INTEGRATING GENETICS INTO ECONOMICS
Integrating Genetics into Economics

Het integreren van genetica in de economie

Thesis

to obtain the degree of Doctor from the Erasmus University Rotterdam by command of the rector magnificus

Prof. dr. F.A. van der Duijn Schouten

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on Friday February 19, 2021 at 13:00 hrs

by

ERIC ARSENE WILLEM SLOB
born in Utrecht.
Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself. - F.H.C. Crick
Acknowledgements

First and foremost, I would like to thank my supervisors, Patrick Groenen, Niels Rietveld, and Roy Thurik for enabling me to do a PhD and being always there for me over the past four years. Together you make the best team I can think of in supervising a PhD thesis. You managed to create a nice and productive environment for me, with Niels as the mastermind behind everything. Patrick, I thank you for your enthusiasm, your patience and the time you always created for me. Despite always being in one important management function or another, I could always knock on your door for advice, or to talk about optimization and/or computational tricks. You taught me a lot during our matrix algebra puzzling sessions, where you were always convinced we could do things smarter, faster and better. Roy, I thank you for being such a good listener, your great advice and improving my writing skills. Each meeting I had with you, I always laughed a lot no matter the situation. Niels, working together with you over the past years has been an absolute pleasure. Not only are you an amazing scientist with a network that helped me a lot, I especially value your kindness, accessibility, availability, and encouragement throughout the past years. I thank you a lot for improving my multi-tasking skills (still not even close to yours), the feedback you would give me so quickly that I would run behind on other projects as I felt I had to send things back to you quickly again, and for always being there for me. I loved our daily coffee sessions which could be about work, politics, the Tour de France or anything else.

I am very grateful to all (former) members of the Organisation, Strategy and Entrepreneurship group and the broader Applied Economics department: Ajay, Bas, Brigitte, Enrico, Fleur, Francesco, Frank, Hans, Jan, Jolanda, Kirsten, Michiel, Nicola, Owen, Peter, Sophie, Stephanie, Teresa, Thomas, Tom, and Zhiling. I am grateful to the members of the MRC Biostatistics Unit for making me feel at home throughout my visit in Cambridge.

I am extremely grateful to have had such a nice group of fellow PhD students at Erasmus School of Economics. I would like to express gratitude for having such a nice office atmosphere with Sam and Joaquim (our adopted secretary to pick up the phone or play basketball). I think we helped each other a lot by
discussing work and doing walks to the coffee machine. Another part of the wolf pack was Gianluca. With the four of us we had a lot of fun and crazy events, going from boat trips to a Great Gatsby themed party to winning the EUR Vital Sport Day 2019 with the golden dream team. Next to these individuals, I would like to thank my fellow PhD students where I could always drop by: Yannis, Thomas, Sara, Sanaz, Sai, Rutger, Plato, Nienke, Merel, Megan, Max, Kristel, Kevin, Indy, Gertjan, Esmee, Didier, Cristian, Caroline, and Annelot.

I would like to thank my co-authors Philip, Philipp, Ronald, Stephen. Philip, thank you for your patience and trying to make me understand our brain a bit better. I have learned a lot by working together with you. Philipp, your ambition and enthusiasm always helped me. I am also grateful that you would always include me in your group at conferences. Ronald, thank you for your companionship, and passion. I learned a lot of things by programming with you. I also appreciated the valuable discussions we had on my other projects and not work-related discussions. Steve, thanks a lot for your hospitality and patience. I greatly value our discussions and I thoroughly enjoyed working together with you. Also, by having me over as a guest I think I learned a lot not only as an academic, but also as a person. I still have fond memories of my time there and am happy to continue working with you.

I would also like to express my gratitude to Dennis Fok, Hans van Kippersluis, and Stephanie von Hinke for reading this dissertation and being part of the doctoral committee.

Next to all these people in my academic world, I would like to thank a lot of friends and family. I would like to thank my friends from the rowing club A.R.S.R. “Skadi” for the dinners, bike rides and holidays we had together. Sander, Wessel, our moving in together was at the time when I started my PhD project. Living together has been a lot of fun. I loved being able to discuss my work (or anything else) and always having such a nice and relaxing atmosphere to come home to. Patrick, I am very grateful to have you around me as a friend since secondary school and still being able to fall back to you. Eva, meeting you in the last year has been a nice unexpected surprise in a busy time. I know that my work has been occasionally difficult for you, so I want to thank you for your patience and support. I know I can always count on you. I want to thank my sister, Elise, for drawing the cover of my dissertation. Elise, Jacqueline, Jan, Robert, I am very thankful for having had such a lovely and stimulating environment to grow up. I know I can always come back to your warmth. I want to thank you for all your kind words and the trust you gave me over the past four years.

Eric Slob
Rotterdam, 27 Augustus 2020
# Table of Contents

Acknowledgements vii  
Table of Contents ix  
List of Figures xiv  
List of Tables xvi  

1. Introduction and conclusion 1  
   1.1. Motivation 2  
   1.2. Research topics 4  
   1.3. Research questions and results 11  
   1.4. Conclusion and implications 16  
   1.5. Individual contributions and publication status per chapter 18  

PART I: MENDELIAN RANDOMIZATION 21  

2. A note on the use of Egger regression in Mendelian randomization studies 23  
   2.1. Introduction 24  
   2.2. Methods 25
# Index

2.3. Conclusion ......................................................... 27

2.A. Approximation of the correlation between the first stage effects and the direct effects in two examples .............................. 29

3. A comparison of robust Mendelian randomization methods using summary data ................................................................. 33

3.1. Introduction ............................................................ 34

3.2. Methods ............................................................... 35

3.3. Results ................................................................. 45

3.4. Discussion ............................................................ 52

3.A. Details of simulation study ............................................ 55

3.B. Outliers according to different methods ............................... 57

# PART II: POLYGENIC RISK SCORES

4. A decade of research on the genetics of entrepreneurship: a review and view ahead ................................................................. 63

4.1. Introduction ............................................................ 64

4.2. The heritability of entrepreneurship .................................... 65

4.3. The molecular genetic analysis of entrepreneurship .................. 66

4.4. Empirical illustration .................................................. 74

4.5. Conclusion: a second decade? ........................................ 76

5. Does the genetic predisposition to smoking moderate the response to tobacco excise taxes? .......................................................... 83

5.1. Introduction ............................................................ 84

5.2. Data description ....................................................... 86

5.3. Methods ............................................................... 89

5.4. Results ................................................................. 90

5.5. Discussion and conclusion ............................................ 93
6. Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence

6.1. Introduction ................................................................. 100
6.2. Data: UK Biobank Imaging Study ....................................... 101
6.3. Methods ................................................................. 103
6.4. Results ................................................................. 105
6.5. Discussion .............................................................. 110
6.A. Method derivation ...................................................... 113
6.B. Data usage for constructing phenotypes ............................. 123
6.C. Data cells used for identification of brain damage ................... 128
6.D. Pipeline .............................................................. 129
6.E. Heritability estimates .................................................. 131

Bibliography 135

Summary 155

Samenvatting 157

About the Author 159

Portfolio 161

ERIM Publications List 165
List of Figures

1.1. Illustrative diagram showing the model assumed for genetic variant $G_j$, with effect $\phi_j$ on the unobserved confounder $U$, effect $\gamma_j$ on exposure $X$, and direct effect $\alpha_j$ on outcome $Y$. The causal effect of the exposure on the outcome is $\theta$. Dotted lines represent possible ways the instrumental variable assumptions could be violated. 7


3.1. Illustrative diagram showing the model assumed for genetic variant $G_j$, with effect $\phi_j$ on the unobserved confounder $U$, effect $\gamma_j$ on exposure $X$, and direct effect $\alpha_j$ on outcome $Y$. The causal effect of the exposure on the outcome is $\theta$. Dotted lines represent possible ways the instrumental variable assumptions could be violated. 36

3.2. Scatter plot of genetic associations with BMI (standard deviation units) and coronary artery disease risk (log odds ratios) for 94 variants taken from the GIANT and CARDIoGRAMplusC4D consortia respectively. 45

3.3. Mean squared errors for the different methods in scenario 2 (directional pleiotropy, InSIDE satisfied) with a null causal effect for 30 variants. Note the vertical axis is on a logarithmic scale. 47

3.4. Mean squared errors for the different methods in scenario 3 (directional pleiotropy, InSIDE violated) with a null causal effect for 30 variants. Note the vertical axis is on a logarithmic scale. 51
3.5. Mean squared error for the different methods in scenario 2 for 10,000 simulations, with directional pleiotropy and InSIDE satisfied with 10 variants. ........................................ 59

3.6. Mean squared error for the different methods in scenario 3 for 10,000 simulations, with directional pleiotropy and InSIDE violated with 10 variants. ........................................ 59

3.7. Mean squared error for the different methods in scenario 2 for 10,000 simulations, with directional pleiotropy and InSIDE satisfied with 100 variants. ........................................ 60

3.8. Mean squared error for the different methods in scenario 3 for 10,000 simulations, with directional pleiotropy and InSIDE violated with 100 variants. ........................................ 60

5.1. The average, minimum and maximum tobacco excise taxes levied per pack of 20 cigarettes in the United States from 1992 to 2014. ........................................ 91

6.1. Spatial mapping of the estimates for SNP-based heritability and genetic correlation across the different brain regions, SNP-based heritability per anatomical area, and genetic correlation table of aggregated anatomical area. ........................................ 107

6.2. Dendogram of the hierarchical clustering of the genetic correlation matrix. ........................................ 108

6.3. Spatial mapping of the genetic correlation between brain regions and the behavioral traits, where blue points represent a negative correlation and red points a positive correlation. ........................................ 109
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Publication status of the chapters.</td>
<td>19</td>
</tr>
<tr>
<td>2.1</td>
<td>Summary association results for 29 SNPs associated with systolic blood pressure (SNPs are ordered as in Table 1 of the study by the International Consortium for Blood Pressure Genome-Wide Association Studies (2011)).</td>
<td>31</td>
</tr>
<tr>
<td>2.2</td>
<td>Summary association results for 31 SNPs associated with plasma urate concentration (SNPs are ordered as in Table S3 of the study by White et al. (2016)).</td>
<td>32</td>
</tr>
<tr>
<td>3.1</td>
<td>Summary comparison of methods.</td>
<td>38</td>
</tr>
<tr>
<td>3.2</td>
<td>Mean, median, standard deviation (SD) of estimates, and Type 1 error/empirical power (%) with 10 genetic variants.</td>
<td>48</td>
</tr>
<tr>
<td>3.3</td>
<td>Mean, median, standard deviation (SD) of estimates, and Type 1 error/empirical power (%) with 30 genetic variants.</td>
<td>49</td>
</tr>
<tr>
<td>3.4</td>
<td>Mean, median, standard deviation (SD) of estimates, and Type 1 error/empirical power (%) with 100 genetic variants.</td>
<td>50</td>
</tr>
<tr>
<td>3.5</td>
<td>Estimates and 95% confidence intervals (CI) for the effect of BMI on coronary artery disease risk from robust methods. Estimates represent log odds ratios for CAD risk per 1 kg/m² increase in BMI.</td>
<td>52</td>
</tr>
</tbody>
</table>
3.6. Genetic variants identified as outliers by the different methods in
the Mendelian Randomization study of the effect of BMI on cardio-
vascular disease risk and other traits the variants are associated
with according to the NHGRI-EBI Catalog. 58

4.1. The association between the polygenic risk scores for traits in the
mental health domain and self-employment (random-effects regres-
sion, \(N_{\text{individual-year}} = 31,927, N_{\text{individual}} = 7,948\)). 76

4.2. In-sample prediction results for self-employment (versus wage work)
for the models with and without polygenic risk scores; observations
in the top 19.9% (percentage of person-year observations reporting
self-employment in the sample) of the predicted values in each model
are classified as self-employed. 77

5.1. Descriptive statistics analysis sample. 90

5.2. Results of the regressions explaining someone’s current smoking status. 92

5.3. Results of the regressions explaining someone’s current smoking
intensity. 93

6.1. UK Biobank phenotype data used in this study, with corresponding
description, measurement units and data fields. 123

6.2. Brain diseases with corresponding data fields in the self report and
ICD10 codes. 129

6.3. The estimated SNP-heritability for the different phenotypes in UK
Biobank using M-GREML. 131
Introduction and conclusion

Abstract

The massive increase in sample size of genetic cohorts, combined with an increase in the collection of data on social-scientific outcomes in these datasets, has made it possible to study many socio-economically relevant individual characteristics from a genetics perspective. In economics, the subfield that studies the genetic architecture of socioeconomic outcomes and preferences is often called genoeconomics. Ultimately, genoeconomics can help economics in four different ways: genes can be used as measures of previous latent variables, genes can uncover biological mechanisms, genes can be used as control variables or instrumental variables, and genes can be used to target policy interventions. In this thesis, I develop and compare some methods that can be used in genoeconomics, and I show through empirical studies how genetically informed study designs can give new insights to economists. The methods developed and compared in this thesis foster the use of genes as instrumental variables and help further the understanding of genetic relationships across socio-economically relevant characteristics. The main empirical applications in this thesis concern smoking behaviour, entrepreneurship, and the structure of the brain. This first chapter provides an overview of the thesis, including a discussion of the research questions it addresses and the implications resulting from the answers to these questions.
1.1 Motivation

Economics is the social science that studies the production, distribution, and consumption of goods and services (Krugman and Wells, 2015). All these activities require choices from so-called economic agents (individuals or organizations), as resources are scarce. Over the past few decades, it has been convincingly shown that all human traits (including preferences) are heritable (Polderman et al., 2015, Turkheimer, 2000). Moreover, significant associations have been found between genetic variants and preferences such as risk aversion (Linnér et al., 2019), health behaviours such as smoking (Gelernter et al., 2015), and indicators of socio-economic status such as educational attainment (Lee et al., 2018). The use of insights from genetics to increase our understanding of how economic agents make their choices is called 'genoeconomics' by Benjamin et al. (2008). In this thesis, I develop and compare methods to foster the further emergence of the field of genoeconomics, and I perform genetically informed empirical analyses to better understand smoking behaviour, entrepreneurship, and the structure of the brain.

In their article, Benjamin et al. (2012a) discuss four promises of how genoeconomics can contribute to economics. The first promise is that genes can be used as a direct measure for a previously latent variable. Sometimes, it can be difficult to measure an individual's preferences. However, in some cases, it is possible to proxy these preferences by using an individual's genetic profile. For example, one can potentially use genetic information to determine whether an individual is likely to be risk averse (Linnér et al., 2019) or to have particular abilities (Lee et al., 2018).

The second promise relates to the uncovering of biological mechanisms using genetic data. Genetic data can be used not only to test existing hypotheses about the biological constitutes of behaviour but also to generate new hypotheses. For example, Benjamin et al. (2012a) discuss an earlier experiment by Kosfeld et al. (2005) showing that individuals who received a dose of the neuropeptide oxytocin exhibit high levels of trusting behaviour. This experiment suggests that oxytocin causally influences trusting behaviour. Using genes that encode the receptor for oxytocin, one can test whether this hypothesis is true. New insights and hypotheses about the biological foundation of behaviour may, however, result from unexpected associations between certain markers in the DNA and individual characteristics. This often occurs in a genome-wide association study (GWAS), in which the trait of interest is associated with a large genome-wide set of genetic variants. In such GWASs, one often finds significant associations between the trait of interest and genetic variants for which the biological function is still poorly understood. As such, it could happen that a GWAS on time-preferences generates
new hypotheses about biological mechanisms influencing human behaviour.

Third, genes can be used as an instrumental variable or as a control variable in empirical models. Using genes as an instrumental variable may help to establish causal effects in cases in which randomization is difficult or unethical. For example, it is arguably unethical to use a form of randomization in which some individuals are not allowed to obtain education to estimate the impact of education on someone’s lifetime salary. However, one could instead use genes that are associated with educational attainment as an instrumental variable to investigate whether education causally influences someone’s salary. As the distribution of genes is random conditional on family fixed effects, it is still possible to make causal interferences if there are significant salary differences between individuals with a high and low genetic endowment for education. Given the heritable nature of human behaviour, genes could also be used as a regular control in order to remove some of the residual variance. This may be particularly useful in an experimental setting in which the recruitment of participants is difficult or costly. Consider for example an experiment in which one is interested in the differences in risk preferences between males and females (these experiments can be costly as the participants usually get a financial reward based on their choices to mimic reality as closely as possible). Because of the heritable nature of risk preferences (Benjamin et al., 2012b, Linnér et al., 2019), controlling for genetic endowments towards risk preferences may lower the residual variance in these experiments and thus, stronger inferences can be obtained. By adding this information, the uncertainty (standard errors) in the sex effect estimates are lower and thus, a smaller sample size is needed for testing the hypothesis.

Fourth, genes could be used for targeting interventions. In medicine, there are already programmes in which individuals with a high genetic risk to develop diseases such as breast cancer are given treatments before they actually develop the disease in order to improve the quality of life of these individuals. Similarly, one could think of using genetic screening for children who are likely to develop dyslexia. We could think of giving these children extra attention in school early on to reduce the difficulties they have with reading compared to their peers.

In this thesis, I contribute to the realization of the four promises outlined by Benjamin et al. (2012a). In the first part of this thesis, related to the third promise of Benjamin et al. (2012a), I look into methods and techniques using genetic markers as instrumental variables. These so-called Mendelian randomization studies constitute Chapters 2 and 3. In the second part, I use so-called polygenic risk scores to describe pathways from genes to entrepreneurship (Chapter 4) and to explain why individuals make different choices in response to an increase in tobacco excise taxes (Chapter 5). This part relates to the second and fourth
promises of Benjamin et al. (2012a). Last, in the third part, I develop a method to understand to what extent traits are genetically related (Chapter 6). With this method, it is possible to estimate what part of a correlation between two traits is shared because they are influenced by the same genetic variants. As such, this chapter contributes to the realization of the first and second promises of Benjamin et al. (2012a).

The remainder of this introductory chapter is organized as follows. In Section 1.2, I will give a short description of the main methods used in genoeconomics and of the chapters in this thesis. The research questions and main findings will be presented in Section 1.3. Next, in Section 1.4, I will address the question of how the chapters in this thesis contribute to the fulfilment of the promises of genoeconomics outlined in the present section. Finally, in Section 1.5 I will discuss my contribution to each chapter, and I give an overview of the publication status of the chapters in this thesis.

1.2 Research Topics

In this section, I provide a brief description of the human genome, and I discuss methods used in genoeconomics to analyse genetic data. Thereafter, I discuss the research topics of my thesis. Parts of this section are taken from chapters 3, 4, and 6 of this thesis.

1.2.1 The human genome

A complete human genome consists of 23 pairs of chromosomes, from which the 23rd pair determines the biological sex of an individual. One of each pair of chromosomes is inherited from the mother, and the other is inherited from the father. A chromosome is composed of two intertwined strands of deoxyribonucleic acid (DNA), each made up of a sequence of nucleotide molecules. There are four different nucleotide molecules in the DNA: adenine, cytosine, thymine, and guanine. Adenine on one strand is always paired with thymine on the other strand, and cytosine is always paired with guanine. These combinations are called base pairs. Every human genome consists of approximately 3 billion base pairs. The stretches of base pairs in the DNA coding of a protein are called genes. There are approximately 20,000 genes in the human genome with varying lengths. A random pair of individuals share approximately 99.9% of their DNA (National Human Genome Research Institute, 2018b), and most genetic differences across population members can be attributed to single nucleotide polymorphisms (SNPs, pronounced “snips”). Therefore, genoeconomists focus primarily on SNPs when analysing heritable genetic variation. A SNP is defined as a location in the DNA
1. Introduction and conclusion

strand at which two different nucleotides are present in the population. Each of the two possible nucleotides is called an allele for that SNP. The allele that is least common in the population is called the minor allele; the other allele is called the major allele. For each SNP, an individual’s genotype is coded as 0, 1 or 2, depending on the number of minor alleles present. Individuals who inherited the same allele from each parent are called homozygous for that SNP (and have genotype 0 or 2), while individuals who inherited different alleles are called heterozygous (and have genotype 1). SNPs can be found in every part of the genome, within genes or in regions in between genes, and may influence the production of proteins. In the human genome, there are approximately 85 million SNPs with a minor allele prevalence of at least 1% (The 1000 Genomes Project Consortium, 2015). When relating so many SNPs $x_{ij}$ (coded as 0, 1, or 2) to a specific outcome $y_i$ in a regression framework such as the following:

$$y_i = \mu + \sum_{j=1}^{J} \beta_j x_{ij} + \varepsilon_i,$$

(1.1)

with intercept $\mu$, SNP effects $\beta_j$ and residual term $\varepsilon_i$, it is evident that this is an overidentified model with fewer individuals $I$ than SNPs $J$ (Benjamin et al., 2012a). For this purpose, two basic approaches have been developed to deal with the overidentification problem. Hypothesis-driven methods such as the candidate gene approach do not consider all $J$ SNPs, and hypothesis-free methods such as the Genome-Wide Association Study consider all $J$ SNPs but not in one model. The candidate gene approach consists of testing a subset of genetic variants for association with the outcome of interest. These genetic variants are selected based on what is known or believed about their biological function (Benjamin et al., 2012a,b, Ebstein et al., 2010). This approach resembles the classic method of justifying and then testing a hypothesis. A clear advantage of this approach is that the interpretation of revealed significant relationships is relatively straightforward. However, it turns out that findings of candidate gene studies often fail in replications of the experiment (Benjamin et al., 2012a,b, Ioannidis, 2005, Rietveld et al., 2014a). In principle, a theoretical framework guides empirical research in reducing the number of hypotheses being tested. However, the analytical rigor that a theory-guided approach provides is not helpful in the context of behavioural genetics because it is difficult to reduce the number of plausible hypotheses purely on theoretical grounds. For instance, 70% of all genes (approximately 14,000) are expressed in the brain (Ramsköld et al., 2009), and for many of these genes (and hence the SNPs within these genes), a seemingly plausible relation between genes and behaviour could be hypothesized ex ante. As a matter of fact, in 2012, the editor of the leading field
journal *Behaviour Genetics* issued an editorial policy on candidate gene studies of behavioural traits that reads “The literature on candidate gene associations is full of reports that have not stood up to rigorous replication” and that went on to say “…it now seems likely that many of the published findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge” (Hewitt, 2012). This editorial policy outlines the strict quality criteria that candidate gene studies must meet to be considered for publication. Most importantly, the editors stressed the importance of sufficient statistical power in genetic discovery studies (Hewitt, 2012). An alternative to the candidate gene study is the GWAS. A GWAS is a hypothesis-free approach to genetic discovery because no prior selection is made on the set of SNPs used in the analysis. To deal with the overidentification problem, a GWAS runs a single regression for every SNP. In a GWAS, a simple regression is performed according to the following simple regression model:

\[ y_i = \mu + x_{ij}b_j + \varepsilon_i, \]  

(1.2)

where \( y_i \) is the value of the phenotype for individual \( i \), \( \mu \) is the intercept, and \( x_{ij} \) is an indicator variable that takes values 0, 1 or 2 if the genotype of individual \( i \) at SNP \( j \) is aa, Aa or AA, respectively. The corresponding allelic effect of SNP \( j \) for each trait is \( b_j \). Hence, millions of regressions are performed in a GWAS. An advantage of the hypothesis-free study design of a GWAS is that it makes the need to correct for multiple testing transparent. If the null hypothesis of no association is true for all these millions of SNPs, one still finds a \( p \)-value < 0.05 for 5% of the SNPs. Therefore, in a GWAS, the significance threshold is set to \( 0.05/1,000,000 = 5 \times 10^{-8} \) (“genome-wide significance”) because of the approximately 1 million independent SNPs in the human genome (adjacent SNPs in the genome are often inherited together). A clear disadvantage of this approach is that GWASs may prioritize SNPs for which the biological function is yet unknown or unclear.

1.2.2 Part I: Mendelian randomization

In this part of the thesis, I investigate how we can use genetic variants identified in a GWAS as being associated with a particular outcome as instrumental variables in empirical models. Because of the genetic nature of these instrumental variables, this technique is called Mendelian randomization (MR). This promising method for making causal inferences is already very often used in medicine and is gaining much traction in economics, for example, to estimate the causal effects of health conditions on healthcare cost (Dixon et al., 2016) and to analyse the
relationship between education and obesity (Böckerman et al., 2017)).

The main rationale of the MR method is as follows. Consider a model for $J$ genetic variants $G_1, G_2, \ldots, G_J$ that are independent in their distributions, a modifiable exposure $X$, an outcome variable $Y$, and a (unobserved) confounder $U$ (a variable that influences both our exposure $X$ and our outcome variable $Y$, as previously described by Palmer et al. (2008) and Bowden et al. (2017b)). I assume that all relationships between the variables are linear and homogeneous without effect modification, meaning that the same causal effect is estimated by any valid instrumental variable (IV) (Didelez and Sheehan, 2007). A visual representation of the model is shown in Figure 1.1.

![Illustrative diagram showing the model assumed for genetic variant $G_j$, with effect $\phi_j$ on the unobserved confounder $U$, effect $\gamma_j$ on exposure $X$, and direct effect $\alpha_j$ on outcome $Y$. The causal effect of the exposure on the outcome is $\theta$. Dotted lines represent possible ways the instrumental variable assumptions could be violated.](image)

The summary-level MR methods considered in this thesis work take the association between a genetic variant and the exposure (beta-coefficient $\hat{\beta}_{Xj}$ and standard error $\sigma_{Xj}$) and the association between the genetic variants and the outcome (beta-coefficient $\hat{\beta}_{Yj}$ and standard error $\sigma_{Yj}$) for each variant $G_j$ as established in a GWAS as input. The causal effect of the exposure on the outcome can be estimated using a single genetic variant $G_j$ by the following ratio method:

$$\hat{\theta}_{Rj} = \frac{\hat{\beta}_{Yj}}{\hat{\beta}_{Xj}}.$$  

(1.3)
The ratio estimate $\hat{\theta}_{Rj}$ is a consistent estimate of the causal effect if variant $G_j$ satisfies the IV assumptions (Didelez and Sheehan, 2007). In case of multiple genetic variants, one can obtain an efficient estimator by taking a weighted combination of the ratio estimates.

However, there are some considerable doubts about whether the assumptions of instrumental variable regression hold in Mendelian randomization studies. In the first chapter of this part (Chapter 2), I study the MR-Egger method that has been developed to verify the robustness of MR estimates. In the second chapter of this part (Chapter 3), I compare nine robust Mendelian randomization methods from a theoretic and empirical viewpoint. In this chapter, I use a simulation study to compare the performance of the various methods.

**Chapter 2: A note on the use of Egger regression in Mendelian randomization studies**

Compared to most studies in economics, where we have only one or a few instruments, we can have dozens or hundreds of instruments when we use SNPs as instruments. This may strengthen the power to detect causal effects. However, given that we do not fully understand the exact function of all these SNPs, there is doubt if all our instruments satisfy the required conditions to be valid. Hence, several robust methods have been developed. One of the robust methods is MR-Egger regression, that tries to adjust for the average “pleiotropic” effect. Pleiotropy means that a genetic variant influences the outcome not only through the exposure and thus, the exclusion restriction of IV regression is violated. By including an intercept in the regression of the first stage effects on the second stage effects, MR-Egger aims to control for possible pleiotropy. MR-Egger is often used as a robustness check in Mendelian randomization studies. In this chapter, I inspect the underlying assumptions for this method and the merits of using this method as a robustness check.

**Chapter 3: A comparison of robust Mendelian randomization methods using summary data**

In the third chapter, I compare nine robust Mendelian randomization methods that rely on summary data. The methods I investigate are the weighted median method, the mode-based estimator, MR-PRESSO, MR-Robust, MR-Lasso, MR-Egger, the contamination mixture, MR-Mix, and MR-RAPS. I compare the methods regarding their theoretical properties and inspect their performance in an extensive simulation model in which some of the instrumental variable assumptions are not met. I also compare the robust methods in an empirical example considering the effect of BMI on coronary artery disease risk.
1.2.3 Part II: Polygenic risk scores

This part of my thesis concerns the use of polygenic risk scores in empirical models. In the fourth chapter, I use polygenic risk scores to describe pathways from genes to entrepreneurship. In the fifth chapter, I use polygenic risk scores as a source of heterogeneity in the response to changes in smoking excise taxes. Below, I will give a short explanation of how one can construct these polygenic risk scores.

GWASs have made it clear that individual SNPs typically explain less than 0.02% of the variance in a behavioural outcome (Chabris et al., 2015). Hence, individually, genetic variants are practically useless for inclusion in empirical studies. However, the tiny explanatory power of individual genetic variants has encouraged researchers to develop methods that combine individual genetic variants into so-called polygenic risk scores with larger explanatory power. A polygenic risk score is a weighted sum of SNPs and is constructed as follows:

\[
P_{GS_i} = \sum_{j=1}^{J} \beta_j x_{ij},
\]

where \(P_{GS_i}\) is the value for the polygenic risk score for individual \(i\), \(\beta_j\) is the regression coefficient of SNP \(j\) from the GWAS, and \(x_{ij}\) is the genotype of individual \(i\) for SNP \(j\) (coded as 0, 1 or 2). This simple approach has been shown to be effective in the out-of-sample prediction of behavioural outcomes. For example, Rietveld et al. (2013) found only three SNPs significantly associated with educational attainment at the genome-wide significance level. Each SNP explained approximately 0.02% of the variance in educational attainment. However, the polygenic risk score based on all SNPs (including the non-significant ones) explained approximately 2.5% of the variance. This percentage increases with the sample size of the GWAS (Dudbridge, 2013). For example, the most recent polygenic risk score for educational attainment now explains 9.4% of the variance (Lee et al., 2018).

Chapter 4: A decade of research on the genetics of entrepreneurship: a review and view ahead

Entrepreneurship has been shown to be heritable. However, there have not been any robust associations found between SNPs and entrepreneurship despite several attempts. Through an extensive literature review I try to answer why we have not yet found any associations. Given that there has been no significant association found at this time, I suggest taking an alternative approach to linking genes to entrepreneurship. Namely, I argue that one should use polygenic risk
scores for a range of traits to investigate the genetic background of entrepreneurship. In an empirical application using data from the US Health and Retirement Study, I explain entrepreneurship using the polygenic risk scores for traits in the mental health domain. Furthermore, I look ahead at how genetics can contribute to the field of entrepreneurship.

Chapter 5: Does the genetic predisposition to smoking moderate the response to tobacco excise taxes?

Tobacco use is one of the leading causes of preventable death. Over the past decades, public policies have been effective in reducing the prevalence of smoking. One of the most often used policy instruments to reduce tobacco consumption is the imposition of excise taxes, as they are easy to implement. However, over the past 20 years, the decrease in tobacco consumption has stalled. Some individuals do not seem to alter their behaviour despite these increases in excise taxes. In this chapter, I show that polygenic risk scores are predictive for smoking behaviour (measured as smoking initiation and smoking intensity). Next, I identify whether there can be a difference in response to increased excise taxes based on these polygenic risk scores.

1.2.4 Part III: Multivariate GREML

In this part of my thesis, I develop a multivariate extension of genome-based restricted maximum likelihood (GREML), which is a method for variance component estimation. With this method, one can estimate what fraction of a trait is heritable and to what extent different traits are genetically related. In addition, I implement the method such that it allows one to perform the estimations in a much more computationally efficient manner than does the current benchmark. Below, I will give the main idea behind variance component estimation. If all genetic variants influencing a trait are known, they can be added into one single model for the trait of interest $y_i$ as follows:

$$y_i = \mu + g_i + \varepsilon_i$$

and

$$g_i = \sum_{k=1}^{m} s_{ik} u_k,$$

where $\mu$ is the intercept, $g_i$ is the total genetic contribution of all SNPs for individual $i$, $m$ is the total number of causal genetic variants, $u_k$ is the scaled effect of causal SNP $k$, and $s_{ik}$ is standardized genotype of individual $i$ at SNP $k$ (that is, $s_{ik} = x_{ik} - 2f_k \sqrt{2f_k (1-f_k)}$ with $f_k$ the frequency of the minor allele at locus $k$). Observe that (1.5) can be rewritten in matrix notation as $\mathbf{y} = \mu \mathbf{1} + \mathbf{g} + \mathbf{\varepsilon}$

10
and \( g = S u \). Now, the variance of \( Y \) can be partitioned as follows:

\[

\text{Var}(y) = \sigma^2_u SS^\top + \sigma^2_e I = \frac{\sigma^2_g}{m} SS^\top + \sigma^2_e I = \sigma^2_g G + \sigma^2_e I, \tag{1.6}
\]

where \( G (= m^{-1}SS^\top) \) is the genetic relationship matrix between pairs of individuals at causal loci. With the equation above, the estimate for SNP-based heritability \( h^2 \) of a trait is \( \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e} \). This model can be extended to a multivariate model, such that the model can estimate heritability and genetic relatedness among traits simultaneously.

Chapter 6: Multivariate GREML reveals shared genetic architecture between brain regions and behavioural traits

To estimate the genetic correlations across multiple traits (>2) using genome-wide data, one typically applies bivariate methods repeatedly. This pairwise bivariate approach has important disadvantages. First, combining pairwise bivariate correlation estimates into a cross-trait correlation matrix does not necessarily yield a positive (semi)-definite correlation matrix. Second, the pairwise bivariate approach does not yield a complete sampling correlation matrix for all parameters of interest. Third, the current bivariate approaches fail to exploit large computational efficiency gains that are possible within a multivariate context. In this study, I propose a novel multivariate method that addresses these three issues under a design with balanced data. The model is parametrized such that the resulting correlation matrix is always positive (semi-)definite. To ensure numerical stability of the method, a quasi-Newton algorithm is used to optimize the log-likelihood. In this chapter, I use the developed method to analyse the genetic structure of the brain using the UK Biobank imaging data. Moreover, I investigate genetic correlations with several behavioural outcomes.

1.3 Research questions and results

The five chapters in this thesis answer six research questions. In the current section, I describe these research questions and present the main results.

How appropriate is MR-Egger analysis as a robustness check in MR studies? (Chapter 2)

Throughout this chapter, I analyse the MR-Egger method from both a theoretical and empirical perspective to answer my research question. The MR-Egger regression relies on the assumption that the strength of the gene-exposure association
(the first stage) is uncorrelated with the strength of the pleiotropic effects across instruments (this is called the instrument strength independent of direct effect (InSIDE) assumption). Since in practice one cannot test whether the InSIDE assumption (the key assumption for MR-Egger that is different from the exclusion restriction used by IVW) holds, one cannot judge which of the two estimates is closer to reality. Hence, using this method as a sole robustness check is prone to unwarranted conclusions. Of course, MR-Egger can be used as a sensitivity check but should be treated as a fallible check in tandem with other analyses to assess the plausibility of the causal effect estimate (Burgess and Thompson, 2017).

What robust Mendelian randomization methods work best when some of the instrumental variable assumptions are violated? (Chapter 3)

In this chapter, I compare nine robust methods for Mendelian randomization based on summary data that can be implemented using standard statistical software. The methods are reviewed in three different ways: by reviewing the theoretical properties, in an extensive simulation study and in an empirical example. From a theoretical point of view, these methods have different consistency assumptions. The three main strategies used to come up with a consistent estimator are to use a consensus approach (weighted median and mode-based estimator), an outlier removal/downweigh approach (MR-PRESSO, MR-Robust, and MR-Lasso), and the modelling approach (MR-Egger, contamination mixture, MR-Mix, and MR-RAPS). Each of these three approaches has its merits depending on the type of violations there may be. In the simulation study, I vary the type of violation and the number of genetic variants used per method. With up to 30% of the instruments being invalid, most methods are able to still come up with correct type 1 errors. Once I increase the percentage of invalid instruments, most methods start to break down. Overall, judging by the mean squared error, the contamination mixture method performs the best. The other methods perform better according to different metrics. In the empirical example, I estimate the effect of body mass index on coronary artery disease risk. In total, I use 94 genome-wide significant variants. In general, most variants suggest a harmful effect of increased BMI on CAD risk. However, there is apparent heterogeneity in the IV estimates from the different genetic variants. All methods, except the MR-Mix method, agree that there is a positive effect of BMI on coronary artery disease risk. Nevertheless, the methods that detect outliers vary in terms of how lenient or strict they are in identifying outliers. Taking this all into consideration, I encourage researchers to use robust methods from all categories (consensus approach, outlier removal/downweigh approach, and the modelling approach) in their empirical applications. For example, an investigator
Why has the identification of robust associations between genetic variants and entrepreneurship been unsuccessful in the last decade? (Chapter 4)

Despite several attempts over the last decade, no significant robust association between a genetic variant and entrepreneurship has been found. Despite working with the required sample size as calculated by Koellinger et al. (2010), Van der Loos et al. (2013) were unable to find any significant associations. The past years of research in behavioural genetics have shown that a single SNP typically explains less than 0.02% of the variance (Chabris et al., 2015, Rietveld et al., 2014a). In hindsight, the effect size estimates used in the power analyses by Koellinger et al. (2010) were too large. This is the reason why Van der Loos et al. (2013) have not been able to find any robust associations. This lack of power due to an insufficient sample size has been the reason why we have not been able to find any robust associations yet. A back-of-the-envelope calculation using the individual variance explained per SNP of 0.02% obtained from (Chabris et al., 2015, Rietveld et al., 2014a) suggests that a sample size of at least 200,000 individuals is required to identify a SNP at a genome-wide significance level with 80% power. Despite the rapidly increasing sample sizes (of mostly medical cohorts), the currently available sample sizes for entrepreneurship in genetic cohorts are still insufficient. This is due to measures for entrepreneurship are often not included in these datasets. Smaller datasets, such as the US Health and Retirement Study, and the English Longitudinal study of Ageing, do include entrepreneurship variables; however, these are still not of sufficient size at the moment to do a GWAS that is sufficiently powered.
Would the identification of associations between genetic variants and entrepreneurship help to advance the field of entrepreneurship research? (Chapter 4)

Benjamin et al. (2012a) outlined four different motives for studying the intersection of genetics and economics (and entrepreneurship as well). Section 1.1 already discusses these promises in detail. First, studies using directly observed genes may reveal the genetic pathways and mechanisms underlying behaviour and may lead to a more complete understanding of entrepreneurial behaviour. Second, these studies have the potential to provide measures for constructs that are difficult to measure empirically. Third, based on someone’s genetic profile, interventions may be channelled. In this vein, entrepreneurship scholars argue that the prediction of entrepreneurial behaviour using genetic data could have practical applications in business and for individual decision-making (Nicolaou et al., 2008a, Nicolaou and Shane, 2010, Shane, 2010). Fourth, genes can be used to enrich otherwise non-genetic models. For example, the inclusion of control variables for genetic endowments may absorb the residual variance in regression models or experimental settings and allow for stronger statistical inference (DiPrete et al., 2018a, Rietveld and Webbink, 2016). In some instances, it could also be possible to infer causal relationships in observational data by using genes as instrumental variables (Van Kippersluis and Rietveld, 2018, Von Hinke et al., 2016). Hence, the use of genes may be instrumental to obtain a better understanding the effects of environmental factors. Regarding the first two promises, I have seen that for behavioural outcomes (such as entrepreneurship), one should not expect values of $R^2$ in excess of 0.02% for individual SNPs. Hence, it is unlikely that such a SNP will provide much information about the mechanisms underlying entrepreneurial behaviour. In contrast to focusing on individual genetic variants, there are good arguments for shifting the attention to polygenic risk scores that summarize the contribution of several genetic variants to a trait.

Regarding the third and fourth promises (the use of genetic information to predict individual behaviour and to enrich otherwise nongenetic models), the current state of the behavioural genetics literature as well as the analyses presented in Chapter 4 make clear that the added value of genetics for entrepreneurship scholars should be thought of in terms of enriching population-level models rather than improving individual-level prediction (Morris et al., 2019). Van der Loos et al. (2013) show that all SNPs together may explain up to 25% of the differences in entrepreneurial behaviour between individuals. Even if one is able to realize this prediction $R^2$, the likelihood of the misclassification of individuals into occupational groups remains great. Hence, early speculations about the use of molecular genetic data for understanding and predicting entrepreneur-
ship (Shane, 2010) remain premature, at a minimum. Even though it may be useful to capture some of the (otherwise residual) variance in polygenic risk scores, the gene-based prediction of individual entrepreneurial behaviour will remain of limited value for individuals and entities such as governments and banks. Nevertheless, capturing residual variance in polygenic risk scores may improve the understanding of the effects of environmental factors. In so-called gene-by-environment (“GxE”) studies (Keller, 2014, Thompson, 2017), polygenic risk scores could also be used to investigate how entrepreneurship results from the interplay between genetic endowments and environmental factors.

**Does the genetic predisposition to smoking moderate the response to tobacco excise taxes? (Chapter 5)**

To answer this research question, I use a restricted version of the US Health and Retirement Study longitudinal data (1992-2014) that includes the postal codes of individuals. I link the individual’s postal codes to the Tax Burden on Tobacco dataset from Orzechowski and Walker (2016) to obtain yearly state-level information about levied tobacco excise taxes. I interact polygenic risk scores for smoking initiation and smoking intensity with state excise tax rates on tobacco. My analyses show that someone’s genetic propensity to smoking moderates the effect of tobacco excise taxes on smoking behaviour, but only along the extensive margin (smoking vs. not smoking). The results along the intensive margin (the amount of tobacco consumed) are inconclusive. Even in a restricted sample of smokers only, I am unable to find significant results along the intensive margin. These findings suggest that excise taxes are an effective method to reduce tobacco usage, even among the group with a high genetic predisposition towards smoking. Even more, those with a high genetic predisposition to smoking respond most strongly to changes in tobacco excise taxes.

**Can a multivariate extension of GREML be formulated such (i) that the resulting estimates yield a valid genetic and environmental covariance matrix (i.e., positive (semi-)definite) and (ii) that the procedure is computationally feasible? (Chapter 6)**

In this chapter, I develop a multivariate extension of GREML. Based on a Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, this method uses an iterative procedure to obtain unbiased estimates of the genetic and environmental variance-covariance matrix for balanced data of $P$ traits observed for $N$ individuals. By changing the parameters over which I optimize to a Cholesky decomposition, I ensure that the variance estimates are positive (semi-)definite. To ensure that the model is computationally feasible, I rewrite the log-likelihood.
and the gradient in terms of the eigen decomposition of an $N \times N$ GRM and transformations of $P \times P$ matrices of parameters. Using this transformation, I am able to reduce the complexity of the problem from the order $O(NP^6)$ to an order of $O(NP^5)$. In an empirical application using $P = 86$ traits from $N = 14,341$ unrelated individuals from the UK Biobank imaging study, I show that the current implementation of our method is computationally feasible. Our method reveals distinct clusters of genetic correlations between brain areas, as well as genetic correlations between brain regions and behavioural traits. The findings fit with how the neuroscience literature considers the development of the brain taking place.

1.4 Conclusion and implications

In this section, I elaborate how the chapters in this thesis contribute to the promises of genoeconomics discussed in Section 1.1. I discuss how the findings of this thesis help the emerging field of genoeconomics and the general field of economics in a broader context. Next, to this, I explain how the methodological contributions of this thesis will eventually help us in empirical applications by using genes as control variables and/or instrumental variables. I also explain how genes can be used to measure predispositions to (mental) diseases and economic outcomes, which may result in targeted interventions to prevent undesired outcomes. Furthermore, I look ahead by discussing directions for future research on the intersection of genes and economics. In chapters 2 and 3, Mendelian randomization methods are analysed and compared to give guidance on what set of robust methods researchers should use to assess the reliability of Mendelian randomization estimates. In the future, once the number of large-scale genetic association studies on economic choices and outcomes has further increased, this review of methods can be used to inform causal inference in economics. There has been much debate about whether genes meet all the requirements to be a valid instrument. This debate is mostly about the validity of the exclusion restriction in empirical applications (Taylor et al., 2014). With the methods studied in these chapters, researchers will be able to make robust interferences even if some genes violate the IV assumptions. These methods will be very useful in the near future, as randomized clinical trials are often difficult or unethical to perform in economics. With the increasing number of genetic variants that are linked to socio-economically relevant characteristics, I believe Mendelian randomization studies will gain even more traction. Nevertheless, there remain some potential sources of bias that robust methods are unable to solve (such as selection bias, population stratification, dynastic effects and assortative mating), but they can be solved by within-family Mendelian randomization studies, as recently suggested.
by Davies et al. (2019). Due to the increased availability of data from related individuals in large cohort studies, this approach will lead to new opportunities to overcome potential sources of biases that may currently hamper Mendelian randomization studies. Chapters 4 and 5 show that polygenic risk scores may help to explain economic choices and outcomes at the population level. It has been known for decades that these choices and outcomes are heritable, but only since the last few years, due to the large amount of publicly available GWAS results, has it been possible to capture these genetic effects with polygenic risk scores. The results in this thesis offer a new way to explain heterogeneity in entrepreneurship and smoking behaviours. However, for individual prediction, the misclassification rate is still very high, and polygenic risk score prediction does not seem promising. Given that polygenic risk scores are only predictive at the population level, considering the use of genes for targeted policy interventions is premature. If we will ever be able to predict sufficiently well at individual level using genetic information (which I doubt), it could not only lead to positive interventions but also to genetic discrimination. Therefore, I believe it is of utmost importance to have ethical discussions about the desirability of individual-level predictions using genes. As such, I consider the current provision of individual genetic prediction profiles by companies such as Leadership Consultants and Goldmen Genetics as premature and threatening. In chapter 6, I develop a method that is able to estimate the genetic correlation between economic choices and outcomes for a large number of traits simultaneously. As soon as a large dataset with a sufficient number of economic choices and outcomes becomes available, this method is available to reveal whether there is genetic overlap between certain traits. The results obtained with this method may help to understand the preferences and decisions of individuals in a more comprehensive manner. Using heritability estimates and genetic correlation for informing policy is not straightforward, as outlined by Goldberger (1979) and Manski (2011). Nevertheless, (co-)heritability estimates are descriptive facts that constrain the set of plausible theories regarding heterogeneity in preferences and abilities. Relatedly, significant heritability estimates for economic outcomes indicate that genetic endowment can bias the effect of environmental variables on outcomes of interest if not adequately controlled for. An example would be that parental genetic endowments influence not only the child’s genotype (which leads to differences in behaviour) but also influences the child’s environmental exposures (through the pathway of the behaviour of the parents). Kong et al. (2018) have shown that this type of “genetic nurture” indeed exists.
1.5 Individual Contributions and Publication Status per Chapter

This section discusses my contributions to each chapter in the present thesis. The current chapter (1), I wrote independently, although I received valuable feedback on drafts of it from my supervisors. The research idea of Chapter 2 came from my daily supervisor, Dr. Rietveld. The first draft of this chapter was written by Dr. Rietveld and myself. I was responsible for the data analysis. Professors Groenen and Thurik had a supervisory role and were responsible for the final checks. During the 2017 Mendelian Randomization Conference in Bristol, I received the reserve poster prize for my presentation of this chapter.

After discussions with Dr. Rietveld about robust Mendelian randomization methods, I came up with the idea for Chapter 3 myself. In the Mendelian Randomization Conference of 2017, the development of new (robust) Mendelian randomization methods was flourishing, and I considered it to be of importance for practical users to have an overview of the different robust methods available. For this project, I decided to team up with Dr. Burgess, who is an expert in Mendelian randomization. Dr. Burgess was happy to host me for a period of three months at the MRC Biostatistics Unit in Cambridge. For this chapter, we came up with a simulation setup together. Thereafter, I performed the extensive simulation study, conducted the empirical analyses, and wrote the first draft of the chapter. Afterwards, Dr. Burgess edited the draft manuscript, and we alternately improved and changed parts of it.

Chapter 4 resulted from intense discussions with Professor Thurik and Dr. Rietveld. Given that no new sufficiently large genetic datasets that include entrepreneurship-related variables had become available in recent years, not much progress had been made regarding the genetic analysis of entrepreneurship since the first GWAS on self-employment in 2013. Dr. Rietveld suggested that we could use the proxy-phenotype approach in the US Health and Retirement Study to circumvent this barrier. I performed the data analysis and was responsible for writing the first draft of this chapter. Afterwards, Dr. Rietveld, Prof. Thurik and I edited the manuscript in several rounds.

I came up with the research idea for Chapter 5 myself. Dr. Rietveld helped me with the data acquisition and the positioning of the paper within the literature. I wrote the first draft of this chapter. Thereafter, Dr. Rietveld and I alternately improved and changed parts of it. The original idea for Chapter 6 came from Dr. de Vlaming. Together with Prof. Groenen, he performed the first derivations of the model. These derivations constituted a chapter in his PhD thesis, which he defended in 2017. At the suggestion of Dr. Rietveld, I joined the research for this project. I started by implementing the method in MATLAB. Thereafter, I devoted
considerable time to fine-tuning the optimization algorithm. I also performed preliminary empirical analyses of the US Health and Retirement Study. Dr. Koellinger was responsible for constructing the UK Biobank brain phenotypes. Together with Dr. de Vlaming, I performed the quality control and empirical analyses using the UK Biobank data. Dr. Jansen was responsible for interpreting the findings in light of the neuroscience literature. I wrote the first draft of this chapter, and together with Prof. Groenen and Dr. Rietveld, I rewrote parts of the initial draft. For the new version of the chapter (not included in this thesis), which is based on a larger sample resulting from a new release of brain imaging data in UK Biobank, I performed the empirical analysis alone. The publication status of each chapter is shown in Table 1.1. This table also shows where I have presented the projects throughout my PhD trajectory.

**Table 1.1 – Publication status of the chapters.**

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Reference</th>
<th>Presentations</th>
<th>Publication status</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>A note on the use of Egger regression in Mendelian randomization studies</td>
<td>Slob, Groenen, Thurik &amp; Rietveld</td>
<td>Bristol (2017)</td>
<td>Published in <em>International Journal of Epidemiology</em></td>
</tr>
<tr>
<td>4</td>
<td>A decade of research on the genetics of entrepreneurship: a review and view ahead</td>
<td>Rietveld, Slob &amp; Thurik</td>
<td></td>
<td>Published in <em>Small Business Economics</em></td>
</tr>
<tr>
<td>5</td>
<td>Does the genetic predisposition to smoking moderate the response to tobacco excise taxes?</td>
<td>Slob &amp; Rietveld</td>
<td>Rotterdam (2019)</td>
<td>Manuscript submitted</td>
</tr>
</tbody>
</table>
I

Mendelian Randomization
A note on the use of Egger regression in Mendelian randomization studies

Eric A.W. Slob, Patrick J.F. Groenen, A. Roy Thurik, Cornelius A. Rietveld

Abstract

A large number of epidemiological studies uses genetic variants as instrumental variables to infer causal relationships. Given that these methods rely on strong assumptions that are not testable, MR-Egger regression has been proposed to correct for pleiotropic effects. In this study, we compare the bias between MR-Egger and the IVW estimate, and look at two empirical examples where we inspect the ‘InSIDE’ assumption. Our findings suggest that the use of MR-Egger as robustness check of IVW estimates is prone to unwarranted conclusions about the causal effect estimate, because in empirical settings the assumption that InSIDE holds is often questionable.

This chapter is based on Slob et al. (2017).
A large number of epidemiological studies uses genetic variants as instrumental variables to infer causal relationships (Smith and Ebrahim, 2003, Burgess et al., 2015). For a genetic variant to be a valid instrument in these so-called Mendelian randomization (MR) studies, three assumptions need to hold: (i) The genetic variant is associated with the exposure of interest (relevance assumption); (ii) The genetic variants should be independent of all confounders (independence assumption); (iii) The genetic variants only effects the outcome through the exposure of interest (exclusion restriction). Without specific knowledge about the biological mechanisms affected by genetic variants, it is virtually impossible to prove that the exclusion restriction holds for a specific genetic variant (Glymour et al., 2012). For example, genetic variants may have pleiotropic effects on both the exposure and the outcome through different biological pathways (Solovieff et al., 2013).

Several methods and techniques have been developed to tackle the possible problem of pleiotropy in Mendelian randomization studies. In this journal, Bowden and colleagues recently proposed to use Egger regression to correct for pleiotropic effects of genetics variants (Bowden et al., 2015). Using simulations they show that MR-Egger provides unbiased estimates of causal effects if pleiotropy is balanced (i.e., the direct effects are uniformly distributed around zero). Also in case of directional pleiotropy (i.e., the direct effects are uniformly distribution around a non-zero value) MR-Egger performs well, but only as long as the instrument-exposure and instrument-outcome associations are independent. This so-called “InSIDE” assumption is a relaxation of the exclusion restriction. MR-Egger produces biased results if the InSIDE assumption does not hold, in particular in a one-sample setting in which values for the instrument-exposure association and the instrument-outcome association are obtained in the same sample. Bowden and colleagues acknowledge this in their appendix: “We conclude that IV analysis with weak instruments in a one-sample setting is troublesome, and that these difficulties are not resolved by the application of MR-Egger regression”.

Nevertheless, MR-Egger is currently often used in epidemiological studies as a robustness check on results obtained with regular Mendelian randomization analysis without proper discussion whether the InSIDE assumption holds. For example, a recent MR study states: “We used a second method of Mendelian randomisation, the Egger method, as a sensitivity analysis if the instrumental variables test result was noteworthy. This method is more robust to potential violations of the standard instrumental variable assumptions. (…) so this method is less susceptible to confounding from potentially pleiotropic variants (…)” (Tyrrell
et al., 2016). This is an incorrect use of MR-Egger, and hence the conclusions about the robustness of the findings are unwarranted in this study.

Another recent study derived the exact bias of the IVW and MR-Egger estimators (Bowden et al., 2017a). This study recognizes that in some settings where the InSIDE assumption does not hold, the bias of the MR-Egger estimator can be larger than the bias of the regular Inverse-Variance Weighting (IVW) estimator. However, no practical conclusions are drawn from this finding. For the purpose of the present note, we draw the following conclusion: We conclude that the use of MR-Egger as robustness check of IVW estimates is prone to unwarranted conclusions about the causal effect estimate, because in empirical settings the assumption that InSIDE holds is often questionable. We will illustrate this conclusions by showing that in two illustrative analyses by Bowden and colleagues (Bowden et al., 2015, 2017a) the InSIDE assumption does not seem to hold, and that it is not possible in these examples to evaluate whether the MR-Egger is less biased than the IVW estimator.

2.2 Methods

Following Bowden and colleagues, we deal with a Mendelian randomization study with $N$ participants (Bowden et al., 2015). For each participant $i$, we measure $J$ genetic variants ($G_{i1}, \ldots, G_{iJ}$), a modifiable exposure ($X_i$), and an outcome ($Y_i$). The genetic variants are assumed to take values 0, 1, or 2, representing the number of alleles of a biallelic single nucleotide polymorphism (SNP). The confounder $U_i$ is a function of the genetic variants and an independent error term ($\varepsilon_{iU}$), but is assumed to be unknown. The exposure $X_i$ is a linear function of the genetic variants, the confounder and an independent error term ($\varepsilon_{iX}$). The outcome $Y_i$ is a linear function of the genetic variants, the exposure, the confounders and an independent error term ($\varepsilon_{iY}$). The causal effect of the exposure on the outcome is $\beta$. $\gamma_j$ represents the effect of the instrument on the exposure. The coefficients $\alpha_j$ for each genetic variant $j$ represent the direct effects of the genetic variants on the outcome that are not mediated by the exposure. The total effect of each variant on the outcome comprises the direct effect ($\alpha_j$), and the indirect effects via the exposure ($\beta \gamma_j$) and the confounder ($\phi_j$). The model described above can be written as:

$$U_i = \sum_{j=1}^{J} \phi_j G_{ij} + \varepsilon_{iU} \quad (2.1)$$

$$X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \varepsilon_{iX} \quad (2.2)$$
\[ Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \beta X_i + U_i + \epsilon_i^Y . \]  

(2.3)

We denote the estimate for the instrument-exposure association by \( \hat{\gamma}_j \) and the estimate for the instrument-outcome association by \( \hat{\Gamma}_j \). With Inverse Variance Weighting (IVW), an estimate for the causal effect \( \hat{\beta}_j \) is obtained by dividing \( \hat{\Gamma}_j \) by \( \hat{\gamma}_j \). This ratio equals \( \beta + (\alpha_j + \phi_j)/(\gamma_j + \phi_j) \) (derivation given in the article by Bowden and colleagues (Bowden et al., 2017a)), and hence the bias in the estimation of \( \beta \) is a function of \( \alpha_j, \phi_j, \) and \( \gamma_j \). With multiple genetic variants, the IVW estimator is a weighted average of the ratio of estimates calculated using each genetic variant in turn. In the article by Bowden and colleagues, the bias of the IVW estimator is derived and is equal to \( \sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma^{-2}_{Y_j} \alpha_j + \phi_j / \sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma^{-2}_{Y_j} \), where \( \sigma_{Y_j} \) is the standard error in the regression of the outcome on the \( j \)th genetic variant (Bowden et al., 2017a). In MR-Egger, the absolute values of \( \hat{\Gamma}_j \) are regressed on the absolute values of \( \hat{\gamma}_j \) in order to estimate \( \beta \). Furthermore, Bowden and colleagues find that the bias in the estimation of \( \beta \) with MR-Egger equals \( (\sigma_\alpha \rho_{\alpha,Y} + (1 + \beta)\sigma_{\phi} \rho_{\phi,Y})/\sigma_Y \), where \( \sigma \) denotes the standard deviation of a parameter and \( \rho \) the correlation (Bowden et al., 2017a). Hence, in MR-Egger the bias is a function of \( \sigma_{\alpha}, \rho_{\alpha,Y}, \beta, \sigma_{\phi}, \rho_{\phi,Y} \) and \( \sigma_Y \) (note that MR-Egger requires \( \sigma_Y > 0 \); this is called the 'Variation in Instrument Strength' assumption by Bowden and colleagues (Bowden et al., 2017a)).

As long as the InSIDE assumptions holds, the bias in MR-Egger is zero if both the sample size and the number of instruments increase to infinity (Bowden et al., 2015). Although Bowden and colleagues point to some empirical evidence that may suggest that the InSIDE assumption holds for some traits (Bowden et al., 2016), in general the assumption is quite strong and – more importantly – very difficult to test, since \( \alpha_j \) is typically unknown. Thus, from a practical point of view, it is important to know in which settings the bias of MR-Egger is really smaller than the bias of IVW. That is, when does the following inequality hold?

\[ |\text{Bias}_{\text{MR-Egger}}| = \left| \frac{\sigma_\alpha \rho_{\alpha,Y} + (1 + \beta)\sigma_{\phi} \rho_{\phi,Y}}{\sigma_Y} \right|? \left| \frac{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma^{-2}_{Y_j} \alpha_j + \phi_j}{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma^{-2}_{Y_j}} \right| = |\text{Bias}_{\text{IVW}}| . \]  

(2.4)

Since there are so many unknown parameters in (2.4), it is hard to assess which of the two biases is the largest in a Mendelian randomization study. At first sight, the left hand side seems smaller, since the bias is mostly based upon covariances and not on real effect sizes. Yet, to show that this is not necessarily the case, we simplify by considering a model where there is no unobserved confounder. In that
2. A note on the use of Egger regression in Mendelian randomization studies

Consider a situation where we have relatively strong instruments that all have approximately similar strength, such that \( \gamma_j \sim \mathcal{N}(0.4, 0.1) \). Let there be some directional pleiotropy with an equal variance that is equal to the instrument variance, such that \( \alpha_j \sim \mathcal{N}(0.1, 0.1) \) and let it be positively correlated with \( \gamma_j \), such that \( \rho_{\alpha, \gamma} = 0.3 \). Now, the expected bias of the MR-Egger estimate is equal to \( 0.1 \times 0.3 / 0.1 = 0.3 \) and the expected bias of the IVW estimate is approximately \( 0.1 / 0.4 = 0.25 \). Hence, in this setting the bias of the MR-Egger estimate is larger than the bias of the IVW estimate.

In empirical research settings, it is hard to evaluate whether the IVW estimator is more biased than the MR-Egger estimator. For example, Bowden et al. (2015) estimate the effect of systolic blood pressure on coronary heart disease risk. With IVW the effect is estimated to be 0.054 (log odds ratio per 1 mmHg change in blood pressure), and with MR-Egger it is estimated to be 0.015 (same units). In the Appendix, we show that the approximated correlation between the first stage effects \( \gamma \) and the direct effect \( \alpha \) is \(-0.26\). Hence, the InSIDE assumption is violated and this makes it impossible to conclude whether the smaller effect estimate obtained with MR-Egger is due to a smaller true effect \( \beta \) or to a change in the bias part of the MR-Egger estimate. In another study, Bowden and colleagues analyze the causal role of plasma urate concentration on coronary heart disease risk (Bowden et al., 2017a). In the appendix, we show in this model the approximated correlation between the first stage effects \( \gamma \) and the direct effects \( \alpha \) is even \(-0.35\). Hence, again it is unclear whether the IVW or the MR-Egger estimate is closer to the true \( \beta \).

2.3 Conclusion

In this note, we showed from a practical point of view that the bias of MR-Egger estimator can be larger than the bias of IVW estimator depending on the parameters in the model. If the InSIDE assumption does not hold, it is clear that the MR-Egger procedure cannot guarantee an estimate that is less biased than the estimate obtained with IVW. The InSIDE assumption is a relaxation of the exclusion restriction, but it is still a strong assumption in itself. From a practical point of view, this makes it almost impossible in empirical settings to judge whether the IVW or MR-Egger estimator is closer to the real value.
of the causal effect, because the validity of the InSIDE assumption cannot be tested without knowing the true causal effect. Hence, we conclude that the use of MR-Egger as sole robustness check of IVW estimates is prone to unwarranted conclusions about the causal effect estimate. Of course, MR-Egger regression can be used as a sensitivity analysis for Mendelian randomization, but should be treated as a fallible check and in tandem with other analyses to assess the plausibility of the causal effect estimate (Burgess and Thompson, 2017). We note that in some cases, bias from violations of the InSIDE assumption can be solved by finding a specific subsample for which the first stage effect does not exist (the effect of the instrument on the exposure is zero). In such a subsample, the direct effect of a SNP can be estimated, and used to correct the causal effect estimate. A recent study in this journal shows that this strategy is able to produce unbiased estimates (Van Kippersluis and Rietveld, 2018).
2.A **Approximation of the correlation between the first stage effects and the direct effects in two examples**

Bowden and colleagues analyse the causal effect on systolic blood pressure on cardiovascular diseases risk using 29 SNPs as instruments (see Table 2.1 for an overview of the SNPs) (Bowden et al., 2017b). We extracted the estimates of the first stage effects $\hat{\gamma}_j$ from Table 1 of the study by the International Consortium for Blood Pressure Genome-Wide Association Studies (International Consortium for Blood Pressure Genome-Wide Association Studies, 2011) and the estimator of the total (reduced form) effect $\hat{\Gamma}_j$ from the summary data of the CARDIoGRAM consortium (Schunkert et al., 2011). We aligned the alleles of the SNPs such that the first stage effect is positive ($\hat{\gamma}_j > 0$ for all $j$). In order to calculate $\rho_{\alpha,\gamma}$, the correlation between $\gamma$ and $\alpha$, we need to approximate $\alpha$. For this, we assume the absence of an unobserved confounder as well as that the reported $\hat{\beta}_{IVW}$ is the true causal effect (thus, $\beta = 0.054$). Using the relation $\Gamma_j = \alpha_j + \beta \gamma_j$, we calculate the direct effect with $\alpha_j = \Gamma_j - \beta \gamma_j$. This gives an approximated correlation $\rho_{\alpha,\gamma} = -0.26$.

We are convinced that this is the best way to approximate $\rho_{\alpha,\gamma}$ because it follows the standard MR model depicted in equations (2.1)-(2.3) in the main text of the present note. If we treat $\hat{\beta}_{MR-Egger}$ as the causal effect and use the same way of calculating the direct effect, we find an approximated correlation $\rho_{\alpha,\gamma} = 0.03$. This is very close to 0, since MR-Egger fits a linear model in which $\hat{\Gamma}_j = \beta_{0E} + \beta_E \hat{\gamma}_j$. Thus, the demeaned direct effects, $\alpha^*_j$ equal the “error” terms in this relation (the average pleiotropy $\bar{\alpha}$ is captured by the intercept $\beta_{0E}$, and only the demeaned effects remain). The OLS estimation procedure “attempts” to put these “residuals” orthogonal to the regressors (in our case the instrument strength, $\hat{\gamma}_j$). Hence, with MR-Egger, the correlation between the estimated $\alpha_j$ and $\gamma_j$ is very close to
0. Nevertheless, Figure 2.1 Panel A shows the approximated $\rho_{a,y}$ for a range of possible causal effects. We observe that $\rho_{a,y}$ approaches 1 when $\beta$ becomes more negative, and it approaches -1 when $\beta$ becomes more positive.

In another study, Bowden and colleagues analyze the causal role of plasma urate concentration on coronary heart disease risk (see Table 2.2 for an overview of the SNPs) (Bowden et al., 2017a). The first stage effects $\hat{\gamma}_j$ are obtained from Table S3 of the study by White et al. (2016) and the total (reduced form) effect $\hat{\Gamma}_j$ from the summary data of the CARDIoGRAM consortium (Schunkert et al., 2011). With the reported $\hat{\beta}_{IVW}$, we find an approximated correlation $\rho_{a,y} = -0.35$. When using $\hat{\beta}_{MR-Egger}$, we obtain an approximated correlation $\rho_{a,y} = -0.04$. The approximated correlation for a range of causal effects is shown in Figure 2.1 Panel B. We observe the same relationship between $\rho_{a,y}$ and $\beta$ as in Figure 2.1 Panel A.

**Figure 2.1** – *The correlation between the instrument strength and direct effect for different causal effect estimates. A: The effect of systolic blood pressure on cardiovascular diseases risk. B: The effect of plasma urate concentrate on coronary heart disease risk.*
Table 2.1 – Summary association results for 29 SNPs associated with systolic blood pressure (SNPs are ordered as in Table 1 of the study by the International Consortium for Blood Pressure Genome-Wide Association Studies (2011)).

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chromosome</th>
<th>Position</th>
<th>Ref allele</th>
<th>Other allele</th>
<th>ˆγ</th>
<th>ˆΓ</th>
<th>ˆα</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2932538</td>
<td>1</td>
<td>113,018,066</td>
<td>G</td>
<td>A</td>
<td>0.388</td>
<td>0.022801</td>
<td>0.001849</td>
</tr>
<tr>
<td>rs13082711</td>
<td>3</td>
<td>27,512,913</td>
<td>C</td>
<td>T</td>
<td>0.315</td>
<td>0.034229</td>
<td>0.017219</td>
</tr>
<tr>
<td>rs419076</td>
<td>3</td>
<td>170,583,580</td>
<td>T</td>
<td>C</td>
<td>0.409</td>
<td>0.019725</td>
<td>-0.002361</td>
</tr>
<tr>
<td>rs13107325</td>
<td>4</td>
<td>103,407,732</td>
<td>C</td>
<td>T</td>
<td>0.981</td>
<td>-0.004938</td>
<td>-0.057912</td>
</tr>
<tr>
<td>rs13139571</td>
<td>4</td>
<td>156,864,963</td>
<td>C</td>
<td>A</td>
<td>0.321</td>
<td>0.066131</td>
<td>0.048797</td>
</tr>
<tr>
<td>rs1173771</td>
<td>5</td>
<td>32,850,785</td>
<td>G</td>
<td>A</td>
<td>0.504</td>
<td>0.033135</td>
<td>0.005919</td>
</tr>
<tr>
<td>rs11953630</td>
<td>5</td>
<td>157,777,980</td>
<td>C</td>
<td>T</td>
<td>0.412</td>
<td>0.031746</td>
<td>0.009498</td>
</tr>
<tr>
<td>rs1799945</td>
<td>6</td>
<td>26,199,158</td>
<td>G</td>
<td>C</td>
<td>0.627</td>
<td>0.005740</td>
<td>-0.028118</td>
</tr>
<tr>
<td>rs805303</td>
<td>6</td>
<td>31,724,345</td>
<td>G</td>
<td>A</td>
<td>0.376</td>
<td>0.040028</td>
<td>0.019724</td>
</tr>
<tr>
<td>rs4373814</td>
<td>10</td>
<td>18,459,978</td>
<td>C</td>
<td>G</td>
<td>0.373</td>
<td>0.015007</td>
<td>-0.005135</td>
</tr>
<tr>
<td>rs932764</td>
<td>10</td>
<td>95,885,930</td>
<td>G</td>
<td>A</td>
<td>0.484</td>
<td>-0.005658</td>
<td>-0.031794</td>
</tr>
<tr>
<td>rs7129220</td>
<td>11</td>
<td>10,307,114</td>
<td>A</td>
<td>G</td>
<td>0.619</td>
<td>0.040449</td>
<td>0.007023</td>
</tr>
<tr>
<td>rs633185</td>
<td>11</td>
<td>100,098,748</td>
<td>G</td>
<td>C</td>
<td>0.565</td>
<td>0.013790</td>
<td>-0.016720</td>
</tr>
<tr>
<td>rs2521501</td>
<td>15</td>
<td>89,238,392</td>
<td>T</td>
<td>A</td>
<td>0.650</td>
<td>0.050966</td>
<td>0.015869</td>
</tr>
<tr>
<td>rs17608766</td>
<td>17</td>
<td>42,368,270</td>
<td>C</td>
<td>T</td>
<td>0.556</td>
<td>0.088020</td>
<td>0.057996</td>
</tr>
<tr>
<td>rs1327235</td>
<td>20</td>
<td>100,917,030</td>
<td>G</td>
<td>A</td>
<td>0.340</td>
<td>0.040494</td>
<td>0.022134</td>
</tr>
<tr>
<td>rs6015450</td>
<td>20</td>
<td>57,184,512</td>
<td>G</td>
<td>A</td>
<td>0.896</td>
<td>0.058069</td>
<td>0.009685</td>
</tr>
<tr>
<td>rs17367504</td>
<td>1</td>
<td>11,785,365</td>
<td>A</td>
<td>G</td>
<td>0.903</td>
<td>0.023070</td>
<td>-0.025691</td>
</tr>
<tr>
<td>rs3774372</td>
<td>3</td>
<td>41,852,418</td>
<td>C</td>
<td>T</td>
<td>0.067</td>
<td>-0.018337</td>
<td>-0.021955</td>
</tr>
<tr>
<td>rs1458038</td>
<td>4</td>
<td>81,383,747</td>
<td>T</td>
<td>C</td>
<td>0.706</td>
<td>0.034433</td>
<td>-0.003691</td>
</tr>
<tr>
<td>rs1813353</td>
<td>10</td>
<td>18,747,454</td>
<td>T</td>
<td>C</td>
<td>0.569</td>
<td>0.042546</td>
<td>0.011820</td>
</tr>
<tr>
<td>rs4590817</td>
<td>10</td>
<td>63,137,559</td>
<td>G</td>
<td>C</td>
<td>0.646</td>
<td>0.035467</td>
<td>0.000583</td>
</tr>
<tr>
<td>rs11191548</td>
<td>10</td>
<td>104,836,168</td>
<td>T</td>
<td>C</td>
<td>1.095</td>
<td>0.109050</td>
<td>0.049920</td>
</tr>
<tr>
<td>rs381815</td>
<td>11</td>
<td>16,858,844</td>
<td>T</td>
<td>C</td>
<td>0.575</td>
<td>0.048819</td>
<td>0.017769</td>
</tr>
<tr>
<td>rs17249754</td>
<td>12</td>
<td>88,584,717</td>
<td>G</td>
<td>A</td>
<td>0.928</td>
<td>-0.044766</td>
<td>-0.094878</td>
</tr>
<tr>
<td>rs3184504</td>
<td>12</td>
<td>110,368,991</td>
<td>T</td>
<td>C</td>
<td>0.598</td>
<td>0.070445</td>
<td>0.038153</td>
</tr>
<tr>
<td>rs10850411</td>
<td>12</td>
<td>113,872,179</td>
<td>T</td>
<td>C</td>
<td>0.354</td>
<td>0.021469</td>
<td>0.002353</td>
</tr>
<tr>
<td>rs1378942</td>
<td>15</td>
<td>72,864,420</td>
<td>C</td>
<td>A</td>
<td>0.613</td>
<td>0.000516</td>
<td>-0.032586</td>
</tr>
<tr>
<td>rs12940887</td>
<td>17</td>
<td>44,757,806</td>
<td>T</td>
<td>C</td>
<td>0.362</td>
<td>0.032434</td>
<td>0.012886</td>
</tr>
<tr>
<td>SNP</td>
<td>Chromosome</td>
<td>Position</td>
<td>Ref allele</td>
<td>Other allele</td>
<td>$\hat{p}$</td>
<td>$\hat{f}$</td>
<td>$\hat{a}$</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>--------------</td>
<td>------------</td>
<td>--------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>rs1471633</td>
<td>1</td>
<td>144,435,096</td>
<td>A</td>
<td>C</td>
<td>0.057</td>
<td>-0.015911</td>
<td>0.006654</td>
</tr>
<tr>
<td>rs1260326</td>
<td>2</td>
<td>27,584,444</td>
<td>T</td>
<td>C</td>
<td>0.069</td>
<td>-0.023890</td>
<td>0.012594</td>
</tr>
<tr>
<td>rs12498742</td>
<td>4</td>
<td>9,553,150</td>
<td>A</td>
<td>G</td>
<td>0.360</td>
<td>-0.011611</td>
<td>-0.047067</td>
</tr>
<tr>
<td>rs2231142</td>
<td>4</td>
<td>89,271,347</td>
<td>T</td>
<td>G</td>
<td>0.190</td>
<td>-0.034145</td>
<td>0.003241</td>
</tr>
<tr>
<td>rs675209</td>
<td>6</td>
<td>7,047,083</td>
<td>T</td>
<td>C</td>
<td>0.056</td>
<td>-0.001211</td>
<td>-0.007852</td>
</tr>
<tr>
<td>rs1165151</td>
<td>6</td>
<td>25,929,595</td>
<td>T</td>
<td>G</td>
<td>0.078</td>
<td>-0.004236</td>
<td>-0.016933</td>
</tr>
<tr>
<td>rs1171614</td>
<td>10</td>
<td>61,139,544</td>
<td>T</td>
<td>C</td>
<td>0.079</td>
<td>0.041327</td>
<td>0.028450</td>
</tr>
<tr>
<td>rs2078267</td>
<td>11</td>
<td>64,090,690</td>
<td>T</td>
<td>C</td>
<td>0.073</td>
<td>0.020768</td>
<td>0.008837</td>
</tr>
<tr>
<td>rs478067</td>
<td>11</td>
<td>64,234,639</td>
<td>A</td>
<td>G</td>
<td>0.026</td>
<td>-0.012873</td>
<td>-0.017177</td>
</tr>
<tr>
<td>rs3741414</td>
<td>12</td>
<td>56,130,316</td>
<td>T</td>
<td>C</td>
<td>0.065</td>
<td>0.001895</td>
<td>-0.008684</td>
</tr>
<tr>
<td>rs11264341</td>
<td>1</td>
<td>153,418,117</td>
<td>T</td>
<td>C</td>
<td>0.050</td>
<td>0.005672</td>
<td>-0.002478</td>
</tr>
<tr>
<td>rs17050272</td>
<td>2</td>
<td>121,022,910</td>
<td>A</td>
<td>G</td>
<td>0.035</td>
<td>0.033048</td>
<td>-0.038752</td>
</tr>
<tr>
<td>rs6770152</td>
<td>3</td>
<td>53,075,254</td>
<td>T</td>
<td>G</td>
<td>0.044</td>
<td>0.005058</td>
<td>-0.002111</td>
</tr>
<tr>
<td>rs17632159</td>
<td>5</td>
<td>72,467,238</td>
<td>C</td>
<td>G</td>
<td>0.039</td>
<td>-0.005710</td>
<td>-0.000647</td>
</tr>
<tr>
<td>rs729761</td>
<td>6</td>
<td>43,912,549</td>
<td>T</td>
<td>G</td>
<td>0.047</td>
<td>0.010178</td>
<td>0.002517</td>
</tr>
<tr>
<td>rs1178977</td>
<td>7</td>
<td>72,494,985</td>
<td>A</td>
<td>G</td>
<td>0.047</td>
<td>0.005351</td>
<td>-0.013012</td>
</tr>
<tr>
<td>rs10480300</td>
<td>7</td>
<td>151,036,938</td>
<td>T</td>
<td>C</td>
<td>0.035</td>
<td>-0.00482</td>
<td>-0.001623</td>
</tr>
<tr>
<td>rs2941484</td>
<td>8</td>
<td>76,641,323</td>
<td>T</td>
<td>C</td>
<td>0.044</td>
<td>0.008357</td>
<td>-0.015528</td>
</tr>
<tr>
<td>rs10821905</td>
<td>10</td>
<td>52,316,099</td>
<td>A</td>
<td>G</td>
<td>0.057</td>
<td>0.010269</td>
<td>-0.019560</td>
</tr>
<tr>
<td>rs642803</td>
<td>11</td>
<td>65,317,196</td>
<td>T</td>
<td>C</td>
<td>0.036</td>
<td>0.039500</td>
<td>0.033632</td>
</tr>
<tr>
<td>rs653178</td>
<td>12</td>
<td>110,492,139</td>
<td>T</td>
<td>C</td>
<td>0.035</td>
<td>0.077493</td>
<td>0.071788</td>
</tr>
<tr>
<td>rs1394125</td>
<td>15</td>
<td>73,946,038</td>
<td>A</td>
<td>G</td>
<td>0.043</td>
<td>-0.027942</td>
<td>0.020933</td>
</tr>
<tr>
<td>rs6598541</td>
<td>15</td>
<td>97,088,658</td>
<td>A</td>
<td>G</td>
<td>0.043</td>
<td>-0.035859</td>
<td>0.028837</td>
</tr>
<tr>
<td>rs7193778</td>
<td>16</td>
<td>68,121,391</td>
<td>T</td>
<td>C</td>
<td>0.046</td>
<td>-0.004131</td>
<td>-0.011629</td>
</tr>
<tr>
<td>rs7188445</td>
<td>16</td>
<td>78,292,488</td>
<td>A</td>
<td>G</td>
<td>0.032</td>
<td>-0.011765</td>
<td>-0.016981</td>
</tr>
<tr>
<td>rs7224610</td>
<td>17</td>
<td>50,719,787</td>
<td>A</td>
<td>C</td>
<td>0.042</td>
<td>0.015338</td>
<td>0.008492</td>
</tr>
<tr>
<td>rs742132</td>
<td>17</td>
<td>25,715,550</td>
<td>A</td>
<td>G</td>
<td>0.054</td>
<td>0.010673</td>
<td>-0.019475</td>
</tr>
<tr>
<td>rs2307394</td>
<td>2</td>
<td>148,432,898</td>
<td>T</td>
<td>C</td>
<td>0.029</td>
<td>0.018612</td>
<td>0.013885</td>
</tr>
<tr>
<td>rs17786744</td>
<td>8</td>
<td>23,832,951</td>
<td>A</td>
<td>G</td>
<td>0.029</td>
<td>0.000176</td>
<td>-0.004551</td>
</tr>
<tr>
<td>rs2079742</td>
<td>17</td>
<td>56,820,479</td>
<td>T</td>
<td>C</td>
<td>0.043</td>
<td>-0.038280</td>
<td>0.031271</td>
</tr>
<tr>
<td>rs164009</td>
<td>17</td>
<td>71,795,264</td>
<td>A</td>
<td>G</td>
<td>0.028</td>
<td>-0.015160</td>
<td>0.010596</td>
</tr>
</tbody>
</table>
A comparison of robust Mendelian randomization methods using summary data

Eric A.W. Slob, Stephen Burgess

Abstract

The number of Mendelian randomization analyses including large numbers of genetic variants is rapidly increasing. This is due to the proliferation of genome-wide association studies, and the desire to obtain more precise estimates of causal effects. Since it is unlikely that all genetic variants will be valid instrumental variables, several robust methods have been proposed. We compare nine robust methods for Mendelian randomization based on summary data that can be implemented using standard statistical software. Methods were compared in three ways: by reviewing their theoretical properties, in an extensive simulation study, and in an empirical example. In the simulation study, the best method, judged by mean squared error was the contamination mixture method. This method had well-controlled Type 1 error rates with up to 50% invalid instruments across a range of scenarios. Other methods performed well according to different metrics. Outlier-robust methods had the narrowest confidence intervals in the empirical example. With isolated exceptions, all methods performed badly when over 50% of the variants were invalid instruments. Our recommendation for investigators is to perform a variety of robust methods that operate in different ways and rely on different assumptions for valid inferences to assess the reliability of Mendelian randomization analyses.

This chapter is based on Slob and Burgess (2020).
3.1 Introduction

Mendelian randomization (MR) uses genetic variants as instrumental variables (IV) to determine whether an observational association between a modifiable exposure (often also called the intermediate variable under study or risk factor) and an outcome is consistent with a causal effect (Davey Smith and Ebrahim, 2003, Smith and Ebrahim, 2004). This approach is less vulnerable to traditional problems of epidemiological studies such as confounding and reverse causality. With the increasing availability of genome-wide association studies that find robust associations between genetic variants and exposures of interest (Zheng et al., 2017, Welter et al., 2014), the potential of this approach is rapidly evolving.

A genetic variant is a valid IV if (i) it is associated with the exposure, (ii) it has no direct effect on the outcome, and (iii) there are no associations between the variant and any potential confounders.

There has been much discussion on the potentials and limitations of MR, as the IV assumptions cannot be fully tested (Davey Smith and Ebrahim, 2003, Glymour et al., 2012, VanderWeele et al., 2014). Violation of the IV assumptions can lead to invalid conclusions in applied investigations. In practice, the exclusion restriction assumption that the proposed instruments (genetic variants) should not have a direct effect on the outcome of interest is debatable, particularly if the biological roles of the genetic variants are insufficiently understood (Glymour et al., 2012, von Hinke et al., 2016).

Some genetic variants are associated with multiple traits (Sivakumaran et al., 2011, Solovieff et al., 2013). This is referred to as pleiotropy. There are two types of pleiotropy. Vertical pleiotropy occurs when a variant is directly associated with the exposure and another trait on the same biological pathway. This does not lead to violation of the IV assumptions provided the only causal pathway from the genetic variant to the outcome passes via the exposure. Horizontal pleiotropy occurs when the second trait is on a different biological pathway, and so there may exist different causal pathways from the variant to the outcome. This would violate the exclusion restriction assumption. To solve the problems that arise due to horizontal pleiotropy, several robust methods for MR have been developed that can provide reliable inferences when some genetic variants violate the IV assumptions, or when genetic variants violate the IV assumptions in a particular way. To our knowledge, a comprehensive review and simulation study to compare the statistical performance of these different methods has not been performed.

To focus our simulation study and compare the most relevant robust methods for applied practice, we concentrate on methods that satisfy two criteria. First, the method requires only summary data on estimates (beta-coefficients and standard errors) of genetic variant–exposure and genetic variant–outcome associations. We
3. A comparison of robust Mendelian randomization methods using summary data

exclude methods that require individual participant data (Kang et al., 2016, Guo et al., 2018, Jiang et al., 2017, Tchetgen Tchetgen et al., 2017), and those that require data on additional variants not associated with the exposure (O’Connor and Price, 2018, DiPrete et al., 2018b). This is because the sharing of individual participant data is often impractical, so that many empirical researchers only have access to summary data, and for fairness, to ensure that all methods are using the same information to make inferences. Secondly, the method must be performed using standard statistical software packages. We exclude methods requiring convergence checks that cannot be easily automated for a simulation study (Berzuini et al., 2020) or are computationally infeasible for large numbers of variants in a reasonable running time (Burgess et al., 2018).

In this article, we review nine robust methods for MR from a theoretical perspective, and evaluate their performance in a simulation study set in a two-sample summary data setting. The methods differ in how they estimate a causal effect of the exposure on the outcome, as well as in the assumptions required for consistent estimation. We consider the weighted median, mode based estimation, MR-PRESSO, MR-Robust, MR-Lasso, MR-Egger, contamination mixture, MR-Mix, and MR-RAPS methods. Some methods take a summarized measure of the variant-specific causal estimates as the overall causal effect estimate (weighted median, and mode based estimation), whereas others remove or downweight outliers (MR-PRESSO, MR-Lasso, and MR-Robust), or attempt to model the distribution of the estimates from invalid IVs (MR-Egger, contamination mixture, MR-Mix, and MR-RAPS). We also consider the performance of the methods in an empirical example to evaluate the causal effect of body mass index on coronary artery disease risk.

This paper is organized as follows. First, we give an overview of the robust methods and compare their theoretical properties. Then, we introduce the simulation framework and applied example to compare their properties in practice. Finally, we discuss the implications of this work for applied practice.

3.2 Methods

Modelling assumptions and summary data

We consider a model as previously described by Palmer et al. (2008) and Bowden et al. (2017b) for $J$ genetic variants $G_1, G_2, \ldots, G_J$ that are independent in their distributions, a modifiable exposure $X$, an outcome variable $Y$, and a confounder $U$. We assume that all relationships between variables are linear and homogeneous without effect modification, meaning that the same causal effect is estimated by any valid IV (Didelez and Sheehan, 2007). A visual representation
of the model is shown in Figure 3.1.

![Figure 3.1 - Illustrative diagram showing the model assumed for genetic variant $G_j$, with effect $\phi_j$ on the unobserved confounder $U$, effect $\gamma_j$ on exposure $X$, and direct effect $\alpha_j$ on outcome $Y$. The causal effect of the exposure on the outcome is $\theta$. Dotted lines represent possible ways the instrumental variable assumptions could be violated.](image)

We assume that summary data are available on genetic associations with the exposure (beta-coefficient $\hat{\beta}_{X_j}$ and standard error $\sigma_{X_j}$) and with the outcome (beta-coefficient $\hat{\beta}_{Y_j}$ and standard error $\sigma_{Y_j}$) for each variant $G_j$.

**Inverse-variance weighted method**

The causal effect of the exposure on the outcome can be estimated using a single genetic variant $G_j$ by the ratio method:

$$\hat{\theta}_{R_j} = \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}.$$  \hfill (3.1)

The ratio estimate $\hat{\theta}_{R_j}$ is a consistent estimate of the causal effect if variant $G_j$ satisfies the IV assumptions (Didelez and Sheehan, 2007). If the uncertainty in the genetic association with the exposure is low, then the standard error of the
ratio estimate $\sigma_{R_j}$ can be approximated as (Thomas et al., 2007):

$$\sigma_{R_j} = \left| \frac{\sigma_{Y_j}}{\hat{\beta}_{X_j}} \right|. \quad (3.2)$$

The individual ratio estimates can be combined to obtain a single more efficient estimate. The optimally-efficient combination of the ratio estimates is referred to as the inverse-variance weighted (IVW) estimate (Burgess et al., 2013):

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{J} \hat{\theta}_{R_j} \sigma_{R_j}^{-2}}{\sum_{j=1}^{J} \sigma_{R_j}^{-2}} \cdot \frac{\sum_{j=1}^{J} \hat{\beta}_{X_j} \sigma_{Y_j}^{-2}}{\sum_{j=1}^{J} \hat{\beta}_{X_j}^2 \sigma_{Y_j}^{-2}}. \quad (3.3)$$

The IVW estimate is equal to the estimate from the two-stage least squares method that is performed using individual participant data (Burgess et al., 2016b). It is a weighted mean of the ratio estimates, where the weights are the inverse-variances of the ratio estimates. The IVW estimate can also be obtained by weighted regression of the genetic associations with the outcome on the genetic associations with the exposure:

$$\hat{\beta}_{Y_j} = \theta \hat{\beta}_{X_j} + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma_{Y_j}^2). \quad (3.4)$$

However, the IVW method has a 0% breakdown point, meaning that if only one genetic variant is not a valid IV, then the estimator is typically biased (Bowden et al., 2016). Bias will be present unless the pleiotropic effects of genetic variants average to zero (balanced pleiotropy) and the pleiotropic effects are independent of the genetic variant–exposure associations (see MR-Egger method below) (Bowden et al., 2017b). With the increasing number of variants used in MR investigations, it is increasingly unlikely that all variants are valid IVs. Hence, it is crucial to consider robust estimation methods despite their lower statistical efficiency (that is, lower power to detect a causal effect).

We proceed to introduce the different robust methods we consider in this study in three categories: consensus methods, outlier-robust methods, and modelling methods. A summary table comparing the methods is presented as Table 3.1.
<table>
<thead>
<tr>
<th>Method</th>
<th>Consistency assumption</th>
<th>Strengths and/or weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted Median</td>
<td>Majority valid</td>
<td>Robust to outliers, sensitive to additional/removal of genetic variants, may be less efficient</td>
</tr>
<tr>
<td>Mode Based Estimation</td>
<td>Plurality valid</td>
<td>Robust to outliers, sensitive to bandwidth parameter and addition/removal of genetic variants, generally conservative</td>
</tr>
<tr>
<td>MR-PRESSO</td>
<td>Outlier-robust</td>
<td>Removes outliers, efficient with valid IVs, very high false positive rate with several invalid IVs</td>
</tr>
<tr>
<td>MR-Robust</td>
<td>Outlier-robust</td>
<td>Downweights outliers, efficient with valid IVs, high false positive rate with several invalid IVs</td>
</tr>
<tr>
<td>MR-Lasso</td>
<td>Outlier-robust</td>
<td>Removes outliers, efficient with valid IVs, high false positive rate with several invalid IVs</td>
</tr>
<tr>
<td>MR-Egger</td>
<td>InSIDE</td>
<td>Sensitive to outliers, sensitive to violations of InSIDE assumption, InSIDE assumption often not plausible, may be less efficient</td>
</tr>
<tr>
<td>Contamination Mixture</td>
<td>Plurality valid</td>
<td>Robust to outliers, sensitive to variance parameter and addition/removal of genetic variants, less conservative than MBE</td>
</tr>
<tr>
<td>MR-Mix</td>
<td>Plurality valid</td>
<td>Robust to outliers, requires large numbers of genetic variants, very high false positive rate in several scenarios</td>
</tr>
<tr>
<td>MR-RAPS</td>
<td>Pleiotropic effects (except outliers) normally distributed about zero</td>
<td>Downweights outliers, sensitive to violations of balanced pleiotropy assumption</td>
</tr>
</tbody>
</table>
A comparison of robust Mendelian randomization methods using summary data

**Consensus methods**

A consensus method is one that takes its causal estimate as a summary measure of the distribution of the ratio estimates. The most straightforward consensus method is the median method. Rather than taking a weighted mean of the ratio estimates as in the IVW method, we take the median of the ratio estimates. The median estimator is consistent (that is, unbiased in large samples) even if up to 50% of the variants are invalid (Bowden et al., 2016). We consider a weighted version of the median method, where the median is taken from a distribution of the ratio estimates in which genetic variants with more precise ratio estimates receive more weight. Here, an unbiased estimate will be obtained if up to 50% of the weight comes from variants that are valid IVs. We refer to this as the ‘majority valid’ assumption.

A related assumption is the ‘plurality valid’ assumption (Guo et al., 2018). In large samples, while ratio estimates for all valid IVs should equal the true causal effect, ratio estimates for invalid IVs will take different values. The ‘plurality valid’ assumption is that, out of all the different values taken by ratio estimates in large samples (we term these the ratio estimands), the true causal effect is the value taken for the largest number of genetic variants (that is, the modal ratio estimand). For example, the plurality assumption would be satisfied if only 40% of the genetic variants are valid instruments, provided that out of the remaining 60% invalid instruments, no larger group with the same ratio estimand exists. This assumption is also referred to as the Zero Modal Pleiotropy Assumption (ZEMPA) (Hartwig et al., 2017).

This assumption is exploited by the mode based estimation (MBE) method (Hartwig et al., 2017). As no two ratio estimates will be identical in finite samples, it is not possible to take the mode of the ratio estimates directly. In the MBE method, a normal density is drawn for each genetic variant centered at its ratio estimate. The spread of this density depends on a bandwidth parameter, and (for the weighted version of the MBE method) the precision of the ratio estimate. A smoothed density function is then constructed by summing these normal densities. The maximum of this distribution is the causal estimate.

As these consensus methods take the median or mode of the ratio estimate distribution as the causal estimate, they are naturally robust to outliers, as the median and mode of a distribution are unaffected by the magnitude of extreme values. However, they are still influenced by outliers, as these variants still contribute to determining the location of the median or mode of a distribution. These methods can also be sensitive to changes in the ratio estimates for variants that contribute to the median or mode, and to the addition and removal of variants from the analysis. Additionally, the methods may not be as efficient as
those that base their estimates on all the genetic variants.

**Outlier-robust methods**

Next, we present three outlier-robust methods. These methods either downweight or remove genetic variants from the analysis that have outlying ratio estimates. They provide consistent estimates under the same assumptions as the IVW method for the set of genetic variants that are not identified as outliers.

In the MR-Pleiotropy Residual Sum and Outlier (MR-PRESSO) method (Verbanck et al., 2018), the IVW method is implemented by regression using all the genetic variants, and the residual sum of squares (RSS) is calculated from the regression equation. The RSS is a heterogeneity measure for the ratio estimates. Then, the IVW method is performed omitting each genetic variant from the analysis in turn. If the RSS decreases substantially compared to a simulated expected distribution, then that variant is removed from the analysis. This procedure is repeated until no further variants are removed from the analysis. The causal estimate is then obtained by the IVW method using the remaining genetic variants.

In MR-Robust, the IVW method is performed by regression, except that instead of using ordinary least squares regression, MM-estimation is used combined with Tukey’s biweight loss function (Burgess et al., 2016a). MM-estimation provides robustness against influential points and Tukey’s loss function provides robustness against outliers. Tukey’s loss function is a truncated quadratic function, meaning that there is a limit in the degree to which an outlier contributes to the analysis (Mosteller and Tukey, 1977). This contrasts with the quadratic loss function used in ordinary least squares regression, which is unbounded, meaning that a single outlier can have an unlimited effect on the IVW estimate.

In MR-Lasso, the IVW regression model is augmented by adding an intercept term for each genetic variant (Burgess et al., 2016a). The IVW estimate is the value of $\theta$ that minimizes:

$$ J = \sum_{j=1}^{J} \sigma_{Y_j}^2 \left( \hat{\beta}_Y_j - \theta \hat{\beta}_X_j \right)^2. $$

(3.5)

In MR-Lasso, we minimize:

$$ J = \sum_{j=1}^{J} \sigma_{Y_j}^2 \left( \hat{\beta}_Y_j - \theta_{0j} - \theta \hat{\beta}_X_j \right)^2 + \lambda \sum_{j=1}^{J} |\theta_{0j}|, $$

(3.6)

where $\lambda$ is a tuning parameter. As the regression equation contains more parameters than there are genetic variants, a lasso penalty term is added for
identification (Windmeijer et al., 2019). The intercept term \( \theta_{0j} \) represents the direct (pleiotropic) effect on the outcome, and should be zero for a valid IV, but will be non-zero for an invalid IV. The causal estimate is then obtained by the IVW method using the genetic variants that had \( \theta_{0j} = 0 \) in equation (3.6). A heterogeneity criterion is used to determine the value of \( \lambda \). Increasing \( \lambda \) means that more of the pleiotropy parameters equal zero and so the corresponding variants are included in the analysis; we increase \( \lambda \) step-by-step until one step before there is more heterogeneity in the ratio estimates for variants included in the analysis than expected by chance alone.

The MR-PRESSO and MR-Lasso methods remove variants from the analysis, whereas MR-Robust downweights variants. These methods will be valuable when there is a small number of genetic variants with heterogeneous ratio estimates, as they will be removed from the analysis or heavily downweighted, and so will not influence the overall estimate. In such a case, these methods are likely to be efficient, as they are based on the IVW method. The methods are less likely to be valuable when there is a larger number of genetic variants that are pleiotropic, particularly if the pleiotropic effects are small in magnitude, and when the average pleiotropic effect of non-outliers is not zero.

**Modelling methods**

Finally, we present four methods that attempt to model the distribution of estimates from invalid IVs or make a specific assumption about the way in which the IV assumptions are violated. The MR-Egger method is performed similarly to the IVW method, except that the regression model contains an intercept term \( \theta_0 \):

\[
\hat{\beta}_Y = \theta_0 + \theta \hat{\beta}_X + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2_j).
\]  

This differs from the MR-Lasso method, as there is only one intercept term, which represents the average pleiotropic effect. The MR-Egger method gives consistent estimates of the causal effect under the Instrument Strength Independent of Direct Effect (InSIDE) assumption, which states that pleiotropic effects of genetic variants must be uncorrelated with genetic variant–exposure association. As the regression model is no longer symmetric to changes in the signs of the genetic association estimates (which result from switching the reference and effect alleles), we first re-orientate the genetic associations before performing the regression by fixing all genetic associations with the exposure to be positive, and correspondingly changing the signs of the genetic associations with the outcome if necessary. The intercept in MR-Egger also provides a test of the IV assumptions. The intercept will differ from zero when either the average pleiotropic effect is
not zero, or the InSIDE assumption is violated. These two conditions (average pleiotropy of zero and InSIDE assumption satisfied) are precisely the conditions required for the IVW estimate to be unbiased.

The contamination mixture method assumes that only some of the genetic variants are valid IVs (Burgess et al., 2019). We construct a likelihood function from the ratio estimates. If a variant is a valid instrument, then its ratio estimate is assumed to be normally distributed about the true causal effect $\theta$ with variance $\sigma_{R_j}^2$. If a variant is not a valid instrument, then its ratio estimate is assumed to be normally distributed about zero with variance $\psi^2 + \sigma_{R_j}^2$, where $\psi^2$ represents the variance of the estimands from invalid IVs. This parameter is specified by the analyst. We then maximize the likelihood over different values of the causal effect $\theta$ and different configurations of valid and invalid IVs. Maximization is performed in linear time by first constructing a profile likelihood as a function of $\theta$, and then maximizing this function with respect to $\theta$. The value of $\theta$ that maximizes the profile likelihood is the causal estimate.

The MR-Mix method (Qi and Chatterjee, 2018) is similar to the contamination mixture method, except that rather than dividing the genetic variants into valid and invalid IVs, the method divides variants into four categories: (1) variants that directly influence the exposure only (valid instruments), and (2) variants that influence the exposure and outcome, (3) that influence the outcome only, and (4) that neither influence the exposure or outcome (invalid instruments). This allows for more flexibility in modelling genetic variants, although potentially leads to more uncertainty in assigning genetic variants to categories.

The MR-Robust Adjusted Profile Score (RAPS) (Zhao et al., 2018) method models the pleiotropic effects of genetic variants directly using a random-effects distribution. The pleiotropic effects are assumed to be normally distributed about zero with unknown variance. Estimates are obtained using a profile likelihood function for the causal effect and the variance of the pleiotropic effect distribution. To provide further robustness to outliers, either Tukey’s biweight loss function or Huber’s loss function (Mosteller and Tukey, 1977) can be used.

Modelling methods are likely to be valuable when the modelling assumptions are correct, but not when the assumptions are incorrect. For example, the MR-Egger method requires the InSIDE assumption to be satisfied to give a consistent estimate. The MR-RAPS method is likely to perform well when pleiotropic effects truly are normally distributed about zero, but less well when they are not. The MR-Mix method is likely to require large numbers of genetic variants in order to correctly classify variants into the different categories. The contamination mixture method is less likely to be affected by modelling assumptions as it does not make such strict assumptions, but it is likely to be sensitive to specification of the
A comparison of robust Mendelian randomization methods using summary data

Simulation study

To compare the performance of these methods in a realistic setting, we perform a simulation study. Full details of the simulation study are given in the Supplementary Material.

For each participant $i$, we simulate data on $J$ genetic variants $G_{i1}, G_{i2}, \ldots, G_{iJ}$, a modifiable exposure $X_i$, an outcome variable $Y_i$, and a confounder $U_i$ (assumed unknown). The confounder is a linear function of the genetic variants and an independent error term $\varepsilon^U_i$. The effect of variant $j$ on the confounder is represented by coefficient $\phi_j$ (this is zero for a valid IV). The exposure is linear in the genetic variants, the confounder and an independent error term $\varepsilon^X_i$. The effect of variant $j$ on the exposure is represented by coefficient $\gamma_j$. The outcome is linear in the genetic variants, exposure, confounders and an independent error term $\varepsilon^Y_i$. The effect of variant $j$ on the outcome is represented by coefficient $\alpha_j$ (again, this is zero for a valid IV). The effect of the exposure on the outcome is represented by $\theta$.

The genetic variants are modelled as single nucleotide polymorphisms (SNPs), with a varying minor allele frequency $\text{maf}_j$, and take values 0, 1 or 2. The minor allele frequencies are drawn from an uniform distribution. The error terms $\varepsilon^U_i$, $\varepsilon^X_i$ and $\varepsilon^Y_i$ each follow an independent normal distribution with mean 0 and unit variance.

We can represent the model mathematically as:

\[
U_i = \sum_{j=1}^{J} \phi_j G_{ij} + \varepsilon^U_i, \\
X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \varepsilon^X_i, \\
Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \theta X_i + U_i + \varepsilon^Y_i,
\]

\[
\text{maf}_j \sim \mathcal{U}(0.1, 0.5), \\
G_{ij} \sim \text{Binomial}(2, \text{maf}_j) \text{ independently}, \\
\varepsilon^U_i, \varepsilon^X_i, \varepsilon^Y_i \sim \mathcal{N}(0, 1) \text{ independently}.
\]

In brief, we consider three scenarios:

1. balanced pleiotropy, InSIDE satisfied – invalid IVs have direct effects on the outcome generated from a normal distribution centered at zero (for invalid instruments $\alpha_j \sim \mathcal{N}(0, 0.15)$, $\phi_j = 0$);
2. directional pleiotropy, InSIDE satisfied – invalid IVs have direct effects on the outcome generated from a normal distribution centered away from zero (for invalid instruments $\alpha_j \sim \mathcal{N}(0.1, 0.075)$, $\phi_j = 0$);

3. directional pleiotropy, InSIDE violated – invalid IVs have direct effects on the outcome generated from a normal distribution centered away from zero, and indirect effects on the outcome via the confounder (for invalid instruments $\alpha_j \sim \mathcal{N}(0.1, 0.075)$, $\phi_j \sim \mathcal{U}(0, 0.1)$).

We simulated data on $J = 10, 30, \text{and } 100$ genetic variants. A portion of the genetic variants were invalid IVs (30%, 50% and 70%), and the direct effects of the variants explain 10% of the variance in the exposure. Summary genetic associations were calculated for the exposure and the outcome on non-overlapping sets of individuals, each consisting of 10,000 individuals (Haycock et al., 2016). This situation is often referred to as two-sample summary data MR (Pierce and Burgess, 2013). We considered situations with a null causal effect ($\theta = 0$) and a positive causal effect ($\theta = 0.2$). In total, 10,000 datasets were generated in each scenario.

Methods can be compared by many metrics, including bias, empirical power, and standard deviation of estimates. We use mean squared error, which is the sum of bias squared plus variance, as the main criterion for comparing methods, as this provides a compromise between bias and precision. However, the relative importance of each metric will depend on the specific features of the application.

**Empirical example: the effect of body mass index on coronary artery disease risk**

We also compare the methods in an empirical example considering the effect of body mass index (BMI) on coronary artery disease (CAD) risk. Since BMI is influenced by several biological mechanisms (Monnereau et al., 2016), it is likely that the exclusion restriction is not satisfied for all associated genetic variants. Hence it is necessary to use robust methods to analyse these data. Additionally, we consider methods that detect outliers (MR-Presso, MR-Robust, MR-Lasso, contamination mixture, MR-Mix, and MR-RAPS), and compare whether the same outliers are detected in each of these methods.

We take 97 genome-wide significant variants associated with BMI from the GIANT consortium (Locke et al., 2015). Associations with BMI are estimated in up to 339,224 participants from this consortium. Associations with coronary artery disease risk are estimated in up to 60,801 CAD cases and 123,504 controls from the CARDIoGRAMplusC4D Consortium (Nikpay et al., 2015). Association estimates for CAD were available for 94 of these variants.
The scatter plot of the genetic associations with BMI and CAD risk is shown in Figure 3.2. While most variants seem to suggest a harmful effect of increased BMI on CAD risk, there is apparent heterogeneity in the IV estimates from each genetic variant individually, as evidenced by Cochran’s $Q$ test ($Q$-statistic = 235.7, $P < 0.001$). Even after removing the five outliers as judged by the MR-PRESSO method, which makes use of the heterogeneity statistic to identify outliers, we still reject the null hypothesis of that the regression model (including an intercept) fits the regression model with no additional variability than would be expected by chance ($Q$-statistic = 125.9, $P = 0.005$). This suggests that some of the variants violate the IV assumptions.

**Figure 3.2** – Scatter plot of genetic associations with BMI (standard deviation units) and coronary artery disease risk (log odds ratios) for 94 variants taken from the GIANT and CARDIoGRAMplusC4D consortia respectively.

### 3.3 Results

Results of the simulation study are presented in Table 3.2 (10 variants), Table 3.3 (30 variants), and Table 3.4 (100 variants). For each scenario, we present the mean, median, and standard deviation of estimates across simulations, and the empirical Type 1 error rate (for a null causal effect) or empirical power (for a positive causal effect) at a 95% confidence level. The empirical Type 1 error rate and empirical power are calculated as the proportion of simulated datasets
where zero was not included in the 95% confidence interval. The mean squared error across simulations for the different methods with a null causal effect is presented in Figure 3.3 (Scenario 2), and Figure 3.4 (Scenario 3) for 30 variants. The corresponding plots for 10 variants (Supplementary Figures 1 and 2) and 100 variants (Supplementary Figures 3 and 4) were broadly similar.

Overall, judging by mean squared error, the contamination mixture method performed best with 30% and 50% invalid variants. In some scenarios, other methods had lower mean squared error with 70% invalid variants. However, with some isolated exceptions, all the methods performed badly with 70% invalid instruments. Coverage for the contamination mixture method was around 10% or less when there were up to 50% invalid variants. This was also true for the MR-Robust method, although that method had slightly lower power to detect a causal effect in some scenarios. Several other methods performed well in particular scenarios.

Amongst consensus methods, estimates from the MBE method were less biased than those from the weighted median method, with lower Type 1 errors. The weighted median method had slightly higher power to detect a causal effect, although comparisons of power lose much of their value when a method has inflated Type 1 error rates. Performance of the MBE method improved as the number of variants increased. Amongst outlier-robust methods, bias was greater for the MR-Robust than the MR-Lasso method. The MR-Lasso method generally had the lower mean squared error when the invalidity was 50% or 70%, but MR-Robust had the lower Type 1 error rates. Performance of the MR-Robust method was better when there were at least 30 genetic variants. MR-PRESSO had biased estimates with inflated Type 1 error rates even with 30% invalid variants, and performed particularly badly as the number of variants increased.

The modelling methods performed well in some scenarios, but less well in others. This is unsurprising, as in some scenarios, consistency assumptions for the methods were satisfied, and in others they were not. The MR-Egger method performed well in terms of Type 1 error rate in Scenarios 1 and 2, where the InSIDE assumption was satisfied. Estimates from the method were generally imprecise with low power. However, power in the MR-Egger method depends on the genetic associations with the exposure varying substantially between variants, which was not the case in the simulation study (Burgess and Thompson, 2017). The contamination mixture method performed well with 30% and 50% valid instruments, with low bias and Type 1 error rates at or below 8% with 10 variants, 10% with 30 variants, and 11% with 100 variants. The MR-Mix method performed badly throughout, with highly inflated Type 1 error rates in almost all scenarios with less than 100 instruments and comparatively low power to detect a
3. A comparison of robust Mendelian randomization methods using summary data

causal effect. It performed slightly better with more genetic variants, although its performance was still worse than other methods. However, the method performed much better in a simulation comparison of methods performed by the authors of the MR-Mix method (Qi and Chatterjee, 2019), in which the data-generating model was more similar to the model assumed by the MR-Mix method. The MR-RAPS method performed well in Scenario 1, where its consistency assumption was satisfied, but less well in other scenarios with inflated Type 1 error rates. Its performance also worsened as more variants were included in the analysis.

**Figure 3.3** – Mean squared errors for the different methods in scenario 2 (directional pleiotropy, InSIDE satisfied) with a null causal effect for 30 variants. Note the vertical axis is on a logarithmic scale.

**Empirical example: The effect of body mass index on coronary artery disease**

Results from the empirical example are shown in Table 3.5. All methods agree that there is a positive effect of BMI on CAD risk, except for the MR-Mix method which gives a wide confidence interval that includes the null. The narrowest confidence intervals are for the outlier-robust methods (MR-Lasso, MR-Robust, MR-PRESSO), followed by the modelling methods except MR-Mix and MR-Egger (contamination mixture, MR-RAPS), then the consensus methods (weighted median, mode based estimation), and finally MR-Egger and MR-Mix.

While the methods that detect outliers varied in terms of how lenient or
### Table 3.2 – Mean, median, standard deviation (SD) of estimates, and Type 1 error/empirical power (%) with 10 genetic variants.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario 1: Balanced pleiotropy, InSIDE satisfied</th>
<th>Scenario 2: Directional pleiotropy, InSIDE satisfied</th>
<th>Scenario 3: Directional pleiotropy, InSIDE violated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean, median, standard deviation (SD)</td>
<td>Mean, median, standard deviation (SD)</td>
<td>Mean, median, standard deviation (SD)</td>
</tr>
<tr>
<td></td>
<td>T1 error</td>
<td>T1 error</td>
<td>T1 error</td>
</tr>
<tr>
<td>Mean</td>
<td>Median</td>
<td>SD</td>
<td>Mean, median, standard deviation (SD)</td>
</tr>
<tr>
<td>MR-PRESSO</td>
<td>0.198 0.200 0.028 0.076 0.061</td>
<td>0.198 0.200 0.028 0.076 0.061</td>
<td>0.198 0.200 0.028 0.076 0.061</td>
</tr>
<tr>
<td>MR-Robust</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
</tr>
<tr>
<td>MR-Lasso</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
</tr>
<tr>
<td>MR-Egger</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
</tr>
<tr>
<td>Contamination Mixture</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
</tr>
<tr>
<td>MR-Mix</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
<td>0.200 0.200 0.019 0.046 0.041</td>
</tr>
</tbody>
</table>

**Abbreviations:** T1 error: Type 1 error.
### Table 3.3 – Mean, median, standard deviation (SD) of estimates, and Type 1 error/empirical power (%) with 30 genetic variants.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean (SD)</th>
<th>Median (SD)</th>
<th>SD (SD)</th>
<th>Power (%)</th>
<th>Mean (SD)</th>
<th>Median (SD)</th>
<th>SD (SD)</th>
<th>Power (%)</th>
<th>Mean (SD)</th>
<th>Median (SD)</th>
<th>SD (SD)</th>
<th>Power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode Based Estimation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MR-PRESSO</td>
<td>0.201</td>
<td>0.200</td>
<td>0.051</td>
<td>0.001</td>
<td>0.201</td>
<td>0.200</td>
<td>0.050</td>
<td>0.001</td>
<td>0.201</td>
<td>0.200</td>
<td>0.050</td>
<td>0.001</td>
</tr>
<tr>
<td>MR-Robust</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>MR-Egger</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>Contamination Mixture</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>MR-Mix</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>MR-RAPS</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>Scenario 2: Directional pleiotropy, InSIDE satisfied</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Mean, median, standard deviation (SD) of estimates, and Type 1 error/empirical power (%) with 30 genetic variants.

### Scenario 3: Directional pleiotropy, InSIDE violated

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean (SD)</th>
<th>Median (SD)</th>
<th>SD (SD)</th>
<th>Power (%)</th>
<th>Mean (SD)</th>
<th>Median (SD)</th>
<th>SD (SD)</th>
<th>Power (%)</th>
<th>Mean (SD)</th>
<th>Median (SD)</th>
<th>SD (SD)</th>
<th>Power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode Based Estimation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MR-PRESSO</td>
<td>0.201</td>
<td>0.200</td>
<td>0.051</td>
<td>0.001</td>
<td>0.201</td>
<td>0.200</td>
<td>0.050</td>
<td>0.001</td>
<td>0.201</td>
<td>0.200</td>
<td>0.050</td>
<td>0.001</td>
</tr>
<tr>
<td>MR-Robust</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>MR-Egger</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>Contamination Mixture</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>MR-Mix</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>MR-RAPS</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>Positive causal effect: θ = +0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: T1 error: Type 1 error.

49
<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted Median</td>
<td>0.200</td>
<td>0.200</td>
<td>0.028</td>
<td>1.000</td>
</tr>
<tr>
<td>Mode Based Estimation</td>
<td>0.199</td>
<td>0.199</td>
<td>0.025</td>
<td>1.000</td>
</tr>
<tr>
<td>MR-PRESSO</td>
<td>0.200</td>
<td>0.200</td>
<td>0.026</td>
<td>1.000</td>
</tr>
<tr>
<td>MR-Robust</td>
<td>0.200</td>
<td>0.200</td>
<td>0.021</td>
<td>1.000</td>
</tr>
<tr>
<td>MR-Lasso</td>
<td>0.200</td>
<td>0.200</td>
<td>0.020</td>
<td>1.000</td>
</tr>
<tr>
<td>MR-Egger</td>
<td>0.200</td>
<td>0.199</td>
<td>0.019</td>
<td>1.000</td>
</tr>
<tr>
<td>Contamination Mixture</td>
<td>0.202</td>
<td>0.202</td>
<td>0.021</td>
<td>1.000</td>
</tr>
<tr>
<td>MR-Mix</td>
<td>0.203</td>
<td>0.200</td>
<td>0.019</td>
<td>1.000</td>
</tr>
<tr>
<td>MR-RAPS</td>
<td>0.201</td>
<td>0.201</td>
<td>0.018</td>
<td>1.000</td>
</tr>
<tr>
<td>Mean Median SD Power</td>
<td>0.201</td>
<td>0.201</td>
<td>0.018</td>
<td>Power</td>
</tr>
</tbody>
</table>
| Scenario 1: Balanced pleiotropy, InSIDE satisfied  
- Weighted Median              | 0.200| 0.200  | 0.028 | Power |
| Mode Based Estimation        | 0.199| 0.199  | 0.025 | Power |
| MR-PRESSO                    | 0.200| 0.200  | 0.026 | Power |
| MR-Robust                    | 0.200| 0.200  | 0.021 | Power |
| MR-Lasso                     | 0.200| 0.200  | 0.020 | Power |
| MR-Egger                     | 0.200| 0.199  | 0.019 | Power |
| Contamination Mixture        | 0.202| 0.202  | 0.021 | Power |
| MR-Mix                       | 0.203| 0.200  | 0.019 | Power |
| MR-RAPS                      | 0.201| 0.201  | 0.018 | Power |
| Mean Median SD Power         | 0.201| 0.201  | 0.018 | Power |
| Scenario 2: Directional pleiotropy, InSIDE satisfied  
- Weighted Median              | 0.200| 0.200  | 0.028 | Power |
| Mode Based Estimation        | 0.199| 0.199  | 0.025 | Power |
| MR-PRESSO                    | 0.200| 0.200  | 0.026 | Power |
| MR-Robust                    | 0.200| 0.200  | 0.021 | Power |
| MR-Lasso                     | 0.200| 0.200  | 0.020 | Power |
| MR-Egger                     | 0.200| 0.199  | 0.019 | Power |
| Contamination Mixture        | 0.202| 0.202  | 0.021 | Power |
| MR-Mix                       | 0.203| 0.200  | 0.019 | Power |
| MR-RAPS                      | 0.201| 0.201  | 0.018 | Power |
| Mean Median SD Power         | 0.201| 0.201  | 0.018 | Power |
| Scenario 3: Directional pleiotropy, InSIDE violated  
- Weighted Median              | 0.200| 0.200  | 0.028 | Power |
| Mode Based Estimation        | 0.199| 0.199  | 0.025 | Power |
| MR-PRESSO                    | 0.200| 0.200  | 0.026 | Power |
| MR-Robust                    | 0.200| 0.200  | 0.021 | Power |
| MR-Lasso                     | 0.200| 0.200  | 0.020 | Power |
| MR-Egger                     | 0.200| 0.199  | 0.019 | Power |
| Contamination Mixture        | 0.202| 0.202  | 0.021 | Power |
| MR-Mix                       | 0.203| 0.200  | 0.019 | Power |
| MR-RAPS                      | 0.201| 0.201  | 0.018 | Power |
| Mean Median SD Power         | 0.201| 0.201  | 0.018 | Power |

Abbreviations: T1 error: Type 1 error.
3. A comparison of robust Mendelian randomization methods using summary data

strictly they identified outliers, they agreed on the order of outliers (Supplementary Table 3). The MR-Robust method was the most lenient, downweighting two variants as outliers. Each subsequent method in order of strictness identified all previously identified variants as outliers. MR-PRESSO excluded the two variants identified by MR-Robust plus an additional three variants. MR-RAPS identified these five plus an additional two variants. MR-Lasso identified an additional three variants, 10 in total. The contamination mixture method identified an additional 14 variants, 24 in total. MR-Mix identified an additional 21 variants, 45 in total. This suggests that any difference between results from outlier-robust methods are likely due to the strictness of outlier detection, rather than due to intrinsic differences in how the different methods select outliers. In several methods, the threshold at which outliers are detected can be varied by the analyst (for example, by varying the penalization parameter $\lambda$ in MR-Lasso, or the significance threshold in MR-PRESSO). In practice, rather than performing different outlier-robust methods, it may be better to concentrate on one method, but vary this threshold. In our example, some of the variants that were the most pleiotropic in terms of their associations with other measured risk factors were only removed from the analysis by the MR-Mix method (Supplementary Table 3).
<table>
<thead>
<tr>
<th>Method</th>
<th>Causal estimate (95% CI)</th>
<th>CI width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted Median</td>
<td>0.376 (0.206, 0.546)</td>
<td>0.340</td>
</tr>
<tr>
<td>Mode Based Estimation</td>
<td>0.382 (0.181, 0.583)</td>
<td>0.402</td>
</tr>
<tr>
<td>MR-PRESSO</td>
<td>0.410 (0.309, 0.511)</td>
<td>0.202</td>
</tr>
<tr>
<td>MR-Robust</td>
<td>0.425 (0.325, 0.526)</td>
<td>0.201</td>
</tr>
<tr>
<td>MR-Lasso</td>
<td>0.442 (0.354, 0.530)</td>
<td>0.176</td>
</tr>
<tr>
<td>MR-Egger</td>
<td>0.481 (0.165, 0.796)</td>
<td>0.631</td>
</tr>
<tr>
<td>(intercept)</td>
<td>-0.003 (-0.011, 0.005)</td>
<td></td>
</tr>
<tr>
<td>Contamination Mixture</td>
<td>0.490 (0.372, 0.602)</td>
<td>0.230</td>
</tr>
<tr>
<td>MR-Mix</td>
<td>0.425 (-0.283, 1.133)</td>
<td>1.416</td>
</tr>
<tr>
<td>MR-RAPS</td>
<td>0.390 (0.308, 0.546)</td>
<td>0.238</td>
</tr>
</tbody>
</table>

### 3.4 Discussion

In this paper, we have provided a review of robust methods for MR, focusing on methods that can be performed using summary data and implemented using standard statistical software. We have divided methods into three categories: consensus methods, outlier-robust methods, and modelling methods. Methods were compared in three ways: by their theoretical properties, including the assumptions required for the method to give a consistent estimate, in an extensive simulation study, and in an empirical investigation.

While the use of robust methods for MR analyses with multiple genetic variants is highly recommended, it is not practical or desirable to perform and report results from every single robust method that has been proposed. Guidance is therefore needed as to which robust methods should be performed in practice. As an example, if an investigator performed the MR-PRESSO, MR-Robust, and MR-Lasso methods, they would have assessed robustness of the result to outliers, but they would not have assessed other potential violations of the IV assumptions. The categorization of methods proposed here is not the only possible division of methods, but we hope it is practically useful. For instance, the contamination mixture and MR-Mix methods make the same ‘plurality valid’ assumption as the MBE method, and so could have been placed in the same category.

The similarity and ubiquity of the ‘outlier-robust’ and ‘majority/plurality valid’ assumptions should encourage investigators to consider methods that make alternative assumptions, such as the MR-Egger method. While the InSIDE assumption is often not plausible (Burgess and Thompson, 2017), the MR-Egger method and the intercept test have value in providing a different route to testing the validity of an MR study. Another potential choice is the constrained
IV method, which uses information on measured confounders to construct a composite IV that is not associated with these confounders (Jiang et al., 2017). This method was not considered in the simulation study, as it requires additional data on confounders and individual participant data. Further methods development is needed to develop robust methods for summary data that make different consistency assumptions.

We encourage researchers to perform robust methods from different categories, and that make varied consistency assumptions. For example, an investigator could perform the weighted median method (majority valid assumption), the contamination mixture method (plurality valid assumption), and the MR-Egger method (InSIDE assumption). If there are a few clear outliers in the data, then an outlier-robust method such as MR-PRESSO (best used with few very distinct outliers) or MR-Robust could also be performed. While we are hesitant to make a definitive recommendation as each method has its own strengths and weaknesses, this set of methods would be a reasonable compromise between performing too few methods and not adequately assessing the IV assumptions, and performing so many methods that clarity is obscured. Another danger of the use of large numbers of methods is the possibility to cherry-pick results, either by an investigator seeking to present their results in a more positive light, or a reader picking the one method that gives a different result (such as the MR-Mix method in our empirical example).

One important limitation of these methods is the assumption that all valid IVs estimate the same causal effect. Particularly for complex exposures such as BMI, it is possible that different genetic variants have different ratio estimates not because they are invalid IVs, but because there are different ways of intervening on BMI that lead to different effects on the outcome. This can be remedied somewhat in methods based on the IVW method by using a random-effects model (Bowden et al., 2017b), or in the contamination mixture method, where causal effects evidenced by different sets of variants will lead to a multimodal likelihood function, and potentially a confidence interval that consists of more than one region.

In summary, while robust methods for MR do not provide a perfect solution to violations of the IV assumptions, they are able to detect such violations and help investigators make more reliable causal inferences. Investigators should perform a range of robust methods that operate in different ways and make different assumptions to assess the robustness of findings from a MR investigation.
Acknowledgements

Eric Slob acknowledges funding from the Stichting Erasmus Trustfonds for his research visit to the MRC Biostatistics Unit.
3.A Details of simulation study

For each participant \( i \), we simulate data on \( J \) genetic variants \( G_{i1}, G_{i2}, \ldots, G_{iJ} \), a modifiable exposure \( X_i \), an outcome variable \( Y_i \), and a confounder \( U_i \) (assumed unknown). The confounder is a linear function of the genetic variants and an independent error term \( \epsilon_i^U \). The effect of variant \( j \) on the confounder is represented by coefficient \( \phi_j \) (this is zero for a valid IV). The exposure is linear in the genetic variants, the confounder and an independent error term \( \epsilon_i^X \). The effect of variant \( j \) on the exposure is represented by coefficient \( \gamma_j \). The outcome is linear in the genetic variants, exposure, confounders and an independent error term \( \epsilon_i^Y \). The effect of variant \( j \) on the outcome is represented by coefficient \( \alpha_j \) (again, this is zero for a valid IV). The effect of the exposure on the outcome is represented by \( \theta \). The genetic variants are modelled as single nucleotide polymorphisms (SNPs), with a varying minor allele frequency maf\(_j\), and take values 0, 1 or 2. The minor allele frequencies are drawn from an uniform distribution (maf\(_j\) \( \sim \mathcal{U}(0.1, 0.5) \)). The error terms \( \epsilon_i^U, \epsilon_i^X \) and \( \epsilon_i^Y \) each follow an independent normal distribution with mean 0 and unit variance.
We can represent the model mathematically as:

\[ U_i = \sum_{j=1}^{J} \phi_j G_{ij} + \epsilon_i^U, \quad (3.9) \]

\[ X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \epsilon_i^X, \quad (3.10) \]

\[ Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \theta X_i + U_i + \epsilon_i^Y, \quad (3.11) \]

\[ \text{maf}_j \sim \mathcal{U}(0.1, 0.5), \quad (3.12) \]

\[ G_{ij} \sim \text{Binomial}(2, \text{maf}_j) \text{ independently,} \quad (3.13) \]

\[ \epsilon_i^U, \epsilon_i^X, \epsilon_i^Y \sim \mathcal{N}(0, 1) \text{ independently.} \quad (3.14) \]

The causal effect of the exposure on the outcome was either taken as null \((\theta = 0)\) or positive \((\theta = 0.2)\). Genetic associations with the exposure \(\gamma_j\) are drawn from a left-sided truncated normal distribution (truncation at 0.15, 0.1, and 0.05, for \(J = 10, 30, \) and 100 respectively). The variance of this distribution is chosen such that the total proportion of variance explained in the exposure by direct effects of the genetic variants is on average 10%. In scenario 3, the overall proportion of variance explained in the exposure by genetic variants is slightly larger, as there is an additional effect of the invalid IVs on the exposure via their effect on the confounder.

For valid IVs, \(\phi_j = 0\) and \(\alpha_j = 0\). For invalid IVs, in scenario 1 (balanced pleiotropy, InSIDE satisfied), the effects of the genetic variants on the outcome are generated from a normal distribution centered at zero \((\alpha_j \sim \mathcal{N}(0, 0.15))\) and genetic effects on the confounder are zero \((\phi_j = 0)\). In scenario 2 (directional pleiotropy, InSIDE satisfied), the effects of the genetic variants on the outcome are generated from a normal distribution centered away from zero \((\alpha_j \sim \mathcal{N}(0.1, 0.075))\) and genetic effects on the confounder are zero \((\phi_j = 0)\). In scenario 3 (directional pleiotropy, InSIDE violated), the direct effects of the genetic variants on the outcome are generated from a normal distribution centered away from zero \((\alpha_j \sim \mathcal{N}(0.1, 0.075))\) and genetic effects on the confounder are generated from a uniform distribution \((\phi_j \sim \mathcal{U}(0, 0.1))\).

Summary genetic association data are calculated by regressing the outcome on each genetic variant in turn. Individual participant data are generated for 10,000 individuals, where we perform the outcome regressions on all these individuals to come to the second stage effect estimates and corresponding standard errors. For the exposure summary genetic associations, we give the true value of the first stage effect with corresponding theoretical standard error (which is given
3. A comparison of robust Mendelian randomization methods using summary data

by \( (\sqrt{N} \times \sqrt{2 \times \text{maf}_j \times (1 - \text{maf}_j)})^{-1} \), where \( N \) is the number of individuals in the first stage GWAS) with again 10,000 individuals. This represents a two-sample Mendelian randomization study. We generated 10,000 simulated datasets for each scenario, and for null and positive causal effects.

Each method is performed using the default options suggested by the authors of the method, either in the corresponding publication, or in the software code recommended by the authors. The weighted median method is performed using inverse-variance weights. The mode-based estimation method is performed using inverse-variance weights, the ‘no measurement error’ assumption, and the default bandwidth setting \( (\phi = 1) \). The MR-PRESSO method is performed using a significance cut-off of \( p < 0.05 \) for determining outliers. The MR-Lasso method is performed using the heterogeneity criterion for selecting the lasso penalty parameter. The contamination mixture method is performed using the standard deviation of the ratio estimates multiplied by 1.5 for the variance parameter. For MR-Mix, we choose an initial value of the probability mass at the null component as 0.6 and the initial value of the variance of the non-null component as \( 1 \times 10^{-5} \). As the method performs a grid search, these decisions should not influence the results. For MR-RAPS, we use the overdispersed robust version with the Huber loss function. All regression models use random-effects.

The mean squared errors of the different methods are presented in Supplementary Figure 3.5 (10 variants, scenario 2), Supplementary Figure 3.6 (10 variants, scenario 3), Supplementary Figure 3.7 (100 variants, scenario 2), and Supplementary Figure 3.8 (100 variants, scenario 3). Note that in each case the vertical axis is on a logarithmic scale. Findings are similar to before among the different scenarios. We observe again that the performance of the mode-based estimator is the best for the consensus-based approach, MR-Robust gets the best result among the outlier-robust methods, and the contamination mixture approach has the best performance among the modelling methods.

3.B Outliers according to different methods
Table 3.6 – Genetic variants identified as outliers by the different methods in the Mendelian Randomization study of the effect of BMI on cardiovascular disease risk and other traits the variants are associated with according to the NHGRI-EBI Catalog of published genome-wide association studies (Buniello et al., 2019) (last accessed on 12 July 2019).

<table>
<thead>
<tr>
<th>Variant</th>
<th>Robust</th>
<th>PRESSO</th>
<th>RAPS</th>
<th>Lasso</th>
<th>Contam mix</th>
<th>Mix</th>
<th>Associated traits</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs11191560</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Waist circumference, Hip circumference</td>
</tr>
<tr>
<td>rs2075650</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Waist circumference, Obesity</td>
</tr>
<tr>
<td>rs2176040</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Hand grip strength</td>
</tr>
<tr>
<td>rs6567160</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Hip circumference</td>
</tr>
<tr>
<td>rs7903146</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Parental longevity</td>
</tr>
</tbody>
</table>
| rs11727676 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | HDL cholesterol levels, Colorectal cancer, Diver-
|            |        |        |      |       |            |     | ticular disease                                                                    |
| rs17024393 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Hip circumference, Waist circumference, Obesity                                   |
| rs11126666 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Hip circumference, Hand grip strength                                               |
| rs13078960 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Alzheimer's disease, C-reactive protein, Age-related macular degeneration, Cere
|            |        |        |      |       |            |     | brospinal fluid levels, Waist-hip ratio, Waist circumference, Longevity, LDL choles
|            |        |        |      |       |            |     | terol, Cognitive decline, Cognitive impairment score, Cerebral amyloid deposition |
| rs9914578  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs1000940  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs1057405  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs11847697 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs12446632 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs12566985 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs16907751 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs205262   | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs2650492  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs2836754  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs3849570  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs4787491  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs492490   | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Hip circumference                                             |
| rs7243357  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs9641123  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs10938397 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs10968576 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs11030104 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs11688816 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs12016871 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs13021737 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs13191362 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs13201877 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs1460676  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs1516725  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs1528435  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs17203016 | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs1726598  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs2287019  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs2820292  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs812091   | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs817334   | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs543874   | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs7164727  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs7599312  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |
| rs7899106  | ✓      | ✓      | ✓    | ✓     | ✓          | ✓   | Waist circumference, Fat-free mass                                                 |

3. A comparison of robust Mendelian randomization methods using summary data

**Figure 3.5** – *Mean squared error for the different methods in scenario 2 for 10,000 simulations, with directional pleiotropy and InSIDE satisfied with 10 variants.*

**Figure 3.6** – *Mean squared error for the different methods in scenario 3 for 10,000 simulations, with directional pleiotropy and InSIDE violated with 10 variants.*
Figure 3.7 – Mean squared error for the different methods in scenario 2 for 10,000 simulations, with directional pleiotropy and InSIDE satisfied with 100 variants.

Figure 3.8 – Mean squared error for the different methods in scenario 3 for 10,000 simulations, with directional pleiotropy and InSIDE violated with 100 variants.
II

Polygenic risk scores
A decade of research on the genetics of entrepreneurship: a review and view ahead

Cornelius A. Rietveld, Eric A.W. Slob, A. Roy Thurik

Abstract

Studies analyzing the heritability of entrepreneurship indicate that explanations for why people engage in entrepreneurship that ignore genes are incomplete. However, despite promises that were solidly backed up with ex-ante power calculations, attempts to identify specific genetic variants underlying the heritable variation in entrepreneurship have until now been unsuccessful. We describe the methodological issues hampering the identification of associations between genetic variants and entrepreneurship, but we also outline why this search will eventually be successful. Nevertheless, we argue that the benefits of using these individual genetic variants for empirical research in the entrepreneurship domain are likely to be small. Instead, the use of summary indices comprising multiple genetic variants, so-called polygenic risk scores, is advocated. In doing so, we stress the caveats associated with applying population-level results to the individual level. By drawing upon the promises of “genoeconomics”, we sketch how the use of genetic information may advance the field of entrepreneurship research.

This chapter is based on Rietveld et al. (2020).
4.1 Introduction

In 2000, the field of psychology concluded the nature-nurture debate to be “over” by posing that all human behavioral traits are heritable (Turkheimer, 2000). This “first law” of behavior genetics is backed by a vast body of literature comprising thousands of heritability studies (Polderman et al., 2015, Turkheimer, 2000). Since 2008, several studies have shown that this law also holds for entrepreneurship (Nicolaou et al., 2008a,b, Nicolaou and Shane, 2010, Shane and Nicolaou, 2015, Van der Loos et al., 2013, Zhang et al., 2009). Inspired by these findings and advances in genetics research, Koellinger et al. (2010) provided a sketchy forecast in this journal of the expected identification of relationships between genetic variants and entrepreneurship. Nevertheless, despite several attempts in the past decade (Nicolaou et al., 2011, Quaye et al., 2012, Van der Loos et al., 2011, 2013, Wernerfelt et al., 2012), no single robust association between a genetic variant and entrepreneurship has been found. Therefore, the first question we address in the present study is “Why has the identification of robust associations between genetic variants and entrepreneurship been unsuccessful in the last decade?” We answer this question from a methodological point of view. In doing so, we also provide a review of the literature in this field of research.

The second question we address is “Would the identification of associations between genetic variants and entrepreneurship help to advance the field of entrepreneurship research?” Despite the unsuccessful attempts so far, we provide methodological and empirical reasons for why we may expect the identification of the first robust associations between genetic variants and entrepreneurship in the not too distant future. Entrepreneurship scholars have argued that the prediction of entrepreneurial behavior using genetic data could have practical applications in business and for individual decision-making (Nicolaou et al., 2008a, Nicolaou and Shane, 2010, Shane, 2010). Moreover, several private companies already offer genetic tests to predict someone’s leadership and managerial qualities. We explain how summary indices of genetic variants (so-called polygenic risk scores) can be used for such prediction analyses, but by drawing on the broader behavior genetics literature, we stress the caveats associated with applying population-level results to the individual level. By relating the promises of “genoeconomics” as outlined by Benjamin et al. (2012a) to entrepreneurship research, we then sketch how we think the use of genetic information may advance the field of entrepreneurship research.

To illustrate the answers to our two research questions, we include an empiri-
4. A decade of research on the genetics of entrepreneurship: a review and view ahead

cal analysis of data from the US Health and Retirement Study. The inclusion of the empirical analyses in this study serves three purposes. First, the results of the analyses show how polygenic risk scores constructed for a range of traits (and not just entrepreneurship) can help to identify regions in the human genome particularly important for entrepreneurial behavior. Second, these analyses illustrate how polygenic risk scores can significantly predict entrepreneurship (even when proxied by the relatively episodic activity of self-employment). Third, we use these analyses to illustrate that the estimated relationships between polygenic risk scores and entrepreneurship at the population level only marginally improve the prediction of entrepreneurial behavior at the individual level.

In the following section, we review the studies providing evidence for the heritability of entrepreneurship. By exploiting family-based relationships rather than molecular genetic information, these studies show that approximately 40% of the differences in entrepreneurial behavior can be explained by genes. In Section 4.3, we review the molecular genetic analyses of entrepreneurship. We provide a comprehensive overview and discussion of the methodological approaches taken to identify relationships between genetic variants and entrepreneurship. Our empirical analyses are introduced and presented in Section 4.4. Finally, Section 4.5 concludes by discussing the added value of genetics for entrepreneurship research.

4.2 The heritability of entrepreneurship

Heritability is a technical term denoting the proportion of observed differences in a trait among individuals from a certain population that is due to the genetic differences among these individuals (Visscher et al., 2008). The main challenge in the estimation of heritability is the statistical separation of the effect of genes from the effect of the family environment on the trait of interest. One way to address this challenge is to compare adoptees with biological children. Using this approach, Lindquist et al. (2015) find that parental entrepreneurship increases the likelihood of children’s entrepreneurship by 60%. In their Swedish sample, they show that post-birth factors (i.e., adoptive parents) are two times more important than pre-birth factors (i.e., biological parents) for explaining entrepreneurial involvement.

Another, more common approach to separating the effect of genes from the effect of the family environment is the comparison of monozygotic and dizygotic twins reared together because the number of available twin samples is much larger than the available samples of adoptees (Knopik et al., 2016). Monozygotic twins are genetically identical; however, dizygotic twins are as genetically similar to each other as regular siblings. Under the assumption that monozygotic and
dizygotic twins are influenced by their family environment to the same extent, it is possible to decompose the variance in a trait into three components: the additive genetic effect, the common environment (family specific) effect, and the unique (individual specific) environment effect. Nicolaou et al. (2008a,b), Nicolaou and Shane (2010), Shane and Nicolaou (2015), Van der Loos et al. (2013), Zhang et al. (2009) use the classical twin study methodology to estimate the heritability of entrepreneurship in American, British, and Swedish samples. These studies draw on a broad range of empirical measures for entrepreneurship, such as self-employment and the number of start-up efforts, and provide general support for the heritability of entrepreneurship. Overall, the heritability estimates are in the neighborhood of 40%, indicating that almost one-half of the differences in entrepreneurship in these countries can be attributed to genetic differences across population members.

Although adoptee and twin studies can establish that genetic factors account for variation in a trait, they do not identify specific genes or the biological pathways through which genes function, because the genetic component is inferred from family relationships rather than observed in these studies. The completion of the sequencing of the human genome at the beginning of the present century (Venter et al., 2001) enabled the identification and measurement of locations in the human genome that differ among population members and hence led to the search for the specific genes underlying the heritable variation in entrepreneurship.

4.3 The molecular genetic analysis of entrepreneurship

4.3.1 The human genome

A complete human genome consists of 23 pairs of chromosomes, from which the 23rd pair determines the biological sex of an individual. One of each pair of chromosomes is inherited from the mother, and the other is inherited from the father. A chromosome is composed of two intertwined strands of deoxyribonucleic acid (DNA), each made up of a sequence of nucleotide molecules. There are four different nucleotide molecules in the DNA: adenine, cytosine, thymine, and guanine. Adenine on one strand is always paired with thymine on the other.

---

2 Nofal et al. (2018) provide a review of the literature about “biology and management”. Studies analyzing entrepreneurship are also included in this overview. All studies related to entrepreneurship in their category “Quantitative genetics” are discussed in this section (besides other studies). All entrepreneurship studies in their category “Molecular Genetics” are discussed in Section 4.3 (again, besides other studies).

3 Nicolaou et al. (2009) use an extended version of the classical twin study to show that the genes influencing the tendency to be an entrepreneur and the genes influencing opportunity recognition partially overlap.
strand, and cytosine is always paired with guanine. These combinations are called base pairs. Every human genome consists of approximately 3 billion base pairs. The stretches of base pairs in the DNA coding of a protein are called genes. There are approximately 20,000 genes in the human genome with varying lengths.

A random pair of individuals shares approximately 99.9% of their DNA (National Human Genome Research Institute, 2018b), and most genetic differences across population members can be attributed to single nucleotide polymorphisms (SNPs, pronounced “snips”). Therefore, behavioral genetics researchers focus primarily on SNPs when analyzing heritable genetic variation. A SNP is defined as a location in the DNA strand at which two different nucleotides are present in the population. Each of the two possible nucleotides is called an allele for that SNP. The allele that is least common in the population is called the minor allele; the other allele is called the major allele. For each SNP, an individual’s genotype is coded as 0, 1 or 2, depending on the number of minor alleles present. Individuals who inherited the same allele from each parent are called homozygous for that SNP (and have genotype 0 or 2), while individuals who inherited different alleles are called heterozygous (and have genotype 1). SNPs can be found in every part of the genome, within genes or in regions in between genes, and may influence the production of proteins.

In the human genome, there are approximately 85 million SNPs with a minor allele prevalence of at least 1% (The 1000 Genomes Project Consortium, 2015). When relating so many SNPs \( x_{ij} \) (coded as 0, 1, or 2) to a specific outcome \( y_i \) in a regression framework such as

\[
y_i = \mu + \sum_{j=1}^{J} \beta_j x_{ij} + \epsilon_i, \tag{4.1}
\]

with intercept \( \mu \), SNP effects \( \beta_j \) and residual term \( \epsilon_i \), it is evident that we have to deal with an overidentified model with fewer individuals \( I \) than SNPs \( J \) (Benjamin et al., 2012a).\(^4\) For this purpose, two basic approaches have been developed to deal with the overidentification problem. Hypothesis-driven methods such as the candidate gene approach do not consider all \( J \) SNPs, and hypothesis-free methods such as the Genome-Wide Association Study (GWAS) consider all \( J \) SNPs but not in one model. We continue by discussing these two basic approaches from a methodological point of view, and we review how they have been used for

---

\(^4\) Advanced statistical methods, such as GREML (genome-based restricted maximum likelihood), use two-step procedures to jointly estimate the explained variance of all SNPs (Yang et al., 2010). With this method, Van der Loos et al. (2013) show that all SNPs in their sample explain 25% of the variance in entrepreneurship. However, such approaches do not identify which individual SNPs are associated with the outcome variable.
unravelling the genetic architecture of entrepreneurship.

4.3.2 *Hypothesis-driven approaches*

The candidate gene approach consists of testing a subset of genetic variants for association with the outcome of interest. These genetic variants are selected based on what is known or believed about their biological function (Benjamin et al., 2012a,b, Ebstein et al., 2010). This approach resembles the classic way of justifying and then testing a hypothesis. A clear advantage of this approach is that the interpretation of revealed significant relationships is relatively straightforward. Adopting this approach, Nicolaou et al. (2011) were the first to report an association between a SNP in the DRD3 gene (a dopamine receptor gene) and entrepreneurial behavior in a British sample. Their selection of candidate SNPs was based on the observation that dopamine receptor genes have been associated with novelty seeking/sensation seeking and attention deficit hyperactivity disorder (ADHD). These traits were reported to be particularly prevalent among entrepreneurs (Nicolaou et al., 2008b, Antshel, 2017). Unfortunately, Van der Loos et al. (2011) failed to replicate this association in a Dutch sample seven times larger than the sample Nicolaou et al. (2011) drew upon.

This non-replication is exemplary for candidate gene studies (Benjamin et al., 2012a,b, Ioannidis, 2005, Rietveld et al., 2014a). In principle, a theoretical framework guides empirical research in reducing the number of hypotheses being tested. However, the analytical rigor that a theory-guided approach provides is not helpful in the context of behavioral genetics because it is difficult to reduce the number of plausible hypotheses purely on theoretical grounds. For instance, 70% of all genes (thus approximately 14,000) are expressed in the brain (Ramsköld et al., 2009), and for many of these genes (and hence the SNPs within these genes), a seemingly plausible relation between genes and behavior — including entrepreneurship — could be hypothesized ex ante. As a matter of fact, in 2012, the editor of the leading field journal *Behavior Genetics* issued an editorial policy on candidate gene studies of behavioral traits that reads “The literature on candidate gene associations is full of reports that have not stood up to rigorous replication” and went on to say “…it now seems likely that many of the published findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge” (Hewitt, 2012). This editorial policy outlines strict quality criteria that candidate gene studies must meet to be considered for publication. Most importantly, the editors stressed the importance of sufficient statistical power in genetic discovery studies (Hewitt, 2012).

Statistical power refers to the probability of rejecting the null hypothesis when it is not true. Statistical power of 80% or higher is generally considered
to be adequate (Ellis, 2010). Low statistical power results in a high chance of false negatives, i.e., non-rejections of the null hypothesis when the alternative hypothesis is true. Even more problematic, because of the winner’s curse, low statistical power also results in the overestimation of effect sizes for significant findings (Benjamin et al., 2018, Button et al., 2013, Wacholder et al., 2004). Statistical power is (among other things) a function of the effect size (of the SNP), the size of the analysis sample, and the significance level adopted. Nicolaou and Shane (2010) report that their identified SNP explained 0.5% of the likelihood of being an entrepreneur. With their sample of 1,335 individuals, they had only 6% power to detect such an effect at $p < 0.05$. Hence, it is not surprising that this finding could not be further replicated (Van der Loos et al., 2013).

4.3.3 Hypothesis-free approaches

Genome-wide association studies

GWAS is a hypothesis-free approach to genetic discovery because no prior selection is made on the set of SNPs used in the analysis. To deal with the overidentification problem, a GWAS runs a single regression for every SNP. Hence, millions of regressions are performed in a GWAS. An advantage of the hypothesis-free study design of GWAS is that it makes the need to correct for multiple testing transparent. If the null hypothesis of no association is true for all these millions of SNPs, one still finds a $p$-value $< 0.05$ for 5% of the SNPs. Therefore, in a GWAS, the significance threshold is set to $0.05/1,000,000 = 5 \times 10^{-8}$ (“genome-wide significance”) because of the approximately 1 million independent SNPs in the human genome (adjacent SNPs in the genome are often inherited together). A clear disadvantage of this approach is that GWASs may prioritize SNPs for which the biological function is yet unknown or unclear. Hence, GWAS usually identifies SNPs that need to be subjected to further analyses to understand the pathways between the SNPs and the outcome. Close collaboration with geneticists and biologists in consortia, such as the Gentrepreneur Consortium (Van der Loos et al., 2012), is important.

In their analysis, Nicolaou and Shane (2010) adopted a significance level of $6 \times 10^{-4}$ to account for the correlation between SNPs. As a result, the power of their analysis was almost zero. To be adequately powered (80%), one would have needed a sample of 3,643 individuals to find an effect of 0.5% (at $p = 6 \times 10^{-4}$).

The working paper by Wernerfelt et al. (2012) reports an association between a genetic polymorphism and entrepreneurship (proxied by the number of companies founded) in a sample of 135 participants of an executive education course at Harvard Business School. It is evident that in such a sample, the same concerns about statistical power hold.

Relatedly, GWAS models usually use a very small number of control variables to capture the full relationship between the SNP and the outcome. For example, Van der Loos et al. (2013) control for only sex, age, and genetic relatedness in their GWAS on self-employment. The use of a small number of control variables causes the interpretation of the estimated effects to be not as straightforward because there may be many pathways through which a SNP influences a behavioral outcome.
The combination of a very stringent significance level and the small effect sizes of individual SNPs implies that large samples are needed to be adequately powered for gene discovery. The typical dataset has only several thousands of observations, and therefore, datasets need to be combined into mega-analyses or meta-analyses. In a mega-analysis, individual-level genetic data are merged and jointly analyzed. However, legal and privacy issues generally make it impossible to pursue this strategy. In a meta-analysis, the summary results of specific analyses are combined. The GWAS meta-analysis approach has enabled an unprecedented surge in genetic discoveries that are consistently replicated (Welter et al., 2014, Visscher et al., 2017b), including the discovery of genetic associations with behavioral outcomes such as educational attainment (Lee et al., 2018, Okbay et al., 2016b, Rietveld et al., 2013), subjective well-being (Okbay et al., 2016a), and more recently preferences such as attitudes toward risk-taking (Linnér et al., 2019). The large sample sizes in these studies \((N = 1,000,000\) in some of them) could be obtained due to the dramatic decline in the cost of genotyping in the last decade (National Human Genome Research Institute, 2018a).

In 2010, Koellinger et al. (2010) calculated that at least 30,000 observations were needed to find a relationship between an individual genetic variant and entrepreneurship at the genome-wide significance level. Quaye et al. (2012) used the GWAS approach in a sample of 3,933 British females to assess whether there are associations between specific SNPs and entrepreneurship. Not surprisingly, because of the small sample size, they did not find SNPs that are significant at the genome-wide significance level. Van der Loos et al. (2013) conducted a large-scale GWAS meta-analysis on entrepreneurship in a combined sample of 53,898 individuals from Europe and the US. Despite the sample size, this study did not find any genome-wide significant SNPs. Moreover, this study found no evidence that any of the genes that were previously suggested in the literature to influence entrepreneurship (Shane, 2010) show significant associations with entrepreneurship. From a statistical point of view, this null-result could have been driven by the attenuation of the effect sizes through the meta-analysis of samples from different countries and with different birth year profiles. However, GWASs from the past few years on other behavioral outcomes indicate that the effect sizes used in the power calculations by Koellinger et al. (2010) were too high.

The past years of research in behavioral genetics showed that individual SNPs typically explain less than 0.02% of the variance in a behavioral outcome (Chabris et al., 2010) and the Social Science Genetic Association Consortium\(^8\), is therefore a prerequisite for the success of GWAS analysis.

\(^8\)https://www.thessgac.org/.
et al., 2015, Rietveld et al., 2014a). These findings imply that a sample of at least 197,984 individuals is needed to identify a SNP at the genome-wide significance level with 80% power. Hence, by now, we know that the GWAS meta-analysis of Van der Loos et al. (2013) was underpowered. Although the availability of genetic data is rapidly increasing, genetic data are collected primarily for medical purposes, and measures for entrepreneurship are not always available in medical datasets. There is progress in the collection of genetic data in surveys with an economic focus (such as the US Health and Retirement Study and the English Longitudinal Study of Ageing), but at this moment, a sufficiently large analysis sample for a GWAS on entrepreneurship is not available.

Nevertheless, the heritability estimates for entrepreneurship and the successful discovery of SNPs related to other behavioral outcomes indicate that we can be confident about the eventual success of a GWAS on entrepreneurship. Visscher et al. (2017b) showed that the number of identified genetic associations in a GWAS is positively related to the size of the (meta-) analysis sample. For example, whereas the first GWAS meta-analysis on educational attainment (N ≈ 100,000) found only three genome-wide significant SNPs (Rietveld et al., 2013), the second one (Okbay et al., 2016b) identified 74 SNPs (N ≈ 300,000), and the third one (Lee et al., 2018) identified 1,271 SNPs (N ≈ 1,100,000). Hence, a GWAS with a sufficiently large sample size — at least four times larger than the sample of ~ 50,000 individuals used by Van der Loos et al. (2013) — will also reveal the SNPs that are associated with entrepreneurship.

**Genetic discovery using proxy traits**

A novel way to boost statistical power in GWASs is the identification of genetic associations using a two-step procedure in the so-called proxy-phenotype method. Rietveld et al. (2014b) introduced this approach to identify genetic associations with cognitive performance. Similar to entrepreneurship, cognitive performance is not often measured in genotyped samples. Therefore, the first step in this method is conducting a large-scale GWAS on a genetically related trait. In the second step, the genetic variants associated with this proxy trait are tested for association with the main trait of interest. In this spirit, Rietveld et al. (2014b) used the results of a GWAS on educational attainment to select 69 independent SNPs, which were then tested for association with cognitive performance. The significance threshold adopted in the second step equals α = 0.05/69 rather than the genome-wide significance threshold of α = 5 × 10^{-8}.

Linnér et al. (2019) used this approach in their GWAS on risk tolerance to study the genetic architecture of related traits, such as self-employment. Based on their main GWAS on risk tolerance, 99 SNPs were selected for further analysis.
regarding their association with entrepreneurship. In the second stage, the discovery GWAS \((N = 50,627)\) results of Van der Loos et al. (2013) were used. Using a more lenient threshold for significance, Linnér et al. (2019) found one SNP that was significantly associated with entrepreneurship. The sign of the effect was in the expected direction, meaning that the SNP was related to higher risk tolerance and a higher likelihood of being an entrepreneur. Linnér et al. (2019) claimed in their supplementary materials that "if the association with rs7387531 is robust, this would be the first genetic variant to be found to be significantly associated with self-employment." However, in the replication sample \((N = 3,271)\) of Van der Loos et al. (2013), the effect of the SNP (rs7387531) was in the opposite direction with \(p > 0.05\), so it seems that the first robust association between a SNP and entrepreneurship is yet to be identified. Nevertheless, this approach illustrates that the genetic analysis of related traits may help to find genetic variants associated with entrepreneurship.

### 4.3.4 Polygenic risk scores

Individual SNPs typically explain less than 0.02% of the variance in a behavioral outcome (Chabris et al., 2015), and the GWAS on self-employment by Van der Loos et al. (2013) has shown that the effects of individual SNPs on entrepreneurship are also small (otherwise they would have been found). Hence, individually, genetic variants are practically useless for use in empirical studies. However, the tiny explanatory power of individual genetic variants has encouraged researchers to develop methods that combine individual genetic variants into so-called polygenic risk scores with larger explanatory power. A polygenic risk score is a weighted sum of SNPs and is constructed as follows:

\[
P_{\text{GS}}_i = \sum_{j=1}^{J} \beta_j x_{ij},
\]

where \(P_{\text{GS}}_i\) is the value for the polygenic risk score for individual \(i\), \(\beta_j\) is the regression coefficient of SNP \(j\) from the GWAS, and \(x_{ij}\) is the genotype of individual \(i\) for SNP \(j\) (coded as 0, 1 or 2). This simple approach has been proven to be effective in the out-of-sample prediction of behavioral outcomes. For example, Rietveld et al. (2013) found only three SNPs significantly associated with educational attainment at the genome-wide significance level. Each SNP ex-

---

9 More advanced methods for constructing polygenic risk scores exist, for example, methods that better deal with the correlation structure across SNPs within the genome (see, e.g., So and Sham (2017) and Vilhjálmsson et al. (2015)). However, the main rationale behind these methods is similar to the basic (still commonly used) approach presented in the main text.
explained approximately 0.02% of the variance in educational attainment. However, the polygenic risk score based on all SNPs (including the non-significant ones) explained approximately 2.5% of the variance. This percentage increased with the sample size of the GWAS. For example, the most recent polygenic risk score for educational attainment now explains 9.4% (Lee et al., 2018). The prediction attempt of Van der Loos et al. (2013) was unsuccessful in the sense that their polygenic risk score for entrepreneurship captured less than an insignificant 0.2% of the variance. Nevertheless, this percentage will increase if the GWAS for entrepreneurship increases in terms of sample size (Dudbridge, 2013).

The weights $\beta_j$ used in the calculation of the polygenic risk score capture almost the full relationship between the SNP and entrepreneurship: the only control variables used in the GWAS on self-employment by Van der Loos et al. (2013) are sex, age, and variables to account for genetic relatedness between individuals. The relationship between someone’s genetic makeup and behavior is assumed to be extremely complex and to run through many (possibly also multiplicative) pathways. Therefore, a “direct” relationship between a SNP and entrepreneurship is unlikely to exist. Many pathways, possibly comprising gene-gene and gene-environment interactions, are likely to explain the relationship between a SNP and behavior. Nevertheless, in a GWAS, these pathways are all included in $\beta_j$ and therefore also in the polygenic risk score. In the spirit of the proxy-phenotype approach used in GWAS (see Section 4.3.3), we can therefore use the polygenic risk scores of traits that we think are in the pathway between some SNPs and entrepreneurship to foster our understanding about the genetic architecture of entrepreneurship.

One obvious example of such a pathway is risk tolerance. The recent GWAS by Linnér et al. (2019) on risk tolerance shows how the polygenic risk score for risk tolerance does indeed predict entrepreneurship out of sample. Although the explanatory power of this polygenic risk score is relatively small, between 0.57 and 1.36 in terms of (pseudo-) $R^2$ for different proxies of entrepreneurship, it contributes significantly to the fit of the model. Moreover, the variance explained is already larger than we may expect it to be for individual SNPs. Risk tolerance may be an obvious trait to investigate when analyzing the pathway between SNPs and entrepreneurship. However, other less obvious traits may also be investigated. For example, earlier research shows that body height is associated with entrepreneurship (Rietveld et al., 2015). The newest polygenic risk score for height explains approximately 34.7% of the variance (Yengo et al., 2018). If the effect of the SNPs explaining entrepreneurship runs through height, we will be able to find an association between the polygenic risk score for body height and entrepreneurship.
Hence, polygenic risk scores constructed for traits other than entrepreneurship may help to identify regions in the human genome that are related to entrepreneurship. Moreover, these genetic summary indices may facilitate the gene-based prediction of entrepreneurship. In the next section, we present empirical analyses that illustrate these two conclusions.

4.4 Empirical Illustration

For our empirical illustration, we draw on data from the US Health and Retirement Study. The HRS is a representative panel of Americans over 50 years old and their spouses. The HRS focuses on a variety of labor markets and health and retirement outcomes. Genetic data were collected from consenting HRS participants between 2006 and 2012 (Health and Retirement Study, 2012). We use the RAND HRS Longitudinal File 2014 (V2) for the data on self-employment (Health and Retirement Study, 2018). This longitudinal data file includes the harmonized biennial data of the HRS (1992-2014). Our dependent variable indicating whether an individual is self-employed or not is derived from the question: “Do you work for someone else, are you self-employed, or what?” The respondents could answer “for someone else” or “self-employed”. If respondents said they were self-employed, they were coded as 1, and if they replied that they worked for someone else, they were coded as 0. Self-employment is the most commonly used measure for entrepreneurship studies drawing on survey data (Parker, 2018), although engagement in self-employment can be episodic. We restrict our analyses to those aged between 50 and 65 to exclude individuals active in the labor market after retirement age. Moreover, following the recommendations of the genotyping center, we restrict the analysis to individuals of recent European descent to preempt bias from unobserved relationships between genetic and environmental factors (Health and Retirement Study, 2012).

For the polygenic risk scores, which are the main independent variables in our regressions, we use the HRS Polygenic Scores 2006-2012 Genetic Data - Release 3 (Health and Study, 2018). In the present illustrative analyses, we use all available polygenic risk scores in this file that relate to mental health. We choose to limit ourselves to the polygenic risk scores of only these traits, as the recent entrepreneurship literature suggests an important link between entrepreneurship and mental health in terms of person-job fit (Benz and Frey, 2008, Stephan, 2018). In total, we analyze 16 different polygenic risk scores. In some polygenic risk scores, there are multiple versions, reflecting the publication of increasingly large GWAS studies on these traits. In these cases, we use the newest polygenic risk score. For some other traits, there are separate scores for males, females and the combined sample of males and females. In these cases, we use the combined score.
our analyses, we control for sex, birth year (dummies for each birth year), and survey waves (dummies for each survey wave). We also control for the first ten principal components of the genetic relationship matrix, as is common in genetic association studies. The latter ten variables control for the genetic aspects of common ancestry that could be spuriously correlated with the polygenic risk scores and the outcome of interest, such as cultural or environmental factors (Rietveld et al., 2014a). To estimate the relationships between self-employment and the polygenic risk scores, we use a linear probability model with random effects (to deal with the time-invariant nature of the polygenic risk scores as well as the longitudinal nature of our data):\(^\text{11}\)

\[
SE_{it} = \sum_{k=1}^{K} \gamma_k PGS_{ik} + \delta Z_{it} + \alpha_i + \epsilon_{it},
\]

(4.3)

where \(SE_{it}\) is the binary variable indicating the self-employment status of individual \(i\) at time \(t\), \(\gamma_k\) is the effect of the polygenic risk score \(PGS_{ik}\) for trait \(k\), \(\delta\) is a vector of coefficients for the vector of control variables \(Z_{it}\), \(\alpha_i\) is an unobserved random variable for individual \(i\), and \(\epsilon_{it}\) is the residual for individual \(i\) at time \(t\).\(^\text{12}\)

Overall, 31,927 (person-year) observations are available from 7,948 different individuals. In this sample, 47% of the individuals are male, the average age is 57.4 years (with standard deviation 4.1), and 19.9% of the person-year observations report self-employment. Table 4.1 displays the estimates of the associations between the different polygenic risk scores and self-employment. We observe that there are six (out of 16) significant associations at the 5% level: the polygenic risk scores for ADHD, autism, bipolar disorder, educational attainment, general cognition, and well-being.\(^\text{13}\) For these traits, an increase of one standard deviation leads to an increase or decrease in the likelihood of being self-employed of approximately 1%. These results indicate that polygenic risk scores can significantly predict entrepreneurship (even when proxied by the relatively episodic

\(^{11}\) We present the results of a linear probability model despite the binary nature of our dependent variable because the interpretation of the regression coefficients in a linear probability model with random effects is more straightforward than in a logit model with random effects. However, we note that this choice does not affect our results from a qualitative point of view. In a logit model with random effects, ADHD, autism, bipolar disorder, educational attainment, and cognition are still significant at \(p < 0.05\). However, the \(p\)-value for well-being (0.062) is slightly above the significance threshold.

\(^{12}\) In the analysis, we estimate the effect of several polygenic risk scores in one single model. As some traits are genetically correlated, such as ADHD and bipolar disorder (Faraone and Larsson, 2019), we also analyze models in which we separately include the polygenic risk scores. From a qualitative point of view, the results are similar to the results presented in the main text.

\(^{13}\) Even with a stringent Bonferroni correction (0.05 divided by the number of polygenic risk scores analyzed), the association with ADHD remains significant.
activity of self-employment) and that genes influencing entrepreneurship are likely to be found in regions in the human genome associated with these six traits.\footnote{For illustration purposes, we analyzed all available mental health related polygenic risk scores in the Health and Retirement Study in the present study. The set of polygenic risk scores includes traits for which the link with entrepreneurship in not always evident. Therefore, future studies may use theoretical or other insights for selecting the most promising candidates from the set of available polygenic risk scores rather than using them all. However, the fact that ADHD is found to be the strongest association in our analyses builds confidence in our approach since there are several nongenetic studies showing a similar link (Verheul et al., 2015, 2016, Antshel, 2017, Wiklund et al., 2017, Lerner et al., 2019). Nevertheless, future studies need to replicate the current findings in independent datasets to investigate their robustness and generalizability.}

**TABLE 4.1 – The association between the polygenic risk scores for traits in the mental health domain and self-employment (random-effects regression, \(N_{\text{individual-year}} = 31,927, N_{\text{individual}} = 7,948\)).**

<table>
<thead>
<tr>
<th>Polygenic risk score</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention deficit hyperactivity disorder</td>
<td>0.017</td>
<td>0.004</td>
<td>0.000</td>
</tr>
<tr>
<td>Anxiety (factor score)</td>
<td>0.001</td>
<td>0.004</td>
<td>0.790</td>
</tr>
<tr>
<td>Autism</td>
<td>-0.013</td>
<td>0.006</td>
<td>0.040</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>0.010</td>
<td>0.005</td>
<td>0.047</td>
</tr>
<tr>
<td>Depressive symptoms</td>
<td>0.007</td>
<td>0.005</td>
<td>0.187</td>
</tr>
<tr>
<td>Educational attainment</td>
<td>0.013</td>
<td>0.005</td>
<td>0.004</td>
</tr>
<tr>
<td>Extraversion</td>
<td>0.007</td>
<td>0.004</td>
<td>0.100</td>
</tr>
<tr>
<td>General cognition</td>
<td>-0.012</td>
<td>0.005</td>
<td>0.010</td>
</tr>
<tr>
<td>Major depressive disorder</td>
<td>-0.005</td>
<td>0.005</td>
<td>0.367</td>
</tr>
<tr>
<td>Mental health (cross disorder)</td>
<td>-0.004</td>
<td>0.007</td>
<td>0.358</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.008</td>
<td>0.006</td>
<td>0.202</td>
</tr>
<tr>
<td>Obsessive compulsive disorder</td>
<td>-0.001</td>
<td>0.004</td>
<td>0.752</td>
</tr>
<tr>
<td>Post-traumatic stress disorder</td>
<td>0.001</td>
<td>0.005</td>
<td>0.860</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>0.005</td>
<td>0.008</td>
<td>0.509</td>
</tr>
<tr>
<td>Well-being</td>
<td>0.010</td>
<td>0.005</td>
<td>0.032</td>
</tr>
</tbody>
</table>

Notes: The regression model includes control variables for sex, age, survey waves, and genetic relatedness. Underlined traits are significant at the 5% level.

At the same time, these results illustrate that the predictive power of these polygenic risk scores is small (although larger than the predictive power of individual SNPs). Compared to that of a model without the polygenic risk scores, the explained variance of this model increased by only 0.42%.\footnote{Individual SNPs typically explain less than 0.02% of the variance in a behavioral outcome (Chabris et al., 2015, Rietveld et al., 2014a).} Table 4.2 shows that, from a prediction point of view (by taking the percentage of person-year observations in our sample in self-employment 19.9% as the classification threshold), the correct individual-level prediction of self-employment status increases only marginally with the current model (0.14% point increase).

### 4.5 Conclusion: a second decade?

The “quest for the entrepreneurial gene” (Van der Loos et al., 2011) is largely motivated by the struggle of scholars to have a better understanding of entrepreneurs
Table 4.2 – In-sample prediction results for self-employment (versus wage work) for the models with and without polygenic risk scores; observations in the top 19.9% (percentage of person-year observations reporting self-employment in the sample) of the predicted values in each model are classified as self-employed.

<table>
<thead>
<tr>
<th>Actual occupation</th>
<th>Predicted occupation based on model without polygenic risk scores</th>
<th>Predicted occupation based on model with polygenic risk scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self-employment</td>
<td>Wage work</td>
</tr>
<tr>
<td>Self-employment</td>
<td>5.75%</td>
<td>14.11%</td>
</tr>
<tr>
<td>Wage work</td>
<td>14.10%</td>
<td>66.04%</td>
</tr>
</tbody>
</table>

and entrepreneurship: what makes entrepreneurs decide to start a business, what motivates them, what makes them successful or fail, and what makes them different from other people? Various research approaches, as well as tools and theories from economics, psychology, and sociology, have been proposed and applied to these questions. However, the answers to “what makes an entrepreneur” remain uncertain and incomplete (Shane and Venkataraman, 2000, Parker, 2018). Empirical evidence that genes may be part of the answer (Nicolaou et al., 2008a,b, 2009, 2011, Shane and Nicolaou, 2015, Van der Loos et al., 2011, 2013, Zhang et al., 2009) has been received by scholars and the media with both hopes and enthusiasm, as well as with skepticism and criticism.

Despite several attempts in the past decade, until now, no robust association between genetic variants and entrepreneurship has been discovered. Our overview and discussion of these works gives a clear answer to our first research question, “Why has the identification of robust associations between genetic variants and entrepreneurship been unsuccessful in the last decade?” Irrespective of whether a hypothesis-driven or hypothesis-free approach was used, genetic discovery studies on entrepreneurship have until now been underpowered. Nevertheless, based on the results of large-scale genetic discovery studies on other behavioral traits (such as educational attainment), we may expect that robust associations between genetic variants and entrepreneurship will be identified if a sufficiently large sample can be gathered. Datasets that contain both genetic data and entrepreneurship information are relatively scarce (Van der Loos et al., 2013), but the advent of large genotyped biobanks such as the UK Biobank (Bycroft et al., 2018) and the Estonian Biobank (Leitsalu et al., 2015) is currently changing the landscape. Hence, a sufficiently powered GWAS on entrepreneurship may soon become feasible.

Because of data constraints, the latest and largest GWAS on entrepreneurship used self-employment as a proxy for entrepreneurship (Van der Loos et al., 2013). With more data becoming available, future GWASs of entrepreneurship may benefit from the analysis of an entrepreneurship measure less episodic in nature, such
as serial or high-performance entrepreneurship. With more precise classification of individuals into occupational groups, the GWAS becomes more powerful and hence the chance to detect associations between individual genetic variants and entrepreneurship becomes larger. Nevertheless, in combination with other GWAS results, the analysis of the relatively heterogeneous self-employment measure may help identify specific underlying types of self-employment. For example, by drawing on GWAS results for schizophrenia and educational attainment, Bansal et al. (2018) reveal that the binary schizophrenia diagnosis aggregates over at least two different subtypes. The first type is associated with high intelligence and bipolar disorder, while the second type is a cognitive disorder that is independent of bipolar disorder. With GWAS results for many publicly available traits, similar analyses may also be interesting to conduct on self-employment to possibly identify unexpected subtypes.

However, rather than directly analyzing entrepreneurship, it is possible to shift attention (at least for the time being) to variables mediating the relationship between genes and entrepreneurship. Examples of such variables that can be measured in large samples include traits such as preferences for risk and uncertainty, confidence, and optimism. In addition to these well-known measures in the world of entrepreneurship research, one may also consider characteristics such as body height, body mass index, and mental disorders (possibly in a hypothesis-free setting). One advantage of this approach is that genetic effects on more proximate outcomes are likely to be stronger and hence easier to detect, for a given sample size, than the genetic effects on distal outcomes, such as entrepreneurship (Rietveld et al., 2014b). By using the proxy-phenotype approach, as discussed in subsection 4.3.3, it will be possible to identify associations with entrepreneurship, for example, by using the (publicly available) GWAS results of Van der Loos et al. (2013) in the second step of the analysis. This approach circumvents to some extent the problem of the currently insufficient sample size needed for a well-powered GWAS on entrepreneurship.

Although a regular GWAS looks only at the linear association between a genetic variant and entrepreneurship, the genetic architecture of entrepreneurship may comprise interactions between two or more genetic variants. Theoretically, it is possible to include cross-products of SNPs as explanatory variables in a GWAS to advance our understanding of the possibly complex biological mechanisms that are associated with entrepreneurship. However, in a hypothesis-free setting, such an approach would also require an even more stringent correction of the significance level (as the number of statistical tests increases exponentially with

---

16 For example, in the GWAS Catalog (https://www.ebi.ac.uk/gwas/).
17 The results of the GWAS on self-employment by Van der Loos et al. (2013) are publicly available via www.thessgac.org.
the number of interacting SNPs). Hence, if we assume the size of the interaction effects is not larger than the effects of individual SNPs, this approach is unlikely to be productive in the distant future because of data limitations. The interaction effect may also be identified with (nonlinear) machine learning techniques. Relatively simple machine learning techniques have been proven to have relatively high predictive power for traits such as human height (Pare et al., 2017, Lello et al., 2018). Despite the massive computational burden of these methods, it is promising to analyze to what extent these techniques are also useful for predicting entrepreneurship. Nevertheless, the biological interpretation of the results obtained with machine learning techniques is arguably even more difficult than that of results obtained with a regular GWAS.

To answer our second research question, “Would the identification of associations between genetic variants and entrepreneurship help to advance the field of entrepreneurship research?”, we relate the promises of “genoeconomics”, as outlined by Benjamin et al. (2012a), to entrepreneurship research in light of the recent development in behavioral genetics. Benjamin et al. (2012a) outlined four main reasons why the genetic analysis of behavioral traits is important and relevant. First, studies using directly observed genes may reveal the genetic pathways and mechanisms underlying behavior and may lead to a more complete understanding of entrepreneurial behavior. For example, as already discussed above in light of the findings of Bansal et al. (2018), it may be possible to identify to what extent different mechanisms and cognitive processes are involved in the identification and exploitation of business opportunities. Second, these studies have the potential to provide measures for constructs that are difficult to measure empirically. Benjamin et al. (2012a) use the example that specific genetic variants can be used as a proxy for the taste for fatty foods. In this spirit, rather than using self-reported measures for entrepreneurial intention, one could draw on the genes related to entrepreneurship. Third, based on someone’s genetic profile, interventions may be channeled. In this vein, entrepreneurship scholars argue that the prediction of entrepreneurial behavior using genetic data could have practical applications in business and for individual decision-making (Nicolaou et al., 2008a, Nicolaou and Shane, 2010, Shane, 2010). Fourth, genes can be used to enrich otherwise nongenetic models. For example, the inclusion of control variables for genetic endowments may absorb the residual variance in regression models or experimental settings and allow for stronger statistical inference (DiPrete et al., 2018b, Rietveld and Webbink, 2016). In some instances, it will also be possible to infer causal relationships in observational data by using genes as instrumental variables (Van Kippersluis and Rietveld, 2018, Von Hinke et al., 2016). Hence, the use of genes may be instrumental for better understanding the
effects of environmental factors.

Regarding the first two promises, we have seen that for behavioral outcomes (such as entrepreneurship), one should not expect values of $R^2$ in excess of 0.02% for individual SNPs. Hence, it is unlikely that such a SNP will provide much information about the mechanisms underlying entrepreneurship behavior. In contrast to focusing on individual genetic variants, there are good arguments for shifting our attention to polygenic risk scores that summarize the contribution of several genetic variants to a trait. A clear advantage of this approach is that polygenic risk scores can be used as regular variables in empirical research, and expertise for working with raw genetic data is not necessary, as some polygenic risk scores are already publicly available (such as in the HRS). In the present absence of a polygenic risk score for entrepreneurship with significant explanatory power, we have to shift our focus to the analysis of polygenic risk scores for entrepreneurship-related traits. By doing so, we also come closer to the common practice in entrepreneurship research of testing particular hypotheses (i.e., particular pathways through which genes influence entrepreneurship). For example, we may hypothesize and test whether the genetic variants contributing to the development of ADHD are also related to entrepreneurship. In this spirit, a polygenic risk score can also serve as a proxy for a trait. For example, Patel et al. (2019) use the polygenic risk score for ADHD to study the influence of ADHD on entrepreneurship and entrepreneurial performance in a sample of individuals for which the diagnosis of ADHD was not available.

Regarding the third and fourth promise (the use of genetic information to predict individual behavior and to enrich otherwise nongenetic models), the current state of the behavioral genetics literature as well as the analyses presented in the present study make clear that the added value of genetics for entrepreneurship scholars should be thought of in terms of enriching population-level models rather than improving individual-level prediction (Morris et al., 2019). Van der Loos et al. (2013) show that all SNPs together may explain up to 25% of differences in entrepreneurial behavior between individuals. Even if we are able to realize this prediction $R^2$, the likelihood of misclassification of individual into occupational groups remains great. Hence, early speculations about the use of molecular genetic data for understanding and predicting entrepreneurship (Shane, 2010) remain premature, at a minimum. Even though it may be useful to capture some of the (otherwise residual) variance in polygenic risk scores, the gene-based

---

18 There is currently an important initiative to make a repository of polygenic risk scores for several datasets. However, the exact time window of this initiative is unknown (Okbay et al., 2018). More (future) data sources can be found through portals such as the database of Genotypes and Phenotypes (dbGaP, Mailman et al. (2007)) and the European Genome-phenome Archive (EGA, Lappalainen et al. (2015)).
prediction of individual entrepreneurial behavior will remain of limited value for individuals and entities such as governments and banks.\textsuperscript{19}

Nevertheless, capturing residual variance in polygenic scores may improve the understanding of the effects of environmental factors. In so-called gene-by-environment (\textquotedbl{}GxE\textquotedbl{}) studies (Keller, 2014, Thompson, 2017), polygenic risk scores could be used to investigate how entrepreneurship results from the interplay between genetic endowments and environmental factors. For example, a recent study argues that cultural factors (as proxied by the taste for alcoholic drinks) may influence how genes shape different types of entrepreneurship (Acs and Lappi, 2019). In general, a good fit between individuals and their occupations has been shown to be important for high levels of productivity (Kristof-Brown et al., 2005). Importantly, the identifiable occurrence of matches and mismatches between an individual and his or her career choices and the possible impact on stress and health was a crucial argument for the medical profession to cooperate with behavioral researchers in the search for the genes associated with entrepreneurship (Koellinger et al., 2010, Van der Loos et al., 2010). Because of the large-scale collections of genetic data and expertise on the biological functioning of genes in the medicine and biology fields, the involvement of researchers in these fields will remain crucial to find associations between genetic variants and entrepreneurship.

In sum, although the attempts to identify specific genetic variants underlying the heritable variation in entrepreneurship have until now been unsuccessful, there is reason to be confident about the eventual success of the \textquoteleft quest for the entrepreneurial gene\textquoteright (Van der Loos et al., 2011). The benefits of using individual genetic variants for empirical research in the entrepreneurship domain are likely to be small. However, the use of polygenic risk scores may promote the realization of the promises of genoeconomics for entrepreneurship research. Although the gene-based prediction of individual entrepreneurial behavior will be of limited value, the use of polygenic risk scores in models may help to increase our understanding of which regions in the genome and which combinations of genetic endowments and environmental circumstances drive entrepreneurship and person-job fit at the population level.

\textbf{Acknowledgements}

The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

\textsuperscript{19Besides, ethical considerations are needed to determine whether such gene-based prediction of entrepreneurship is actually desirable.}
Does the genetic predisposition to smoking moderate the response to tobacco excise taxes?
Eric A.W. Slob and Cornelius A. Rietveld

Abstract

Tobacco use is one of the leading causes of preventable death. While some public policies have been effective in reducing the smoking prevalence in the United States, high tobacco excise taxes do not appear to deter all individuals from starting smoking nor to make all smokers stop. Here, we analyze whether someone’s genetic predisposition to smoking may explain why people smoke despite high tobacco excise taxes. For this purpose, we interact polygenic risk scores for smoking initiation and smoking intensity with state excise tax rates on tobacco. Our analyses exploiting longitudinal data (1992-2014) from the US Health and Retirement Study show that someone’s genetic propensity to smoking moderates the effect of tobacco excise taxes on smoking behaviour, but only along the extensive margin (smoking vs. not smoking). The results along the intensive margin (the amount of tobacco consumed) are inconclusive.

This chapter is currently under review at a journal.
5.1 Introduction

Tobacco use is the leading preventable cause of death in the world, with over 7 millions deaths per year (World Health Organization, 2017). In the United States, over 480,000 deaths per year are attributable to smoking (US Department of Health and Human Services, 2014). Tobacco use has been shown to be quite addictive and hence, quitting is often a tough battle characterized by heavy withdrawal symptoms (Benowitz, 2008). As a prime instrument to influence smoking behaviour, governments impose excise taxes on tobacco. Over the past 50 years, the median price of cigarettes has increased from 0.30$ per pack up to 5.70$ (US Department of Health and Human Services, 2014). In the same period, cigarette consumption per capita decreased from 4000 to about 1000 per year. Although this decrease cannot entirely be explained by the increase in tobacco excise taxes, as for example public awareness about the detrimental effects of smoking also increased in this period, there is considerable evidence about the effectiveness of raising tobacco excise taxes for reducing smoking (Chaloupka and Warner, 2000, Institute of Medicine, 2007, MacLean et al., 2016). However, the decrease in smoking consumption has stalled in the past 20 years (Orzechowski and Walker, 2016).

Tobacco excise taxes are identical for each member of a society, and a possible explanation for the stabilizing smoking prevalence may be that for some individuals it is more difficult than for others to stop smoking. For example, studies have shown that demand elasticities for tobacco differ between males and females (Yen, 2005) and across ethnicities (Kandel et al., 2004). Moreover, behavioural preferences such as risk aversion (Barsky et al., 1997, Anderson and Mellor, 2008) and someone’s health status influence smoking behaviour (Jones, 1994, Lahiri and Song, 2000, Clark and Etilé, 2002). There is also clear evidence that heavy smokers react differently to tobacco excise taxes than less heavy smokers (Nesson, 2017), although the precise mechanism explaining these elasticity differences is not known. In the present study, we analyze whether someone’s genetic predisposition to smoking moderates the response to tobacco excise taxes.

Several studies have shown that the heritability of smoking behaviour ranges between 31-60 % (Bidwell et al., 2016), indicating that genes explain a considerable proportion of the variation in smoking in a population possibly through their effect on nicotine dependency. It has also been shown that environmental circumstances such as state policies impact the heritability of smoking: The heritability of smoking is relatively low in states with relatively high excise taxes on tobacco and in those with greater controls on cigarette advertising and the vending machines (Boardman, 2009). Recent large-scale genetic association studies have found more than 500 genetic variants underlying the heritable
variation in smoking behaviour (Erzurumluoglu et al., 2019, Liu et al., 2019). Fletcher shows that individuals carrying one of these genetic variants respond differently to excise tobacco taxes than those not carrying this genetic variant (Fletcher, 2012). Hence, such a gene-environment interaction may explain why certain individuals smoke and others do not.

However, a follow-up study by Fontana (Fontana, 2015) using the same genetic variant shows that Fletcher's gene-environment interaction could be a spurious association explained by the effects of population stratification. Population stratification entails an association between genetic subpopulations in a population and environmental conditions, such as cultural and social norms (Rietveld et al., 2014a). Besides, recent studies have shown that the predictive power of individual genetic variants is limited, often below 0.02% for behavioural outcomes including smoking (Chabris et al., 2015). Hence, low statistical power may be another reason for why Fontana (Fontana, 2015) could not replicate the results of Fletcher (Fletcher, 2012).

To deal with the limited predictive power of genetic variants, approaches have been developed to combine multiple genetic variants into a composite measure. The most often adopted approach is the construction of so-called polygenic risk scores (PGSs) (Dudbridge, 2013). To construct a PGS, all genetic variants in a sample are summed up in a weighted fashion in which each weight is proportional to the strength of the association between the genetic variant and an outcome variable as estimated in a genome-wide association study (GWAS) (International Schizophrenia Consortium, 2009). For example, a recent study shows that polygenic scores explain about 4% of the variance in smoking behaviour (smoking vs. not smoking, and the number of cigarettes consumed per day) out of sample (Liu et al., 2019). A polygenic score not only makes one well powered for out of sample prediction, but also enables more powerful gene-by-environment interaction analysis. By using polygenic risk scores, Fontana shows that the interaction between someone's genetic predisposition (as captured by the polygenic scores for educational attainment and smoking intensity) and tobacco excise taxes is insignificant in a model explaining the intensity of cigarette consumption (Fontana, 2015).

The present study addresses the same question as Fontana, but goes beyond the study by Fontana (Fontana, 2015) in three ways. First, we use a set of polygenic scores more directly related to smoking behaviour than Fontana does. That is, we use polygenic scores for smoking initiation and smoking intensity whereas Fontana uses polygenic scores for educational attainment and smoking intensity. Second, through the inclusion of additional data from the two most recent waves of data collection from the US Health and Retirement Study, our
analysis has more statistical power than Fontana’s analysis. Third, next to analyzing the intensive margin (the amount of cigarettes per day), we also analyze the extensive margin (smoking vs. not smoking). This is important, because there is severe misreporting by smokers regarding the amount of tobacco they consume (Gorber et al., 2009).

This study provides the first robust evidence of the existence of a gene-environment (GxE) interaction influencing smoking behaviour. Establishing a GxE interaction is often complicated by the fact that individuals with a certain genetic predisposition may self-select into certain environments (Jencks, 1980). In this study, we overcome bias from such a gene-environment correlation by exploiting exogenous variation in the level of tobacco excise rates across states and years. Our results suggest that individuals with a higher genetic propensity for smoking respond more heavily to a change in excise taxes compared to individuals with a lower genetic propensity. Still, the results show that someone’s genetic propensity to smoking moderates the effect of tobacco excise taxes on smoking behaviour only along the extensive margin (smoking vs. not smoking). The results along the intensive margin (the number of cigarettes consumed per day) are inconclusive.

5.2 Data description

The data used in this study are derived from the US Health and Retirement Study (HRS) (Juster and Suzman, 1995). The HRS is a longitudinal survey consisting of approximately 20,000 individuals who were surveyed biennially since 1992. The respondents in the survey are a representative sample of Americans over age 50 and their spouses. The HRS aims to analyze the health and behaviour of individuals approaching or just after retirement. Therefore, the dataset includes information about for example work status, pension plans, income, health insurance, physical health and functioning, cognitive functioning, and health behaviours including drinking and smoking (for an overview see Karp (2007)). From 2006 onwards, the study started to collect genetic data from their respondents. In the present study, we exploit data collected in the waves from 1992 up to 2014 (12 waves in total) which have been harmonized by the RAND Corporation.

5.2.1 Smoking behaviour

The main outcome in the present study is smoking behaviour. The current study uses three different measures of tobacco use that are available in the HRS and which capture different dimensions of smoking behaviour. The first question
‘Have you ever smoked cigarettes?’ is used to determine whether an individual ever smoked cigarettes. With the second question, ‘Do you smoke cigarettes now?’, it is determined whether an individual is a current smoker. If the individual answers the second question with ‘yes’, the respondent is asked the question ‘About how many cigarettes or packs do you usually smoke in a day now?’. Based on the response to this question, the number of cigarettes consumed per day is determined.

5.2.2 State-level excise tobacco taxes

The Tax Burden on Tobacco dataset (Orzechowski and Walker, 2016) provides us information about the tax levied by the state on each purchased pack of cigarettes (based on the state and federal tax in each year). These data were merged with the HRS data, based on confidential data about the state the HRS respondent currently lives in. As the HRS contains biennial survey data, we use the tax levied in the year prior to each survey. For consistency with prior studies and to facilitate the interpretation of effects as proportional changes in consumption, the tax levels are logarithmically transformed (Adda and Cornaglia, 2006, Fletcher, 2012).

5.2.3 Polygenic scores

Polygenic scores are used to analyze whether the response to tobacco excise taxes is moderated by someone’s genetic predisposition to smoking. Most genetic differences across individuals in a population can be attributed to single nucleotide polymorphisms (SNPs). A SNP is a location in the DNA strand at which two different nucleotides can be present in the population. For each SNP, an individual’s genotype is coded as a 0, 1 or 2, depending on the number of reference nucleotides present. Individuals who inherited the same nucleotide from each parent are called homozygous for that SNP (and have genotype 0 or 2), while individuals who inherited different nucleotides are called heterozygous (and have genotype 1). Polygenic risk scores reflect the combined additive influence of SNPs on a particular outcome.

To construct a polygenic score, SNPs are summed up in a weighted fashion. The weights reflect the strength of the relationship between a SNP and the outcome of interest, as estimated in a GWAS. In a GWAS, for each SNP the following model is estimated:

$$y_i = \mu + \gamma mgim + \delta z_i + \nu_i,$$

(5.1)
where $y_i$ is the outcome of interest for individual $i$, $\mu$ is an intercept, $\gamma_m$ is the additive effect of SNP $g_{im}$, $z_i$ is a vector of control variables (e.g., sex and age), and $\nu_i$ is the residual. Using the effect size estimates $\gamma_m$ from (5.1), the polygenic score is constructed as:

$$G_i = \sum_{m=1}^{M} \gamma_m g_{im},$$  \hspace{1cm} (5.2)

where $G_i$ represents the value of the polygenic score for individual $i$, $M$ is the total number of SNPs included in the construction of the polygenic score, $\gamma_m$ is the additive effect size of SNP $m$ taken as estimated in the GWAS and $g_{im}$ is the genotype of individual $i$ at locus $m$ (measured as 0, 1 or 2).

The HRS provides polygenic scores for public distribution based on several recently conducted large-scale GWASs (Ware et al., 2018). In this study, we use two polygenic score to measure someone’s genetic predisposition to smoking behaviour. The first polygenic score is based on the results of a GWAS on smoking initiation, and measure someone’s genetic predisposition to start smoking. The second polygenic score is based on the results of a GWAS with the number of cigarettes smoked per day as dependent variable. As such, the second score reflects someone’s genetic predisposition to heavy smoking. Hence, the first polygenic score reflects the genetic predisposition for smoking on the extensive margin, and the second one reflects the genetic predisposition for smoking on the intensive margin. The weights for constructing the polygenic scores come from the GWAS conducted by the Tobacco and Genetics Consortium (Tobacco and Genetics Consortium, 2010). To facilitate an easy interpretation of the results, the polygenic scores are standardized such that they have mean 0 and a standard deviation of 1 in the analysis sample. Higher values reflect a higher genetic predisposition to smoking behaviour.

5.2.4 Covariates

For comparability purposes, the choice of individual level control variables is based on the studies by Fletcher and Fontana (Fletcher, 2012, Fontana, 2015). We include an individual’s gender as a covariate, to control for differences between males and females. Furthermore, we add the individual’s birth year to account for possible age specific differences in smoking behaviour. Furthermore, we added birth year squared, to account for possible non-linearity in age effects. We account for the socio-economic status of the respondent by including individual income (as imputed by the RAND Corporation, see (Hurd et al., 2016)) and years of education (self-reported by participants) in the model.
5. Does the genetic predisposition to smoking moderate the response to tobacco excise taxes?

Although Fontana controls for the change in health status in his models, we abstain from it because of possible endogeneity issues. Compared to Fletcher’s model, we do not control for race/ethnicity because we restrict our sample to individuals of recent European ancestry. This is a recommendation of the genotyping center, as this restriction pre-empts possible bias from unobserved relationships between genetic and environmental factors (i.e., population stratification, (Weir, 2012)). To deal with subtle forms of population stratification in the analysis sample, we include the first 10 genetic principal components of the genetic relationship matrix as control variables (Ware et al., 2018). It has been shown that the inclusion of principal components solves the problem of population stratification adequately in the HRS (Rietveld et al., 2014a).

Finally, we include both state dummies and wave dummies to account for differences across states and over time.

5.3 Methods

To test for the presence of an effect of the interaction between someone’s genetic predisposition to smoking behaviour and tobacco excise taxes on smoking outcomes, we use a moderation framework for both smoking initiation and smoking intensity. The baseline regression for smoking initiation is given by:

\[ S_{ist} = \alpha_0 + \alpha_1 \text{Tax}_{st} + \alpha_2 G_i + \alpha_3 G_i \text{Tax}_{st} + \alpha_4 X_{ist} + S_s + D_t + \varepsilon_{ist}, \]  

(5.3)

where \( S_{ist} \) is a binary variable indicating whether individual \( i \) residing in state \( s \) in year \( t \) smokes or not, \( \text{Tax}_{st} \) represents the cigarette tax in state \( s \) at year \( t \), and \( G_i \) is the value of the polygenic score for individual \( i \). \( X_{ist} \) represents the vector of individual-level control variables. The \( \alpha \)'s represent the corresponding effect size estimates for these variables. The vectors \( S_s \) and \( D_t \) are vectors for state and year fixed effects. Lastly, \( \varepsilon_{ist} \) denotes the error term. Despite the binary nature of \( S_{ist} \), we estimate the model using linear regression to make the interpretation of the coefficient more straightforward and to avoid the difficulties surrounding the estimation of interaction effects in non-linear models.

The response to the taxes in terms of tobacco consumption is estimated by:

\[ C_{ist} = \beta_0 + \beta_1 \text{Tax}_{st} + \beta_2 G_i + \beta_3 G_i \text{Tax}_{st} + \beta_4 X_{ist} + S_s + D_t + \tau_{ist}, \]  

(5.4)

where \( C_{ist} \) denotes the number of cigarettes smoked per day by individual \( i \) at time \( t \) in state \( s \). In this regression, the \( \beta \)'s are the effect size estimates and \( \tau_{ist} \) is the residual term. We estimate this model both in the full sample and in the subsample of smokers, because non-smokers are not likely to start smoking after
an increase of excise taxes.

5.4 Results

Table 5.1 contains the descriptive statistics of the analysis sample. It contains information about the full sample and the subsample of current smokers. Static variables are constant over the waves of data collection, dynamic variables can take different values over time. Not surprisingly, the means of the polygenic scores for smoking behaviour as well as the smoking prevalence and the number of cigarettes smoked per day are relatively high in the subsample of current smokers. Besides these differences, there are only small differences between the full sample and the subsample of smokers with respect to birth year and income. Figure 5.1 shows that there is a gradual increase of tobacco excise taxes over time and that there is considerable variation across states regarding the level of tobacco excise taxes imposed.

**Table 5.1 – Descriptive statistics analysis sample.**

<table>
<thead>
<tr>
<th>Static variables</th>
<th>Full sample</th>
<th>Subsample of current smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>Female</td>
<td>0.570</td>
<td>0.495</td>
</tr>
<tr>
<td>Birth Year</td>
<td>1941</td>
<td>11.939</td>
</tr>
<tr>
<td>Years of education</td>
<td>13.265</td>
<td>2.529</td>
</tr>
<tr>
<td>PGS Smoking Initiation</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>PGS Smoking Intensity</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Dynamic variables</td>
<td>Mean</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>Currently smoking</td>
<td>0.139</td>
<td>0.346</td>
</tr>
<tr>
<td>Ever smoked</td>
<td>0.575</td>
<td>0.494</td>
</tr>
<tr>
<td>Cigarettes per day</td>
<td>2.398</td>
<td>7.394</td>
</tr>
<tr>
<td>Income</td>
<td>16,323</td>
<td>42,721</td>
</tr>
<tr>
<td>Married</td>
<td>0.663</td>
<td>0.473</td>
</tr>
<tr>
<td>Individuals</td>
<td>12,089</td>
<td>2,643</td>
</tr>
<tr>
<td>Observations</td>
<td>98,605</td>
<td>13,642</td>
</tr>
</tbody>
</table>

*Notes: Std. Dev. = Standard deviation.*

Table 5.2 present the results of the model explaining whether an individual is currently smoking. Column 1 shows that state-level tobacco excise taxes are negatively associated with the dependent variable, and that the polygenic score for smoking initiation is positively associated with smoking. Both these results are in line with expectations. In terms of effect sizes, an increase of excise taxes by 1% reduces the likelihood of smoking by about 6 percentage points, and an increase of one standard deviation in the polygenic risk score increases the chance of smoking by about 2 percentage points.
5. Does the genetic predisposition to smoking moderate the response to tobacco excise taxes?

**Figure 5.1 – The average, minimum and maximum tobacco excise taxes levied per pack of 20 cigarettes in the United States from 1992 to 2014.**

In Column 2, the interaction between the state-level tobacco excise taxes and the polygenic score for smoking initiation has been added to the model. This interaction term is significantly negative, indicating that high excise taxes on tobacco make those with a high genetic predisposition for smoking less likely to smoke. Column 4 shows that upon inclusion of state and wave fixed effects, the coefficient for the tobacco excise taxes becomes insignificant. This change can be explained by the fact that tobacco taxes within a state tend to increase in a monotonic fashion. These dynamics are absorbed by the state and wave dummies. However, the interaction term between the polygenic score and the tobacco excise taxes remains statistically significant in Column 3.

Table 5.3 presents the results of the regressions explaining someone’s smoking intensity (the intensive margin, in terms of cigarettes per day). In Column 1 (Full sample) and Column 4 (Subsample of current smokers), tobacco excise taxes are significantly negatively associated with the number of cigarettes smoked per day. In terms of effect sizes, an increase of excise taxes by 1% reduces cigarette consumption by 1.46 cigarettes per day in the full sample, and 3.29 cigarettes in the sample of current smokers. The polygenic score is again predictive of smoking behaviour (one standard deviation increase in the polygenic score leads to an increase in consumption of 0.17 cigarettes per day in the full sample and
TABLE 5.2 – Results of the regressions explaining someone's current smoking status.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log(Tax)</td>
<td>-0.060***</td>
<td>-0.060***</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>PGS\textsubscript{Smoking Initiation}</td>
<td>0.020***</td>
<td>0.021***</td>
<td>0.021***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>Log(Tax) × PGS\textsubscript{Smoking Initiation}</td>
<td>-0.008*</td>
<td>-0.010**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.028***</td>
<td>-0.028***</td>
<td>-0.029***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>Birth Year</td>
<td>-0.000</td>
<td>0.000</td>
<td>-0.088</td>
</tr>
<tr>
<td></td>
<td>(0.078)</td>
<td>(0.079)</td>
<td>(0.077)</td>
</tr>
<tr>
<td>Birth Year\textsuperscript{2}</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>Income (in USD 1,000)</td>
<td>-0.000***</td>
<td>-0.000***</td>
<td>-0.000***</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>Years of Education</td>
<td>-0.018***</td>
<td>-0.018***</td>
<td>-0.017***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Married</td>
<td>-0.086***</td>
<td>-0.086***</td>
<td>-0.090***</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>State &amp; Wave Dummies</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>97,984</td>
<td>97,984</td>
<td>97,984</td>
</tr>
<tr>
<td>Individuals</td>
<td>12,058</td>
<td>12,058</td>
<td>12,058</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.0737</td>
<td>0.0740</td>
<td>0.0850</td>
</tr>
</tbody>
</table>

Notes: Standard errors in parentheses (clustered by state and individual); Coefficients for the principal components are not reported, but available upon request from the authors; * \( p < 0.05 \), ** \( p < 0.01 \), *** \( p < 0.001 \).

an increase of 0.47 in the subsample of current smokers). Column 2 and Column 5 show that the interaction effect is negative but insignificant. When comparing the results in the full sample with those in the subsample of current smokers, we observe that the effect sizes are relatively large in the latter subsample. The estimates suggest that current smokers are more receptive to differences in taxes. This could be explained by the fact that smokers are able to reduce their smoking intensity, whereas in the full sample the non-smokers are not likely to change their smoking behaviour (i.e., to start smoking). When adding state and wave dummies (Column 3 and 6), the effects of the taxes are again rendered insignificant.

In sum, the present results suggest that the interaction between the state-level tobacco excise taxes and someone’s genetic predisposition to smoking impacts whether someone smokes or not (the extensive margin), but not the number of cigarettes someone consumes (the intensive margin).
5. Does the genetic predisposition to smoking moderate the response to tobacco excise taxes?

### Table 5.3 – Results of the regressions explaining someone’s current smoking intensity.

<table>
<thead>
<tr>
<th></th>
<th>Full sample</th>
<th>Subsample of current smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td></td>
<td>(3)</td>
<td>(3)</td>
</tr>
<tr>
<td>Log(Tax)</td>
<td>-1.460**</td>
<td>-1.459***</td>
</tr>
<tr>
<td></td>
<td>(0.098)</td>
<td>(0.098)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.146)</td>
</tr>
<tr>
<td></td>
<td>-3.285***</td>
<td>-3.262***</td>
</tr>
<tr>
<td></td>
<td>(0.314)</td>
<td>(0.309)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.457)</td>
</tr>
<tr>
<td>PGS Smoking Intensity</td>
<td>0.165**</td>
<td>0.172**</td>
</tr>
<tr>
<td></td>
<td>(0.058)</td>
<td>(0.060)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.061)</td>
</tr>
<tr>
<td></td>
<td>0.467*</td>
<td>0.455*</td>
</tr>
<tr>
<td></td>
<td>(0.213)</td>
<td>(0.205)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.217)</td>
</tr>
<tr>
<td>Log(Tax) × PGS Smoking Intensity</td>
<td>-0.078</td>
<td>-0.080</td>
</tr>
<tr>
<td></td>
<td>(0.059)</td>
<td>(0.058)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.200)</td>
</tr>
<tr>
<td></td>
<td>-0.297</td>
<td>-0.317</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.197)</td>
</tr>
<tr>
<td>Female</td>
<td>-0.953***</td>
<td>-0.953***</td>
</tr>
<tr>
<td></td>
<td>(0.100)</td>
<td>(0.099)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.101)</td>
</tr>
<tr>
<td></td>
<td>-3.556***</td>
<td>-3.562***</td>
</tr>
<tr>
<td></td>
<td>(0.393)</td>
<td>(0.393)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.403)</td>
</tr>
<tr>
<td>Birth Year</td>
<td>0.994</td>
<td>0.985</td>
</tr>
<tr>
<td></td>
<td>(1.569)</td>
<td>(1.568)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.535)</td>
</tr>
<tr>
<td></td>
<td>13.490</td>
<td>13.340</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.091)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.102)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.197)</td>
</tr>
<tr>
<td>Birth Year²</td>
<td>-0.000</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td></td>
<td>-0.003</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.002)</td>
</tr>
<tr>
<td>Income (in USD 1,000)</td>
<td>-0.005***</td>
<td>-0.005***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.002)</td>
</tr>
<tr>
<td></td>
<td>-0.008</td>
<td>-0.007</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.006)</td>
</tr>
<tr>
<td>Years of Education</td>
<td>-0.347***</td>
<td>-0.347***</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.024)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.023)</td>
</tr>
<tr>
<td></td>
<td>-0.310**</td>
<td>-0.309**</td>
</tr>
<tr>
<td></td>
<td>(0.091)</td>
<td>(0.090)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.088)</td>
</tr>
<tr>
<td>Married</td>
<td>-1.568***</td>
<td>-1.569***</td>
</tr>
<tr>
<td></td>
<td>(0.125)</td>
<td>(0.125)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.124)</td>
</tr>
<tr>
<td></td>
<td>-0.588</td>
<td>-0.594</td>
</tr>
<tr>
<td></td>
<td>(0.303)</td>
<td>(0.302)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.295)</td>
</tr>
<tr>
<td>State &amp; Wave Dummies</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>97,955</td>
<td>97,955</td>
</tr>
<tr>
<td></td>
<td>97,955</td>
<td>97,955</td>
</tr>
<tr>
<td></td>
<td>13,491</td>
<td>13,491</td>
</tr>
<tr>
<td></td>
<td>13,491</td>
<td>13,491</td>
</tr>
<tr>
<td>R²</td>
<td>0.0597</td>
<td>0.0598</td>
</tr>
<tr>
<td></td>
<td>0.0726</td>
<td>0.0648</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0651</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0974</td>
</tr>
</tbody>
</table>

**Notes:** Standard errors in parentheses (clustered by state and individual); Coefficients for the principal components are not reported, but available upon request from the authors; * p < 0.05, ** p < 0.01, *** p < 0.001.

### 5.5 Discussion and conclusion

The present study shows that someone’s genetic predisposition to smoking behaviour moderates the impact of tobacco excises taxes on tobacco usage. However, this interaction does not have a meaningful impact on the total number of cigarettes consumed. These findings suggest that excise taxes are an effective method to reduce tobacco usage, even among the group with a high genetic predisposition towards smoking. Even more, those with a high genetic predisposition to smoking respond most strongly to changes in tobacco excise taxes.

Although Fletcher (Fletcher, 2012) was the first to show that only individuals with a certain genetic variant respond to increases in excise taxes, Fontana provided evidence that population stratification was driving these initial results (Fontana, 2015). However, based on a weighted combination of multiple (approximately 700,000) genetic variants, i.e., a polygenic score, in the present study we do find again a significant interaction effect along the inclusive margin for smoking. The sample restriction to individuals of European ancestry and the in-
clusion of principal components makes that the present findings are not likely to be driven by (subtle forms of) population stratification. As such, the present findings contribute to the literature analyzing heterogeneity in smoking behaviour (Nesson, 2017). In line with the findings of Fontana (Fontana, 2015), we do not find a significant impact of the interaction between the genetic predisposition to heavy smoking and excise taxes on someone’s smoking intensity. Even in the subsample of current smokers, we could not detect such an effect.

Considering all findings jointly, it seems puzzling that the interaction between the genetic predisposition to smoking and tobacco excise taxes does impact the decision to smoke but not the total number of cigarettes consumed. A possible solution for this puzzle could be that there is considerable misreporting in the number of cigarettes smoked, making that the estimated results along the intensive margin are less precise. In the data, the reported number of cigarettes consumed is often a multiple of 5 which does indeed suggest there may be considerable measurement error present in this variable. Therefore, we caution that the results along the intensive margin are currently inconclusive.

The present study is not without limitations. Importantly, HRS participants are only surveyed every two years. In the analyses, we therefore used the excise taxes one year before each smoking measurement. This may be less suitable if the response time to increases in excise taxes differs among individuals. Also, individuals who live close to the border of a state could purchase their tobacco in the neighbouring state with a lower excise tax on tobacco (Chiou and Muehlegger, 2008). In our analyses, we cannot rule out whether this is driving our results but we note that this may also be another reason for why the results along the intensive margin are inconclusive. Another limitation of the current sample is that it is a representative sample of older Americans only. As over time only the most addicted individuals are likely to remain smoking, this set of individuals might be particularly insensitive to change in excise taxes. At the same time, the prevalence of smoking in this age cohort is relatively high making it particularly suitable to study smoking behaviour. Therefore, we consider the replication of the present results in a younger sample to be particularly relevant. Finally, the polygenic scores used in this study are predictive of smoking behaviour, but, as outlined by Dudbridge (Dudbridge, 2013), the predictive power of a polygenic score is (amongst others) a function of the GWAS sample size. A larger GWAS sample would lead to more accurate effect estimates and hence a more predictive polygenic score. When such polygenic scores become available, it may be worthwhile to repeat the present analyses to see whether also genetic heterogeneity in responses to excise tax can be detected along the intensive margin for smoking.
5. Does the genetic predisposition to smoking moderate the response to tobacco excise taxes?

From a policy perspective, our findings suggest that there is genetic heterogeneity in response to excise taxes. Individuals with a high genetic predisposition towards smoking respond stronger to tobacco excise taxes compared to individuals with a lower genetic predisposition. Further research is needed to understand what exactly makes that those with a low genetic propensity for smoking to respond relatively mildly to changes in tobacco excise taxes. Possibly, the nature of their smoking behaviour (e.g., recreational use) differs from those with a high genetic predisposition for smoking (who may be more likely to be addicted to smoking). If so, the present study suggests that different policies for genetically different types of individuals are needed to bring down smoking in the population. Moreover, if true, then the present results provide a possible explanation for why the decrease in cigarette consumption stalled over the past years: The reasons for why current smokers keep smoking the number of cigarettes they have been smoking are difficult to manipulate by further increases of tobacco excise taxes.

ACKNOWLEDGEMENTS

We thank the National Institute on Aging (U01 AG009740), the Health and Retirement Study (HRS application number HRS RDA 2019-025) staff, and the HRS participants.
III

**Multivariate GREML**
Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence

Ronald de Vlaming, Eric A.W. Slob, Philip R. Jansen, Philipp D. Koellinger, Patrick J.F. Groenen, Cornelius A. Rietveld

Abstract

Global grey matter volume and the sizes of several specific brain areas are positively associated with intelligence in human populations. Although current methods are able to estimate bivariate genetic correlations between two of such traits, they are not able to estimate genetic correlations across more than two traits simultaneously. We developed a multivariate linear mixed model and optimization procedure to simultaneously analyze the genetic correlations among 76 brain regions and 10 behavioral outcomes, including intelligence and educational attainment. Compared to the existing bivariate approach, our method is faster and able to guarantee the internal consistency of the estimated genetic correlation matrix. Based on a sample of 14,341 unrelated individuals from the UK Biobank, we find genetically distinct clusters across brain areas, one of the 'older' part of the brain (cerebellum and brain stem), and one of the newer part of the brain (the neocortex). These findings suggest that our current way of thinking about the brain makes sense from a genetics perspective.

This chapter is based on Slob et al. (2018).
6.1 Introduction

Through Genome-wide Association Studies (GWASs), thousands of single-nucleotide polymorphisms (SNPs) have been associated with a range of human traits (Buniello et al., 2019, Visscher et al., 2017a). Still, together these SNPs do not fully account for the twin heritability of traits. Genome-based Restricted Maximum Likelihood (GREML) estimation has been developed to estimate the proportion of the variation in a trait that can be explained by a scan of SNPs (the so called SNP-heritability) across the whole genome using observed genetic similarities among unrelated individuals in a population (Yang et al., 2010). The bivariate extensions of this method enabled the estimation of the genetic correlation between two traits (Lee et al., 2012). One often combines the estimates of pairwise combinations of traits into a multivariate genetic correlation matrix in case one is interested in the genetic correlation across more than two traits (e.g., Power and Pluess (2015)). However, this ‘pairwise bivariate’ approach may result in a genetic correlation matrix which is not internally consistent (i.e., it may not be positive semidefinite). Next to this, the corresponding standard errors of this ‘pairwise bivariate’ genetic correlation matrix do not fully reflect the structure of the data. We develop a multivariate extension of the GREML method which guarantees the internal consistency of the resulting genetic correlation matrix, and which produces corresponding standard errors does reflect the full data structure.

To deal with the computational complexity of the model, we developed an improved optimization procedure. With a precomputed eigendecomposition of the individual-by-individual genomic-relatedness matrix, the computational complexity of our method is of the order $NT^5$ (where $N$ denotes the number of individuals and $T$ the number of traits). For comparison, the bivariate GREML approach has a computational complexity of the order $NT^6$. Central in the optimization procedure is the transformation of the vector of correlated traits into a new vector of uncorrelated traits. As a result, our procedure improves over the ‘pairwise bivariate’ approach by guaranteeing that standard errors across traits are correct by taking into account the full data structure. Hence, the results of methods using a multivariate genetic correlation matrix as input, such as genomic structural equation modeling (Grotzinger et al., 2019), may improve when the results of the method proposed here are used as input compared to when one uses results of this ‘pairwise bivariate’ method.

We used multivariate GREML to analyze the shared genetic architecture of the human brain (76 cortical and subcortical structures) and 10 behavioral traits. Recent GWASs (Zhao et al., 2019, Grasby et al., 2020) reported several significant genetic correlations between some grey matter volumes and neu-
6. Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence

roperties/cognitive and behavioral traits using bivariate LD-score regression (Bulik-Sullivan et al., 2015). Whereas we analyze more brain regions than these studies do (we include both cortical and subcortical areas), the main advantage of our approach is that we do not need to take the intermediate GWAS step and hence our inferences are stronger given a certain sample size (see Ni et al. (2018) for a comparison between LD score regression and GREML).

Our findings suggest that the evolutionary more conserved areas of our brain are more genetically determined compared to the evolutionary more recent areas of the brain, as the heritability of the older parts (cerebellar and subcortical structure) is much higher compared to the heritability of the new parts of the brain (frontal part of the cortex). Furthermore, our multivariate GREML identifies two clusters in the brain based on genetic correlations: one in the subcortical areas of the brain (cerebellar and subcortical structures) and one in the cortical area (including frontal, temporal, and parietal lobes). Our results confirm that the current way of partitioning the human brain into broad anatomical areas closely follows the genetic pattern across the regions. When we link these genetic correlations to our behavioral traits, we can confirm previous found relationships, such as a relation between the cerebellum and visual spatial memory. Furthermore, we confirm the strong genetic similarity between intelligence and educational attainment. Next, we find a link between cerebral atrophy and alcohol consumption. Lastly, a link is found between subjective well-being and the temporo-occipital part of the middle temporal gyrus.

6.2 Data: UK Biobank Imaging Study

In total, we analyzed 86 phenotypes from UK Biobank (UKB) participants of European ancestry. Among these phenotypes, are 74 measures of relative grey matter volume of brain parts obtained using T1-weighted structural imaging. These volume measures are derived from UKB Category 110. In the Supplementary Material, a full description of variables is given. For each subcortical region, the left and right volume were added and divided by the total volume to come to relative volumes. Next to these brain measures, we have the total volume of both grey and white matter, the total volume of grey matter only, and some other anthropometric measures, such as height, body mass index. Lastly, we have some behavioral traits: IQ, educational attainment, visual memory, reaction time, neuroticism, subjective well-being, depression, and alcohol consumption. We restrict our sample such that we exclude individuals with brain damage.

---

1 The original measurements were performed with a standard Siemens Skyra 3T running VD13A SP4 (as of October 2015) with a standard Siemens 32-channel RF receive head coil. The full UK Biobank imaging protocol can be found at http://biobank.ox.ac.uk/crystal/refer.cgi?id=1977
(identified using the medical records). Next, we restrict our sample such that we have fully balanced data. The full pipeline and quality control protocols can be found in the Supplementary Materials 4.

6.2.1 Phenotypic covariates

Several control variables are included in the models to account for spurious correlations across the phenotypes. We opt for a more conventional set of confounds. This means we correct for age, age squared, age cubed, sex, sex × age, sex × age squared, sex × age cubed, batch number (used as a dummy variable) and an intercept.

Furthermore, for the IQ-measurement we employ some additional covariates. Some participants only did the touch screen based test once, whereas others did multiple touch based tests and even a web based tests. The selection for moments of measurement and number of tests appears to be non-random. Next to that, there seems to be a learning effect in these sort of tests, where individuals who did the test before score higher compared to individuals who did the test for the first time. Hence, we add covariates for the average IQ-measurement to take into account learning effects and participation across waves. Last, as there are effects of being at a certain age at the moment of testing, we use the age at the moment of testing.

6.2.2 Genetic covariates

To measure genetic relatedness between individuals, we use the genomic relatedness matrix (Yang et al., 2010). To calculate the GRM, we restrict our SNPs to all HapMap 3 SNPs (International HapMap 3 Consortium, 2010) and the SNPs that have imputation quality of over 0.9. This imputation threshold considered to be a quite conservative approach (Verma et al., 2014). Furthermore, we restrict SNPs to have at least a minor allele frequency of 1%, a missingness per marker of 5%, an Hardy-Weinberg-Equilibrium-test p-value below 0.001. These are all standard quality control filters. Furthermore, we exclude individuals with a SNP-missing value of 5% and over. For individuals that are too closely related, we drop one of each pair. We use a relationship cutoff of 0.025, this maximum relatedness approximately corresponds to cousins two to three times removed.

Next to these corrections to our phenotypes, we also do corrections on the genetic measurement. As the genotyping of participants in UKB was done in batches on different platforms, there can be differences in imputation quality based on these different batches and platforms. After doing an F-test, it turns out that there is no difference in model fit if we use batch dummies or only a
6. Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence platform dummy (P-value = 1.00). Hence, we opt to only correct for the platform dummies.

Moreover, to control for subtle forms of population stratification in the analysis sample, we include the leading 20 genetic principal components to account for population structure (Browning and Browning, 2011). Furthermore, we correct for the long range-LD regions (Price et al., 2008) as identified by Linnér et al. (2019).

6.3 Methods

The most frequently used method to find associations between individual SNPs and a quantitative trait of interest is a genome-wide association study (GWAS). In a GWAS, a simple regression is performed in the following simple regression model

$$y_i = \mu + x_{ik}b_k + \varepsilon_i,$$  \hspace{1cm} (6.1)

where $y_i$ is the value of the phenotype for individual $i$, $\mu$ is the intercept, $x_{ik}$ is an indicator variable that takes values 0, 1 or 2 if the genotype of individual $i$ at SNP $k$ is aa, Aa or AA, respectively. The corresponding allelic effect of SNP $k$ for trait is $b_k$, and finally a residual term $\varepsilon_i$ identically and independently distributed as $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_e)$, where $\sigma^2_e$ the residual variance for the trait. If all causal variants are known, they can be added into one single model for the trait:

$$y_i = \mu + g_i + \varepsilon_i \quad \text{and} \quad g_i = \sum_{k=1}^{m} s_{ik}u_k,$$  \hspace{1cm} (6.2)

where $g_i$ is the total genetic contribution of all SNPs for individual $i$, $m$ is the total number of causal loci, $u_k$ is the scaled effect of causal SNP $k$, and $s_{ik}$ is standardized genotype of individual $i$ at SNP $k$ (that is, $s_{ik} = x_{ik} - 2f_k/\sqrt{2f_k(1-f_k)}$ with $f_k$ the frequency of the minor allele at locus $k$). Observe that (6.2) can be rewritten in matrix notation as $y = \mu 1 + g + \varepsilon$ and $g = Su$. Now, $u$ is treated as a random effect that follows the distribution $u \sim \mathcal{N}(0, \sigma^2_u I)$, where $\sigma^2_u$ is the variance of causal effects. As a result, the distribution of the total genetic contribution is multivariate normally distributed as $\mathbf{g} \sim \mathcal{N}(0, \sigma^2_u \mathbf{SS}^\top)$. Now, $\sigma^2_g$ ($= m\sigma^2_u$) can be interpreted as the variance of the total additive genetic effects. The variance-covariance matrix of $y$ becomes

$$\text{Var}(y) = \sigma^2_u \mathbf{SS}^\top + \sigma^2_e \mathbf{I} = \sigma^2_g \mathbf{SS}^\top + \sigma^2_e \mathbf{I} = \sigma^2_g \mathbf{G} + \sigma^2_e \mathbf{I},$$  \hspace{1cm} (6.3)
where \( G = m^{-1}S\mathbf{S}^\top \) is the genetic relationship matrix between pairs of individuals at causal loci. With the equation above, we can estimate the SNP-based heritability \( h^2 \) of a trait as \( \sigma_g^2/(\sigma_g^2 + \sigma_e^2) \).

However, in practice the causal variants are unknown and hence we cannot obtain the genetic relationship matrix \( G \) directly. This \( G \) matrix can be approximated by applying the same formula to a genome-wide sample of SNPs \( X^* \) instead of \( S \), that is,

\[
A = M^{-1}X^*X^{*\top},
\]

where \( M \) is the total number of SNPs used in the standardized genome-wide sample \( X^* \), such that \( x^*_{ij} = (x_{ij} - 2f_i)/\sqrt{2f_i(1-f_i)} \) where again \( f_i \) is the minor allele frequency at SNP \( i \). There are more efficient ways of creating this approximation, see e.g. Yang et al. (2010) and VanRaden (2008).

From here, we extend the previous univariate model in (6.1) to a multivariate model with \( T \) different phenotypes \( y_t \). Then, we can model not only the genetic and environmental variances for each of the \( T \) phenotypes but also their covariances in the \( T \times T \) matrices \( V_G \) and \( V_E \). Let \( y \) now be the stacked vector of all \( T \) vectors \( y_t \), that is, \( y^\top = [y_1^\top, \ldots, y_T^\top]^\top \). The multivariate model is

\[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_T
\end{bmatrix} =
\begin{bmatrix}
Z^* & 0 & \cdots & 0 \\
0 & Z^* & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & Z^*
\end{bmatrix}
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_T
\end{bmatrix} +
\begin{bmatrix}
X^* & 0 & \cdots & 0 \\
0 & X^* & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & X^*
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_T
\end{bmatrix} +
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_T
\end{bmatrix},
\]

where \( Z^* \) is the \( N \times P \) incidence matrix containing the \( P \) covariates for \( N \) individuals, \( y_i \) is a \( P \) vector of fixed effects for trait \( t \) (hence we allow trait-specific covariates), \( X^* \) is the \( N \times M \) incidence matrix containing all SNPs, \( \beta_i \) is an \( M \) vector of additive genetic values for trait \( t \), and \( \epsilon_t \) represents the residual for trait \( t \). Model (6.5) can equivalently be written as:

\[
y = Z\gamma + X\beta + \epsilon,
\]

where \( \epsilon \sim \mathcal{N}(0, V_E \otimes I_N) \) and \( \beta \sim \mathcal{N}(0, V_G \otimes I_P) \). Here, \( \otimes \) denotes of the Kronecker product and \( I_j \) is the \( j \times j \) identity matrix. The main interest here lies in the genetic covariance matrix \( V_G \), of which the off-diagonal elements give information about the shared genetic architecture between the different traits. Note that for \( V_G \) and \( V_E \) to be covariance matrices, they must be constrained to be positive
semidefinite.

To estimate the the genetic and environment covariance matrices $V_G$ and $V_E$ in the model of equation (6.6), we use a maximum likelihood approach. As an optimization method, we employ the quasi-Newton approach of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. For this algorithm, computationally efficient expressions are needed for the log-likelihood, the gradient, and the average information matrix. These expressions can be found in Supplementary Material 1. We ensure the positive semidefiniteness of the genetic correlation matrix (this is a need for internal consistency between the correlations of the random variables) by optimizing over the triangular elements of the Cholesky decomposition of the variance-covariance matrix.

### 6.4 Results

From the UK Biobank sample (Sudlow et al., 2015), we selected 14,341 unrelated individuals. We then estimated their pairwise genetic relationships using all autosomal SNPs that pass our quality control protocol and retained 14,341 individuals whose pairwise relationship was estimated to be less than 0.025. We fitted a multivariate mixed linear model to 86 phenotypes using our BFGS algorithm.

From the proportions of phenotypic variance explained by the autosomal SNPs (Figure 6.1 a-b), we observe that the SNP-based heritabilities are relatively high in the cerebellar and subcortical structures of the brain (mean SNP $h^2$: 33% and 31% with average standard error of 5%) and lower in the frontal part of the cortex (mean SNP $h^2$: 23% with average standard error of 5%). A full table of heritability estimates with corresponding standard errors can be found in Supplementary Materials 5. This suggests that the evolutionary more conserved areas of our brain are more heritable determined compared to more evolutionary recent areas of the brain (Rakic, 2009), such as the neocortex.

Next, we investigated the extent to which cortical and subcortical areas are influenced by the same genetic factors. By estimating the genetic correlations among the relative brain volumes, we observe a cluster of correlations in the subcortical areas of the brain (cerebellar and subcortical structures, Figure 6.1 c-d). In addition, we identify a second cluster that captures correlations between cortical correlations, including the frontal, temporal, and parietal lobes.

To identify whether there are specific clusters within the observed genetic correlation matrix, We used Ward’s method for hierarchical clustering to create a dendogram (Figure 6.2). We observe five different clusters in the genetic correlation matrix. The first cluster represents the frontal cortex of the brain, the second the cerebellum, the third the brain stem, and the last two are a mixture of
temporal and occipital parts. This suggests that the current way of partitioning the human brain into these broad anatomical areas closely follows the genetic differences observed across the regions (Standring, 2015), as the current regions share a similar genetic architecture.

In order to find clearer links between the behavioral traits and the brain regions, we create spatial mappings of the genetic correlation with the brain regions for each behavioral trait. Here we run into the issue that due to our limited sample size, our standard errors are quite large for the genetic correlation estimates (average standard error of 0.17). Hence, we will look mostly at effect sizes instead of significance. Still, our results should be interpreted with great care due to this limitation. Regarding the genetic correlations between regions and behavioral traits, we observe a fair positive genetic correlation ($\rho = 0.25$ with standard error 0.17) between the Cerebellum VIIIb and visual spatial memory (Figure 6.3 a). This relationship has been suggested in earlier phenotypic studies (Molinari et al., 2004). For intelligence and educational attainment, the strongest correlations are found in the frontal lobe region (Figure 6.3 b-c). The strong genetic similarity between intelligence and educational attainment has been established in earlier studies (Allegrini et al., 2019). We find a strong negative genetic correlation between the cerebellum and the number of alcoholic drinks consumed per week (Figure 6.3 d, average correlation $\rho = -0.08$, strongest observed genetic correlation $\rho = -0.23$ with Vermis VI with standard error 0.16), which fits with earlier findings that cerebral atrophy is a common feature in alcoholics (Luo, 2015). For depression, the strongest genetic correlation is found in the cerebellum (Figure 6.3 e). For subjective well-being, we confirm the previously observed strong link to the temporooccipital part of the Middle Temporal Gyrus (Figure 6.3 f)(Song et al., 2019).

\footnote{Our findings do not suggest a direction of effect here, as it could be that consuming more alcohol leads to a reduced cerebellum size, or that a smaller cerebellum leads to individuals consuming more alcoholic drinks. Previous findings suggest that alcoholic abuse causes cerebral atrophy, hence that seems to be the most likely explanation.}
Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence

Figure 6.1 – Spatial mapping of the estimates for SNP-based heritability and genetic correlation across the different brain regions, SNP-based heritability per anatomical area, and genetic correlation table of aggregated anatomical area. 

a. Spatial mapping of the SNP-based heritability of the different brain regions, where blue dots represent a low heritability and yellow represents a high heritability. 

b. Boxplot of the SNP-based heritability per brain region. 

c. Spatial mapping of the genetic correlation across different anatomical regions in the brain, where blue vertices represent a negative correlation and red vertices a positive correlation. 

d. Average genetic correlation between the different anatomical areas in the brain.
**Figure 6.2** – Clusters are identified using Ward’s method with a D2 ward for hierarchical clustering. This method minimizes the error sum of squares by agglomerating clusters each step (see Kaufmann and Roussseeuw (1990) for more details). The colors represent the different clusters, where the number of clusters is identified using the elbow method.
Figure 6.3 – Spatial mapping of the genetic correlation between brain regions and the behavioral traits, where blue points represent a negative correlation and red points a positive correlation. a, Spatial mapping of the genetic correlation between visual spatial memory (measured using a pairs-matching test) and the different brain regions. b, Spatial mapping of the genetic correlation between intelligence (measured using a fluid intelligence score) and the different brain regions. c, Spatial mapping of the genetic correlation between Educational attainment (measured using years of education) and the different brain regions. d, Spatial mapping of the genetic correlation between Drinks per week (measured on a logarithmic scale) and the different brain regions. e, Spatial mapping of the genetic correlation between depression score (measured by logarithm of first PC of depression intensity and frequency) and the different brain regions. f, Spatial mapping of the genetic correlation between subjective well-being (measured by average happiness over time) and the different brain regions.
Our multivariate GREML method enabled us to reveal distinct clusters of genetic correlations between brain areas as well as genetic correlations between brain regions and behavioral traits. We find that there are strong differences in heritability across the anatomical areas in the brain, where the more central anatomical areas have higher heritability compared to the outer parts of the brain. The behavioral traits have even lower heritability. Our findings confirm that the current way we think of brain anatomy makes sense from a genetic perspective. Next to this, we find strong genetic correlation between several behavioral traits and different anatomical areas of the brain.

To ensure our findings were not realised by data ascertainment or spurious associations, several quality control measures were taken. We carefully adjusted phenotypes for systematic differences, such as age, and sex, and applied thorough quality control to the SNP data (see Supplementary Material 4 for the full pipeline). We restricted our sample such only unrelated individuals of recent European ancestry were included. To deal with subtle population stratification, we performed our REML analysis by fitting the first 20 principal component of the genetic relatedness matrix as covariates. Next to this, we corrected for the different platforms used for genotyping the study participants. Given the conservative approach taken in our SNP and individual selection, our results are unlikely to be biased by population stratification. However, one could opt for an even more conservative approach as suggested by Alfaro-Almagro et al. (2020). Their suggested covariates (322 principal components) are able to soak up 99% of the variance in control variables. However, we opt to not take this approach as it is unclear what exactly these covariates capture. Furthermore, it could lead to Berkson’s paradox, which can happen when one adjusts two independent variables for a potential confound that was actually a consequence of the independent variables (Berkson, 1946, Zhang, 2008, Pearl et al., 2009). This leads to a spurious association between the independent variables that can be incorrectly induced.

To confirm that our results are not only a reflection of the proximity of brain regions (brain regions physically close could be more strongly correlated), we reran the multivariate GREML model with covariates for the distance between the different regions (results available upon request from the authors). Although the resulting genetic correlations are somewhat smaller (∼25% smaller on average), also after correcting for proximity the main patterns are still observed. This suggests that our found patterns are not purely driven by proximity between regions.

In our main analyses, we analyzed relative brain volumes (i.e. relative
volume of grey matter in Frontal Medical Cortex). Previous findings suggest that the size of the brain is positively associated with intelligence (Nave et al., 2019). Therefore, it might be worthwhile to do the same analyses for absolute volumes instead of relative volumes, as this relative measure is unable to pick up differences in total size of the brain. With the new UK Biobank imaging release, we plan to do this new analysis.

Since some regions in the brain have more sub regions, it could be that our results may be somewhat driven by this difference, as it places more weight on the higher represented regions compared to regions with fewer sub regions. To get a grasp of how this influences our analysis, a robustness check was performed where all the sub regions were aggregated into one larger region. This did not have strong effect on the correlations between the brain regions and behavioral traits.

Genetic correlations are informative about the genetic overlap between sets of traits (Lynch and Walsh, 1998). The popularity of GREML as a method to estimate SNP-based heritability, and the importance of the estimation of genetic correlation, makes us to conclude that the method proposed here may help to advance our understanding of the multivariate genetic nature of human traits.

Acknowledgements

The UKB has obtained ethical approval covering this study from the National Research Ethics Committee (reference no. 11/NW/0382). This research has been conducted using the UK Biobank Resource under application number 11425. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative (NWO Call for Compute Time EINF-403 to E.A.W. Slob).
Appendices

6.A Method derivation

A maximum likelihood approach is taken to estimate the genetic and environment covariance matrices $V_G$ and $V_E$ in the model of equation (6). In this section, efficient expressions are derived that are fundamental to make their estimation computationally feasible for reasonably large data sets. As an optimization method, we employ the quasi-Newton approach of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. For this algorithm, computationally efficient expressions are needed for the log-likelihood, the gradient, and the average information matrix. To arrive at the algorithm and derive standard deviations of the estimates, the following steps are taken:

1. Remove the effect of the covariates $Z$.

2. Reparametrize the covariance matrices $V_G$ and $V_E$ to guarantee their positive semi-definiteness.

3. Express the log likelihood.

4. Maximize the likelihood in the unknown covariance matrices $V_G$ and $V_E$ by the BFGS algorithm.

5. Formulate efficient expressions for the gradients, the determinants, and the log likelihood needed for the BFGS algorithm.


7. Derive expressions for the standard errors of the covariance matrices and heritability estimates using a delta method.
The multivariate GREML model from the main text is:

\[ y = Z\gamma + X\beta + \epsilon, \]  

(6.7)

where \( \epsilon \sim \mathcal{N}(0, \mathbf{V}_E \otimes \mathbf{I}_N) \) and \( \beta \sim \mathcal{N}(0, \mathbf{V}_G \otimes \mathbf{I}_M) \) and the vectors \( y \) and \( \epsilon \) are of length \( NT \). Throughout these derivations, we will make use of the fundamental property of linear combinations of multivariate normal distributed vectors, that is, if vector \( \delta \sim \mathcal{N}(\mu, \Sigma) \), then the linear combination \( C\delta + m \sim \mathcal{N}(C\mu + m, C\Sigma C^\top) \).

Applying this result to the last two terms of (6.7) implies that the linear sum \( X\beta + \epsilon \) is also normally distributed, that is, \( X\beta + \epsilon \sim \mathcal{N}(0, X(\mathbf{V}_G \otimes \mathbf{I}_M)X^\top + \mathbf{V}_E \otimes \mathbf{I}_N) \).

**Step 1. Removing the effects of the covariates \( Z \)** The main interest in this paper is in estimating the genetic and environmental covariance matrices \( \mathbf{V}_G \) and \( \mathbf{V}_E \), not the weight vectors \( \gamma \) of the covariates. The same covariance matrices can be obtained by removing the effects of the covariates \( Z \) by premultiplying the phenotype vector \( y \) by the anti-projection matrix \( \mathbf{I} - Z(Z^\top Z)^{-1}Z^\top \). Additional computational efficiency can be obtained by using the sparsity in our variance-covariance structure.

Consider the singular value decomposition of \((\mathbf{I} - Z(Z^\top Z)^{-1}Z^\top)X = \mathbf{P}\Phi\mathbf{Q}^\top\) with \( \mathbf{P} \) the \( NT \times MT \) orthonormal matrix of left singular vectors, \( \Phi \) the \( MT \times MT \) diagonal matrix with nonnegative singular values, and \( \mathbf{Q} \) the \( MT \times MT \) orthonormal matrix of right singular vectors. Then, we can write

\[
\tilde{y} = \mathbf{P}^\top\left(\mathbf{I} - Z(Z^\top Z)^{-1}Z^\top\right)y = \Phi\mathbf{Q}^\top\beta + \mathbf{P}^\top\epsilon \\
\sim \mathcal{N}(0, \Phi\mathbf{Q}^\top(\mathbf{V}_G \otimes \mathbf{I}_M)\mathbf{Q}\Phi + \mathbf{P}^\top(\mathbf{V}_E \otimes \mathbf{I}_N)\mathbf{P}).
\]  

(6.8)

Next, we will look into the partitioned block structure of \( X \) and \( Z \) to simplify the variance of \( \tilde{y} \) in (6.8). We start by inspecting the antiprojection matrix of \( Z \)

\[
\mathbf{I} - Z(Z^\top Z)Z^\top = \begin{bmatrix}
\mathbf{I} - Z^* (Z^\top Z^*) Z^*^\top & 0 & \cdots & 0 \\
0 & \mathbf{I} - Z^* (Z^\top Z^*) Z^*^\top & \cdots & 0 \\
0 & 0 & \ddots & \vdots \\
0 & 0 & \cdots & \mathbf{I} - Z^* (Z^\top Z^*) Z^*^\top
\end{bmatrix}.
\]  

(6.9)

Because \( X \) has a similar block structure as the antiprojection matrix of \( Z \), more efficient expressions are possible. Denote the singular value decomposition of
6. Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence

\[
\begin{bmatrix}
I - Z^* (Z^* Z^*) \tilde{X}^* \\
\end{bmatrix} \tilde{X}^* \text{ be given by}
\]

\[
\left( I - Z^* (Z^* Z^*) Z^* \right) \tilde{X}^* = P^* \Phi^* Q^* \tilde{X}^*.
\]

(6.10)

This gives us easy expressions for

\[
P = \begin{bmatrix}
P^* & 0 & \ldots & 0 \\
0 & P^* & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & P^*
\end{bmatrix} = I_T \otimes P^*,
\]

(6.11)

\[
\Phi = \begin{bmatrix}
\Phi^* & 0 & \ldots & 0 \\
0 & \Phi^* & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \Phi^*
\end{bmatrix} = I_T \otimes \Phi^*,
\]

(6.12)

\[
Q = \begin{bmatrix}
Q^* & 0 & \ldots & 0 \\
0 & Q^* & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & Q^*
\end{bmatrix} = I_T \otimes Q^*.
\]

(6.13)

With these expressions, we can rewrite the variance of \( \tilde{y} \):

\[
\text{Var}(\tilde{y}) = \Phi Q^* (V_G \otimes I_M) Q \Phi + P^* (V_E \otimes I_N) P
\]

\[
= (I_T \otimes \Phi^*) (I_T \otimes Q^*) (V_G \otimes I_M) \left( I_T \otimes Q^* \right) (I_T \otimes \Phi^*)
\]

\[
+ (I_T \otimes P^*) (V_E \otimes I_N) (I_T \otimes P^*)
\]

\[
= (I_T \otimes \Phi^*) (V_G \otimes I_M) (I_T \otimes \Phi^*) + (V_E \otimes I_N)
\]

\[
= V_G \otimes \Phi^* + V_E \otimes I_N.
\]

(6.14)

Hence, \( \tilde{y} \sim \mathcal{N} \left( 0, V_G \otimes \Phi^* + V_E \otimes I_N \right) \).

**Step 2. Ensuring \( V_G \) and \( V_E \) to be positive semi-definite** For \( V_G \) and \( V_E \) to be covariance matrices, they need to be symmetric and positive semi-definite. This requirement is enforced by reparametrizing these matrices by their Cholesky decomposition, that is \( V_G = \Gamma_G \Gamma_G^T \) and \( V_E = \Gamma_E \Gamma_E^T \) where \( \Gamma_G \) and \( \Gamma_E \) are lower
Step 3. Expression of the log likelihood  To find estimates of $V_G$ and $V_E$, we have to maximize the corresponding log-likelihood (where we denote the $T(T+1)$ vector of all parameters in $\Gamma_G$ and $\Gamma_E$ by $\theta$) up to a constant:

$$
\log l(\theta) = -\frac{1}{2} \left( \log \left| V_G \otimes \Phi^* + V_E \otimes I_N \right| + \bar{y}^T \left( V_G \otimes \Phi^* + V_E \otimes I_N \right)^{-1} \bar{y} \right)
$$

$$
= -\frac{1}{2} \left( N T \log(2\pi) + \sum_{i=1}^{N} \log \left| \phi_{ii} V_G + V_E \right| + \sum_{i=1}^{N} \bar{y}_i^T \left( \phi_{ii} V_G + V_E \right)^{-1} \bar{y}_i \right)
$$

$$
= -\frac{1}{2} NT \log(2\pi) + \sum_{i=1}^{N} \log \left| \phi_{ii} \Gamma_G \Gamma_G^T + \Gamma_E \Gamma_E^T \right|
$$

$$
= -\frac{1}{2} \sum_{i=1}^{N} \bar{y}_i^T \left( \phi_{ii} \Gamma_G \Gamma_G^T + \Gamma_E \Gamma_E^T \right)^{-1} \bar{y}_i,
$$

where $\bar{y}_i^T$ is row $i$ of the $N \times T$ matrix $\bar{Y}$ which is the unstacked version of $\bar{y}$. This likelihood is also described in Yang et al. (2010), and is based upon previous work by Casella and Searle (1985), Searle et al. (1992), and Harville (1977). This approach is known as restricted maximum likelihood (REML).

Step 4. Maximize the likelihood by the BFGS algorithm  To maximize the likelihood, we use the quasi-Newton approach of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Nocedal and Wright, 2006). The reason for doing so is that BFGS only needs gradient and function evaluations and that the updates can be computed relatively fast. BFGS is a quasi-Newton method where each update takes the form:

$$
\theta_{k+1} = \theta_k + \alpha_k p_k,
$$

where $\alpha$ is the step size of the line search, $k$ is an iteration counter, and $p_k$ is the search direction defined by:

$$
p_k = -B_k^{-1} \nabla \log l(\theta_k),
$$
6. Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence

With ∇ log l(θ) the gradient of log l(θ) and the approximation of the inverse of the Hessian B_{k+1}^{-1} is defined by:

\[ s_k = \theta_{k+1} - \theta_k = \alpha_k p_k, \]
\[ d_k = \nabla \log l(\theta_{k+1}) - \nabla \log l(\theta_k), \]
\[ \rho_k = (s_k^T d_k)^{-1}, \]
\[ B_{k+1}^{-1} = (I - \rho_k s_k d_k^T)B_k^{-1}(I - \rho_k d_k s_k^T) + \rho_k s_k s_k^T. \]

Then, the BFGS algorithm is defined by:

1. Given start \( \theta_0 \), convergence tolerance \( \varepsilon > 0 \), and \( B_0^{-1} = I \).
2. \( k \leftarrow 0 \).
3. While \( \| \nabla \log l(\theta_k) \| > \varepsilon \).
4. Compute search direction \( p_k = -B_k^{-1} \nabla \log l(\theta_k) \).
5. Set \( \theta_{k+1} = \theta_k + \alpha_k p_k \) where \( \alpha_k \) is obtained by a Golden section line search.
6. Compute \( B_{k+1}^{-1} \) by (6.21).
7. \( k \leftarrow k + 1 \).
8. End while.

**Step 5a. Efficient expressions for evaluating log likelihood** The BFGS algorithm needs to do a line search and evaluate the log likelihood at several points along the line \( \theta_k + \alpha_k p_k \) for different \( \alpha > 0 \). We chose to use the Golden section to do so (Fletcher, 1991). Therefore, fast and efficient computations are needed to do so.

Denote by \( \mathbf{V} \) the variance of \( \tilde{y} \), that is \( \mathbf{V} = \mathbf{V}_G \otimes \Phi^* \otimes \mathbf{V}_E \otimes \mathbf{I}_N \). Observe that this matrix has a block diagonal structure, where each block \( \mathbf{V}_i = \phi_{ii}^* \mathbf{V}_G + \mathbf{V}_E \). Let the eigenvalue decomposition of \( \mathbf{V}_E \) be given by \( \mathbf{DAD}^T \), where \( \mathbf{DD}^T = \mathbf{D}^T \mathbf{D} = \mathbf{I}_T \). Using this eigendecomposition, we can rewrite \( \mathbf{V}_i \) as:

\[ \mathbf{V}_i = \mathbf{D} \Lambda^{1/2} \left( \phi_{ii}^* \mathbf{V}_G + \mathbf{I}_T \right) \Lambda^{1/2} \mathbf{D}^T, \]  
(6.22)

\[ \tilde{\mathbf{V}}_G = \Lambda^{1/2} \mathbf{D}^T \mathbf{V}_G \mathbf{D} \Lambda^{-1/2}. \]
(6.23)

Let \( \mathbf{RKR}^T \) denote the eigenvalue decomposition of \( \tilde{\mathbf{V}}_G \). Then, we can rewrite \( \mathbf{V}_i \) as:

\[ \mathbf{V}_i = \mathbf{D} \Lambda^{1/2} \mathbf{R} \left( \phi_{ii}^* \mathbf{K} + \mathbf{I}_T \right) \mathbf{R}^T \Lambda^{1/2} \mathbf{D}^T. \]
(6.24)
The corresponding inverse of \( V_i \) is now also straightforward:

\[
V_i^{-1} = D \Lambda^{-1/2} R \left( \phi_{ii}^2 K + I_T \right)^{-1} R^\top \Lambda^{-1/2} D^\top. \tag{6.25}
\]

Using this, we will now derive a quick expression for the determinant of \( V \):

\[
|V| = \prod_{i=1}^{N} |V_i|
\]

\[
= \prod_{i=1}^{N} \left( |D\Lambda^{1/2} R \left( \phi_{ii}^2 K + I_T \right) R^\top \Lambda^{1/2} D^\top| \right)
\]

\[
= \prod_{i=1}^{N} \left( |D| \left| \Lambda^{1/2} \right| |R| \left| \phi_{ii}^2 K + I_T \right| |R^\top| \left| \Lambda^{1/2} \right| |D^\top| \right)
\]

\[
= \prod_{i=1}^{N} \left( \left| \phi_{ii}^2 K + I_T \right| |R^\top| |R| \left| \Lambda^{1/2} \right| |D^\top| \right)
\]

\[
= \prod_{i=1}^{N} \left( \left| \phi_{ii}^2 K + I_T \right| |R R^\top| \left| \Lambda \right| |D^\top D| \right)
\]

\[
= \prod_{i=1}^{N} \left( \left| \phi_{ii}^2 K + I_T \right| |R R^\top| \left| \Lambda \right| \right)
\]

\[
= \prod_{i=1}^{N} \left( \left| \phi_{ii}^2 K + I_T \right| \prod_{t=1}^{T} \left| \lambda_t \right| \right)
\]

\[
= \prod_{i=1}^{N} \left( \left| \phi_{ii}^2 K + I_T \right| \prod_{t=1}^{T} \left| \lambda_t \right| \right)
\]

where \( \kappa_t \) is the \( t \)-th diagonal entry of \( K \), and \( \lambda_t \) is defined analogously with respect to \( \Lambda \). Now, the log determinant of \( V \) is given by:

\[
\log |V| = N \sum_{t=1}^{T} \log (\lambda_t) + \sum_{i=1}^{N} \sum_{t=1}^{T} \log \left( \phi_{ii}^2 \kappa_t + 1 \right). \tag{6.27}
\]

The third part of the log-likelihood in (6.15) can be made computationally easier by plugging in the inverse of \( V_i \):

\[
\sum_{i=1}^{N} \tilde{y}_i^\top \left( \phi_{ii}^2 V_G + V_E \right)^{-1} \tilde{y}_i
\]

\[
= \sum_{i=1}^{N} \tilde{y}_i^\top D \Lambda^{-1/2} R \left( \phi_{ii}^2 K + I_T \right)^{-1} R^\top \Lambda^{-1/2} D^\top \tilde{y}_i
\]

\[
= \sum_{i=1}^{N} \tilde{y}_i^\top F^\top \left( \phi_{ii}^2 K + I_T \right)^{-1} F \tilde{y}_i \tag{6.28}
\]

with \( F = R^\top \Lambda^{-1/2} D^\top \). Now, by combining the terms, we have a computationally
6. Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence

Efficient expression for the log-likelihood in (6.15):

\[
\log l(\theta) = -\frac{1}{2} N \sum_{t=1}^{T} \log(\lambda_t) + \sum_{i=1}^{N} \sum_{t=1}^{T} \log\left(\phi_{ii}^2 \kappa_t + 1\right) \\
+ \sum_{i=1}^{N} \tilde{y}_i^\top D \Lambda^{-1/2} R \left(\phi_{ii}^2 K + I_T\right)^{-1} R^\top \Lambda^{-1/2} D^\top \tilde{y}_i. \tag{6.29}
\]

Note that \(\left(\phi_{ii}^2 K + I_T\right)^{-1}\) is a diagonal matrix with easily computed diagonal elements \((\phi_{ii}^2 \kappa_t + 1)^{-1}\).

Based on the above, to evaluate \(\log l(\theta)\) with \(\theta = \theta_k + \alpha p_k\) for a given value of \(\alpha\), the following steps lead to a computationally efficient evaluation.

1. Compute the eigen decomposition of \(V_E = D \Lambda D^\top\).
2. Compute \(\tilde{V}_G = \Lambda^{-1/2} D^\top V_G D \Lambda^{-1/2}\) by (6.23).
3. Compute the eigen decomposition \(\tilde{V}_G = RKR^\top\).
4. Compute \(F = R^\top \Lambda^{1/2} D^\top\).
5. Compute the loglikelihood as:

\[
\log l(\theta) = -\frac{1}{2} N \sum_{t=1}^{T} \log(\lambda_t) + \sum_{i=1}^{N} \sum_{t=1}^{T} \log\left(\phi_{ii}^2 \kappa_t + 1\right) \\
+ \sum_{i=1}^{N} \sum_{t=1}^{T} (\phi_{ii}^2 \kappa_t + 1)^{-1} \left(\sum_{s=1}^{T} f_{is} \tilde{y}_{is}\right)^2. \tag{6.30}
\]

The order of the number of operations to compute the eigen decompositions in Steps 1 and 3 is \(O(T^3)\) (Demmel, 1997), and from the expressions it may be verified that those in Steps 2 and 4 are \(O(T^2)\), and in Step 5 \(O(NT^2)\). As in our applications \(N \gg T\), the computational costs for a single function evaluation are of the order \(O(NT^2)\).

**Step 5b. Efficient expressions for evaluating the gradient of the log likelihood** In the BFGS algorithm, we need to evaluate the gradient of the log-likelihood, \(\nabla \log l(\theta_k)\). For ease of notation we will drop the subscript \(k\) in this section. Then, the gradient \(\nabla \log l(\theta_k)\) has elements \(\partial \log l(\theta_k)/\partial \theta_\ell\). For a given parameter \(\theta_\ell\) in set of parameters \(\theta\) (using index \(\ell\) without loss of generality),
the gradient of the log-likelihood is given by:

$$\frac{\partial \log l(\theta_k)}{\partial \theta_\ell} = \frac{1}{2} \bar{y}^T V^{-1} \frac{\partial V}{\partial \theta_\ell} V^{-1} \bar{y} - \frac{1}{2} \text{tr}\left(V^{-1} \frac{\partial V}{\partial \theta_\ell}\right)$$

$$= \frac{1}{2} \sum_{i=1}^{N} r_i^T \frac{\partial V_i}{\partial \theta_\ell} r_i - \frac{1}{2} \sum_{i=1}^{N} \text{tr}\left(V_i^{-1} \frac{\partial V_i}{\partial \theta_\ell}\right),$$

where \(r_i = V_i^{-1} \bar{y}_i = F_i^T \left(\phi_{ii}^2 K + I_T\right)^{-1} \bar{F} \bar{y}_i\). Let us now consider an expression for \(\frac{\partial V_i}{\partial \theta_\ell}\). In case \(\theta_\ell\) pertains to an element in \(\Gamma_G\), then:

$$\frac{\partial V_i}{\partial \theta_\ell} = \phi_{ii}^2 \left(\Gamma_G \frac{\partial \Gamma_G^T}{\partial \theta_\ell}\right)^T + \left(\Gamma_G \frac{\partial \Gamma_G^T}{\partial \theta_\ell}\right) = \phi_{ii}^2 V_\ell. \quad (6.31)$$

Note that this derivative only differs across observations in scale (i.e., scaled by \(\phi_{ii}^2\) for \(i = 1, \ldots, N\)). Also, \(\frac{\partial \Gamma_G^T}{\partial \theta_\ell}\) has a special form, that is, it is a \(T \times T\) matrix with all values being zero except a one in the position that corresponds to \(\theta_\ell\) in \(\Gamma_G\). Suppose that \(\theta_\ell\) refers to element \(j, j'\) of \(\Gamma_G\) with \(j \geq j'\) as \(\Gamma_G\) is a lower triangular matrix. Then, \(\left(\Gamma_G \frac{\partial \Gamma_G^T}{\partial \theta_\ell}\right)\) results in a \(T \times T\) matrix of zeros except in column \(j\) which is a copy of column \(j'\) of \(\Gamma_G^T\). This allows us to conveniently express:

$$\frac{1}{2} \bar{y}^T V^{-1} \frac{\partial V}{\partial \theta_\ell} V^{-1} \bar{y} = \sum_{i=1}^{N} \phi_{ii}^2 \bar{y}_i^T F_i \left(\phi_{ii}^2 K + I_T\right)^{-1} \bar{F}_i \bar{V}_\ell \bar{F}^T \left(\phi_{ii}^2 K + I_T\right)^{-1} \bar{F} \bar{y}_i,$$

and

$$-\frac{1}{2} \sum_{i=1}^{N} \text{tr}\left(V_i^{-1} \frac{\partial V_i}{\partial \theta_\ell}\right) = -\frac{1}{2} \sum_{i=1}^{N} \phi_{ii}^2 \text{tr}\left(\left(\phi_{ii}^2 K + I_T\right)^{-1} \bar{F}_i \bar{V}_\ell \bar{F}^T \left(\phi_{ii}^2 K + I_T\right)^{-1} \bar{F} \bar{y}_i\right).$$

If we define \(c_{i,\theta_\ell} = (\frac{\partial V_i}{\partial \theta_\ell})r_i\), we can reduce the expression of the gradient to:

$$\frac{\partial \log l(\theta_k)}{\partial \theta_\ell} = \frac{1}{2} \sum_{i=1}^{N} r_i^T c_{i,\theta_\ell} - \frac{1}{2} \sum_{i=1}^{N} \text{tr}\left(V_i^{-1} \frac{\partial V_i}{\partial \theta_\ell}\right). \quad (6.32)$$

Similarly, when \(\theta_\ell\) pertains to environment components, we have that:

$$\frac{\partial V_i}{\partial \theta_\ell} = \left(\Gamma_E \frac{\partial \Gamma_E^T}{\partial \theta_\ell}\right)^T + \left(\Gamma_E \frac{\partial \Gamma_E^T}{\partial \theta_\ell}\right). \quad (6.33)$$

These derivatives of the environment term are equal across observations.

When estimating a vector of parameters, we can use the Hessian matrix, the matrix of second partials of the log-likelihood, to approximate the variance of our parameters. Element $ij$ is given by:

$$H_{ij} = \frac{\partial^2 \log l(\theta)}{\partial \theta_i \partial \theta_j}.$$  \hfill (6.34)

$H(\theta_{ML})$ refers to the Hessian matrix evaluated at our optimum $\theta_{ML}$ and gives us a measure of the curvatures of our log-likelihood function at the optimum. Given that calculating this Hessian $H$ is a computational burden, we instead opt for a different approach using the Fisher information matrix $I(\theta_{ML})$, the negative of the expected value of the Hessian matrix, that is:

$$I(\theta_{ML}) = -E[H(\theta_{ML})].$$  \hfill (6.35)

This Fisher information matrix gives a measure of the multidimensional curvature of the log-likelihood. Alternatively, one can calculate it using the expected value of the outer product of the gradient of the log-likelihood

$$I(\theta) = E \left[ \left( \frac{\partial \log l(\theta)}{\partial \theta} \right) \left( \frac{\partial \log l(\theta)}{\partial \theta} \right)^\top \right].$$  \hfill (6.36)

Now, the covariance matrix of our maximum likelihood estimates is simply the inversion of the information matrix:

$$\text{Var}(\hat{\theta}_{ML}) = J^{-1}(\theta_{ML}).$$  \hfill (6.37)

As we calculate the Fisher information matrix using the outer product of the gradient, we have to calculate the gradient once more in this optimum. Now, we can use (6.36) and (6.37) to calculate the standard errors of our parameter estimates.

Step 7. Derive standard errors for correlation and heritability estimates  The estimation procedure returns the parametrization in terms of $\theta$, that is, in the Cholesky decompositions of $V_G$ and $V_E$. Practically it may be more interesting to consider instead the covariance matrices $V_G$ and $V_E$, correlation matrix, and the heritability estimates. In this section, the appropriate standard
errors for these transformations are presented using the delta method.

The delta method states that for some function \( g(\hat{\theta}_{\text{ML}}) \) the function in the optimum \( \hat{\theta}_{\text{ML}} \) is distributed as:

\[
g(\hat{\theta}_{\text{ML}}) \sim N\left( g(\hat{\theta}_{\text{ML}}), \nabla g(\hat{\theta}_{\text{ML}})^\top \mathcal{J}^{-1}(\hat{\theta}_{\text{ML}}) \nabla g(\hat{\theta}_{\text{ML}}) \right),
\]

where \( \nabla g(\theta) \) is the gradient of \( g() \) with respect to \( \theta \). Below, the functions \( g(\theta) \) and their gradients are defined to find the estimates of the standard errors for the heritability, genetic variance-covariance matrix, and genetic correlation matrix. Note that the derivations for the environmental variance-covariance matrix and its correlation matrix are equivalent to that its genetic counterpart and is therefore omitted.

Let \( v_{ij}^G(\theta) \) be element \( i,j \) of the genetic covariance matrix \( V_G \). Then:

\[
v_{ij}^G(\theta) = \theta^\top A_i^G A_j^G \theta,
\]

where \( A_i^G \) is a matrix with zeros and ones such that \( A_i^G \theta \) selects the row \( i \) elements of \( \Gamma_G \). The gradient of \( v_{ij}^G(\theta) \) is given by:

\[
\nabla v_{ij}^G(\theta) = \left( A_i^G A_j^G + A_j^G A_i^G \right) \theta.
\]

Using (6.39) and the gradient derived above and substitute \( g(\theta) \) in (6.38) by \( v_{ij}^G(\theta) \) gives the variance of \( v_{ij}^G(\theta) \) and its square root the standard deviation.

The heritability of trait \( i \) can be written as:

\[
h_i^2(\theta) = \frac{v_{ii}^G(\theta)}{v_{ii}^G(\theta) + v_{ii}^E(\theta)}.
\]

Using the chain rule, we obtain the corresponding gradient for the heritability of trait \( i \) is:

\[
\nabla h_i^2(\theta) = \frac{2(1 - h_i^2(\theta))}{v_{ii}^G(\theta) + v_{ii}^E(\theta)} A_i^G A_i^\top \theta - \frac{2h_i^2(\theta)}{v_{ii}^G(\theta) + v_{ii}^E(\theta)} A_i^E A_i^\top \theta.
\]

The standard deviation of \( h_i^2(\theta) \) can be obtained in an analogue way as for \( v_{ij}^G(\theta) \) using \( \nabla h_i^2(\theta) \) and the delta method.

The genetic correlation matrix \( R_G \) has elements

\[
r_{ij}(\theta) = \frac{v_{ij}(\theta)}{v_{ii}^{1/2}(\theta)v_{jj}^{1/2}(\theta)},
\]

where \( v_{ij}(\theta) \) is the covariance between traits \( i \) and \( j \), and \( v_{ii}^{1/2}(\theta) \) and \( v_{jj}^{1/2}(\theta) \) are the standard deviations of traits \( i \) and \( j \), respectively.
where we have dropped the superscript $G$ for notational simplicity. Then, its gradient becomes

$$\nabla r_{ij}(\theta) = \frac{1}{\sqrt{v_{ii}(\theta)} \sqrt{v_{jj}(\theta)}} \left( A_i^\top A_j + A_j^\top A_i \right) \theta - \frac{r_{ij}(\theta)}{\sqrt{v_{ii}(\theta)} \sqrt{v_{jj}(\theta)}} \left( \sqrt{v_{ii}(\theta)} A_i \theta + \sqrt{v_{jj}(\theta)} A_j \theta \right).$$

Again, from $\nabla r_{ij}(\theta)$ and the delta method the standard deviation for $r_{ij}(\theta)$ can be obtained.

6.B DATA USAGE FOR CONSTRUCTING PHENOTYPES

Table 6.1 provides an overview of the phenotypes used in our study.

**Table 6.1 – UK Biobank phenotype data used in this study, with corresponding description, measurement units and data fields.**

<table>
<thead>
<tr>
<th>Trait</th>
<th>Description</th>
<th>Measurement units</th>
<th>UK Biobank data field</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>Logarithm average body mass index for all measurements</td>
<td>Kg/m$^2$</td>
<td>21001</td>
</tr>
<tr>
<td>Depression Score</td>
<td>Logarithm of first PC of depression intensity and frequency</td>
<td>NA</td>
<td>2050, 2060, 4609, 4620, 5375, 5386, 2090, 2100</td>
</tr>
<tr>
<td>Drinks consumed</td>
<td>Logarithm drinks per week</td>
<td>Number of units of alcohol per week</td>
<td>1558, 1568, 1578, 1588, 1598, 1608, 4407, 4418, 4429, 4440, 4451, 4462, 5364</td>
</tr>
<tr>
<td>Educational attainment</td>
<td>Highest self-reported schooling degree converted to US-schooling year equivalents using ISCED categories</td>
<td>years</td>
<td>6138</td>
</tr>
<tr>
<td>Grey matter in Amygdala</td>
<td>Volume of grey matter in Amygdala (left+right)</td>
<td>mm$^3$</td>
<td>25888, 25889</td>
</tr>
<tr>
<td>Grey matter in Angular Gyrus</td>
<td>Volume of grey matter in Angular Gyrus</td>
<td>mm$^3$</td>
<td>25822, 25823</td>
</tr>
<tr>
<td>Grey matter in Brain-Stem</td>
<td>Volume of grey matter in Brain-Stem</td>
<td>mm$^3$</td>
<td>25892</td>
</tr>
<tr>
<td>Grey matter in Caudate</td>
<td>Volume of grey matter in Caudate (left+right)</td>
<td>mm$^3$</td>
<td>25880, 25881</td>
</tr>
<tr>
<td>Grey matter in Central Opercular Cortex</td>
<td>Volume of grey matter in Central Opercular Cortex (left+right)</td>
<td>mm$^3$</td>
<td>25864, 25865</td>
</tr>
<tr>
<td>Trait</td>
<td>Description</td>
<td>Measurement units</td>
<td>UK Biobank data field</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Grey matter in Cingulate Gyrus, (ad)</td>
<td>Volume of grey matter in Cingulate Gyrus, anterior division (left+right)</td>
<td>mm$^3$</td>
<td>25838, 25839</td>
</tr>
<tr>
<td>Grey matter in Cingulate Gyrus, (pd)</td>
<td>Volume of grey matter in Cingulate Gyrus, posterior division (left+right)</td>
<td>mm$^3$</td>
<td>25840, 25841</td>
</tr>
<tr>
<td>Grey matter in Crus I Cerebellum</td>
<td>Volume of grey matter in Crus I Cerebellum (left-right)</td>
<td>mm$^3$</td>
<td>25900, 25902</td>
</tr>
<tr>
<td>Grey matter in Crus I Cerebellum (vermis)</td>
<td>Volume of grey matter in Crus I Cerebellum (vermis)</td>
<td>mm$^3$</td>
<td>25901</td>
</tr>
<tr>
<td>Grey matter in Crus II Cerebellum</td>
<td>Volume of grey matter in Crus II Cerebellum (vermis)</td>
<td>mm$^3$</td>
<td>25903, 25905</td>
</tr>
<tr>
<td>Grey matter in Crus II Cerebellum (left-right)</td>
<td>Volume of grey matter in Crus II Cerebellum (left-right)</td>
<td>mm$^3$</td>
<td>25844, 25845</td>
</tr>
<tr>
<td>Grey matter in Cuneal Cortex</td>
<td>Volume of grey matter in Cuneal Cortex (left-right)</td>
<td>mm$^3$</td>
<td>25830, 25831</td>
</tr>
<tr>
<td>Grey matter in Frontal Medial Cortex</td>
<td>Volume of grey matter in Frontal Medial Cortex (left-right)</td>
<td>mm$^3$</td>
<td>25862, 25863</td>
</tr>
<tr>
<td>Grey matter in Frontal Operculum Cortex</td>
<td>Volume of grey matter in Frontal Operculum Cortex (left+right)</td>
<td>mm$^3$</td>
<td>25846, 25847</td>
</tr>
<tr>
<td>Grey matter in Frontal Orbital Cortex</td>
<td>Volume of grey matter in Frontal Orbital Cortex (left-right)</td>
<td>mm$^3$</td>
<td>25782, 25783</td>
</tr>
<tr>
<td>Grey matter in Frontal Pole</td>
<td>Volume of grey matter in Frontal Pole (left+right)</td>
<td>mm$^3$</td>
<td>25870, 25871</td>
</tr>
<tr>
<td>Grey matter in Heschl's Gyrus</td>
<td>Volume of grey matter in Heschl's Gyrus (includes H1 and H2) (left+right)</td>
<td>mm$^3$</td>
<td>25886, 25887</td>
</tr>
<tr>
<td>Grey matter in Hippocampus</td>
<td>Volume of grey matter in Hippocampus (left+right)</td>
<td>mm$^3$</td>
<td>25893, 25894</td>
</tr>
<tr>
<td>Grey matter in I-IV Cerebellum</td>
<td>Volume of grey matter in I-IV Cerebellum (left+right)</td>
<td>mm$^3$</td>
<td>25792, 25793</td>
</tr>
<tr>
<td>Grey matter in Inferior Frontal Gyrus, po</td>
<td>Volume of grey matter in Inferior Frontal Gyrus, pars opercularis (left+right)</td>
<td>mm$^3$</td>
<td>25790, 25790</td>
</tr>
<tr>
<td>Grey matter in Inferior Frontal Gyrus, pt</td>
<td>Volume of grey matter in Inferior Frontal Gyrus, pars triangularis (left+right)</td>
<td>mm$^3$</td>
<td></td>
</tr>
</tbody>
</table>
6. **Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence**

<table>
<thead>
<tr>
<th>Trait</th>
<th>Description</th>
<th>Measurement units</th>
<th>UK Biobank data field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grey matter in Inferior Temporal Gyrus, (tp)</td>
<td>Volume of grey matter in Inferior Temporal Gyrus, temporooccipital part (left+right)</td>
<td>mm$^3$</td>
<td>25812, 25813</td>
</tr>
<tr>
<td>Grey matter in Inferior Temporal Gyrus, (ad)</td>
<td>Volume of grey matter in Inferior Temporal Gyrus, anterior division (left+right)</td>
<td>mm$^3$</td>
<td>25808, 25808</td>
</tr>
<tr>
<td>Grey matter in Inferior Temporal Gyrus, (pd)</td>
<td>Volume of grey matter in Inferior Temporal Gyrus, posterior division (left+right)</td>
<td>mm$^3$</td>
<td>25810, 25811</td>
</tr>
<tr>
<td>Grey matter in Insular Cortex</td>
<td>Volume of grey matter in Insular Cortex (left+right)</td>
<td>mm$^3$</td>
<td>25784, 25785</td>
</tr>
<tr>
<td>Grey matter in Intracalcarine Cortex</td>
<td>Volume of grey matter in Intracalcarine Cortex (left+right)</td>
<td>mm$^3$</td>
<td>25828, 25829</td>
</tr>
<tr>
<td>Grey matter in IX Cerebellum</td>
<td>Volume of grey matter in IX Cerebellum (left+right)</td>
<td>mm$^3$</td>
<td>25915, 25917</td>
</tr>
<tr>
<td>Grey matter in Juxtapositional Lobule Cortex</td>
<td>Volume of grey matter in Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex) (left+right)</td>
<td>mm$^3$</td>
<td>25832, 25833</td>
</tr>
<tr>
<td>Grey matter in Lateral Occipital Cortex, (id)</td>
<td>Volume of grey matter in Lateral Occipital Cortex, inferior division (left+right)</td>
<td>mm$^3$</td>
<td>25826, 25827</td>
</tr>
<tr>
<td>Grey matter in Lateral Occipital Cortex, (sd)</td>
<td>Volume of grey matter in Lateral Occipital Cortex, superior division (left+right)</td>
<td>mm$^3$</td>
<td>25824, 25825</td>
</tr>
<tr>
<td>Grey matter in Lingual Gyrus</td>
<td>Volume of grey matter in Lingual Gyrus (left+right)</td>
<td>mm$^3$</td>
<td>25852, 25853</td>
</tr>
<tr>
<td>Grey matter in Middle Frontal Gyrus</td>
<td>Volume of grey matter in Middle Frontal Gyrus (left+right)</td>
<td>mm$^3$</td>
<td>25788, 25789</td>
</tr>
<tr>
<td>Grey matter in Middle Temporal Gyrus, (tp)</td>
<td>Volume of grey matter in Middle Temporal Gyrus, temporooccipital part (left+right)</td>
<td>mm$^3$</td>
<td>25806, 25807</td>
</tr>
<tr>
<td>Grey matter in Middle Temporal Gyrus, (ad)</td>
<td>Volume of grey matter in Middle Temporal Gyrus, anterior division (left+right)</td>
<td>mm$^3$</td>
<td>25802, 25803</td>
</tr>
<tr>
<td>Trait</td>
<td>Description</td>
<td>Measurement units</td>
<td>UK Biobank data field</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Grey matter in Middle Temporal Gyrus, (pd)</td>
<td>Volume of grey matter in Middle Temporal Gyrus, posterior division (left+right)</td>
<td>mm$^3$</td>
<td>25804, 25805</td>
</tr>
<tr>
<td>Grey matter in Occipital Fusiform Gyrus</td>
<td>Volume of grey matter in Occipital Fusiform Gyrus (left+right)</td>
<td>mm$^3$</td>
<td>25860, 25861</td>
</tr>
<tr>
<td>Grey matter in Occipital Pole</td>
<td>Volume of grey matter in Occipital Pole (left+right)</td>
<td>mm$^3$</td>
<td>25876, 25877</td>
</tr>
<tr>
<td>Grey matter in Pallidum</td>
<td>Volume of grey matter in Pallidum (left+right)</td>
<td>mm$^3$</td>
<td>25884, 25884</td>
</tr>
<tr>
<td>Grey matter in Paracingulate Gyrus</td>
<td>Volume of grey matter in Paracingulate Gyrus (left+right)</td>
<td>mm$^3$</td>
<td>25836, 25837</td>
</tr>
<tr>
<td>Grey matter in Parahippocampal Gyrus, (ad)</td>
<td>Volume of grey matter in Parahippocampal Gyrus, anterior division (left+right)</td>
<td>mm$^3$</td>
<td>25848, 25849</td>
</tr>
<tr>
<td>Grey matter in Parahippocampal Gyrus, (pd)</td>
<td>Volume of grey matter in Parahippocampal Gyrus, posterior division (left+right)</td>
<td>mm$^3$</td>
<td>25850, 25851</td>
</tr>
<tr>
<td>Grey matter in Parietal Operculum Cortex</td>
<td>Volume of grey matter in Parietal Operculum Cortex (left+right)</td>
<td>mm$^3$</td>
<td>25866, 25867</td>
</tr>
<tr>
<td>Grey matter in Planum Polare</td>
<td>Volume of grey matter in Planum Polare (left+right)</td>
<td>mm$^3$</td>
<td>25868, 25869</td>
</tr>
<tr>
<td>Grey matter in Planum Temporale</td>
<td>Volume of grey matter in Planum Temporale (left+right)</td>
<td>mm$^3$</td>
<td>25872, 25783</td>
</tr>
<tr>
<td>Grey matter in Postcentral Gyrus</td>
<td>Volume of grey matter in Postcentral Gyrus (left+right)</td>
<td>mm$^3$</td>
<td>25814, 25815</td>
</tr>
<tr>
<td>Grey matter in Precentral Gyrus</td>
<td>Volume of grey matter in Precentral Gyrus (left+right)</td>
<td>mm$^3$</td>
<td>25794, 25795</td>
</tr>
<tr>
<td>Grey matter in Precuneous Cortex</td>
<td>Volume of grey matter in Precuneous Cortex (left+right)</td>
<td>mm$^3$</td>
<td>25842, 25843</td>
</tr>
<tr>
<td>Grey matter in Putamen</td>
<td>Volume of grey matter in Putamen (left+right)</td>
<td>mm$^3$</td>
<td>25882, 25883</td>
</tr>
<tr>
<td>Grey matter in Subcallosal Cortex</td>
<td>Volume of grey matter in Subcallosal Cortex (left+right)</td>
<td>mm$^3$</td>
<td>25834, 25835</td>
</tr>
<tr>
<td>Grey matter in Superior Frontal Gyrus</td>
<td>Volume of grey matter in Superior Frontal Gyrus (left)</td>
<td>mm$^3$</td>
<td>25786</td>
</tr>
</tbody>
</table>
6. Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence

<table>
<thead>
<tr>
<th>Trait</th>
<th>Description</th>
<th>Measurement units</th>
<th>UK Biobank data field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grey matter in Superior Parietal Lobule</td>
<td>Volume of grey matter in Superior Parietal Lobule (left+right)</td>
<td>mm$^3$</td>
<td>25816, 25817</td>
</tr>
<tr>
<td>Grey matter in Superior Temporal Gyrus, (ad)</td>
<td>Volume of grey matter in Superior Temporal Gyrus, anterior division (left+right)</td>
<td>mm$^3$</td>
<td>25798, 25799</td>
</tr>
<tr>
<td>Grey matter in Superior Temporal Gyrus, (pd)</td>
<td>Volume of grey matter in Superior Temporal Gyrus, posterior division (left+right)</td>
<td>mm$^3$</td>
<td>25800, 25801</td>
</tr>
<tr>
<td>Grey matter in Suprachalcarine Cortex</td>
<td>Volume of grey matter in Suprachalcarine Cortex (left+right)</td>
<td>mm$^3$</td>
<td>25874, 25875</td>
</tr>
<tr>
<td>Grey matter in Supramarginal Gyrus, (ad)</td>
<td>Volume of grey matter in Supramarginal Gyrus, anterior division (left+right)</td>
<td>mm$^3$</td>
<td>25818, 25819</td>
</tr>
<tr>
<td>Grey matter in Supramarginal Gyrus, (pd)</td>
<td>Volume of grey matter in Supramarginal Gyrus, posterior division (left+right)</td>
<td>mm$^3$</td>
<td>25820, 25821</td>
</tr>
<tr>
<td>Grey matter in Temporal Fusiform Cortex, (ad)</td>
<td>Volume of grey matter in Temporal Fusiform Cortex, anterior division (left+right)</td>
<td>mm$^3$</td>
<td>25854, 25855</td>
</tr>
<tr>
<td>Grey matter in Temporal Fusiform Cortex, (pd)</td>
<td>Volume of grey matter in Temporal Fusiform Cortex, posterior division (left+right)</td>
<td>mm$^3$</td>
<td>25856, 25857</td>
</tr>
<tr>
<td>Grey matter in Temporal Occipital Fusiform Cortex</td>
<td>Volume of grey matter in Temporal Occipital Fusiform Cortex (left+right)</td>
<td>mm$^3$</td>
<td>25858, 25859</td>
</tr>
<tr>
<td>Grey matter in Temporal Pole</td>
<td>Volume of grey matter in Temporal Pole (left+right)</td>
<td>mm$^3$</td>
<td>25796, 25797</td>
</tr>
<tr>
<td>Grey matter in Thalamus</td>
<td>Volume of grey matter in Thalamus (left+right)</td>
<td>mm$^3$</td>
<td>25878, 25879</td>
</tr>
<tr>
<td>Grey matter in V Cerebellum</td>
<td>Volume of grey matter in V Cerebellum (left+right)</td>
<td>mm$^3$</td>
<td>25895, 25896</td>
</tr>
<tr>
<td>Grey matter in Ventral Striatum</td>
<td>Volume of grey matter in Ventral Striatum (left+right)</td>
<td>mm$^3$</td>
<td>25890, 25891</td>
</tr>
<tr>
<td>Grey matter in VI Cerebellum</td>
<td>Volume of grey matter in VI Cerebellum (left+right)</td>
<td>mm$^3$</td>
<td>25897, 25899</td>
</tr>
</tbody>
</table>
## Trait Description Measurement units UK Biobank data field

<table>
<thead>
<tr>
<th>Trait</th>
<th>Description</th>
<th>Measurement units</th>
<th>UK Biobank data field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grey matter in VI Cerebellum</td>
<td>Volume of grey matter in VI Cerebellum (vermis)</td>
<td>mm³</td>
<td>25898</td>
</tr>
<tr>
<td>Grey matter in VIIb Cerebellum</td>
<td>Volume of grey matter in VIIb Cerebellum (left+right)</td>
<td>mm³</td>
<td>25906, 25908</td>
</tr>
<tr>
<td>Grey matter in VIIb Cerebellum</td>
<td>Volume of grey matter in VIIb Cerebellum (vermis)</td>
<td>mm³</td>
<td>25907</td>
</tr>
<tr>
<td>Grey matter in VIIIa Cerebellum</td>
<td>Volume of grey matter in VIIIa Cerebellum (left+right)</td>
<td>mm³</td>
<td>25909, 25911</td>
</tr>
<tr>
<td>Grey matter in VIIIb Cerebellum</td>
<td>Volume of grey matter in VIIIb Cerebellum (vermis)</td>
<td>mm³</td>
<td>25910</td>
</tr>
<tr>
<td>Grey matter in VIIIb Cerebellum</td>
<td>Volume of grey matter in VIIIb Cerebellum (left+right)</td>
<td>mm³</td>
<td>25912, 25914</td>
</tr>
<tr>
<td>Grey matter in VIIIb Cerebellum</td>
<td>Volume of grey matter in VIIIb Cerebellum (vermis)</td>
<td>mm³</td>
<td>25913</td>
</tr>
<tr>
<td>Grey matter in X Cerebellum</td>
<td>Volume of grey matter in X Cerebellum (left+right)</td>
<td>mm³</td>
<td>25918, 25920</td>
</tr>
<tr>
<td>Grey matter in X Cerebellum</td>
<td>Volume of grey matter in X Cerebellum (vermis)</td>
<td>mm³</td>
<td>25919</td>
</tr>
<tr>
<td>IQ</td>
<td>Standardized fluid intelligence score</td>
<td>correct-answers</td>
<td>20016, 20191</td>
</tr>
<tr>
<td>Reaction Time</td>
<td>Standardized reaction time</td>
<td>milliseconds</td>
<td>20023</td>
</tr>
<tr>
<td>Standing height</td>
<td>Standing height</td>
<td>cm</td>
<td>50</td>
</tr>
<tr>
<td>Subjective Well-being</td>
<td>Subjective well-being: In general how happy are you? (Average value over time)</td>
<td>NA</td>
<td>4526, 20458</td>
</tr>
<tr>
<td>Visual memory</td>
<td>Log standardized visual memory score</td>
<td>NA</td>
<td>399, 20132</td>
</tr>
<tr>
<td>Volume of brain</td>
<td>Volume of brain, grey+white matter</td>
<td>mm³</td>
<td>25010</td>
</tr>
</tbody>
</table>

### 6.C Data cells used for identification of brain damage

To make sure our find patterns are not due to individuals with brain diseases or surgical damage, we remove all individuals with brain diseases. In Table 6.2 we have listed the brain diseases with corresponding ICD10 codes used as exclusion...
6. **Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence**

6.1 Table 6.2 – *Brain diseases with corresponding data fields in the self report and ICD10 codes.*

<table>
<thead>
<tr>
<th>Disease</th>
<th>UK Biobank data field</th>
<th>ICD10 code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dementia or Alzheimer’s disease</td>
<td>1263</td>
<td>F01, F02, G30</td>
</tr>
<tr>
<td>Parkinson’s disease</td>
<td>1262</td>
<td>G20, G21</td>
</tr>
<tr>
<td>Chronic degenerative neurological</td>
<td>1258</td>
<td>G23, G31, G32</td>
</tr>
<tr>
<td>Guillain-Barré syndrome</td>
<td>1256</td>
<td>G610</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>1261</td>
<td>G35</td>
</tr>
<tr>
<td>Other demyelinating disease</td>
<td>1397</td>
<td>G37</td>
</tr>
<tr>
<td>Stroke or ischaemic stroke</td>
<td>1081</td>
<td>G463, G464, I64, I694</td>
</tr>
<tr>
<td>Brain cancer</td>
<td>1031</td>
<td>I60, I61, I62, I691, I692, I693</td>
</tr>
<tr>
<td>Brain haemorrhage</td>
<td>1491</td>
<td>G35</td>
</tr>
<tr>
<td>Brain/intracranial abscess</td>
<td>1245</td>
<td>G060, G07</td>
</tr>
<tr>
<td>Cerebral aneurysm</td>
<td>1425</td>
<td>I671, Q282, Q283</td>
</tr>
<tr>
<td>Cerebral palsy</td>
<td>1433</td>
<td>G80, A521, A504, I64</td>
</tr>
<tr>
<td>Encephalitis</td>
<td>1246</td>
<td>A83, A86, B011, B020, B262, A85, B004, B582, A84, B050, B941, G04, A321, G05</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>1264</td>
<td>G40, F803</td>
</tr>
<tr>
<td>Head injury</td>
<td>1266</td>
<td>S07, T040</td>
</tr>
<tr>
<td>Infections of the nervous system</td>
<td>1244</td>
<td>A80, A81, A82, A83, A84, A85, A86, A87, A88, A89</td>
</tr>
<tr>
<td>Ischaemic stroke</td>
<td>1583</td>
<td>G45</td>
</tr>
<tr>
<td>Meningeal cancer</td>
<td>1031</td>
<td>C70, C793</td>
</tr>
<tr>
<td>Meningioma (benign)</td>
<td>1659</td>
<td>D33, D32</td>
</tr>
<tr>
<td>Meningitis</td>
<td>1247</td>
<td>G03, A170, A171, A203, G01, G02, G00, G07</td>
</tr>
<tr>
<td>Motor Neuron Disease (ALS)</td>
<td>1259</td>
<td>G122</td>
</tr>
<tr>
<td>Neurological injury / trauma</td>
<td>1240</td>
<td>G45</td>
</tr>
<tr>
<td>Spina bifida</td>
<td>1524</td>
<td>Q05, Q760</td>
</tr>
<tr>
<td>Subdural haematoma</td>
<td>1083</td>
<td>P100</td>
</tr>
<tr>
<td>Subarachnoid haemorrhage</td>
<td>1086</td>
<td>I60, S066, P103</td>
</tr>
<tr>
<td>Transient ischaemic attack</td>
<td>1082</td>
<td>G45</td>
</tr>
</tbody>
</table>

6.6 Pipeline

The pipeline used to get to our results can be found below:

1. Convert geno-pheno link file to stata format
2. Merge phenotype file with file from step 1
3. Export list over overlapping individuals in geno and pheno data
4. Copy HM3 UKB PLINK data
5. Update FID in FAM files
6. Use PLINK to keep only relevant individuals from step 3

7. Merge across chromosomes

8. Generate list of SNP IDs with imputation quality > 0.9.

9. Use PLINK to extract only SNPS with high imputation quality
   - This leaves 600k directly genotyped SNPs included in HM3 + additional SNPs imputed accurately

10. Regular QC: MAF 0.01, MIND 0.05, GENO 0.05, HWE 0.001

11. Construct GRM, apply relatedness cutoff of 0.025, and inspect lead PCs

12. Drop long-range LD regions from risk GWAS

13. Construct new GRM, apply relatedness cutoff of 0.025 using PLINK, and inspect lead PCs

14. Export new binary (or gzipped) GRM to Python

15. Keep only phenotype data for individuals in GRM from step 13 (after rel.cutoff)

16. Drop additional individuals with possible brain damage

17. Curate phenotype data, including generating genotyping-platform dummy
   - Exclude individuals with too much missingness on phenotypes, and vice versa We have opted for balanced data only for now; N=14,341

18. Put pheno-covar observations in same order as GRM using Bash

19. Import data in python & residualise phenotypes w.r.t. covariates
   - Covariates = sex, age, age², age³, sex x age, sex x age², sex x age³, intercept, batch dummies
   - Covariates for IQ: replace age by IQage (age at moment of assessment), and include dummies for participation in various waves.

20. Transform GRM: take $A^* = \mathbf{MAM}$, where $\mathbf{M}$ is anti-projection matrix based on platform dummy and intercept, and recompute eigenvalue decomposition: $A^* = [\mathbf{P1}, \mathbf{P2}] \text{diag}(D1, D2)[\mathbf{P1}, \mathbf{P2}]^\top$. Store GRM, eigenvalue decomposition, and Y.

21. Calculate values for warm start with bivariate GREML

22. Run M-GREML on relative brain volume + genetic covariates
6.E HERITABILITY ESTIMATES

In Table 6.3, the estimated SNP-heritabilities are shown for all the different phenotypes.

<table>
<thead>
<tr>
<th>Trait</th>
<th>$h^2$</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>0.561</td>
<td>0.046</td>
</tr>
<tr>
<td>log(BMI)</td>
<td>0.257</td>
<td>0.049</td>
</tr>
<tr>
<td>Years of Education</td>
<td>0.209</td>
<td>0.049</td>
</tr>
<tr>
<td>Visual memory</td>
<td>0.147</td>
<td>0.048</td>
</tr>
<tr>
<td>Reaction time</td>
<td>0.128</td>
<td>0.050</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.159</td>
<td>0.049</td>
</tr>
<tr>
<td>Subjective well-being</td>
<td>0.085</td>
<td>0.049</td>
</tr>
<tr>
<td>Depressive symptoms</td>
<td>0.117</td>
<td>0.050</td>
</tr>
<tr>
<td>Log(drinks per week)</td>
<td>0.160</td>
<td>0.051</td>
</tr>
<tr>
<td>Absolute volume of grey matter</td>
<td>0.361</td>
<td>0.049</td>
</tr>
<tr>
<td>Absolute volume of grey and white matter</td>
<td>0.412</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Frontal Pole</td>
<td>0.255</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Insular Cortex</td>
<td>0.460</td>
<td>0.050</td>
</tr>
<tr>
<td>Relative volume of grey matter in Middle Frontal Gyrus</td>
<td>0.239</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Inferior Frontal Gyrus, pars triangularis</td>
<td>0.182</td>
<td>0.050</td>
</tr>
<tr>
<td>Relative volume of grey matter in Inferior Frontal Gyrus, pars opercularis</td>
<td>0.138</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Precentral Gyrus</td>
<td>0.229</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Temporal Pole</td>
<td>0.315</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Superior Temporal Gyrus, anterior division</td>
<td>0.248</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Superior Temporal Gyrus, posterior division</td>
<td>0.244</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Middle Temporal Gyrus, anterior division</td>
<td>0.205</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Middle Temporal Gyrus, posterior division</td>
<td>0.235</td>
<td>0.050</td>
</tr>
<tr>
<td>Relative volume of grey matter in Middle Temporal Gyrus, temporooccipital part</td>
<td>0.252</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Inferior Temporal Gyrus, anterior division</td>
<td>0.235</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Inferior Temporal Gyrus, posterior division</td>
<td>0.232</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Inferior Temporal Gyrus, temporooccipital part</td>
<td>0.224</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Postcentral Gyrus</td>
<td>0.245</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Superior Parietal Lobule</td>
<td>0.165</td>
<td>0.048</td>
</tr>
<tr>
<td>Trait</td>
<td>$\lambda^2$</td>
<td>Standard error</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Relative volume of grey matter in Supramarginal Gyrus. anterior division</td>
<td>0.170</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Supramarginal Gyrus. posterior division</td>
<td>0.158</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Angular Gyrus</td>
<td>0.178</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Lateral Occipital Cortex. superior division</td>
<td>0.167</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Lateral Occipital Cortex. inferior division</td>
<td>0.188</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Intracalcarine Cortex</td>
<td>0.446</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Frontal Medial Cortex</td>
<td>0.178</td>
<td>0.050</td>
</tr>
<tr>
<td>Relative volume of grey matter in Juxtapositional Lobule Cortex (formerly Supplementary M)</td>
<td>0.175</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Subcallosal Cortex</td>
<td>0.391</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Paracingulate Gyrus</td>
<td>0.264</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Cingulate Gyrus. anterior division</td>
<td>0.216</td>
<td>0.046</td>
</tr>
<tr>
<td>Relative volume of grey matter in Cingulate Gyrus. posterior division</td>
<td>0.208</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Precuneous Cortex</td>
<td>0.287</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Cuneal Cortex</td>
<td>0.238</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Frontal Orbital Cortex</td>
<td>0.299</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Parahippocampal Gyrus. anterior division</td>
<td>0.300</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Parahippocampal Gyrus. posterior division</td>
<td>0.334</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Lingual Gyrus</td>
<td>0.275</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Temporal Fusiform Cortex. anterior division</td>
<td>0.262</td>
<td>0.050</td>
</tr>
<tr>
<td>Relative volume of grey matter in Temporal Fusiform Cortex. posterior division</td>
<td>0.314</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Temporal Occipital Fusiform Cortex</td>
<td>0.245</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Occipital Fusiform Gyrus</td>
<td>0.237</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Frontal Operculum Cortex</td>
<td>0.236</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Central Opercular Cortex</td>
<td>0.296</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Parietal Operculum Cortex</td>
<td>0.300</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Planum Polare</td>
<td>0.297</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Heschls Gyrus (includes H1 and H2)</td>
<td>0.353</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Planum Temporale</td>
<td>0.308</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Supracalcarine Cortex</td>
<td>0.277</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Occipital Pole</td>
<td>0.347</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Thalamus</td>
<td>0.346</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Caudate</td>
<td>0.239</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Putamen</td>
<td>0.355</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Pallidium</td>
<td>0.162</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Hippocampus</td>
<td>0.436</td>
<td>0.047</td>
</tr>
</tbody>
</table>
6. Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence

<table>
<thead>
<tr>
<th>Trait</th>
<th>$h^2$</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative volume of grey matter in Amygdala</td>
<td>0.305</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Ventral Striatum</td>
<td>0.391</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in I-IV Cerebellum</td>
<td>0.372</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in V Cerebellum</td>
<td>0.374</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in VI Cerebellum</td>
<td>0.347</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Crus I Cerebellum</td>
<td>0.426</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Crus II Cerebellum</td>
<td>0.401</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in VIIb Cerebellum</td>
<td>0.373</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in VIIIa Cerebellum</td>
<td>0.319</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in VIIIb Cerebellum</td>
<td>0.300</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in IX Cerebellum</td>
<td>0.370</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in X Cerebellum</td>
<td>0.377</td>
<td>0.049</td>
</tr>
<tr>
<td>Relative volume of grey matter in Superior Frontal Gyrus</td>
<td>0.190</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Brain-Stem</td>
<td>0.241</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Vermis VI Cerebellum</td>
<td>0.336</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Vermis Crus I Cerebellum</td>
<td>0.092</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Vermis Crus II Cerebellum</td>
<td>0.318</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Vermis VIIb Cerebellum</td>
<td>0.290</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Vermis VIIIa Cerebellum</td>
<td>0.337</td>
<td>0.047</td>
</tr>
<tr>
<td>Relative volume of grey matter in Vermis VIIIb Cerebellum</td>
<td>0.352</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Vermis IX Cerebellum</td>
<td>0.311</td>
<td>0.048</td>
</tr>
<tr>
<td>Relative volume of grey matter in Vermis X Cerebellum</td>
<td>0.264</td>
<td>0.048</td>
</tr>
<tr>
<td>IQ</td>
<td>0.292</td>
<td>0.047</td>
</tr>
</tbody>
</table>
Bibliography


Summary

Over the past decades, it has been convincingly shown that all human traits (including preferences) are heritable. The use of insights from genetics to increase our understanding of how economic agents make their choices is called “genoeconomics”. In this thesis, I develop and compare methods to foster the further emergence of the field of genoeconomics and I perform genetically informed empirical analyses to better understand smoking behaviour, entrepreneurship, and the structure of the brain.

The first chapter of this thesis discusses four ways through which genoeconomics can contribute to economics: (i) genes can be used as a direct measure for a previously latent variable, (ii) genes can uncover biological mechanisms leading to differences in economic behaviour, (iii) genes can be used as control or instrumental variables in economic models, and (iv) genes could be used for targeting interventions. The present thesis contributes to the realization of these promises by analyzing how genes can be used as instrumental variables (Part 1: Chapters 2 and 3), investigating how genes help to explain why individuals engage in entrepreneurship and why individuals make different choices in response to an increase in tobacco excise taxes (Part 2: Chapters 4 and 5), and developing a method to better understand the shared genetic architecture of economic behaviour and other traits (Part 3: Chapter 6).

In the first part of my thesis (Chapters 2 and 3), I compare methods that use genetic variants as instrumental variables. In situations in which it is difficult or unethical to perform a randomized controlled trial, these so-called Mendelian randomization studies may help to infer causal relationships. Because of severe concerns about whether the assumptions underlying Mendelian randomization studies hold in practice, several robust Mendelian randomization methods have been developed. In these chapters, I review their merits from a theoretical point of view and I use a simulation study to compare their empirical performance in
order to give clear recommendations to applied researchers using these methods.

In the second part of my thesis (Chapters 4 and 5), I empirically investigate how genes may help to explain economic behaviour. Individual genetic variants typically explain only a small fraction of the variance in behaviour, and therefore I advocate the use of weighted combinations of genetic variants – polygenic risk scores – in these chapters. In Chapter 4, I use polygenic risk scores to explain why individuals engage in entrepreneurship. Most interestingly, I find that genetic variants associated with Attention Deficit/Hyperactivity Disorder are associated with entrepreneurship. In Chapter 5, I show that polygenic risk scores are predictive for smoking behaviour (measured as smoking initiation and smoking intensity). Moreover, my analyses show that someone’s genetic propensity to smoking moderates the effect of tobacco excise taxes on smoking behavior, but only along the extensive margin (smoking vs. not smoking).

In the third part of my thesis (Chapter 6), I develop a multivariate version of Genome-based Restricted Maximum Likelihood (GREML) estimation. With this method, one can estimate what fraction of a trait is heritable and to what extent different traits are genetically related. Multivariate GREML improves over existing bivariate methods by ensuring the internally consistency of the resulting multivariate correlation matrix and by decreasing the computational burden required for parameter estimation. By employing this method using data from the UK Biobank Imaging study, I study genetic correlations across brain regions and behavioural outcomes. By doing so, I show that the method can effectively deal with large datasets.
Samenvatting

Over de afgelopen decennia is het duidelijk geworden dat alle eigenschappen (inclusief voorkeuren) bij mensen erfelijk zijn. Het gebruik van de inzichten uit genetica om onze kennis van hoe economische agenten handelen te vergroten wordt “genoeconomie” genoemd. In dit proefschrift ontwikkel en vergelijk ik methoden om de opkomst van het veld genoeconomie te bevorderen. Tevens doe ik een aantal genetisch geïnformeerde empirische analyses om onze kennis van rookgedrag, ondernemerschap en de structuur van het brein te vergroten.

Het eerste hoofdstuk van dit proefschrift beschrijft vier manieren hoe genoeconomie kan bijdragen aan economie: (i) genen kunnen een directe maatstaf zijn voor voorheen latente variabelen, (ii) genen kunnen biologische mechanismen blootleggen die leiden tot verschillen in economisch gedrag, (iii) genen kunnen worden gebruikt als controle of instrumentele variabele in economische modellen, and (iv) genen kunnen gebruikt worden voor gerichte interventies. Het huidige proefschrift draagt bij aan de realisatie van deze manieren door te bestuderen hoe genen gebruikt kunnen worden als instrumentele variabelen (Deel 1: Hoofdstuk 2 en 3), bestuderen hoe genen helpen om uit te leggen waarom individuen ondernemer worden en waarom individuen individuen verschillende keuzes maken als reactie op een verhoging van tabaksaccijnzen (Deel 2: Hoofdstuk 4 en 5), en het ontwikkelen van een methode om een beter inzicht te krijgen in de gedeelde genetische structuur van economisch gedrag en andere eigenschappen (Deel 3: Hoofdstuk 6).

In het eerste deel van mijn proefschrift (Hoofdstuk 2 en 3), vergelijk ik methoden die genetische varianten gebruiken als instrumentele variabelen. In situaties waar een traditioneel gerandomiseerd onderzoek lastig of niet mogelijk is, kunnen deze zogenaamde Mendelianaanse randomisatie studies helpen om te schatten of er een causaal verband is. Door ernstige zorgen over of bepaalde aannames die ten grondslag liggen aan Mendelianaanse randomisatie, zijn er
verschillende robuuste Mendeliaanse randomisatie methoden voorgesteld. In deze hoofdstukken bekijk ik de voordelen van deze verschillende methoden en vergelijk ik ze onderling om duidelijke richtlijnen te geven voor wetenschappers die deze methodes gebruiken.

In het tweede deel van mijn proefschrift (Hoofdstukken 4 en 5), bestudeer ik empirisch hoe genen kunnen worden gebruikt om gedrag te verklaren. Individuele genetische varianten verklaren over het algemeen slechts een kleine fractie van de variantie in gedrag, en daarom pleit ik voor het gebruik van gewogen combinaties van genetische varianten - polygene risicoscores - in deze hoofdstukken. In Hoofdstuk 4 gebruik ik deze polygene risicoscores om te verklaren waarom individuen beginnen aan ondernemerschap. Het meest opmerkelijke is dat ik vind dat genen die geassocieerd zijn met ADHD (Attention Deficit/Hyperactivity Disorder) ook geassocieerd zijn met ondernemerschap. In Hoofdstuk 5 laat ik zien dat de polygene risicoscores voorspellend zijn voor voor rookgedrag (gemeten als initiatie en intensiteit van tabaksconsumptie). Daarnaast laten mijn analyses zien dat een individu’s genetische aanleg voor rookgedrag een moderator is voor het effect van tabaksaccijnzen op rookgedrag, maar enkel op de extensieve marge (roken tegenover niet roken).

In het derde deel van mijn proefschrift (Hoofdstuk 6), ontwikkel ik een multivariate versie van GREML (Genome-based restricted maximum likelihood). Met deze methode kan geschat worden welk deel van een eigenschap erfelijk is en in hoeverre verschillende eigenschappen genetisch verwant zijn. Multivariate GREML verbetert de huidige bivariate methodiek door te garanderen dat de geschatte multivariate correlatie matrix intern consistent is en door de computatienele belasting te verlagen die nodig is om het model te schatten. Door deze methodiek toe te passen op data van de UK Biobank Imaging Study, bestudeer ik genetische correlaties tussen brein regio’s en gedragsuitkomsten. Hiermee laat ik zien dat deze methode efficiënt om kan gaan met grote datasets.
About the Author

Eric Arsène Willem Slob was born on the 7th of April in 1994 in Utrecht, The Netherlands. In 2016, he obtained the degree of Master of Science (MSc) in Econometrics and Management Science at the Erasmus School of Economics, Erasmus University Rotterdam. In 2016, Eric started as a Ph.D. candidate under the supervision of professor Dr. Patrick J.F. Groenen, Dr. Cornelius A. Rietveld, professor Dr. A. Roy Thurik. He carried out his research within the Department of Applied Economics at the Erasmus School of Economics as a member of the Erasmus Research Institute of Management and the Erasmus University Rotterdam Institute for Behavior and Biology. In 2018 Eric visited the Medical Research Council Biostatistics Unit at the University of Cambridge for a period of three months. During his research visit he was supervised by professor Dr. Stephen Burgess.

Eric’s research focuses on the methodological integration of genetics into economics. His work has been published in the following peer-reviewed journals: *International Journal of Epidemiology, Genetic Epidemiology*, and *Small Business Economics*. He has presented his work, amongst others, at meetings of the *Behavior Genetics Association*, and the *Mendelian Randomization conference*. Eric will continue his career as a research associate at the University of Cambridge.
Portfolio

Peer-Reviewed Publications


Working Papers


Grants and Prizes

- 2020 Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), Call for Compute Time, EINF-403.

Refereed Articles Submitted To

- Genetic Epidemiology
- Health Psychology
- International Journal of Epidemiology
- Small Business Economics

Teaching Activities

- Economics & Genetics: Teaching assistant
- Economics of Entrepreneurship: Teaching assistant
- Internship Supervisor
- Small Business Economics: Teaching assistant
- Thesis supervisions: Supervised various bachelor and master’s theses

PhD Courses and Certificates

- Advanced Econometrics II (Tinbergen Institute)
- Advanced Statistical Methods (Erasmus Research Institute of Management)
- Cambridge Certificate of Proficiency in English (Cambridge ESOL examinations)
- Economics of Entrepreneurship (Erasmus School of Economics)
- EDEN Doctoral Seminar on Methods, Techniques and Theories in Entrepreneurship and Innovation (European Institute for Advanced Studies in Management)
- Health Economics (Tinbergen Institute)
- Introduction in Genome-Wide Data Analysis (Tinbergen Institute)
- Micro Economics (Erasmus Research Institute of Management)
- Publishing Strategy (Erasmus Research Institute of Management)
- Summer Institute in Social Science Genomics (Russell Sage Foundation)
CONFERENCES, WORKSHOPS, AND MEETINGS

- Meeting of the Behavior Genetics Association (Online, 2020)
- Norface Grant Meeting (Bristol, United Kingdom, 2019)
- Mendelian Randomization Conference (Bristol, United Kingdom, 2019)
- Mendelian Randomization Symposium (Cambridge, United Kingdom, 2019)
- MRC Biostatistics Unit Together (Cambridge, United Kingdom, 2018)
- Meeting of the Behavior Genetics Association (Boston, USA, 2018)
- Econometric Institute PhD Conference (Rotterdam, The Netherlands, 2018)
- Mendelian Randomization Conference (Bristol, United Kingdom, 2017)

EDUCATION

- **Master of Science in Econometrics and Management Science** (2014-2016, Erasmus University Rotterdam)
  - Specialization in Econometrics

- **Bachelor of Science in Econometrics and Management Science** (2011-2014, Erasmus University Rotterdam)
  - Major in Econometrics
  - ESE Bachelor Honours Class
ERIM Publications List

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://repub.eur.nl/pub. ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam (EUR).

Dissertations in the last four years


