pileggi, Raquel Rabionet, Iolanda Riba-Llena, Marta Ribasés, Jose R. Romero, Jaume Roquer, Anthony G. Rudd, Antti-Pekka Sarin, Ralhan Sarju, Chloe Sarnowski, Makoto Sasaki, Claudia L. Satizabal, Mamoru Satoh, Naveed Sattar, Norie Sawada, Gerli Sibolt, Ásgeir Sigurdsson, Albert Smith, Kenji Sobue, Carolina Soriano-Tárraga, Tara Stanne, O. Colin Stine, David J. Stott, Konstantin Strauch, Takako Takai, Hideo Tanaka, Kozo Tanno, Alexander Teumer, Liisa Tomppo, Nuria P. Torres-Aguila, Emmanuel Touze, Shoichiro Tsugane, Andre G. Uitterlinden, Einar M. Valdimarsson, Sven J. van der Lee, Henry Völzke, Kenji Wakai, David Weir, Stephen R. Williams, Charles D. A. Wolfe, Quenna Wong, Huichun Xu, Taiki Yamaji, Dharambir K. Sanghera, Olle Melander, Christina Jern, Daniel Strbian, Israel Fernandez-Cadenas, W. T. Longstreth, Arndt Rolfs, Jun Hata, Daniel Woo, Jonathan Rosand, Guillaume Pare, Jemma C. Hopewell, Danish Saleheen, Kari Stefansson, Bradford B. Worrall, Steven J. Kittner, Sudha Seshadri, Myriam Fornage, Hugh S. Markus, Joanna M. M. Howson, Yoichiro Kamatani, Stephanie Debette, Martin Dichgans, Dharambir K. Sanghera, Olle Melander, Christina Jern, Daniel Strbian, Israel Fernandez-Cadenas, W. T. Longstreth, Arndt Rolfs, Jun Hata, Daniel Woo, Jonathan Rosand, Guillaume Pare, Jemma C. Hopewell, Danish Saleheen, Kari Stefansson, Bradford B. Worrall, Steven J. Kittner, Sudha Seshadri, Myriam Fornage, Hugh S. Markus, Joanna M. M. Howson, Yoichiro Kamatani, Stephanie Debette, and Martin Dichgans. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature Genetics*, 50:524–537,

4 2018.

[46] Arjun K Manrai, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375:655–665, 8 2016. doi: 10.1056/NEJMsa1507092.

[47] Davide Marnetto, Katri Pärna, Kristi Läll, Ludovica Molinaro, Francesco Montinaro, Toomas Haller, Mait Metspalu, Reedik Mägi, Krista Fischer, and Luca Pagani. Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nature Communications*, 11:1628, 2020.

[48] Nina Mars, Sini Kerminen, Yen-Chen A. Feng, Masahiro Kanai, Kristi Läll, Laurent F. Thomas, Anne Heidi Skogholt, Pietro della Briotta Parolo, Benjamin M. Neale, Jordan W. Smoller, Maiken E. Gabrielsen, Kristian Hveem, Reedik Mägi, Koichi Matsuda, Yukinori Okada, Matti Pirinen, Aarno Palotie, Andrea Ganna, Alicia R. Martin, and Samuli Ripatti. Genome-wide risk prediction of common diseases across ancestries in one million people. *Cell Genomics*, 2:100118, 4 2022.

[49] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100:635–649, 2017.

[50] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51:584–591, 2019.

[51] Enock Matovu, Bruno Bucheton, John Chisi, John Enyaru, Christiane Hertz-Fowler, Mathurin Koffi, Annette Macleod, Dieuodonne Mumba, Issa Sidibe, Gustave Simo, Martin Simuunza, Bongani Mayosi, Rajkumar Ramesar, Nicola Mulder, Stephen Ogendo, Ana Olga Mocumbi, Christopher Hugo-Hamman, Okechukwu Ogah, Ahmed El Sayed, Charles Mondo, John Musuku, Mark Engel, Jantina De Vries, Maia Lesosky, Gasnat Shaboodien, Heather Cordell, Guillaume Paré, Bernard Keavney, Ayesha Motala, Eugene Sobngwi, Jean Claude Mbanya, Branwen Hennig, Naby Balde, Moffat Nyirenda, John Oli, Clement Adebamowo, Naomi Levitt, Mary Mayige, Saidi Kapiga, Pontiano Kaleebu, Manjinder Sandhu, Liam Smeeth, Mark McCarthy, and Charles Rotimi. Enabling the genomic revolution in africa. *Science*, 344:1346–1348, 6 2014.

[52] Hakhamanesh Mostafavi, Jeffrey P. Spence, Sahin Naqvi, and Jonathan K. Pritchard. Limited overlap of eqtls and gwas hits due to systematic differences in discovery. *bioRxiv*, 2022.

[53] Carla Márquez-Luna, Po-Ru Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, The SIGMA Type 2 Diabetes Consortium, and Alkes L Price. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology*, 41:811–823, 12 2017. https://doi.org/10.1002/gepi.22083.

[54] Beat Neuenschwander, Michael Branson, and David J. Spiegelhalter. A note on the power prior. *Statistics in Medicine*, 28, 12 2009.

[55] Cathy O'Neil. *Weapons of math destruction*. Penguin Books, Harlow, England, 2017.

[56] Michael D. Osterman, Tyler G. Kinzy, and Jessica N. Cooke Bailey. Polygenic risk scores. *Current Protocols*, 1, 5 2021.

[57] Roshni A. Patel, Shaila A. Musharoff, Jeffrey P. Spence, Harold Pimentel, Catherine Tcheandjieu, Hakhamanesh Mostafavi, Nasa Sinnott-Armstrong, Shoa L. Clarke, Courtney J. Smith, Peter P. Durda, Kent D. Taylor, Russell Tracy, Yongmei Liu, W. Craig Johnson, Francois Aguet, Kristin G. Ardlie, Stacey Gabriel, Josh Smith, Deborah A. Nickerson, Stephen S. Rich, Jerome I. Rotter, Philip S. Tsao, Themistocles L. Assimes, and Jonathan K. Pritchard. Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *The American Journal of Human Genetics*, 109(7):1286–1297, 2022.

[58] Alice B. Popejoy and Stephanie M. Fullerton. Genomics is failing on diversity. *Nature*, 538, 10 2016.

[59] Jonathan K Pritchard and Molly Przeworski. Linkage disequilibrium in humans: Models and data. *The American Journal of Human Genetics*, 69:1–14, 2001.

[60] Laura M Raffield, Tin Louie, Tamar Sofer, Deepti Jain, Eli Ipp, Kent D Taylor, George J Papanicolaou, Larissa Avilés-Santa, Leslie A Lange, Cathy C Laurie, Matthew P Conomos, Timothy A Thornton, Yii-Der Ida Chen, Qibin Qi, Scott Cotler, Bharat Thyagarajan, Neil Schneiderman, Jerome I Rotter, Alex P Reiner, and Henry J Lin. Genome-wide association study of iron traits and relation to diabetes in the hispanic community health study/study of latinos (hchs/sol): potential genomic intersection of iron and glucose regulation? *Human Molecular Genetics*, 26:1966–1978, 5 2017.

[61] Sulev Reisberg, Tatjana Iljasenko, Kristi Läll, Krista Fischer, and Jaak Vilo. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLOS ONE*, 12:e0179238–, 7 2017.

[62] Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Yong Min Ahn, Kazufumi Akiyama, Makoto Arai, Ji Hyun Baek, Wei J. Chen, Young-Chul Chung, Gang Feng, Kumiko Fujii, Stephen J. Glatt, Kyooseob Ha, Kotaro Hattori, Teruhiko Higuchi, Akitoyo Hishimoto, Kyung Sue Hong, Yasue Horiuchi, Hai-Gwo Hwu, Masashi Ikeda, Sayuri Ishiwata, Masanari Itokawa, Nakao Iwata, Eun-Jeong Joo, Rene S. Kahn, Sung-Wan Kim, Se Joo Kim, Se Hyun Kim, Makoto Kinoshita, Hiroshi Kunugi, Agung Kusumawardhani, Jimmy Lee, Byung Dae Lee, Heon-Jeong Lee, Jianjun Liu, Ruize Liu, Xiancang Ma, Woojae Myung, Shusuke Numata, Tetsuro Ohmori, Ikuo Otsuka, Yuji Ozeki, Sibylle G. Schwab, Wenzhao Shi, Kazutaka Shimoda, Kang Sim, Ichiro Sora, Jinsong Tang, Tomoko Toyota, Ming Tsuang, Dieter B. Wildenauer, Hong-Hee Won, Takeo Yoshikawa, Alice Zheng, Feng Zhu, Lin He, Akira Sawa, Alicia R. Martin, Shengying Qin, Hailiang Huang, and Tian Ge. Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics*, 54:573–580, 5 2022.

[63] Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Stanley Global Asia Initiatives, Lin He, Akira Sawa, Alicia R. Martin, Shengying Qin, Hailiang Huang, and Tian Ge. Improving polygenic prediction in ancestrally diverse populations. *medRxiv*, 2021.

[64] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19:491–504, 8 2018.

[65] Fredrick R Schumacher, Ali Amin Al Olama, Sonja I Berndt, Sara Benlloch, Mahbubl Ahmed, Edward J Saunders, Tokhir Dadaev, Daniel Leongamornlert, Ezequiel Anokian, Clara Cieza-Borrella, Chee Goh, Mark N Brook, Xin Sheng, Laura Fachal, Joe Dennis, Jonathan Tyrer, Kenneth Muir, Artitaya Lophatananon, Victoria L Stevens, Susan M Gapstur, Brian D Carter, Catherine M Tangen, Phyllis J Goodman, Ian M Thompson, Jyotsna Batra, Suzanne Chambers, Leire Moya, Judith Clements, Lisa Horvath, Wayne Tilley, Gail P Risbridger, Henrik Gronberg, Markus Aly, Tobias Nordström, Paul Pharoah, Nora Pashayan, Johanna Schleutker, Teuvo L J Tammela, Csilla Sipeky, Anssi Auvinen, Demetrius Albanes, Stephanie Weinstein, Alicja Wolk, Niclas Håkansson, Catharine M L West, Alison M Dunning, Neil Burnet, Lorelei A Mucci, Edward Giovannucci, Gerald L Andriole, Olivier Cussenot, Géraldine Cancel-Tassin, Stella Koutros, Laura E Beane Freeman, Karina Dalsgaard Sorensen, Torben Falck Orntoft, Michael Borre, Lovise Maehle, Eli Marie Grindedal, David E Neal, Jenny L Donovan, Freddie C Hamdy, Richard M Martin, Ruth C Travis, Tim J Key, Robert J Hamilton, Neil E Fleshner, Antonio Finelli, Sue Ann Ingles, Mariana C Stern, Barry S Rosenstein, Sarah L Kerns, Harry Ostrer, Yong-Jie Lu, Hong-Wei Zhang, Ninghan Feng, Xueying Mao, Xin Guo, Guomin Wang, Zan Sun, Graham G Giles, Melissa C Southey, Robert J MacInnis, Liesel M FitzGerald, Adam S Kibel, Bettina F Drake, Ana Vega, Antonio Gómez-Caamaño, Robert Szulkin, Martin Eklund, Manolis Kogevinas, Javier Llorca, Gemma Castaño-Vinyals, Kathryn L Penney, Meir Stampfer, Jong Y Park, Thomas A Sellers, Hui-Yi Lin, Janet L Stanford, Cezary Cybulski, Dominika Wokolorczyk, Jan Lubinski, Elaine A Ostrander, Milan S Geybels, Børge G Nordestgaard, Sune F Nielsen, Maren Weischer, Rasmus Bisbjerg, Martin Andreas Røder, Peter Iversen, Hermann Brenner, Katarina Cuk, Bernd Holleczek, Christiane Maier, Manuel Luedeke, Thomas Schnoeller, Jeri Kim, Christopher J Logothetis, Esther M John, Manuel R Teixeira, Paula Paulo, Marta Cardoso, Susan L Neuhausen, Linda Steele, Yuan Chun Ding, Kim De Ruyck, Gert De Meerleer, Piet Ost, Azad Razack, Jasmine Lim, Soo-Hwang Teo, Daniel W Lin, Lisa F Newcomb, Davor Lessel, Marija Gamulin, Tomislav Kulis, Radka Kaneva, Nawaid Usmani, Sandeep Singhal, Chavdar Slavov, Vanio Mitev, Matthew Parliament, Frank Claessens, Steven Joniau, Thomas Van den Broeck, Samantha Larkin, Paul A Townsend, Claire Aukim-Hastie, Manuela Gago-Dominguez, Jose Esteban Castelao, Maria Elena Martinez, Monique J Roobol, Guido Jenster, Ron H N van Schaik, Florence Menegaux, Thérèse Truong, Yves Akoli Koudou, Jianfeng Xu, Kay-Tee Khaw, Lisa Cannon-Albright, Hardev Pandha, Agnieszka Michael, Stephen N Thibodeau, Shannon K McDonnell, Daniel J Schaid, Sara Lindstrom, Constance Turman, Jing Ma, David J Hunter, Elio Riboli, Afshan Siddiq, Federico Canzian, Laurence N Kolonel, Loic Le Marchand, Robert N Hoover, Mitchell J Machiela, Zuxi Cui, Peter Kraft, Christopher I Amos, David V Conti, Douglas F Eas-

ton, Fredrik Wiklund, Stephen J Chanock, Brian E Henderson, Zsofia Kote-Jarai, Christopher A Haiman, Rosalind A Eeles, The Profile Study, Australian Prostate Cancer BioResource (APCB), The IMPACT Study, Canary PASS Investigators, Breast, Prostate Cancer Cohort Consortium (BPC3), The PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, Cancer of the Prostate in Sweden (CAPS), Prostate Cancer Genome wide Association Study of Uncommon Susceptibility Loci (PEGASUS), The Genetic Associations, and Mechanisms in Oncology (GAME-ON)/Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) Consortium. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature Genetics*, 50:928–936, 2018.

[66] Seth A Sharp, Stephen S Rich, Andrew R Wood, Samuel E Jones, Robin N Beaumont, James W Harrison, Darius A Schneider, Jonathan M Locke, Jess Tyrrell, Michael N Weedon, William A Hagopian, and Richard A Oram. Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes Care*, 42:200–207, 2019.

[67] Huwenbo Shi, Kathryn S Burch, Ruth Johnson, Malika K Freund, Gleb Kichaev, Nicholas Mancuso, Astrid M Manuel, Natalie Dong, and Bogdan Pasaniuc. Localizing components of shared transethnic genetic architecture of complex traits from gwas summary data. *The American Journal of Human Genetics*, 106:805–817, 2020.

[68] Michelle J. Sternthal, Natalie Slopen, and David R. Williams. Racial disparities in health. *Du Bois Review: Social Science Research on Race*, 8:95–113, 4 2011.

[69] Jeremy M G Taylor, Kyuseong Choi, and Peisong Han. Data integration: exploiting ratios of parameter estimates from a reduced external model. *Biometrika*, 4 2022.

[70] Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, pages 1–14, 6 2022.

[71] Sarah A Tishkoff and Brian C Verrelli. Patterns of human genetic diversity: Implications for human evolutionary history and disease. *Annual Review of Genomics and Human Genetics*, 4:293–340, 9 2003. doi: 10.1146/annurev.genom.4.070802.110226.

[72] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19:581–590, 2018.

[73] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[74] Yogasudha Veturi, Gustavo de los Campos, Nengjun Yi, Wen Huang, Ana I Vazquez, and Brigitte Kühnel. Modeling heterogeneity in the genetic architecture of ethnically diverse groups using random effect interaction models. *Genetics*, 211:1395–1407, 4 2019.

[75] Bjarni J. Vilhjálmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E. Kenny, Mikkel H. Schierup, Philip De Jager, Nikolaos A. Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M. Visscher, Peter Kraft, Nick Patterson, Alkes L. Price, Stephan Ripke, Benjamin M. Neale, Aiden Corvin, James T.R. Walters, Kai-How Farh, Peter A. Holmans, Phil Lee, Brendan Bulik-Sullivan, David A. Collier, Hailiang Huang, Tune H. Pers, Ingrid Agartz, Esben Agerbo, Margot Albus, Madeline Alexander, Farooq Amin, Silviu A. Bacanu, Martin Begemann, Richard A. Belliveau, Judit Bene, Sarah E. Bergen, Elizabeth Bevilacqua, Tim B. Bigdeli, Donald W. Black, Richard Bruggeman, Nancy G. Buccola, Randy L. Buckner, William Byerley, Wiepke Cahn, Guiqing Cai, Dominique Campion, Rita M. Cantor, Vaughan J. Carr, Noa Carrera, Stanley V. Catts, Kimberly D. Chambert, Raymond C.K. Chan, Ronald Y.L. Chen, Eric Y.H. Chen, Wei Cheng, Eric F.C. Cheung, Siow Ann Chong, C. Robert Cloninger, David Cohen, Nadine Cohen, Paul Cormican, Nick Craddock, James J. Crowley, David Curtis, Michael Davidson, Kenneth L. Davis, Franziska Degenhardt, Jurgen Del Favero, Lynn E. DeLisi, Ditte Demontis, Dimitris Dikeos, Timothy Dinan, Srdjan Djurovic, Gary Donohoe, Elodie Drapeau, Jubao Duan, Frank Dudbridge, Naser Durmishi, Peter Eichhammer, Johan Eriksson, Valentina Escott-Price, Laurent Essioux, Ayman H. Fanous, Martilias S. Farrell, Josef Frank, Lude Franke, Robert Freedman, Nelson B. Freimer, Marion Friedl, Joseph I. Friedman, Menachem Fromer, Giulio Genovese, Lyudmila Georgieva, Elliot S. Gershon, Ina Giegling, Paola Giusti-Rodrguez, Stephanie Godard, Jacqueline I. Goldstein, Vera Golimbet, Srihari Gopal, Jacob Gratten, Jakob Grove, Lieuwe de Haan, Christian Hammer, Marian L. Hamshere, Mark Hansen, Thomas Hansen, Vahram Haroutunian, Annette M. Hartmann, Frans A. Henskens, Stefan Herms, Joel N. Hirschhorn, Per Hoffmann, Andrea Hofman, Mads V. Hollegaard, David M. Hougaard, Masashi Ikeda, Inge Joa, Antonio Julia, Rene S. Kahn, Luba Kalaydjieva, Sena Karachanak-Yankova, Juha Karjalainen, David Kavanagh, Matthew C. Keller, Brian J. Kelly, James L. Kennedy, Andrey Khrunin, Yunjung Kim, Janis Klovins, James A. Knowles, Bettina Konte, Vaidutis Kucinskas, Zita Ausrele Kucinskiene, Hana Kuzelova-Ptackova, Anna K. Kahler, Claudine Laurent, Jimmy Lee Chee Keong, S. Hong Lee, Sophie E. Legge, Bernard Lerer, Miaoxin Li, Tao Li, Kung-Yee Liang, Jeffrey Lieberman, Svetlana Limborska, Carmel M. Loughland, Jan Lubinski, Jouko Lnnqvist, Milan Macek, Patrik K.E. Magnusson, Brion S. Maher, Wolfgang Maier, Jacques Mallet, Sara Marsal, Manuel Mattheisen, Morten Mattingsdal, Robert W. McCarley, Colm McDonald, Andrew M. McIntosh, Sandra Meier, Carin J. Meijer, Bela Melegh, Ingrid Melle, Raquelle I. Mesholam-Gately, Andres Metspalu, Patricia T. Michie, Lili Milani, Vihra Milanova, Younes Mokrab, Derek W. Morris, Ole Mors, Preben B. Mortensen, Kieran C. Murphy, Robin M. Murray, Inez Myin-Germeys, Bertram Mller-Myhsok, Mari Nelis, Igor Nenadic, Deborah A. Nertney, Gerald Nestadt, Kristin K. Nicodemus, Liene Nikitina-Zake, Laura Nisenbaum, Annelie Nordin, Eadbhard O'Callaghan, Colm O'Dushlaine, F. Anthony O'Neill, Sang-Yun Oh, Ann Olincy, Line Olsen, Jim Van Os, Christos Pantelis, George N. Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T. Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O. Perkins, Olli Pietilinen, Jonathan Pimm, Andrew J. Pocklington, John Powell, Alkes Price, Ann E. Pulver, Shaun M. Purcell, Digby Quested, Henrik B. Rasmussen, Abraham Reichen-

berg, Mark A. Reimers, Alexander L. Richards, Joshua L. Roffman, Panos Roussos, Douglas M. Ruderfer, Veikko Salomaa, Alan R. Sanders, Ulrich Schall, Christian R. Schubert, Thomas G. Schulze, Sibylle G. Schwab, Edward M. Scolnick, Rodney J. Scott, Larry J. Seidman, Jianxin Shi, Engilbert Sigurdsson, Teimuraz Silagadze, Jeremy M. Silverman, Kang Sim, Petr Slominsky, Jordan W. Smoller, Hon-Cheong So, Chris C.A. Spencer, Eli A. Stahl, Hreinn Stefansson, Stacy Steinberg, Elisabeth Stogmann, Richard E. Straub, Eric Strengman, Jana Strohmaier, T. Scott Stroup, Mythily Subramaniam, Jaana Suvisaari, Dragan M. Svrakic, Jin P. Szatkiewicz, Erik Sderman, Srinivas Thirumalai, Draga Toncheva, Paul A. Tooney, Sarah Tosato, Juha Veijola, John Waddington, Dermot Walsh, Dai Wang, Qiang Wang, Bradley T. Webb, Mark Weiser, Dieter B. Wildenauer, Nigel M. Williams, Stephanie Williams, Stephanie H. Witt, Aaron R. Wolen, Emily H.M. Wong, Brandon K. Wormley, Jing Qin Wu, Hualin Simon Xi, Clement C. Zai, Xuebin Zheng, Fritz Zimprich, Naomi R. Wray, Kari Stefansson, Peter M. Visscher, Rolf Adolfsson, Ole A. Andreassen, Douglas H.R. Blackwood, Elvira Bramon, Joseph D. Buxbaum, Anders D. Børglum, Sven Cichon, Ariel Darvasi, Enrico Domenici, Hannelore Ehrenreich, Tonu Esko, Pablo V. Gejman, Michael Gill, Hugh Gurling, Christina M. Hultman, Nakao Iwata, Assen V. Jablensky, Erik G. Jonsson, Kenneth S. Kendler, George Kirov, Jo Knight, Todd Lencz, Douglas F. Levinson, Qingqin S. Li, Jianjun Liu, Anil K. Malhotra, Steven A. McCarroll, Andrew McQuillin, Jennifer L. Moran, Preben B. Mortensen, Bryan J. Mowry, Markus M. Nthen, Roel A. Ophoff, Michael J. Owen, Aarno Palotie, Carlos N. Pato, Tracey L. Petryshen, Danielle Posthuma, Marcella Rietschel, Brien P. Riley, Dan Rujescu, Pak C. Sham, Pamela Sklar, David St. Clair, Daniel R. Weinberger, Jens R. Wendland, Thomas Werge, Mark J. Daly, Patrick F. Sullivan, Michael C. O'Donovan, Peter Kraft, David J. Hunter, Muriel Adank, Habibul Ahsan, Kristiina Aittomäki, Laura Baglietto, Sonja Berndt, Carl Blomquist, Federico Canzian, Jenny Chang-Claude, Stephen J. Chanock, Laura Crisponi, Kamila Czene, Norbert Dahmen, Isabel dos Santos Silva, Douglas Easton, A. Heather Eliassen, Jonine Figueroa, Olivia Fletcher, Montserrat Garcia-Closas, Mia M. Gaudet, Lorna Gibson, Christopher A. Haiman, Per Hall, Aditi Hazra, Rebecca Hein, Brian E. Henderson, Albert Hofman, John L. Hopper, Astrid Irwanto, Mattias Johansson, Rudolf Kaaks, Muhammad G. Kibriya, Peter Lichtner, Sara Lindström, Jianjun Liu, Eiliv Lund, Enes Makalic, Alfons Meindl, Hanne Meijers-Heijboer, Bertram Müller-Myhsok, Taru A. Muranen, Heli Nevanlinna, Petra H. Peeters, Julian Peto, Ross L. Prentice, Nazneen Rahman, María José Sánchez, Daniel F. Schmidt, Rita K. Schmutzler, Melissa C. Southey, Rulla Tamimi, Ruth Travis, Clare Turnbull, Andre G. Uitterlinden, Rob B. van der Luijt, Quinten Waisfisz, Zhaoming Wang, Alice S. Whittemore, Rose Yang, and Wei Zheng. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97, 10 2015.

[76] Karani S Vimaleswaran, Diane J Berry, Chen Lu, Emmi Tikkanen, Stefan Pilz, Linda T Hiraki, Jason D Cooper, Zari Dastani, Rui Li, Denise K Houston, Andrew R Wood, Karl Michaëlsson, Liesbeth Vandenput, Lina Zgaga, Laura M Yerges-Armstrong, Mark I McCarthy, Josée Dupuis, Marika Kaakinen, Marcus E Kleber, Karen Jameson, Nigel Arden, Olli Raitakari, Jorma Viikari, Kurt K Lohman, Luigi Ferrucci, Håkan Melhus, Erik Ingelsson, Liisa Byberg, Lars Lind, Mattias Lorentzon, Veikko Salomaa, Harry Campbell, Malcolm Dunlop, Braxton D Mitchell, Karl-Heinz Herzig, Anneli Pouta, Anna-Liisa Hartikainen, the Genetic Investigation of Anthropometric Traits (GIANT) consortium, Elizabeth A Streeten,

Evropi Theodoratou, Antti Jula, Nicholas J Wareham, Claes Ohlsson, Timothy M Frayling, Stephen B Kritchevsky, Timothy D Spector, J Brent Richards, Terho Lehtimäki, Willem H Ouwehand, Peter Kraft, Cyrus Cooper, Winfried März, Chris Power, Ruth J F Loos, Thomas J Wang, Marjo-Riitta Järvelin, John C Whittaker, Aroon D Hingorani, and Elina Hyppönen. Causal relationship between obesity and vitamin d status: Bi-directional mendelian randomization analysis of multiple cohorts. *PLOS Medicine*, 10:e1001383–, 2 2013.

[77] Benjamin F Voight, Gina M Peloso, Marju Orho-Melander, Ruth Frikke-Schmidt, Maja Barbalic, Majken K Jensen, George Hindy, Hilma Hólm, Eric L Ding, Toby Johnson, Heribert Schunkert, Nilesh J Samani, Robert Clarke, Jemma C Hopewell, John F Thompson, Mingyao Li, Gudmar Thorleifsson, Christopher Newton-Cheh, Kiran Musunuru, James P Pirruccello, Danish Saleheen, Li Chen, Alexandre F R Stewart, Arne Schillert, Unnur Thorsteinsdottir, Gudmundur Thorgeirsson, Sonia Anand, James C Engert, Thomas Morgan, John Spertus, Monika Stoll, Klaus Berger, Nicola Martinelli, Domenico Girelli, Pascal P McKeown, Christopher C Patterson, Stephen E Epstein, Joseph Devaney, Mary-Susan Burnett, Vincent Mooser, Samuli Ripatti, Ida Surakka, Markku S Nieminen, Juha Sinisalo, Marja-Liisa Lokki, Markus Perola, Aki Havulinna, Ulf de Faire, Bruna Gigante, Erik Ingelsson, Tanja Zeller, Philipp Wild, Paul I W de Bakker, Olaf H Klungel, Anke-Hilse Maitland van der Zee, Bas J M Peters, Anthonius de Boer, Diederick E Grobbee, Pieter W Kamphuisen, Vera H M Deneer, Clara C Elbers, N Charlotte Onland-Moret, Marten H Hofker, Cisca Wijmenga, W M Monique Verschuren, Jolanda M A Boer, Yvonne T van der Schouw, Asif Rasheed, Philippe Frossard, Serkalem Demissie, Cristen Willer, Ron Do, Jose M Ordovas, Gonçalo R Abecasis, Michael Boehnke, Karen L Mohlke, Mark J Daly, Candace Guiducci, Noël P Burtt, Aarti Surti, Elena Gonzalez, Shaun Purcell, Stacey Gabriel, Jaume Marrugat, John Peden, Jeanette Erdmann, Patrick Diemert, Christina Willenborg, Inke R König, Marcus Fischer, Christian Hengstenberg, Andreas Ziegler, Ian Buysschaert, Diether Lambrechts, Frans Van de Werf, Keith A Fox, Nour Eddine El Mokhtari, Diana Rubin, Jürgen Schrezenmeir, Stefan Schreiber, Arne Schäfer, John Danesh, Stefan Blankenberg, Robert Roberts, Ruth McPherson, Hugh Watkins, Alistair S Hall, Kim Overvad, Eric Rimm, Eric Boerwinkle, Anne Tybjaerg-Hansen, L Adrienne Cupples, Muredach P Reilly, Olle Melander, Pier M Mannucci, Diego Ardissino, David Siscovick, Roberto Elosua, Kari Stefansson, Christopher J O'Donnell, Veikko Salomaa, Daniel J Rader, Leena Peltonen, Stephen M Schwartz, David Altshuler, and Sekar Kathiresan. Plasma hdl cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*, 380:572–580, 2012.

[78] Ying Wang, Jing Guo, Guiyan Ni, Jian Yang, Peter M Visscher, and Loic Yengo. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications*, 11:3865, 2020.

[79] Omer Weissbrod, Masahiro Kanai, Huwenbo Shi, Steven Gazal, Wouter Peyrot, Amit Khera, Yukinori Okada, The Biobank, Japan Project, Alicia Martin, Hilary Finucane, and Alkes L Price. Leveraging fine-mapping and non-european training data to improve trans-ethnic polygenic risk scores. *medRxiv*, 2021.

[80] Omer Weissbrod, Masahiro Kanai, Huwenbo Shi, Steven Gazal, Wouter J. Peyrot, Amit V. Khera, Yukinori Okada, Koichi Matsuda, Yuji Yamanashi, Yoichi Furukawa, Takayuki

Morisaki, Yoshinori Murakami, Yoichiro Kamatani, Kaori Muto, Akiko Nagai, Wataru Obara, Ken Yamaji, Kazuhisa Takahashi, Satoshi Asai, Yasuo Takahashi, Takao Suzuki, Nobuaki Sinozaki, Hiroki Yamaguchi, Shiro Minami, Shigeo Murayama, Kozo Yoshimori, Satoshi Nagayama, Daisuke Obata, Masahiko Higashiyama, Akihide Masumoto, Yukihiro Koretsune, Alicia R. Martin, Hilary K. Finucane, and Alkes L. Price. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nature Genetics*, 54:450–458, 4 2022.

[81] David R. Williams, Jourdyn A. Lawrence, Brigette A. Davis, and Cecilia Vu. Understanding how discrimination can affect health. *Health Services Research*, 54:1374–1388, 12 2019.

[82] Genevieve L Wojcik, Mariaelisa Graff, Katherine K Nishimura, Ran Tao, Jeffrey Haessler, Christopher R Gignoux, Heather M Highland, Yesha M Patel, Elena P Sorokin, Christy L Avery, Gillian M Belbin, Stephanie A Bien, Iona Cheng, Sinead Cullina, Chani J Hodonsky, Yao Hu, Laura M Huckins, Janina Jeff, Anne E Justice, Jonathan M Kocarnik, Unhee Lim, Bridget M Lin, Yingchang Lu, Sarah C Nelson, Sung-Shim L Park, Hannah Poisner, Michael H Preuss, Melissa A Richard, Claudia Schurmann, Veronica W Setiawan, Alexandra Sockell, Karan Vahi, Marie Verbanck, Abhishek Vishnu, Ryan W Walker, Kristin L Young, Niha Zubair, Victor Acuña-Alonso, Jose Luis Ambite, Kathleen C Barnes, Eric Boerwinkle, Erwin P Bottinger, Carlos D Bustamante, Christian Caberto, Samuel Canizales-Quinteros, Matthew P Conomos, Ewa Deelman, Ron Do, Kimberly Doheny, Lindsay Fernández-Rhodes, Myriam Fornage, Benyam Hailu, Gerardo Heiss, Brenna M Henn, Lucia A Hindorff, Rebecca D Jackson, Cecelia A Laurie, Cathy C Laurie, Yuqing Li, Dan-Yu Lin, Andres Moreno-Estrada, Girish Nadkarni, Paul J Norman, Loreall C Pooler, Alexander P Reiner, Jane Romm, Chiara Sabatti, Karla Sandoval, Xin Sheng, Eli A Stahl, Daniel O Stram, Timothy A Thornton, Christina L Wassel, Lynne R Wilkens, Cheryl A Winkler, Sachi Yoneyama, Steven Buyske, Christopher A Haiman, Charles Kooperberg, Loic Le Marchand, Ruth J F Loos, Tara C Matise, Kari E North, Ulrike Peters, Eimear E Kenny, and Christopher S Carlson. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570:514–518, 2019.

[83] Cheng Zheng, Sayan Dasgupta, Yuxiang Xie, Asad Haris, and Ying Qing Chen. On data enriched logistic regression, 2019.

[84] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, 9:e1003264, 2 2013.

[85] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The Annals of Applied Statistics*, 11, 9 2017.