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 ARSÈNE WILLEM SLOB
born in Utrecht.

Erasmus University Rotterdam

Doctoral Committee

Promotor: Prof.dr. A.R. Thurik
Prof.dr. P.J.F. Groenen

Other members: Prof.dr. D. Fok
Prof.dr. S.M.L. von Hinke
Prof.dr. J.L.W. van Kippersluis

Copromotor: Dr. C.A. Rietveld

Erasmus Research Institute of Management - ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam.
Internet: http://www.erim.eur.nl.

ERIM Electronic Series Portal: https://repub.eur.nl

ERIM PhD Series in Research in Management, 517
ERIM reference number: EPS-2021-517-S&E
ISBN 978-90-5892-596-1

Design: E.A.W. Slob.
Cover Artwork: © Elise Slob

This publication (cover and interior) is printed by Tuijtel on recycled paper, BalanceSilk®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Ecolabel, FSC®C007225.
More info: https://www.tuijtel.com.

FSC
www.fsc.org
RECYCLED
Paper made from recycled material
FSC® C007225

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

List of Figures

List of Tables

1

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 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:

$$y_i = \mu + \sum_{j=1}^{J} \beta_j x_{ij} + \varepsilon_i, \tag{1.1}$$

with intercept μ, SNP effects β_j and residual term ε_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:

$$y_i = \mu + x_{ij} b_j + \varepsilon_i, \tag{1.2}$$

where y_i is the value of the phenotype for individual i, μ 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.

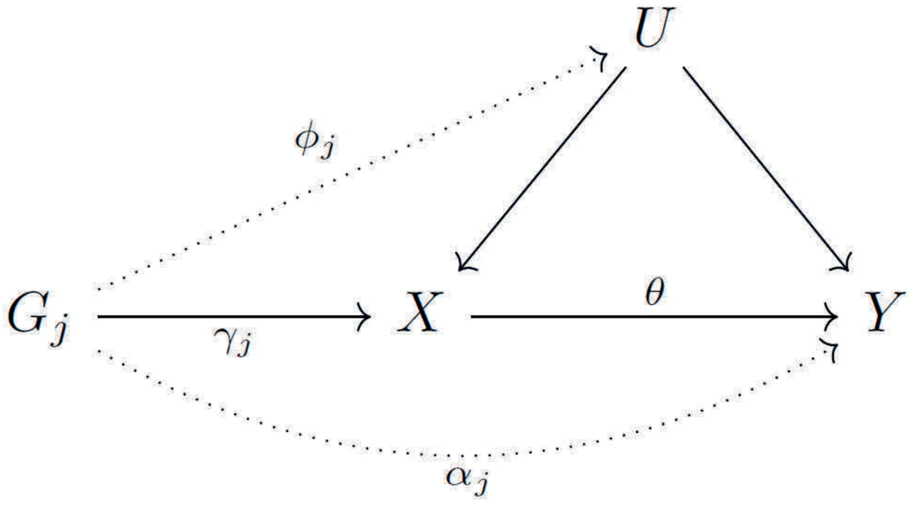

FIGURE 1.1 – *Illustrative diagram showing the model assumed for genetic variant G_j, with effect ϕ_j on the unobserved confounder U, effect γ_j on exposure X, and direct effect α_j on outcome Y. The causal effect of the exposure on the outcome is θ. Dotted lines represent possible ways the instrumental variable assumptions could be violated.*

The summary-level MR methods considered in this thesis work take the association between a genetic variant and the exposure (beta-coefficient $\hat{\beta}_{X_j}$ and standard error σ_{X_j}) and the association between the genetic variants and the outcome (beta-coefficient $\hat{\beta}_{Y_j}$ and standard error σ_{Y_j}) 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:

$$\hat{\theta}_{R_j} = \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}. \tag{1.3}$$

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). 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:

$$PGS_i = \sum_{j=1}^{J} \beta_j x_{ij}, \tag{1.4}$$

where PGS_i is the value for the polygenic risk score for individual i, β_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:

$$y_i = \mu + g_i + \varepsilon_i \quad \text{and} \quad g_i = \sum_{k=1}^{m} g_i = \sum_{k=1}^{m} s_{ik} u_k, \tag{1.5}$$

where μ 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 $\mathbf{y} = \mu\mathbf{1} + \mathbf{g} + \boldsymbol{\varepsilon}$

and $\mathbf{g} = \mathbf{Su}$. Now, the variance of Y can be partitioned as follows:

$$\mathrm{Var}(\mathbf{y}) = \sigma_u^2 \mathbf{SS}^\top + \sigma_e^2 \mathbf{I} = \frac{\sigma_g^2}{m} \mathbf{SS}^\top + \sigma_e^2 \mathbf{I} = \sigma_g^2 \mathbf{G} + \sigma_e^2 \mathbf{I}, \tag{1.6}$$

where $\mathbf{G}$ ($= m^{-1}\mathbf{SS}^\top$) 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 $\sigma_g^2/(\sigma_g^2 + \sigma_e^2)$. 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 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 analysis (best used with few very distinct outliers) or MR-Robust analysis could also be performed. While I am 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 thus not adequately assessing the IV assumptions and performing so many methods that the clarity is obscured.

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 entrepreneurship 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 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 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.*

Chapter	**Title**	**Reference**	**Presentations**	**Publication status**
2	A note on the use of Egger regression in Mendelian randomization studies	Slob, Groenen, Thurik & Rietveld	Bristol (2017)	Published in *International Journal of Epidemiology*
3	A comparison of robust Mendelian randomization methods using summary data	Slob & Burgess	Cambridge (2018, 2019), Rotterdam (2019), Bristol (2019)	Published in *Genetic Epidemiology*
4	A decade of research on the genetics of entrepreneurship: a review and view ahead	Rietveld, Slob & Thurik		Published in *Small Business Economics*
5	Does the genetic predisposition to smoking moderate the response to tobacco excise taxes?	Slob & Rietveld	Rotterdam (2019)	Manuscript submitted
6	Multivariate GREML finds shared genetic architecture of 76 brain traits and intelligence	De Vlaming, Slob, Jansen, Koellinger, Groenen & Rietveld	Boston (2018), Rotterdam (2017, 2019), Online (2020)	Manuscript in preparation

I

Mendelian randomization

2

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).

2.1 Introduction

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 (ε_i^U), 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 (ε_i^X). The outcome Y_i is a linear function of the genetic variants, the exposure, the confounders and an independent error term (ε_i^Y). The causal effect of the exposure on the outcome is β. γ_j represents the effect of the instrument on the exposure. The coefficients α_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 (α_j), and the indirect effects via the exposure ($\beta\gamma_j$) and the confounder (ϕ_j). The model described above can be written as:

$$U_i = \sum_{j=1}^{J} \phi_j G_{ij} + \varepsilon_i^U \tag{2.1}$$

$$X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \varepsilon_i^X \tag{2.2}$$

$$Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \beta X_i + U_i + \varepsilon_i^Y \ . \tag{2.3}$$

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 β is a function of α_j, ϕ_j, and γ_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 $\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Y_j}^{-2}((\alpha_j + \phi_j)/(\gamma_j + \phi_j))/\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Y_j}^{-2}$, where σ_{Y_j} is the standard error in the regression of the outcome on the jth 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 β. Furthermore, Bowden and colleagues find that the bias in the estimation of β with MR-Egger equals $(\sigma_\alpha \rho_{\alpha,\gamma} + (1+\beta)\sigma_\phi \rho_{\phi,\gamma})/\sigma_\gamma$, where σ denotes the standard deviation of a parameter and ρ the correlation (Bowden et al., 2017a). Hence, in MR-Egger the bias is a function of σ_α, $\rho_{\alpha,\gamma}$, β, σ_ϕ, $\rho_{\phi,\gamma}$ and σ_γ (note that MR-Egger requires $\sigma_\gamma > 0$; 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 α_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?

$$\left|\text{Bias}_{\text{MR-Egger}}\right| = \left|\frac{\sigma_\alpha \rho_{\alpha,\gamma} + (1+\beta)\sigma_\phi \rho_{\phi,\gamma}}{\sigma_\gamma}\right| \overset{?}{<} \left|\frac{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Y_j}^{-2} \frac{\alpha_j + \phi_j}{\gamma_j + \phi_j}}{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Y_j}^{-2}}\right| = |\text{Bias}_{\text{IVW}}| \ . \tag{2.4}$$

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 case, (2.4) reduces to

$$\left|\frac{\sigma_\alpha \rho_{\alpha,\gamma}}{\sigma_\gamma}\right| \overset{?}{<} \left|\frac{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Y_j}^{-2} \frac{\alpha_j}{\gamma_j}}{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Y_j}^{-2}}\right|. \tag{2.5}$$

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 γ_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 γ and the direct effect α 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 β 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 γ and the direct effects α is even -0.35. Hence, again it is unclear whether the IVW or the MR-Egger estimate is closer to the true β.

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 practival 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).

Appendices

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 γ and α, we need to approximate α. 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 $\hat{\beta}_{\text{MR-Egger}}$ 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, α^*_j equal the "error" terms in this relation (the average pleiotropy $\bar{\alpha}$ is captured by the intercept β_{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 α_j and γ_j is very close to

0. Nevertheless, Figure 2.1 Panel A shows the approximated $\rho_{\alpha,\gamma}$ for a range of possible causal effects. We observe that $\rho_{\alpha,\gamma}$ approaches 1 when β becomes more negative, and it approaches -1 when β 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 $\rho_{\alpha,\gamma} = -0.35$. When using $\hat{\beta}_{\text{MR-Egger}}$, we obtain an approximated correlation $\rho_{\alpha,\gamma} = -0.04$. The approximated correlation for a range of causal effects is shown in Figure 2.1 Panel B. We observe the same relationship between $\rho_{\alpha,\gamma}$ and β as in Figure 2.1 Panel A.

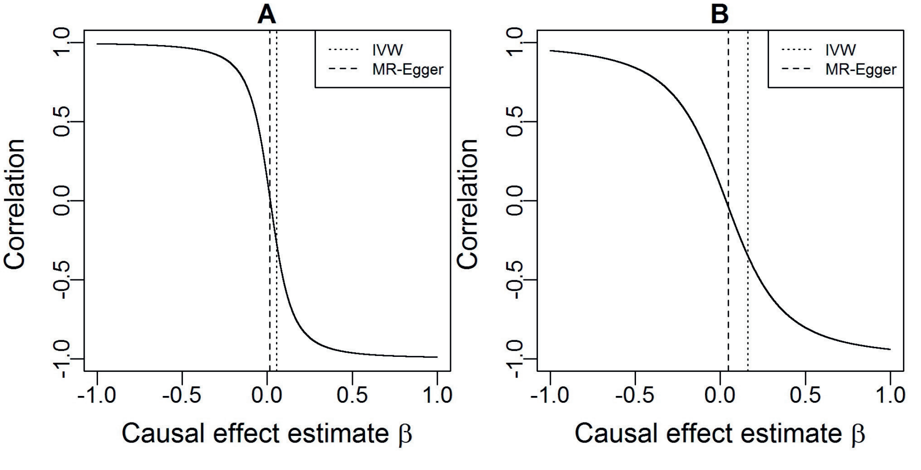

FIGURE 2.1 – *The correlation between the instrument strength and direct effect for different causal effect estimates. A: The effect of systolic blood pressure on cardiovascular diseases risk. B: The effect of plasma urate concentrate on coronary heart disease risk.*

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)).*

SNP	Chromosome	Position	Ref allele	Other allele	$\hat{\gamma}$	$\hat{\Gamma}$	$\hat{\alpha}$
rs2932538	1	113,018,066	G	A	0.388	0.022801	0.001849
rs13082711	3	27,512,913	C	T	0.315	0.034229	0.017219
rs419076	3	170,583,580	T	C	0.409	0.019725	-0.002361
rs13107325	4	103,407,732	C	T	0.981	-0.004938	-0.057912
rs13139571	4	156,864,963	C	A	0.321	0.066131	0.048797
rs1173771	5	32,850,785	G	A	0.504	0.033135	0.005919
rs11953630	5	157,777,980	C	T	0.412	0.031746	0.009498
rs1799945	6	26,199,158	G	C	0.627	0.005740	-0.028118
rs805303	6	31,724,345	G	A	0.376	0.040028	0.019724
rs4373814	10	18,459,978	C	G	0.373	0.015007	-0.005135
rs932764	10	95,885,930	G	A	0.484	-0.005658	-0.031794
rs7129220	11	10,307,114	A	G	0.619	0.040449	0.007023
rs633185	11	100,098,748	G	C	0.565	0.013790	-0.016720
rs2521501	15	89,238,392	T	A	0.650	0.050969	0.015869
rs17608766	17	42,368,270	C	T	0.556	0.088020	0.057996
rs1327235	20	100,917,030	G	A	0.340	0.040494	0.022134
rs6015450	20	57,184,512	G	A	0.896	0.058069	0.009685
rs17367504	1	11,785,365	A	G	0.903	0.023070	-0.025691
rs3774372	3	41,852,418	C	T	0.067	-0.018337	-0.021955
rs1458038	4	81,383,747	T	C	0.706	0.034433	-0.003691
rs1813353	10	18,747,454	T	C	0.569	0.042546	0.011820
rs4590817	10	63,137,559	G	C	0.646	0.035467	0.000583
rs11191548	10	104,836,168	T	C	1,095	0.109050	0.049920
rs381815	11	16,858,844	T	C	0.575	0.048819	0.017769
rs17249754	12	88,584,717	G	A	0.928	-0.044766	-0.094878
rs3184504	12	110.368,991	T	C	0.598	0.070445	0.038153
rs10850411	12	113,872,179	T	C	0.354	0.021469	0.002353
rs1378942	15	72,864,420	C	A	0.613	0.000516	-0.032586
rs12940887	17	44,757,806	T	C	0.362	0.032434	0.012886

TABLE 2.2 – *Summary association results for* 31 *SNPs associated with plasma urate concentration (SNPs are ordered as in Table S*3 *of the study by White et al. (2016)).*

SNP	Chromosome	Position	Ref allele	Other allele	$\hat{\gamma}$	$\hat{\Gamma}$	$\hat{\alpha}$
rs1471633	1	144,435,096	A	C	0.057	-0.015911	0.006654
rs1260326	2	27,584,444	T	C	0.069	-0.023890	0.012594
rs12498742	4	9,553,150	A	G	0.360	-0.011611	-0.047067
rs2231142	4	89,271,347	T	G	0.190	-0.034145	0.003241
rs675209	6	7,047,083	T	C	0.056	-0.001211	-0.007852
rs1165151	6	25,929,595	T	G	0.078	-0.004236	-0.016933
rs1171614	10	61,139,544	T	C	0.079	0.041327	0.028450
rs2078267	11	64,090,690	T	C	0.073	0.020768	0.008837
rs478607	11	64,234,639	A	G	0.026	-0.012873	-0.017177
rs3741414	12	56,130,316	T	C	0.065	0.001895	-0.008684
rs11264341	1	153,418,117	T	C	0.050	0.005672	-0.002478
rs17050272	2	121,022,910	A	G	0.035	0.033048	-0.038752
rs6770152	3	53,075,254	T	G	0.044	0.005058	-0.002111
rs17632159	5	72,467,238	C	G	0.039	-0.005710	-0.000647
rs729761	6	43,912,549	T	G	0.047	0.010178	0.002517
rs1178977	7	72,494,985	A	G	0.047	0.005351	-0.013012
rs10480300	7	151,036,938	T	C	0.035	-0.004082	-0.001623
rs2941484	8	76,641,323	T	C	0.044	0.008357	-0.015528
rs10821905	10	52,316,099	A	G	0.057	0.010269	-0.019560
rs642803	11	65,317,196	T	C	0.036	0.039500	0.033632
rs653178	12	110,492,139	T	C	0.035	0.077493	0.071788
rs1394125	15	73,946,038	A	G	0.043	-0.027942	0.020933
rs6598541	15	97,088,658	A	G	0.043	-0.035859	0.028837
rs7193778	16	68,121,391	T	C	0.046	-0.004131	-0.011629
rs7188445	16	78,292,488	A	G	0.032	-0.011765	-0.016981
rs7224610	17	50,719,787	A	C	0.042	0.015338	0.008492
rs742132	6	25,715,550	A	G	0.054	0.010673	-0.019475
rs2307394	2	148,432,898	T	C	0.029	0.018612	0.013885
rs17786744	8	23,832,951	A	G	0.029	0.000176	-0.004551
rs2079742	17	56,820,479	T	C	0.043	-0.038280	0.031271
rs164009	17	71,795,264	A	G	0.028	-0.015160	0.010596

3

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 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.

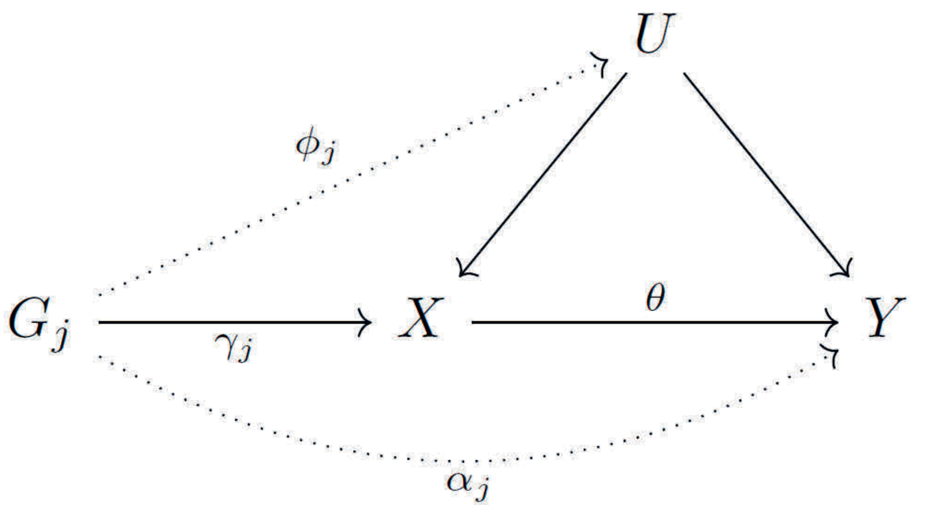

FIGURE 3.1 – *Illustrative diagram showing the model assumed for genetic variant G_j, with effect ϕ_j on the unobserved confounder U, effect γ_j on exposure X, and direct effect α_j on outcome Y. The causal effect of the exposure on the outcome is θ. Dotted lines represent possible ways the instrumental variable assumptions could be violated.*

We assume that summary data are available on genetic associations with the exposure (beta-coefficient $\hat{\beta}_{X_j}$ and standard error σ_{X_j}) and with the outcome (beta-coefficient $\hat{\beta}_{Y_j}$ and standard error σ_{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:

$$\hat{\theta}_{R_j} = \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}. \tag{3.1}$$

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 σ_{R_j} can be approximated as (Thomas et al., 2007):

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

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):

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

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:

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

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 3.1 – *Summary comparison of methods.*

Method	Consistency assumption	Strengths and/or weaknesses
Weighted Median	Majority valid	Robust to outliers, sensitive to additional/removal of genetic variants, may be less efficient
Mode Based Estimation	Plurality valid	Robust to outliers, sensitive to bandwidth parameter and addition/removal of genetic variants, generally conservative
MR-PRESSO	Outlier-robust	Removes outliers, efficient with valid IVs, very high false positive rate with several invalid IVs
MR-Robust	Outlier-robust	Downweights outliers, efficient with valid IVs, high false positive rate with several invalid IVs
MR-Lasso	Outlier-robust	Removes outliers, efficient with valid IVs, high false positive rate with several invalid IVs
MR-Egger	InSIDE	Sensitive to outliers, sensitive to violations of InSIDE assumption, InSIDE assumption often not plausible, may be less efficient
Contamination Mixture	Plurality valid	Robust to outliers, sensitive to variance parameter and addition/removal of genetic variants, less conservative than MBE
MR-Mix	Plurality valid	Robust to outliers, requires large numbers of genetic variants, very high false positive rate in several scenarios
MR-RAPS	Pleiotropic effects (except outliers) normally distributed about zero	Downweights outliers, sensitive to violations of balanced pleiotropy assumption

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 θ that minimizes:

$$\sum_{j=1}^{J} \sigma_{Y_j}^{-2} \left(\hat{\beta}_{Y_j} - \theta \, \hat{\beta}_{X_j} \right)^2 . \tag{3.5}$$

In MR-Lasso, we minimize:

$$\sum_{j=1}^{J} \sigma_{Y_j}^{-2} \left(\hat{\beta}_{Y_j} - \theta_{0j} - \theta \, \hat{\beta}_{X_j} \right)^2 + \lambda \sum_{j=1}^{J} | \, \theta_{0j} \, |, \tag{3.6}$$

where λ 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 θ_{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 λ. Increasing λ means that more of the pleiotropy parameters equal zero and so the corresponding variants are included in the analysis; we increase λ 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 θ_0:

$$\hat{\beta}_{Y_j} = \theta_0 + \theta\,\hat{\beta}_{X_j} + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2_{Y_j}). \tag{3.7}$$

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 θ with variance $\sigma^2_{R_j}$. If a variant is not a valid instrument, then its ratio estimate is assumed to be normally distributed about zero with variance $\psi^2 + \sigma^2_{R_j}$, where ψ^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 θ and different configurations of valid and invalid IVs. Maximization is performed in linear time by first constructing a profile likelihood as a function of θ, and then maximizing this function with respect to θ. The value of θ 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 correct 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 variance parameter.

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 ε_i^U. The effect of variant j on the confounder is represented by coefficient ϕ_j (this is zero for a valid IV). The exposure is linear in the genetic variants, the confounder and an independent error term ε_i^X. The effect of variant j on the exposure is represented by coefficient γ_j. The outcome is linear in the genetic variants, exposure, confounders and an independent error term ε_i^Y. The effect of variant j on the outcome is represented by coefficient α_j (again, this is zero for a valid IV). The effect of the exposure on the outcome is represented by θ. 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. The error terms ε_i^U, ε_i^X and ε_i^Y each follow an independent normal distribution with mean 0 and unit variance.

We can represent the model mathematically as:

$$
\begin{aligned}
U_i &= \sum_{j=1}^{J} \phi_j G_{ij} + \varepsilon_i^U, \\
X_i &= \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \varepsilon_i^X, \\
Y_i &= \sum_{j=1}^{J} \alpha_j G_{ij} + \theta X_i + U_i + \varepsilon_i^Y, \\
\text{maf}_j &\sim \mathcal{U}(0.1, 0.5), \\
G_{ij} &\sim \text{Binomial}(2, \text{maf}_j) \text{ independently}, \\
\varepsilon_i^U, \varepsilon_i^X, \varepsilon_i^Y &\sim \mathcal{N}(0,1) \text{ independently}.
\end{aligned} \tag{3.8}
$$

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, 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.

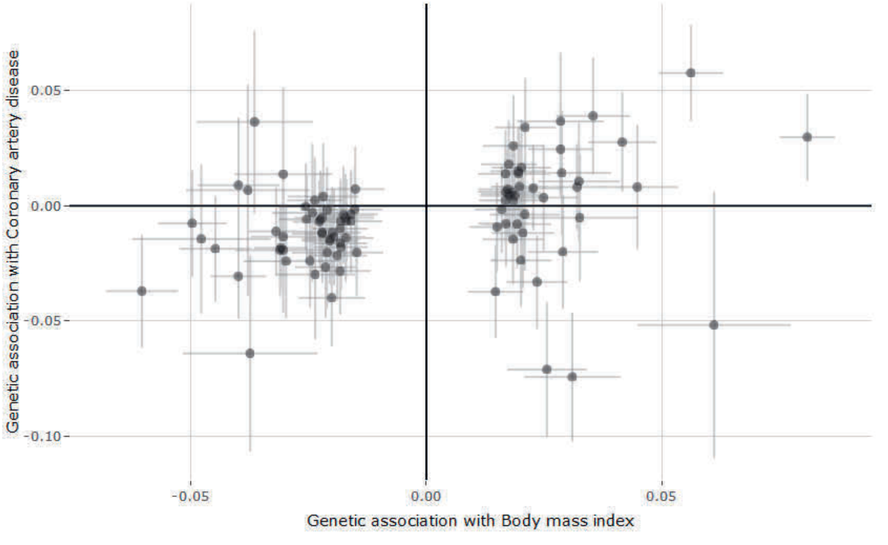

FIGURE 3.2 – *Scatter plot of genetic associations with BMI (standard deviation units) and coronary artery disease risk (log odds ratios) for 94 variants taken from the GIANT and CARDIoGRAMplusC4D consortia respectively.*

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 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.

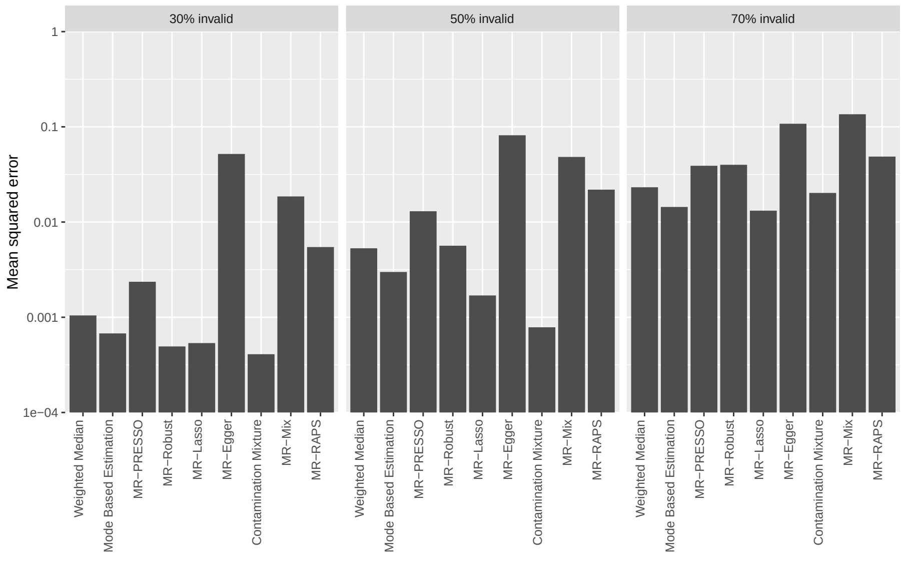

FIGURE 3.3 – *Mean squared errors for the different methods in scenario 2 (directional pleiotropy, InSIDE satisfied) with a null causal effect for 30 variants. Note the vertical axis is on a logarithmic scale.*

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.*

	30% invalid				50% invalid				70% invalid			
Null causal effect: $\theta = 0$												
Method	Mean	Median	SD	T1 error	Mean	Median	SD	T1 error	Mean	Median	SD	T1 error
Scenario 1: Balanced pleiotropy, InSIDE satisfied												
Weighted Median	0.000	0.000	0.071	0.139	0.002	0.001	0.132	0.276	0.002	0.000	0.223	0.481
Mode Based Estimation	0.000	0.000	0.101	0.111	0.002	0.000	0.151	0.268	0.002	0.001	0.224	0.619
MR-PRESSO	0.000	0.000	0.111	0.122	-0.001	0.000	0.178	0.154	0.000	0.001	0.239	0.174
MR-Robust	0.000	0.000	0.029	0.110	0.001	0.001	0.127	0.076	0.002	0.002	0.224	0.104
MR-Lasso	0.001	0.000	0.048	0.042	0.000	0.000	0.088	0.076	0.004	0.001	0.183	0.156
MR-Egger	0.007	0.004	0.419	0.093	0.005	0.008	0.563	0.097	0.006	0.014	0.684	0.098
Contamination Mixture	0.000	0.000	0.025	0.052	0.000	0.000	0.077	0.069	0.002	0.000	0.379	0.126
MR-Mix	0.000	0.000	0.274	0.225	-0.001	0.000	0.431	0.292	0.000	0.000	0.561	0.356
MR-RAPS	0.000	-0.001	0.106	0.039	0.001	0.000	0.172	0.062	0.001	0.000	0.226	0.083
Scenario 2: Directional pleiotropy, InSIDE satisfied												
Weighted Median	0.013	0.006	0.060	0.140	0.036	0.016	0.108	0.287	0.084	0.036	0.175	0.500
Mode Based Estimation	0.007	0.001	0.081	0.114	0.020	0.006	0.122	0.264	0.059	0.030	0.180	0.585
MR-PRESSO	0.028	0.013	0.079	0.132	0.069	0.031	0.133	0.168	0.122	0.071	0.182	0.214
MR-Robust	0.003	0.002	0.031	0.106	0.042	0.023	0.105	0.084	0.115	0.094	0.169	0.152
MR-Lasso	0.008	0.005	0.044	0.056	0.024	0.012	0.082	0.125	0.075	0.035	0.161	0.283
MR-Egger	0.001	-0.006	0.329	0.093	0.000	-0.013	0.408	0.091	-0.005	-0.012	0.477	0.095
Contamination Mixture	0.000	0.001	0.025	0.059	0.003	0.001	0.056	0.078	0.060	0.006	0.281	0.137
MR-Mix	0.045	0.016	0.200	0.247	0.084	0.023	0.301	0.331	0.144	0.050	0.399	0.443
MR-RAPS	0.039	0.030	0.082	0.053	0.081	0.071	0.128	0.095	0.130	0.119	0.165	0.152
Scenario 3: Directional pleiotropy, InSIDE violated												
Weighted Median	0.022	0.011	0.071	0.179	0.073	0.030	0.137	0.384	0.135	0.080	0.188	0.599
Mode Based Estimation	0.013	0.002	0.090	0.132	0.044	0.011	0.148	0.317	0.094	0.051	0.192	0.621
MR-PRESSO	0.047	0.023	0.095	0.155	0.113	0.063	0.153	0.223	0.179	0.147	0.185	0.301
MR-Robust	0.004	0.002	0.032	0.106	0.069	0.040	0.121	0.109	0.169	0.152	0.171	0.216
MR-Lasso	0.013	0.008	0.050	0.073	0.050	0.024	0.108	0.203	0.122	0.067	0.180	0.415
MR-Egger	0.049	0.024	0.326	0.098	0.066	0.042	0.411	0.097	0.048	0.034	0.464	0.096
Contamination Mixture	0.000	0.000	0.025	0.060	0.005	0.001	0.061	0.080	0.079	0.009	0.273	0.163
MR-Mix	0.064	0.026	0.207	0.283	0.125	0.040	0.304	0.375	0.196	0.080	0.391	0.529
MR-RAPS	0.062	0.050	0.091	0.085	0.132	0.118	0.132	0.182	0.188	0.180	0.160	0.262
Positive causal effect: $\theta = +0.2$												
Method	Mean	Median	SD	Power	Mean	Median	SD	Power	Mean	Median	SD	Power
Scenario 1: Balanced pleiotropy, InSIDE satisfied												
Weighted Median	0.201	0.200	0.069	0.979	0.201	0.200	0.131	0.939	0.200	0.201	0.221	0.877
Mode Based Estimation	0.198	0.200	0.102	0.983	0.192	0.199	0.156	0.945	0.183	0.193	0.235	0.867
MR-PRESSO	0.199	0.200	0.106	0.860	0.202	0.201	0.166	0.734	0.200	0.202	0.232	0.564
MR-Robust	0.200	0.200	0.033	0.953	0.201	0.201	0.129	0.506	0.199	0.200	0.225	0.282
MR-Lasso	0.200	0.200	0.052	0.962	0.201	0.201	0.091	0.906	0.198	0.200	0.189	0.774
MR-Egger	0.199	0.201	0.442	0.166	0.199	0.199	0.549	0.122	0.197	0.193	0.660	0.113
Contamination Mixture	0.200	0.200	0.028	0.997	0.202	0.201	0.074	0.959	0.228	0.204	0.399	0.704
MR-Mix	0.210	0.203	0.242	0.562	0.219	0.205	0.370	0.612	0.224	0.210	0.522	0.644
MR-RAPS	0.200	0.200	0.108	0.538	0.201	0.202	0.168	0.309	0.197	0.201	0.228	0.222
Scenario 2: Directional pleiotropy, InSIDE satisfied												
Weighted Median	0.214	0.207	0.060	0.991	0.240	0.216	0.114	0.978	0.285	0.242	0.175	0.952
Mode Based Estimation	0.205	0.201	0.081	0.983	0.219	0.204	0.129	0.961	0.248	0.226	0.180	0.917
MR-PRESSO	0.225	0.213	0.072	0.945	0.267	0.232	0.129	0.849	0.319	0.274	0.177	0.729
MR-Robust	0.204	0.203	0.034	0.954	0.244	0.225	0.109	0.646	0.315	0.301	0.168	0.555
MR-Lasso	0.209	0.206	0.047	0.985	0.225	0.213	0.085	0.971	0.274	0.239	0.161	0.926
MR-Egger	0.200	0.188	0.323	0.215	0.199	0.187	0.407	0.153	0.196	0.187	0.462	0.133
Contamination Mixture	0.201	0.201	0.030	0.997	0.206	0.201	0.085	0.968	0.286	0.210	0.307	0.823
MR-Mix	0.252	0.228	0.175	0.613	0.291	0.240	0.265	0.664	0.353	0.276	0.367	0.738
MR-RAPS	0.238	0.229	0.080	0.825	0.285	0.275	0.127	0.675	0.329	0.322	0.164	0.595
Scenario 3: Directional pleiotropy, InSIDE violated												
Weighted Median	0.225	0.212	0.074	0.994	0.272	0.233	0.137	0.985	0.339	0.287	0.185	0.975
Mode Based Estimation	0.211	0.201	0.092	0.983	0.239	0.211	0.147	0.961	0.290	0.252	0.189	0.940
MR-PRESSO	0.243	0.223	0.086	0.925	0.307	0.262	0.149	0.835	0.379	0.342	0.182	0.759
MR-Robust	0.205	0.204	0.036	0.945	0.271	0.244	0.122	0.651	0.372	0.353	0.168	0.651
MR-Lasso	0.216	0.210	0.053	0.991	0.250	0.226	0.109	0.981	0.326	0.274	0.179	0.964
MR-Egger	0.248	0.225	0.330	0.245	0.266	0.242	0.408	0.183	0.251	0.236	0.458	0.155
Contamination Mixture	0.201	0.201	0.029	0.996	0.209	0.202	0.082	0.970	0.317	0.217	0.318	0.850
MR-Mix	0.274	0.240	0.180	0.654	0.327	0.260	0.260	0.713	0.405	0.319	0.357	0.788
MR-RAPS	0.263	0.251	0.090	0.872	0.329	0.316	0.134	0.797	0.389	0.378	0.158	0.759

Abbrevations: 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.*

	30% invalid				50% invalid				70% invalid			
Null causal effect: $\theta = 0$												
Method	Mean	Median	SD	T1 error	Mean	Median	SD	T1 error	Mean	Median	SD	T1 error
Scenario 1: Balanced pleiotropy, InSIDE satisfied												
Weighted Median	0.000	0.000	0.033	0.085	-0.001	0.000	0.066	0.168	-0.002	-0.002	0.134	0.333
Mode Based Estimation	0.000	0.000	0.029	0.052	0.000	0.000	0.063	0.127	0.000	-0.001	0.136	0.494
MR-PRESSO	0.000	0.000	0.052	0.208	-0.001	0.000	0.091	0.276	-0.002	0.000	0.145	0.351
MR-Robust	0.000	0.000	0.023	0.069	0.000	0.000	0.075	0.024	-0.001	-0.004	0.172	0.054
MR-Lasso	0.000	-0.001	0.025	0.038	0.000	0.000	0.036	0.061	-0.001	0.000	0.081	0.111
MR-Egger	0.004	0.003	0.319	0.068	0.006	0.002	0.400	0.073	-0.010	-0.008	0.464	0.074
Contamination Mixture	0.000	0.000	0.022	0.062	0.000	0.000	0.030	0.078	-0.002	0.001	0.177	0.127
MR-Mix	0.000	0.000	0.141	0.052	0.000	0.000	0.215	0.053	0.002	0.000	0.321	0.036
MR-RAPS	-0.001	-0.001	0.077	0.019	0.000	-0.003	0.132	0.041	-0.002	-0.004	0.178	0.055
Scenario 2: Directional pleiotropy, InSIDE satisfied												
Weighted Median	0.011	0.009	0.031	0.100	0.031	0.021	0.066	0.235	0.083	0.048	0.127	0.438
Mode Based Estimation	0.001	0.000	0.026	0.049	0.006	0.003	0.054	0.132	0.040	0.026	0.113	0.454
MR-PRESSO	0.024	0.016	0.042	0.230	0.071	0.047	0.089	0.424	0.145	0.119	0.134	0.584
MR-Robust	0.003	0.002	0.022	0.065	0.034	0.026	0.067	0.030	0.149	0.140	0.133	0.159
MR-Lasso	0.004	0.003	0.023	0.058	0.014	0.011	0.039	0.135	0.061	0.039	0.097	0.340
MR-Egger	0.004	-0.004	0.228	0.073	0.001	-0.005	0.285	0.074	-0.002	-0.008	0.328	0.071
Contamination Mixture	0.001	0.001	0.020	0.064	0.001	0.001	0.028	0.085	0.015	0.003	0.141	0.140
MR-Mix	0.018	0.006	0.135	0.078	0.041	0.010	0.216	0.107	0.096	0.010	0.355	0.119
MR-RAPS	0.046	0.042	0.058	0.051	0.110	0.105	0.099	0.160	0.179	0.175	0.129	0.273
Scenario 3: Directional pleiotropy, InSIDE violated												
Weighted Median	0.022	0.017	0.042	0.168	0.067	0.040	0.095	0.401	0.156	0.114	0.155	0.668
Mode Based Estimation	0.002	0.001	0.033	0.057	0.016	0.006	0.073	0.172	0.077	0.048	0.140	0.531
MR-PRESSO	0.050	0.035	0.061	0.397	0.132	0.108	0.114	0.653	0.232	0.216	0.146	0.816
MR-Robust	0.004	0.004	0.023	0.052	0.059	0.045	0.080	0.041	0.224	0.216	0.136	0.335
MR-Lasso	0.008	0.007	0.025	0.086	0.033	0.024	0.054	0.267	0.123	0.089	0.130	0.597
MR-Egger	0.092	0.074	0.234	0.105	0.099	0.090	0.277	0.091	0.094	0.089	0.312	0.084
Contamination Mixture	0.000	0.001	0.020	0.062	0.002	0.002	0.029	0.093	0.026	0.005	0.156	0.166
MR-Mix	0.029	0.010	0.141	0.095	0.056	0.010	0.220	0.139	0.125	0.020	0.327	0.154
MR-RAPS	0.082	0.075	0.068	0.174	0.172	0.165	0.103	0.415	0.256	0.251	0.124	0.591
Positive causal effect: $\theta = +0.2$												
Method	Mean	Median	SD	Power	Mean	Median	SD	Power	Mean	Median	SD	Power
Scenario 1: Balanced pleiotropy, InSIDE satisfied												
Weighted Median	0.200	0.200	0.035	0.998	0.201	0.200	0.066	0.978	0.202	0.202	0.135	0.908
Mode Based Estimation	0.199	0.199	0.032	0.997	0.197	0.198	0.062	0.982	0.187	0.193	0.143	0.870
MR-PRESSO	0.199	0.200	0.050	0.983	0.200	0.200	0.089	0.928	0.202	0.202	0.142	0.846
MR-Robust	0.200	0.200	0.025	0.997	0.200	0.199	0.077	0.668	0.203	0.204	0.170	0.271
MR-Lasso	0.200	0.200	0.026	1.000	0.200	0.200	0.038	0.996	0.201	0.201	0.080	0.942
MR-Egger	0.200	0.199	0.311	0.149	0.209	0.211	0.396	0.120	0.196	0.196	0.462	0.102
Contamination Mixture	0.201	0.201	0.023	1.000	0.201	0.200	0.032	0.997	0.215	0.203	0.194	0.943
MR-Mix	0.209	0.200	0.141	0.606	0.211	0.200	0.233	0.793	0.182	0.170	0.353	0.200
MR-RAPS	0.199	0.199	0.075	0.644	0.201	0.202	0.131	0.345	0.202	0.204	0.177	0.231
Scenario 2: Directional pleiotropy, InSIDE satisfied												
Weighted Median	0.212	0.210	0.033	1.000	0.232	0.222	0.065	0.998	0.289	0.255	0.132	0.989
Mode Based Estimation	0.200	0.199	0.031	0.998	0.205	0.203	0.052	0.989	0.236	0.224	0.116	0.950
MR-PRESSO	0.223	0.216	0.042	1.000	0.267	0.247	0.083	0.999	0.344	0.319	0.134	0.995
MR-Robust	0.203	0.203	0.025	0.999	0.237	0.229	0.070	0.821	0.353	0.344	0.135	0.731
MR-Lasso	0.204	0.204	0.025	1.000	0.216	0.213	0.041	1.000	0.266	0.244	0.101	0.994
MR-Egger	0.202	0.194	0.222	0.217	0.197	0.188	0.277	0.150	0.204	0.197	0.331	0.126
Contamination Mixture	0.201	0.201	0.022	1.000	0.203	0.203	0.034	0.999	0.234	0.206	0.193	0.969
MR-Mix	0.230	0.210	0.141	0.461	0.263	0.220	0.232	0.518	0.328	0.230	0.378	0.502
MR-RAPS	0.248	0.244	0.059	0.969	0.307	0.303	0.099	0.881	0.381	0.376	0.131	0.837
Scenario 3: Directional pleiotropy, InSIDE violated												
Weighted Median	0.225	0.219	0.045	1.000	0.270	0.244	0.096	1.000	0.361	0.320	0.158	0.998
Mode Based Estimation	0.202	0.201	0.039	0.995	0.215	0.206	0.072	0.986	0.270	0.245	0.137	0.963
MR-PRESSO	0.247	0.234	0.058	1.000	0.326	0.302	0.108	1.000	0.429	0.415	0.146	0.999
MR-Robust	0.206	0.205	0.026	0.997	0.265	0.251	0.084	0.781	0.427	0.419	0.137	0.838
MR-Lasso	0.209	0.208	0.027	1.000	0.235	0.226	0.056	1.000	0.326	0.293	0.131	0.999
MR-Egger	0.289	0.269	0.231	0.316	0.305	0.295	0.276	0.250	0.297	0.293	0.314	0.201
Contamination Mixture	0.201	0.202	0.023	1.000	0.204	0.203	0.036	0.999	0.248	0.209	0.208	0.974
MR-Mix	0.241	0.215	0.150	0.505	0.288	0.223	0.248	0.593	0.362	0.250	0.366	0.546
MR-RAPS	0.281	0.274	0.068	0.990	0.371	0.365	0.102	0.976	0.459	0.454	0.125	0.974

Abbrevations: T1 error: Type 1 error.

TABLE 3.4 – *Mean, median, standard deviation (SD) of estimates, and Type 1 error/empirical power (%) with 100 genetic variants.*

	30% invalid				50% invalid				70% invalid			
Null causal effect: $\theta = 0$												
Method	Mean	Median	SD	T1 error	Mean	Median	SD	T1 error	Mean	Median	SD	T1 error
Scenario 1: Balanced pleiotropy, InSIDE satisfied												
Weighted Median	0.000	0.000	0.025	0.069	-0.001	0.000	0.041	0.124	0.000	0.000	0.077	0.234
Mode Based Estimation	0.000	0.000	0.024	0.038	0.000	0.000	0.035	0.082	0.000	0.000	0.084	0.333
MR-PRESSO	0.000	0.000	0.025	0.134	0.000	0.001	0.047	0.224	0.000	-0.001	0.083	0.313
MR-Robust	0.000	0.000	0.020	0.052	0.000	0.001	0.053	0.024	0.000	-0.001	0.126	0.044
MR-Lasso	0.000	0.000	0.019	0.042	0.000	0.000	0.029	0.072	0.000	0.000	0.055	0.120
MR-Egger	-0.001	-0.001	0.195	0.067	-0.001	0.000	0.252	0.069	-0.003	-0.005	0.296	0.065
Contamination Mixture	0.000	0.000	0.019	0.064	0.000	0.000	0.029	0.088	0.002	0.000	0.211	0.136
MR-Mix	0.000	0.000	0.075	0.038	-0.001	0.000	0.072	0.024	0.000	0.000	0.058	0.000
MR-RAPS	0.000	-0.001	0.053	0.016	-0.001	0.000	0.095	0.036	0.000	-0.003	0.133	0.052
Scenario 2: Directional pleiotropy, InSIDE satisfied												
Weighted Median	0.013	0.012	0.023	0.105	0.033	0.029	0.039	0.258	0.087	0.071	0.084	0.537
Mode Based Estimation	0.000	0.000	0.020	0.037	0.004	0.003	0.030	0.089	0.034	0.030	0.067	0.351
MR-PRESSO	0.022	0.018	0.026	0.294	0.071	0.062	0.056	0.628	0.162	0.150	0.096	0.856
MR-Robust	0.004	0.004	0.018	0.051	0.042	0.038	0.047	0.040	0.193	0.189	0.100	0.425
MR-Lasso	0.004	0.004	0.017	0.077	0.020	0.018	0.029	0.242	0.076	0.066	0.067	0.617
MR-Egger	0.001	-0.003	0.143	0.062	-0.002	-0.005	0.180	0.059	0.003	0.001	0.210	0.058
Contamination Mixture	0.000	0.001	0.017	0.061	0.001	0.001	0.025	0.090	0.018	0.005	0.160	0.156
MR-Mix	0.005	0.000	0.074	0.034	0.004	0.000	0.072	0.035	0.006	0.000	0.070	0.007
MR-RAPS	0.058	0.056	0.042	0.142	0.140	0.138	0.072	0.435	0.233	0.232	0.097	0.663
Scenario 3: Directional pleiotropy, InSIDE violated												
Weighted Median	0.027	0.025	0.027	0.258	0.077	0.065	0.062	0.619	0.184	0.163	0.116	0.881
Mode Based Estimation	0.001	0.001	0.021	0.042	0.010	0.008	0.035	0.120	0.065	0.054	0.087	0.465
MR-PRESSO	0.053	0.047	0.040	0.658	0.152	0.142	0.079	0.943	0.276	0.270	0.103	0.993
MR-Robust	0.007	0.007	0.019	0.054	0.078	0.071	0.059	0.080	0.292	0.289	0.099	0.805
MR-Lasso	0.010	0.009	0.018	0.153	0.049	0.043	0.041	0.575	0.165	0.151	0.097	0.927
MR-Egger	0.119	0.114	0.148	0.162	0.141	0.138	0.178	0.159	0.123	0.124	0.194	0.114
Contamination Mixture	0.001	0.001	0.017	0.069	0.003	0.003	0.026	0.107	0.021	0.009	0.135	0.199
MR-Mix	0.007	0.000	0.073	0.037	0.008	0.000	0.070	0.034	0.007	0.000	0.068	0.006
MR-RAPS	0.104	0.101	0.049	0.545	0.224	0.221	0.076	0.896	0.330	0.327	0.090	0.976
Positive causal effect: $\theta = +0.2$												
Method	Mean	Median	SD	Power	Mean	Median	SD	Power	Mean	Median	SD	Power
Scenario 1: Balanced pleiotropy, InSIDE satisfied												
Weighted Median	0.200	0.200	0.028	1.000	0.201	0.201	0.043	0.996	0.201	0.200	0.078	0.939
Mode Based Estimation	0.199	0.199	0.025	1.000	0.199	0.199	0.036	0.998	0.192	0.192	0.085	0.908
MR-PRESSO	0.200	0.200	0.026	1.000	0.201	0.200	0.047	0.993	0.201	0.200	0.083	0.934
MR-Robust	0.200	0.200	0.021	1.000	0.201	0.202	0.055	0.896	0.200	0.200	0.126	0.373
MR-Lasso	0.200	0.200	0.020	1.000	0.201	0.201	0.031	1.000	0.200	0.200	0.057	0.986
MR-Egger	0.200	0.200	0.199	0.212	0.199	0.200	0.248	0.146	0.206	0.206	0.298	0.130
Contamination Mixture	0.202	0.202	0.021	1.000	0.203	0.204	0.031	1.000	0.228	0.206	0.253	0.977
MR-Mix	0.203	0.200	0.091	0.979	0.191	0.200	0.105	0.873	0.028	0.000	0.103	0.001
MR-RAPS	0.201	0.201	0.054	0.880	0.201	0.199	0.095	0.504	0.201	0.202	0.133	0.332
Scenario 2: Directional pleiotropy, InSIDE satisfied												
Weighted Median	0.214	0.213	0.025	1.000	0.237	0.233	0.043	1.000	0.290	0.275	0.086	1.000
Mode Based Estimation	0.199	0.199	0.023	1.000	0.203	0.203	0.033	1.000	0.229	0.226	0.071	0.986
MR-PRESSO	0.222	0.219	0.027	1.000	0.271	0.263	0.056	1.000	0.362	0.351	0.095	1.000
MR-Robust	0.205	0.205	0.020	1.000	0.247	0.243	0.051	0.995	0.395	0.390	0.100	0.982
MR-Lasso	0.205	0.205	0.019	1.000	0.223	0.220	0.032	1.000	0.281	0.270	0.071	1.000
MR-Egger	0.201	0.198	0.144	0.325	0.203	0.199	0.182	0.229	0.201	0.199	0.213	0.187
Contamination Mixture	0.202	0.202	0.019	1.000	0.204	0.204	0.028	1.000	0.256	0.211	0.271	0.995
MR-Mix	0.208	0.200	0.094	0.641	0.211	0.200	0.098	0.899	0.055	0.000	0.136	0.061
MR-RAPS	0.260	0.257	0.043	1.000	0.342	0.340	0.073	0.997	0.434	0.431	0.098	0.994
Scenario 3: Directional pleiotropy, InSIDE violated												
Weighted Median	0.230	0.228	0.029	1.000	0.282	0.271	0.065	1.000	0.389	0.369	0.116	1.000
Mode Based Estimation	0.201	0.200	0.024	1.000	0.210	0.208	0.039	0.999	0.263	0.252	0.088	0.988
MR-PRESSO	0.252	0.246	0.039	1.000	0.349	0.339	0.077	1.000	0.474	0.468	0.100	1.000
MR-Robust	0.209	0.209	0.022	1.000	0.287	0.280	0.063	0.986	0.495	0.492	0.097	0.999
MR-Lasso	0.212	0.212	0.021	1.000	0.254	0.248	0.045	1.000	0.372	0.359	0.096	1.000
MR-Egger	0.321	0.314	0.146	0.640	0.343	0.339	0.180	0.534	0.327	0.323	0.194	0.420
Contamination Mixture	0.203	0.203	0.020	1.000	0.206	0.205	0.031	1.000	0.268	0.217	0.269	0.995
MR-Mix	0.211	0.205	0.094	0.730	0.211	0.200	0.101	0.900	0.058	0.000	0.150	0.033
MR-RAPS	0.306	0.303	0.050	1.000	0.426	0.424	0.077	1.000	0.531	0.529	0.088	1.000

Abbrevations: T1 error: Type 1 error.

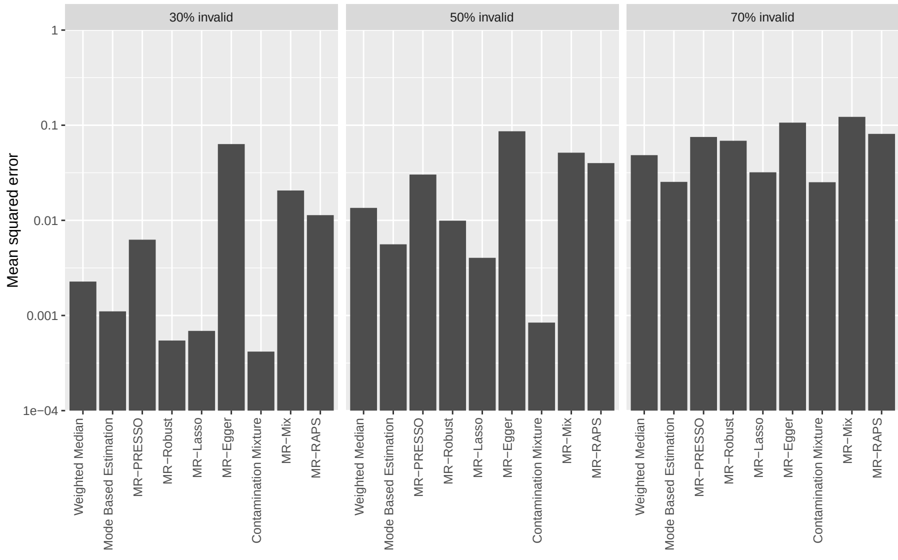

FIGURE 3.4 – *Mean squared errors for the different methods in scenario 3 (directional pleiotropy, InSIDE violated) with a null causal effect for 30 variants. Note the vertical axis is on a logarithmic scale.*

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 λ 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 3.5 – *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/m^2 increase in BMI.*

Method	Causal estimate (95% CI)	CI width
Weighted Median	0.376 (0.206, 0.546)	0.340
Mode Based Estimation	0.382 (0.181, 0.583)	0.402
MR-PRESSO	0.410 (0.309, 0.511)	0.202
MR-Robust	0.425 (0.325, 0.526)	0.201
MR-Lasso	0.442 (0.354, 0.530)	0.176
MR-Egger	0.481 (0.165, 0.796)	0.631
(intercept)	-0.003 (-0.011, 0.005)	
Contamination Mixture	0.490 (0.372, 0.602)	0.230
MR-Mix	0.425 (-0.283, 1.133)	1.416
MR-RAPS	0.390 (0.308, 0.546)	0.238

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 not 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.

Appendices

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 ε_i^U. The effect of variant j on the confounder is represented by coefficient ϕ_j (this is zero for a valid IV). The exposure is linear in the genetic variants, the confounder and an independent error term ε_i^X. The effect of variant j on the exposure is represented by coefficient γ_j. The outcome is linear in the genetic variants, exposure, confounders and an independent error term ε_i^Y. The effect of variant j on the outcome is represented by coefficient α_j (again, this is zero for a valid IV). The effect of the exposure on the outcome is represented by θ. 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 ($\text{maf}_j \sim \mathcal{U}(0.1, 0.5)$). The error terms ε_i^U, ε_i^X and ε_i^Y each follow an independent normal distribution with mean 0 and unit variance.

We can represent the model mathematically as:

$$U_i = \sum_{j=1}^{J} \phi_j G_{ij} + \varepsilon_i^U, \tag{3.9}$$

$$X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \varepsilon_i^X, \tag{3.10}$$

$$Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \theta X_i + U_i + \varepsilon_i^Y, \tag{3.11}$$

$$\text{maf}_j \sim \mathscr{U}(0.1, 0.5), \tag{3.12}$$

$$G_{ij} \sim \text{Binomial}(2, \text{maf}_j) \text{ independently}, \tag{3.13}$$

$$\varepsilon_i^U, \varepsilon_i^X, \varepsilon_i^Y \sim \mathscr{N}(0,1) \text{ independently}. \tag{3.14}$$

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 γ_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 ($\alpha_j \sim \mathscr{N}(0, 0.15)$) 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 ($\alpha_j \sim \mathscr{N}(0.1, 0.075)$) 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 ($\alpha_j \sim \mathscr{N}(0.1, 0.075)$) and genetic effects on the confounder are generated from a uniform distribution ($\phi_j \sim \mathscr{U}(0, 0.1)$).

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 by $(\sqrt{N} * \sqrt{2 * \text{maf}_j * (1 - \text{maf}_j)})^{-1}$, 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 1×10^{-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).*

Variant	Robust	PRESSO	RAPS	Lasso	Contam mix	Mix	Associated traits
rs11191560	✓	✓	✓	✓	✓	✓	
rs2075650	✓	✓	✓	✓	✓	✓	Waist circumference, Hip circumference
rs2176040		✓	✓	✓	✓	✓	Waist circumference, Obesity
rs6567160		✓	✓	✓	✓	✓	Hand grip strength
rs7903146		✓	✓	✓	✓	✓	Hip circumference
rs11727676			✓	✓	✓	✓	Parental longevity
rs17024393			✓	✓	✓	✓	
rs11126666				✓	✓	✓	
rs13078960				✓	✓	✓	HDL cholesterol levels, Colorectal cancer, Diverticular disease
rs9914578				✓	✓	✓	
rs1000940					✓	✓	
rs11057405					✓	✓	Waist circumference, Hip circumference
rs11847697					✓	✓	
rs12446632					✓	✓	
rs12566985					✓	✓	
rs16907751					✓	✓	
rs205262					✓	✓	
rs2650492					✓	✓	
rs2836754					✓	✓	Hip circumference, Waist circumference, Obesity
rs3849570					✓	✓	
rs4787491					✓	✓	
rs492400					✓	✓	Hip circumference, Hand grip strength
rs7243357					✓	✓	
rs9641123					✓	✓	
rs10938397						✓	Alzheimer's disease, C-reactive protein, Age-related macular degeneration, Cerebrospinal fluid levels, Waist-hip ratio, Waist circumference, Longevity, LDL cholesterol, Total cholesterol, Cognitive decline, Cognitive impairment score, Cerebral amyloid deposition
rs10968576						✓	Waist Circumference
rs11030104						✓	
rs11688816						✓	Waist circumference, Waist-hip ratio
rs12016871						✓	Waist circumference, Intelligence
rs13021737						✓	Waist circumference
rs13191362						✓	Crohn's disease
rs13201877						✓	Waist circumference, Fat-free mass
rs1460676						✓	Intelligence
rs1516725						✓	Waist circumference
rs1528435						✓	Feeling nervous
rs17203016						✓	
rs2176598						✓	Body fat percentage, Hip circumference
rs2287019						✓	Body fat percentage, Waist circumference, Fat-free mass
rs2820292						✓	
rs3810291						✓	
rs3817334						✓	
rs543874						✓	
rs7164727						✓	Type 2 diabetes, Hip circumference, Waist circumference, Peak insulin response, Clinical laboratory measurements, Metabolic syndrome, Glycated hemoglobin levels, Schizophrenia, Systolic blood pressure, Pulse pressure, Fasting blood insulin, Fasting blood glucose, Proinsulin levels
rs7599312						✓	Lean body mass
rs7899106						✓	

Abbrevations: Robust: MR-Robust, PRESSO: MR-PRESSO, RAPS: MR-RAPS, Lasso: MR-:Lasso, Contam mix: Contamination Mixture, Mix: MR-Mix.

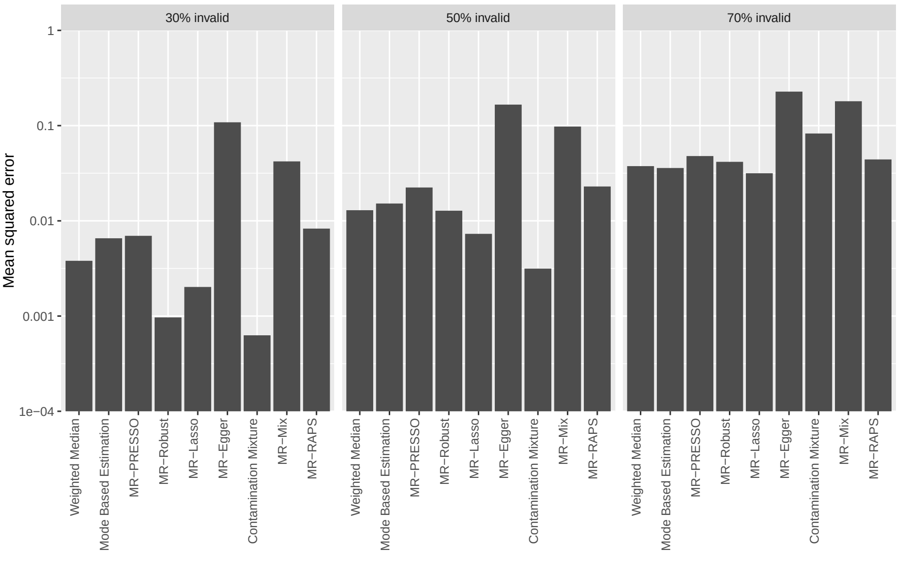

FIGURE 3.5 – *Mean squared error for the different methods in scenario 2 for 10 000 simulations, with directional pleiotropy and InSIDE satisfied with 10 variants.*

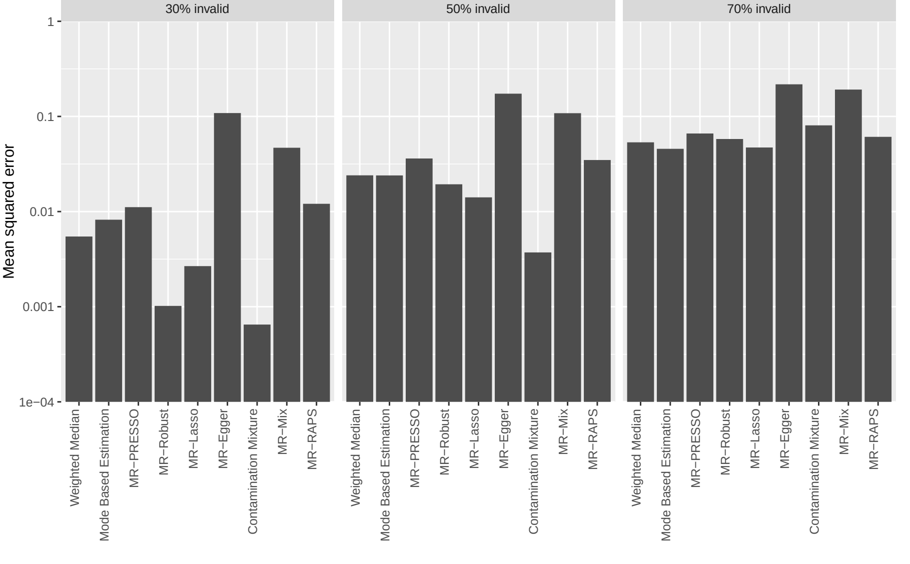

FIGURE 3.6 – *Mean squared error for the different methods in scenario 3 for 10 000 simulations, with directional pleiotropy and InSIDE violated with 10 variants.*

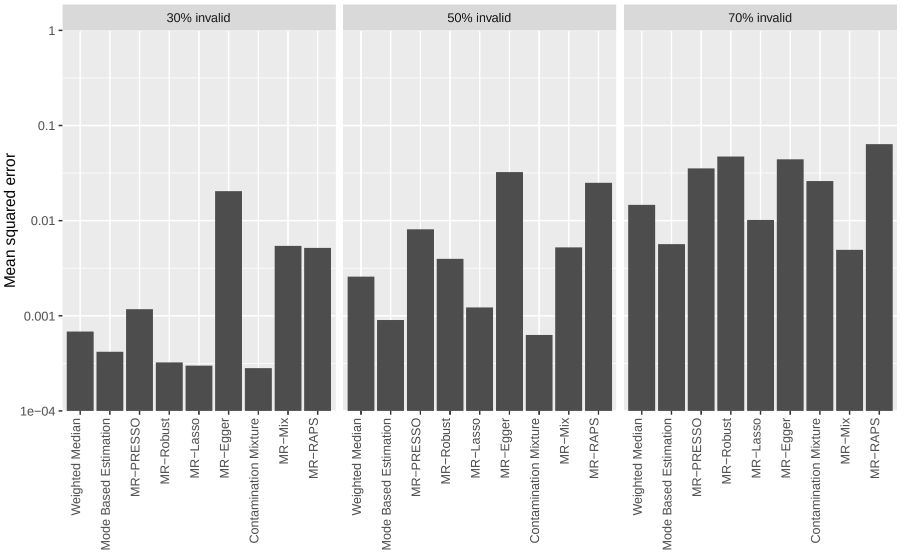

FIGURE 3.7 – *Mean squared error for the different methods in scenario 2 for 10 000 simulations, with directional pleiotropy and InSIDE satisfied with 100 variants.*

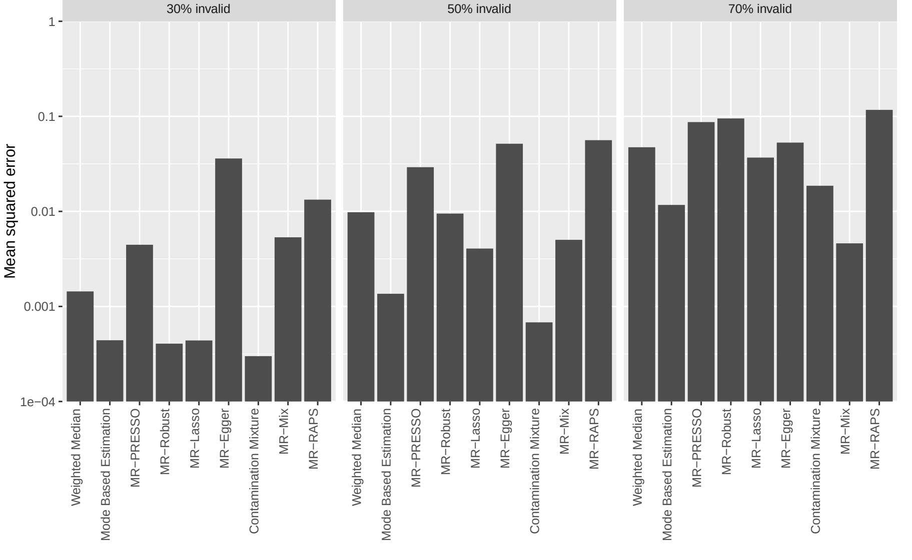

FIGURE 3.8 – *Mean squared error for the different methods in scenario 3 for 10 000 simulations, with directional pleiotropy and InSIDE violated with 100 variants.*

II

Polygenic risk scores

4

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[1]. 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-

[1]For example, such tests are provided by Leadership Consultants (https://leapership.com/shop/karmagene-dna-based-personality-test/) and Goldmen Genetics (https://goldmen.eu/).

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[2]. 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.[3]

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 23$^{\text{th}}$ 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

$$y_i = \mu + \sum_{j=1}^{J} \beta_j x_{ij} + \varepsilon_i, \tag{4.1}$$

with intercept μ, SNP effects β_j and residual term ε_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 $p < 0.05$.[5] Hence, it is not surprising that this finding could not be further replicated (Van der Loos et al., 2013).[6]

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.[7] 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

[5] In their analysis, Nicolaou and Shane (2010) adopted a significance level of 6×10^{-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 $p = 6 \times 10^{-4}$).

[6] 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.

[7] 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.

et al., 2010) and the Social Science Genetic Association Consortium[8], is therefore a prerequisite for the success of GWAS analysis.

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

[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 ($N \approx 100,000$) found only three genome-wide significant SNPs (Rietveld et al., 2013), the second one (Okbay et al., 2016b) identified 74 SNPs ($N \approx 300,000$), and the third one (Lee et al., 2018) identified 1,271 SNPs ($N \approx 1,100,000$). 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 $\alpha = 0.05/69$ rather than the genome-wide significance threshold of $\alpha = 5 \times 10^{-8}$.

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:[9]

$$PGS_i = \sum_{j=1}^{J} \beta_j x_{ij}, \tag{4.2}$$

where PGS_i is the value for the polygenic risk score for individual i, β_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.

plained 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 β_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 β_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 some 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[10]. 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

[10]For 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):[11]

$$SE_{it} = \sum_{k=1}^{K} \gamma_k PGS_{ik} + \boldsymbol{\delta} \boldsymbol{Z}_{it} + \alpha_i + \varepsilon_{it}, \tag{4.3}$$

where SE_{it} is the binary variable indicating the self-employment status of individual i at time t, γ_k is the effect of the polygenic risk score PGS_{ik} for trait k, $\boldsymbol{\delta}$ is a vector of coefficients for the vector of control variables $\boldsymbol{Z}_{it}$, α_i is an unobserved random random variable for individual i, and ε_{it} is the residual for individual i at time t.[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.[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.[14]

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

Polygenic risk score	**Coefficient**	**Standard error**	***p-value***
Attention deficit hyperactivity disorder	0.017	0.004	0.000
Anxiety (factor score)	0.001	0.004	0.796
Autism	-0.013	0.006	0.049
Bipolar disorder	0.010	0.005	0.047
Depressive symptoms	0.007	0.005	0.187
Educational attainment	0.013	0.005	0.004
Extraversion	0.007	0.004	0.100
General cognition	-0.012	0.005	0.010
Major depressive disorder	-0.005	0.005	0.367
Mental health (cross disorder)	-0.004	0.007	0.558
Neuroticism	0.008	0.006	0.202
Obsessive compulsive disorder	-0.001	0.004	0.752
Post-traumatic stress disorder	0.001	0.005	0.900
Schizophrenia	0.005	0.008	0.509
Well-being	0.010	0.005	0.032

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%.[15] 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

[14] 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.

[15] 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 – *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.*

Actual occupation	Predicted occupation based on model without polygenic risk scores		Predicted occupation based on model with polygenic risk scores	
	Self-employment	Wage work	Self-employment	Wage work
Self-employment	5.75%	14.11%	5.82%	14.04%
Wage work	14.10%	66.04%	14.03%	66.11%

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 Venkatamaran, 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,[16] 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.[17] 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).[18] 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.[19]

Nevertheless, capturing residual variance in polygenic 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 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 "quest for the entrepreneurial gene" (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.

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.

[19] Besides, ethical considerations are needed to determine whether such gene-based prediction of entrepreneurship is actually desirable.

5

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 preferenes 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:

$$y_i = \mu + \gamma_m g_{im} + \delta z_i + \nu_i, \tag{5.1}$$

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

$$G_i = \sum_{m=1}^{M} \gamma_m g_{im}, \tag{5.2}$$

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, γ_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.

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:

$$S_{ist} = \alpha_0 + \alpha_1 \text{Tax}_{st} + \alpha_2 G_i + \alpha_3 G_i \text{Tax}_{st} + \alpha_4 X_{ist} + S_s + D_t + \varepsilon_{ist}, \qquad (5.3)$$

where S_{ist} is a binary variable indicating whether individual i residing in state s in year t smokes or not, 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 α'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, ε_{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:

$$C_{ist} = \beta_0 + \beta_1 \text{Tax}_{st} + \beta_2 G_i + \beta_3 G_i \text{Tax}_{st} + \beta_4 X_{ist} + S_s + D_t + \tau_{ist}, \qquad (5.4)$$

where C_{ist} denotes the number of cigarettes smoked per day by individual i at time t in state s. In this regression, the β's are the effect size estimates and τ_{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.*

	Full sample		Subsample of current smokers	
Static variables	Mean	Std. Dev.	Mean	Std. Dev.
Female	0.570	0.495	0.563	0.496
Birth Year	1941	11.939	1944	10.473
Years of education	13.265	2.529	12.565	2.349
$PGS_{\text{Smoking Initiation}}$	0.000	1.000	0.140	0.964
$PGS_{\text{Smoking Intensity}}$	0.000	1.000	0.094	1.005
Dynamic variables	Mean	Std. Dev.	Mean	Std. Dev.
Currently smoking	0.139	0.346	1.000	0.000
Ever smoked	0.575	0.494	1.000	0.000
Cigaretes per day	2.398	7.394	17.277	11.698
Income	16,323	42,721	16,053	30,775
Married	0.663	0.473	0.616	0.486
Individuals	12,089		2,643	
Observations	98,605		13,642	

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.

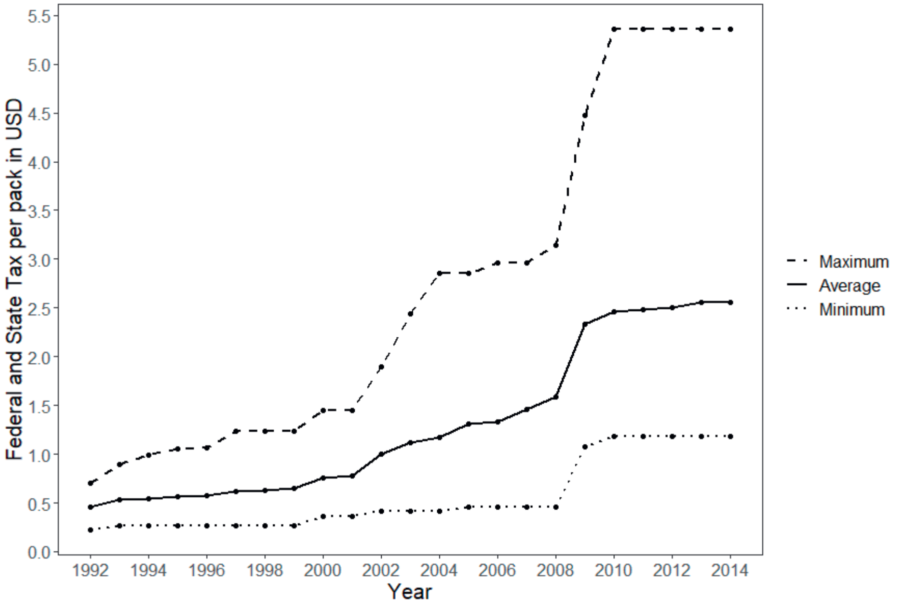

FIGURE 5.1 – *The average, minimum and maximum tobacco excise taxes levied per pack of 20 cigarettes in the United States from 1992 to 2014.*

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.*

	(1)	(2)	(3)
Log(Tax)	-0.060***	-0.060***	0.003
	(0.004)	(0.004)	(0.004)
$PGS_{\text{Smoking Initiation}}$	0.020***	0.021***	0.021***
	(0.004)	(0.004)	(0.004)
Log(Tax) × $PGS_{\text{Smoking Initiation}}$		-0.008*	-0.010**
		(0.003)	(0.003)
Female	-0.028***	-0.028***	-0.029***
	(0.004)	(0.005)	(0.005)
Birth Year	-0.000	0.000	-0.088
	(0.078)	(0.079)	(0.077)
Birth Year^2	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Income (in USD 1,000)	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)
Years of Education	-0.018***	-0.018***	-0.017***
	(0.001)	(0.001)	(0.001)
Married	-0.086***	-0.086***	-0.090***
	(0.007)	(0.007)	(0.007)
State & Wave Dummies	No	No	Yes
Observations	97,984	97,984	97,984
Individuals	12,058	12,058	12,058
R^2	0.0737	0.0740	0.0850

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).

TABLE 5.3 – *Results of the regressions explaining someone's current smoking intensity.*

	Full sample			Subsample of current smokers		
	(1)	(2)	(3)	(3)	(5)	(6)
Log(Tax)	-1.460***	-1.459***	0.024	-3.285***	-3.262***	-0.111
	(0.098)	(0.098)	(0.146)	(0.314)	(0.309)	(0.457)
$PGS_{\text{Smoking Intensity}}$	0.165**	0.172**	0.180**	0.467*	0.455*	0.495*
	(0.058)	(0.060)	(0.061)	(0.213)	(0.205)	(0.217)
Log(Tax) × $PGS_{\text{Smoking Intensity}}$		-0.078	-0.080		-0.297	-0.317
		(0.059)	(0.058)		(0.200)	(0.197)
Female	-0.953***	-0.953***	-0.991***	-3.556***	-3.562***	-3.825***
	(0.100)	(0.099)	(0.101)	(0.393)	(0.393)	(0.403)
Birth Year	0.994	0.985	-1.106	13.490	13.340	12.200
	(1.569)	(1.568)	(1.535)	(7.091)	(7.102)	(7.197)
Birth Year^2	-0.000	-0.000	0.000	-0.003	-0.003	-0.003
	(0.000)	(0.000)	(0.000)	(0.002)	(0.002)	(0.002)
Income (in USD 1,000)	-0.005***	-0.005***	-0.007**	-0.008	-0.007	-0.012*
	(0.002)	(0.002)	(0.002)	(0.005)	(0.005)	(0.006)
Years of Education	-0.347***	-0.347***	-0.340***	-0.310**	-0.309**	-0.295**
	(0.024)	(0.024)	(0.023)	(0.091)	(0.090)	(0.088)
Married	-1.568***	-1.569***	-1.659***	-0.588	-0.594	-0.973**
	(0.125)	(0.125)	(0.124)	(0.303)	(0.302)	(0.295)
State & Wave Dummies	No	No	Yes	No	No	Yes
Observations	97,955	97,955	97,955	13,491	13,491	13,491
Individuals	12,058	12,058	12,058	2,614	2,614	2,614
R^2	0.0597	0.0598	0.0726	0.0648	0.0651	0.0974

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 inclusion 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.

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

6

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 neuropsychiatric/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 suggests 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[1]. 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.ctsu.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 base 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 platform dummy (P-value = 1.000). 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

$$y_i = \mu + x_{ik} b_k + \varepsilon_i, \tag{6.1}$$

where y_i is the value of the phenotype for individual i, μ 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 ε_i identically and independently distributed as $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_e^2\right)$, where σ_e^2 the residual variance for the trait. If all causal variants are known, they can be added into one single model for the trait:

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

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 $\mathbf{y} = \mu\mathbf{1} + \mathbf{g} + \boldsymbol{\varepsilon}$ and $\mathbf{g} = \mathbf{Su}$. Now, $\mathbf{u}$ is treated as a random effect that follows the distribution $\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \sigma_u^2\mathbf{I}\right)$, where σ_u^2 is the variance of causal effects. As a result, the distribution of the total genetic contribution is multivariate normally distributed as $\mathbf{g} \sim \mathcal{N}\left(\mathbf{0}, \sigma_u^2\mathbf{SS}^\top\right)$. Now, σ_g^2 $(= m\sigma_u^2)$ can be interpreted as the variance of the total additive genetic effects. The variance-covariance matrix of $\mathbf{y}$ becomes

$$\operatorname{Var}(\mathbf{y}) = \sigma_u^2\mathbf{SS}^\top + \sigma_e^2\mathbf{I} = \frac{\sigma_g^2}{m}\mathbf{SS}^\top + \sigma_e^2\mathbf{I} = \sigma_g^2\mathbf{G} + \sigma_e^2\mathbf{I}, \tag{6.3}$$

where $\mathbf{G}$ $(= m^{-1}\mathbf{S}\mathbf{S}^\top)$ 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 $\mathbf{G}$ directly. This $\mathbf{G}$ matrix can be approximated by applying the same formula to a genome-wide sample of SNPs $\mathbf{X}^*$ instead of $\mathbf{S}$, that is,

$$\mathbf{A} = M^{-1}\mathbf{X}^*\mathbf{X}^{*\top}, \tag{6.4}$$

where M is the total number of SNPs used in the standardized genome-wide sample $\mathbf{X}^*$, 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 $\mathbf{y}_t$. 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 $\mathbf{V}_G$ and $\mathbf{V}_E$. Let $\mathbf{y}$ now be the stacked vector of all T vectors $\mathbf{y}_t$, that is, $\mathbf{y}^\top = [\mathbf{y}_1^\top, \ldots, \mathbf{y}_T^\top]^\top$. The multivariate model is

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^* & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}^* & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{Z}^* \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_T \end{bmatrix} + \begin{bmatrix} \mathbf{X}^* & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^* & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{X}^* \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_T \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_T \end{bmatrix}, \tag{6.5}$$

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

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6.6}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_E \otimes \mathbf{I}_N)$ and $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_G \otimes \mathbf{I}_P)$. Here, $\otimes$ denotes of the Kronecker product and $\mathbf{I}_j$ is the $j \times j$ identity matrix. The main interest here lies in the genetic covariance matrix $\mathbf{V}_G$, of which the off-diagonal elements give information about the shared genetic architecture between the different traits. Note that for $\mathbf{V}_G$ and $\mathbf{V}_E$ to be covariance matrices, they must be constrained to be positive semidefinite.

To estimate the the genetic and environment covariance matrices $\mathbf{V}_G$ and $\mathbf{V}_E$ 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) [2]. 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).

[2] 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.

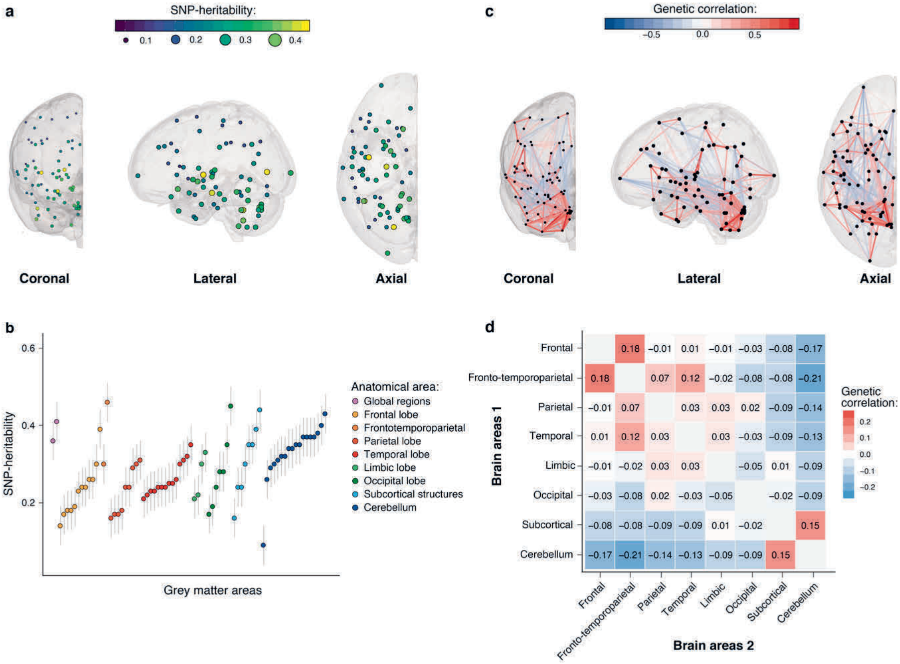

FIGURE 6.1 – *Spatial mapping of the estimates for SNP-based heritability and genetic correlation across the different brain regions, SNP-based heritability per anatomical area, and genetic correlation table of aggregated anatomical area. a, Spatial mapping of the SNP-based heritability of the different brain regions, where blue dots represents 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.*

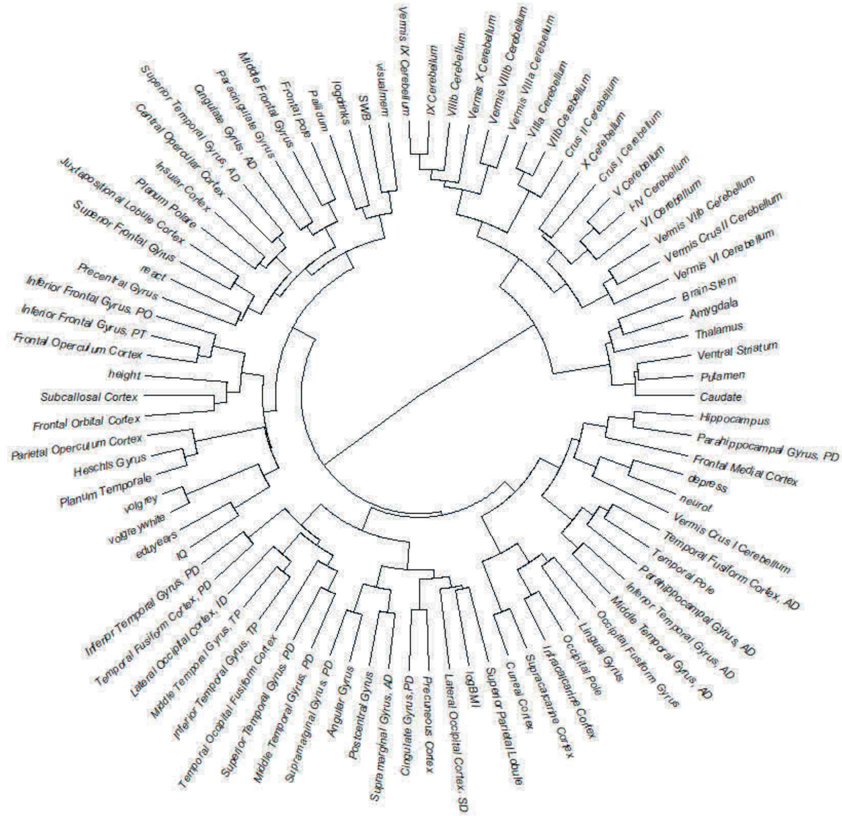

FIGURE 6.2 – *Clusters are identified using Ward's method with a D2 ward for hierarchical clustering. This method minimizes the error sum of squares by agglomerating clusters each step (see Kaufmann and Rousseeuw (1990) for more details). The colors represent the different clusters, where the number of clusters is identified using the elbow method.*

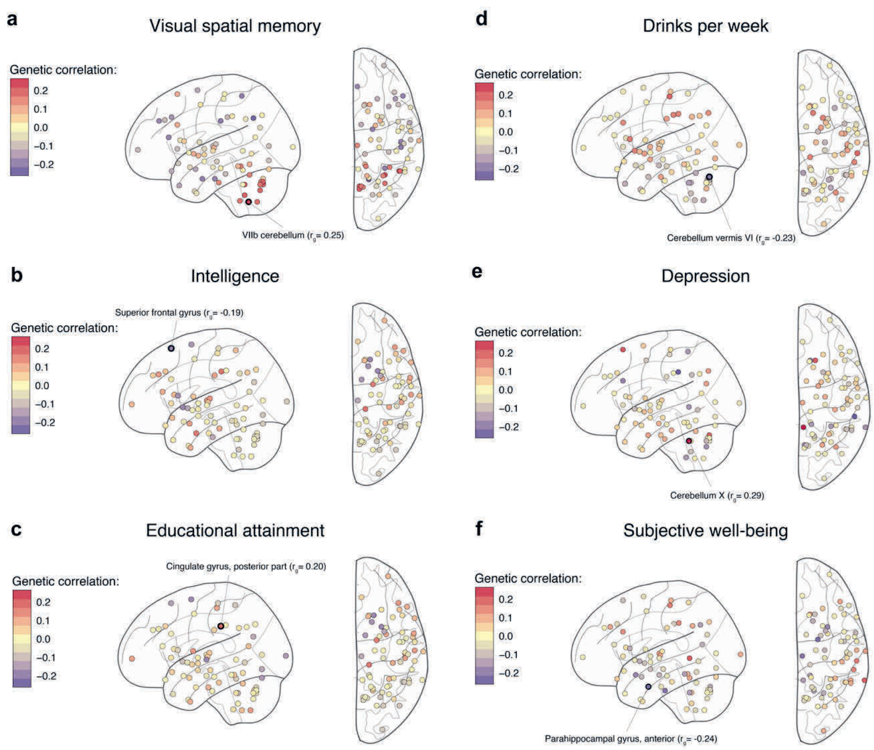

FIGURE 6.3 – *Spatial mapping of the genetic correlation between brain regions and the behavioral traits, where blue points represent a negative correlation and red points a positive correlation. a, Spatial mapping of the genetic correlation between visual spatial memory (measured using a pairs-matching 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.*

6.5 Discussion

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 behavioural 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 behavioural 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 ($\approx 25\%$ 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 $\mathbf{V}_G$ and $\mathbf{V}_E$ 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 $\mathbf{Z}$.
2. Reparametrize the covariance matrices $\mathbf{V}_G$ and $\mathbf{V}_E$ to guarantee their positive semi-definiteness.
3. Express the log likelihood.
4. Maximize the likelihood in the unknown covariance matrices $\mathbf{V}_G$ and $\mathbf{V}_E$ by the BFGS algorithm.
5. Formulate efficient expressions for the gradients, the determinants, and the log likelihood needed for the BFGS algorithm.
6. Estimate the Fisher Information matrix for the variances and genetic correlation matrix.
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:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6.7}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_E \otimes \mathbf{I}_N)$ and $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_G \otimes \mathbf{I}_M)$ and the vectors $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ 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 $\boldsymbol{\delta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the linear combination $\mathbf{C}\boldsymbol{\delta} + \mathbf{m} \sim \mathcal{N}(\mathbf{C}\boldsymbol{\mu} + \mathbf{m}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)$. Applying this result to the last two terms of (6.7) implies that the linear sum $\mathbf{X}\beta + \boldsymbol{\varepsilon}$ is also normally distributed, that is, $\mathbf{X}\beta + \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{X}(\mathbf{V}_G \otimes \mathbf{I}_M)\mathbf{X}^\top + \mathbf{V}_E \otimes \mathbf{I}_N)$.

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 $\boldsymbol{\gamma}$ of the covariates. The same covariance matrices can be obtained by removing the effects of the covariates $\mathbf{Z}$ by premultiplying the phenotype vector $\mathbf{y}$ by the anti-projection matrix $\mathbf{I} - \mathbf{Z}\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top$. Additional computational efficiency can be obtained by using the sparsity in our variance-covariance structure.

Consider the singular value decomposition of $\left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\right)\mathbf{X} = \mathbf{P}\boldsymbol{\Phi}\mathbf{Q}^\top$ with $\mathbf{P}$ the $NT \times MT$ orthonormal matrix of left singular vectors, $\boldsymbol{\Phi}$ 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

$$\begin{aligned}\tilde{\mathbf{y}} &= \mathbf{P}^\top\left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\right)\mathbf{y} = \boldsymbol{\Phi}\mathbf{Q}^\top\boldsymbol{\beta} + \mathbf{P}^\top\boldsymbol{\varepsilon} \\ &\sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Phi}\mathbf{Q}^\top(\mathbf{V}_G \otimes \mathbf{I}_M)\mathbf{Q}\boldsymbol{\Phi} + \mathbf{P}^\top(\mathbf{V}_E \otimes \mathbf{I}_N)\mathbf{P}\right).\end{aligned} \tag{6.8}$$

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

$$\mathbf{I} - \mathbf{Z}\left(\mathbf{Z}^\top\mathbf{Z}\right)\mathbf{Z}^\top = \begin{bmatrix} \mathbf{I} - \mathbf{Z}^*\left(\mathbf{Z}^{*\top}\mathbf{Z}^*\right)\mathbf{Z}^{*\top} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \mathbf{Z}^*\left(\mathbf{Z}^{*\top}\mathbf{Z}^*\right)\mathbf{Z}^{*\top} & \dots & \mathbf{0} \\ | & | & \ddots & | \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} - \mathbf{Z}^*\left(\mathbf{Z}^{*\top}\mathbf{Z}^*\right)\mathbf{Z}^{*\top} \end{bmatrix}. \tag{6.9}$$

Because $\mathbf{X}$ has a similar block structure as the antiprojection matrix of $\mathbf{Z}$, more efficient expressions are possible. Denote the singular value decomposition of

$\left(\mathbf{I}-\mathbf{Z}^*\left(\mathbf{Z}^{*\top}\mathbf{Z}^*\right)\mathbf{Z}^{*\top}\right)\mathbf{X}^*$ be given by

$$\left(\mathbf{I}-\mathbf{Z}^*\left(\mathbf{Z}^{*\top}\mathbf{Z}^*\right)\mathbf{Z}^{*\top}\right)\mathbf{X}^* = \mathbf{P}^*\mathbf{\Phi}^*\mathbf{Q}^{*\top}. \tag{6.10}$$

This gives us easy expressions for

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}^* & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}^* & \dots & \mathbf{0} \\ | & | & \ddots & | \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{P}^* \end{bmatrix} = \mathbf{I}_T \otimes \mathbf{P}^*, \tag{6.11}$$

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}^* & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}^* & \dots & \mathbf{0} \\ | & | & \ddots & | \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Phi}^* \end{bmatrix} = \mathbf{I}_T \otimes \mathbf{\Phi}^*, \tag{6.12}$$

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}^* & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}^* & \dots & \mathbf{0} \\ | & | & \ddots & | \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{Q}^* \end{bmatrix} = \mathbf{I}_T \otimes \mathbf{Q}^*. \tag{6.13}$$

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

$$\begin{aligned} \mathrm{Var}(\tilde{\mathbf{y}}) &= \mathbf{\Phi}\mathbf{Q}^\top(\mathbf{V}_G \otimes \mathbf{I}_M)\mathbf{Q}\mathbf{\Phi} + \mathbf{P}^\top(\mathbf{V}_E \otimes \mathbf{I}_N)\mathbf{P} \\ &= \left(\mathbf{I}_T \otimes \mathbf{\Phi}^*\right)\left(\mathbf{I}_T \otimes \mathbf{Q}^*\right)(\mathbf{V}_G \otimes \mathbf{I}_M)\left(\mathbf{I}_T \otimes \mathbf{Q}^{*\top}\right)\left(\mathbf{I}_T \otimes \mathbf{\Phi}^*\right) \\ &+ \left(\mathbf{I}_T \otimes \mathbf{P}^{*\top}\right)(\mathbf{V}_E \otimes \mathbf{I}_N)\left(\mathbf{I}_T \otimes \mathbf{P}^*\right) \\ &= \left(\mathbf{I}_T \otimes \mathbf{\Phi}^*\right)(\mathbf{V}_G \otimes \mathbf{I}_M)\left(\mathbf{I}_T \otimes \mathbf{\Phi}^*\right) + (\mathbf{V}_E \otimes \mathbf{I}_N) \\ &= \mathbf{V}_G \otimes \mathbf{\Phi}^{*^2} + \mathbf{V}_E \otimes \mathbf{I}_N. \end{aligned} \tag{6.14}$$

Hence, $\tilde{\mathbf{y}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_G \otimes \mathbf{\Phi}^{*^2} + \mathbf{V}_E \otimes \mathbf{I}_N\right)$.

Step 2. Ensuring $\mathbf{V}_G$ and $\mathbf{V}_E$ to be positive semi-definite For $\mathbf{V}_G$ and $\mathbf{V}_E$ 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 $\mathbf{V}_G = \mathbf{\Gamma}_G\mathbf{\Gamma}_G^\top$ and $\mathbf{V}_E = \mathbf{\Gamma}_E\mathbf{\Gamma}_E^\top$ where $\mathbf{\Gamma}_G$ and $\mathbf{\Gamma}_E$ are lower triangular matrices.

Step 3. Expression of the log likelihood To find estimates of $\mathbf{V}_G$ and $\mathbf{V}_E$, we have to maximize the corresponding log-likelihood (where we denote the $T(T+1)$ vector of all parameters in $\boldsymbol{\Gamma}_G$ and $\boldsymbol{\Gamma}_E$ by $\boldsymbol{\theta}$) up to a constant:

$$\begin{aligned}\log l(\boldsymbol{\theta}) &= -\frac{1}{2}\left(\log\left|\mathbf{V}_G \otimes \boldsymbol{\Phi}^{*^2} + \mathbf{V}_E \otimes \mathbf{I}_N\right| + \tilde{\mathbf{y}}^\top \left(\mathbf{V}_G \otimes \boldsymbol{\Phi}^{*^2} + \mathbf{V}_E \otimes \mathbf{I}_N\right)^{-1} \tilde{\mathbf{y}}\right) \\ &= -\frac{1}{2}\left(NT\log(2\pi) + \sum_{i=1}^{N} \log\left|\boldsymbol{\phi}_{ii}^{*^2}\mathbf{V}_G + \mathbf{V}_E\right| + \sum_{i=1}^{N} \tilde{\mathbf{y}}_i^\top \left(\boldsymbol{\phi}_{ii}^{*^2}\mathbf{V}_G + \mathbf{V}_E\right)^{-1} \tilde{\mathbf{y}}_i\right) \\ &= -\frac{1}{2}NT\log(2\pi) + \sum_{i=1}^{N} \log\left|\boldsymbol{\phi}_{ii}^{*^2}\boldsymbol{\Gamma}_G\boldsymbol{\Gamma}_G^\top + \boldsymbol{\Gamma}_E\boldsymbol{\Gamma}_E^\top\right| \\ &-\frac{1}{2}\sum_{i=1}^{N} \tilde{\mathbf{y}}_i^\top \left(\boldsymbol{\phi}_{ii}^{*^2}\boldsymbol{\Gamma}_G\boldsymbol{\Gamma}_G^\top + \boldsymbol{\Gamma}_E\boldsymbol{\Gamma}_E^\top\right)^{-1} \tilde{\mathbf{y}}_i,\end{aligned} \tag{6.15}$$

where $\tilde{\mathbf{y}}_i^\top$ is row i of the $N \times T$ matrix $\widetilde{\mathbf{Y}}$ which is the unstacked version of $\tilde{\mathbf{y}}$. 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:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \mathbf{p}_k, \tag{6.16}$$

where α is the step size of the line search, k is an iteration counter, and $\mathbf{p}_k$ is the search direction defined by:

$$\mathbf{p}_k = -\mathbf{B}_k^{-1} \nabla \log l(\boldsymbol{\theta}_k), \tag{6.17}$$

with $\nabla \log l(\boldsymbol{\theta})$ the gradient of $\log l(\boldsymbol{\theta})$ and the approximation of the inverse of the Hessian $\mathbf{B}_{k+1}^{-1}$ is defined by:

$$\mathbf{s}_k = \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = \alpha_k \mathbf{p}_k, \tag{6.18}$$

$$\mathbf{d}_k = \nabla \log l(\boldsymbol{\theta}_{k+1}) - \nabla \log l(\boldsymbol{\theta}_k), \tag{6.19}$$

$$\rho_k = (\mathbf{s}_k^\top \mathbf{d}_k)^{-1}, \tag{6.20}$$

$$\mathbf{B}_{k+1}^{-1} = (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{d}_k^\top)\mathbf{B}_k^{-1}(\mathbf{I} - \rho_k \mathbf{d}_k \mathbf{s}_k^\top) + \rho_k \mathbf{s}_k \mathbf{s}_k^\top. \tag{6.21}$$

Then, the BFGS algorithm is defined by:

1. Given start $\boldsymbol{\theta}_0$, convergence tolerance $\varepsilon > 0$, and $\mathbf{B}_0^{-1} = \mathbf{I}$.
2. $k \leftarrow 0$.
3. While $\|\nabla \log l(\boldsymbol{\theta}_k)\| > \varepsilon$.
4. Compute search direction $\mathbf{p}_k = -\mathbf{B}_k^{-1}\nabla \log l(\boldsymbol{\theta}_k)$.
5. Set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \mathbf{p}_k$ where α_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 $\boldsymbol{\theta}_k + \alpha \mathbf{p}_k$ 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 $\tilde{\mathbf{y}}$, that is $\mathbf{V} = \mathbf{V}_G \otimes \boldsymbol{\Phi}^{*^2} + \mathbf{V}_E \otimes \mathbf{I}_N$. Observe that this matrix has a block diagonal structure, where each block $\mathbf{V}_i = \boldsymbol{\phi}_{ii}^{*^2}\mathbf{V}_G + \mathbf{V}_E$. Let the eigenvalue decomposition of $\mathbf{V}_E$ be given by $\mathbf{D}\boldsymbol{\Lambda}\mathbf{D}^\top$, where $\mathbf{D}\mathbf{D}^\top = \mathbf{D}^\top\mathbf{D} = \mathbf{I}_T$. Using this eigendecomposition, we can rewrite $\mathbf{V}_i$ as:

$$\mathbf{V}_i = \mathbf{D}\boldsymbol{\Lambda}^{1/2}\left(\boldsymbol{\phi}_{ii}^{*^2}\widetilde{\mathbf{V}}_G + \mathbf{I}_T\right)\boldsymbol{\Lambda}^{1/2}\mathbf{D}^\top \text{ , where} \tag{6.22}$$

$$\widetilde{\mathbf{V}}_G = \boldsymbol{\Lambda}^{-1/2}\mathbf{D}^\top\mathbf{V}_G\mathbf{D}\boldsymbol{\Lambda}^{-1/2}. \tag{6.23}$$

Let $\mathbf{R}\mathbf{K}\mathbf{R}^\top$ denote the eigenvalue decomposition of $\widetilde{\mathbf{V}}_G$. Then, we can rewrite $\mathbf{V}_i$ as:

$$\mathbf{V}_i = \mathbf{D}\boldsymbol{\Lambda}^{1/2}\mathbf{R}\left(\boldsymbol{\phi}_{ii}^{*^2}\mathbf{K} + \mathbf{I}_T\right)\mathbf{R}^\top\boldsymbol{\Lambda}^{1/2}\mathbf{D}^\top. \tag{6.24}$$

The corresponding inverse of $\mathbf{V}_i$ is now also straightforward:

$$\mathbf{V}_i^{-1} = \mathbf{D}\boldsymbol{\Lambda}^{-1/2}\mathbf{R}\left(\boldsymbol{\phi}_{ii}^{*2}\mathbf{K}+\mathbf{I}_T\right)^{-1}\mathbf{R}^\top\boldsymbol{\Lambda}^{-1/2}\mathbf{D}^\top. \tag{6.25}$$

Using this, we will now derive a quick expression for the determinant of $\mathbf{V}$:

$$\begin{aligned}
|\mathbf{V}| &= \prod_{i=1}^{N} |\mathbf{V}_i| \\
&= \prod_{i=1}^{N}\left(\left|\mathbf{D}\boldsymbol{\Lambda}^{1/2}\mathbf{R}\left(\boldsymbol{\phi}_{ii}^{*2}\mathbf{K}+\mathbf{I}_T\right)\mathbf{R}^\top\boldsymbol{\Lambda}^{1/2}\mathbf{D}^\top\right|\right) \\
&= \prod_{i=1}^{N}\left(|\mathbf{D}|\left|\boldsymbol{\Lambda}^{1/2}\right||\mathbf{R}|\left|\boldsymbol{\phi}_{ii}^{*2}\mathbf{K}+\mathbf{I}_T\right|\left|\mathbf{R}^\top\right|\left|\boldsymbol{\Lambda}^{1/2}\right|\left|\mathbf{D}^\top\right|\right) \\
&= \prod_{i=1}^{N}\left(\left|\boldsymbol{\phi}_{ii}^{*2}\mathbf{K}+\mathbf{I}_T\right|\left|\mathbf{R}^\top\right||\mathbf{R}|\left|\boldsymbol{\Lambda}^{1/2}\right|\left|\boldsymbol{\Lambda}^{1/2}\right|\left|\mathbf{D}^\top\right||\mathbf{D}|\right) \\
&= \prod_{i=1}^{N}\left(\left|\boldsymbol{\phi}_{ii}^{*2}\mathbf{K}+\mathbf{I}_T\right|\left|\mathbf{R}\mathbf{R}^\top\right||\boldsymbol{\Lambda}|\left|\mathbf{D}^\top\mathbf{D}\right|\right) \\
&= \prod_{i=1}^{N}\left(\prod_{t=1}^{T}\left(\boldsymbol{\phi}_{ii}^{*2}\kappa_t+1\right)|\mathbf{I}_T|\prod_{t=1}^{T}\lambda_t\,|\mathbf{I}_T|\right) \\
&= \prod_{i=1}^{N}\left(\prod_{t=1}^{T}\left(\boldsymbol{\phi}_{ii}^{*2}\kappa_t+1\right)\prod_{t=1}^{T}\lambda_t\right),
\end{aligned} \tag{6.26}$$

where κ_t is the t-th diagonal entry of $\mathbf{K}$, and λ_t is defined analogously with respect to $\boldsymbol{\Lambda}$. Now, the log determinant of $\mathbf{V}$ is given by:

$$\log|\mathbf{V}| = N\sum_{t=1}^{T}\log(\lambda_t)+\sum_{i=1}^{N}\sum_{t=1}^{T}\log\left(\boldsymbol{\phi}_{ii}^{*2}\kappa_t+1\right). \tag{6.27}$$

The third part of the log-likelihood in (6.15) can be made computationally easier by plugging in the inverse of $\mathbf{V}_i$:

$$\begin{aligned}
&\sum_{i=1}^{N}\widetilde{\mathbf{y}}_i^\top\left(\boldsymbol{\phi}_{ii}^{*2}\mathbf{V}_G+\mathbf{V}_E\right)^{-1}\widetilde{\mathbf{y}}_i \\
&\qquad = \sum_{i=1}^{N}\widetilde{\mathbf{y}}_i^\top\mathbf{D}\boldsymbol{\Lambda}^{-1/2}\mathbf{R}\left(\boldsymbol{\phi}_{ii}^{*2}\mathbf{K}+\mathbf{I}_T\right)^{-1}\mathbf{R}^\top\boldsymbol{\Lambda}^{-1/2}\mathbf{D}^\top\widetilde{\mathbf{y}}_i \\
&\qquad = \sum_{i=1}^{N}\widetilde{\mathbf{y}}_i^\top\mathbf{F}^\top\left(\boldsymbol{\phi}_{ii}^{*2}\mathbf{K}+\mathbf{I}_T\right)^{-1}\mathbf{F}\widetilde{\mathbf{y}}_i
\end{aligned} \tag{6.28}$$

with $\mathbf{F}=\mathbf{R}^\top\boldsymbol{\Lambda}^{-1/2}\mathbf{D}^\top$. Now, by combining the terms, we have a computationally efficient expression for the log-likelihood in (6.15):

$$\log l(\boldsymbol{\theta}) = -\frac{1}{2}N\sum_{t=1}^{T}\log(\lambda_t) + \sum_{i=1}^{N}\sum_{t=1}^{T}\log\left(\boldsymbol{\phi}_{ii}^{*2}\kappa_t + 1\right) + \sum_{i=1}^{N}\widetilde{\mathbf{y}}_i^{\top}\mathbf{D}\boldsymbol{\Lambda}^{-1/2}\mathbf{R}\left(\boldsymbol{\phi}_{ii}^{*2}\mathbf{K}+\mathbf{I}_T\right)^{-1}\mathbf{R}^{\top}\boldsymbol{\Lambda}^{-1/2}\mathbf{D}^{\top}\widetilde{\mathbf{y}}_i. \tag{6.29}$$

Note that $\left(\boldsymbol{\phi}_{ii}^{*2}\mathbf{K}+\mathbf{I}_T\right)^{-1}$ is a diagonal matrix with easily computed diagonal elements $(\boldsymbol{\phi}_{ii}^{*2}\kappa_t + 1)^{-1}$.

Based on the above, to evaluate $\log l(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = \boldsymbol{\theta}_k + \alpha\mathbf{p}_k$ for a given value of α, the following steps lead to a computationally efficient evaluation.

1. Compute the eigen decomposition of $\mathbf{V}_E = \mathbf{D}\boldsymbol{\Lambda}\mathbf{D}^{\top}$.

2. Compute $\widetilde{\mathbf{V}}_G = \boldsymbol{\Lambda}^{-1/2}\mathbf{D}^{\top}\mathbf{V}_G\mathbf{D}\boldsymbol{\Lambda}^{-1/2}$ by (6.23).

3. Compute the eigen decomposition $\widetilde{\mathbf{V}}_G = \mathbf{R}\mathbf{K}\mathbf{R}^{\top}$.

4. Compute $\mathbf{F} = \mathbf{R}^{\top}\boldsymbol{\Lambda}^{1/2}\mathbf{D}^{\top}$.

5. Compute the loglikelihood as:

$$\log l(\boldsymbol{\theta}) = -\frac{1}{2}N\sum_{t=1}^{T}\log(\lambda_t) + \sum_{i=1}^{N}\sum_{t=1}^{T}\log\left(\boldsymbol{\phi}_{ii}^{*2}\kappa_t + 1\right) + \sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{\phi}_{ii}^{*2}\kappa_t + 1)^{-1}\left(\sum_{s=1}^{T} f_{ts}\widetilde{y}_{is}\right)^2. \tag{6.30}$$

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, $\nabla\log l(\boldsymbol{\theta}_k)$. For ease of notation we will drop the subscript k in this section. Then, the gradient $\nabla\log l(\boldsymbol{\theta}_k)$ has elements $\partial\log l(\boldsymbol{\theta}_k)/\partial\boldsymbol{\theta}_{\ell}$. For a given parameter $\boldsymbol{\theta}_{\ell}$ in set of parameters $\boldsymbol{\theta}$ (using index ℓ without loss of generality), the gradient of the log-likelihood is given by:

$$\begin{aligned}
\frac{\partial \log l(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_\ell} &= \tfrac{1}{2}\widetilde{\mathbf{y}}^\top \mathbf{V}^{-1} \tfrac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_\ell} \mathbf{V}^{-1} \widetilde{\mathbf{y}} - \tfrac{1}{2} \mathrm{tr}\left(\mathbf{V}^{-1} \tfrac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_\ell}\right) \\
&= \tfrac{1}{2} \sum_{i=1}^{N} \mathbf{r}_i^\top \tfrac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_\ell} \mathbf{r}_i - \tfrac{1}{2} \sum_{i=1}^{N} \mathrm{tr}\left(\mathbf{V}_i^{-1} \tfrac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_\ell}\right),
\end{aligned}$$

where $\mathbf{r}_i = \mathbf{V}_i^{-1}\widetilde{\mathbf{y}}_i = \mathbf{F}^\top \left(\boldsymbol{\phi}_{ii}^{*^2}\mathbf{K} + \mathbf{I}_T\right)^{-1} \mathbf{F}\widetilde{\mathbf{y}}_i$. Let us now consider an expression for $\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_1}$. In case $\boldsymbol{\theta}_\ell$ pertains to an element in $\boldsymbol{\Gamma}_G$, then:

$$\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_\ell} = \boldsymbol{\phi}_{ii}^{*^2} \left(\left(\boldsymbol{\Gamma}_G \frac{\partial \boldsymbol{\Gamma}_G^\top}{\partial \boldsymbol{\theta}_\ell} \right)^\top + \left(\boldsymbol{\Gamma}_G \frac{\partial \boldsymbol{\Gamma}_G^\top}{\partial \boldsymbol{\theta}_\ell} \right) \right) = \boldsymbol{\phi}_{ii}^{*^2} \mathbf{V}_\ell. \tag{6.31}$$

Note that this derivative only differs across observations in scale (i.e., scaled by $\boldsymbol{\phi}_{ii}^{*^2}$ for $i = 1,\ldots,N$). Also, $\partial \boldsymbol{\Gamma}_G^\top / \partial \boldsymbol{\theta}_\ell$ 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 $\boldsymbol{\theta}_\ell$ in $\boldsymbol{\Gamma}_G$. Suppose that $\boldsymbol{\theta}_\ell$ refers to element j, j' of $\boldsymbol{\Gamma}_G$ with $j \geq j'$ as $\boldsymbol{\Gamma}_G$ is a lower triangular matrix. Then, $\left(\boldsymbol{\Gamma}_G \frac{\partial \boldsymbol{\Gamma}_G^\top}{\partial \boldsymbol{\theta}_\ell}\right)$ results in a $T \times T$ matrix of zeros except in column j which is a copy of column j' of $\boldsymbol{\Gamma}_G^\top$. This allows us to conveniently express:

$$\frac{1}{2}\widetilde{\mathbf{y}}^\top \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_\ell} \mathbf{V}^{-1} \widetilde{\mathbf{y}} = \sum_{i=1}^{N} \boldsymbol{\phi}_{ii}^{*^2} \widetilde{\mathbf{y}}_i^\top \mathbf{F}^\top \left(\boldsymbol{\phi}_{ii}^{*^2}\mathbf{K} + \mathbf{I}_T\right)^{-1} \mathbf{F}\mathbf{V}_\ell \mathbf{F}^\top \left(\boldsymbol{\phi}_{ii}^{*^2}\mathbf{K} + \mathbf{I}_T\right)^{-1} \mathbf{F}\widetilde{\mathbf{y}}_i,$$

and

$$-\frac{1}{2}\sum_{i=1}^{N} \mathrm{tr}\left(\mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_\ell}\right) = -\tfrac{1}{2} \sum_{i=1}^{N} \boldsymbol{\phi}_{ii}^{*^2} \mathrm{tr}\left(\left(\boldsymbol{\phi}_{ii}^{*^2}\mathbf{K} + \mathbf{I}_T\right)^{-1} \mathbf{F}\mathbf{V}_\ell \mathbf{F}^\top\right).$$

If we define $\mathbf{c}_{i,\boldsymbol{\theta}_\ell} = (\partial \mathbf{V}_i / \partial \boldsymbol{\theta}_\ell)\mathbf{r}_i$, we can reduce the expression of the gradient to:

$$\frac{\partial \log l(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_\ell} = \frac{1}{2}\sum_{i=1}^{N} \mathbf{r}_i^\top \mathbf{c}_{i,\boldsymbol{\theta}_\ell} - \frac{1}{2}\sum_{i=1}^{N} \mathrm{tr}\left(\mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_\ell}\right). \tag{6.32}$$

Similarly, when $\boldsymbol{\theta}_\ell$ pertains to environment components, we have that:

$$\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_\ell} = \left(\boldsymbol{\Gamma}_E \frac{\partial \boldsymbol{\Gamma}_E^\top}{\partial \boldsymbol{\theta}_\ell} \right)^\top + \left(\boldsymbol{\Gamma}_E \frac{\partial \boldsymbol{\Gamma}_E^\top}{\partial \boldsymbol{\theta}_\ell} \right). \tag{6.33}$$

These derivatives of the environment term are equal across observations.

Step 6. Estimating the Fisher Information Matrix The equations in the next paragraph rely on the work by Fisher (1922), who introduced the method of maximum likelihood and the maximum likelihood convergence theorem (Billingsley, 1995, p. 357). Here, we use the overview by (Lynch and Walsh, 1998, p. 853-856).

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:

$$\mathbf{H}_{ij} = \frac{\partial^2 \log l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}. \tag{6.34}$$

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

$$\mathbf{I}(\boldsymbol{\theta}_{\mathrm{ML}}) = -\mathbb{E}[\mathbf{H}(\boldsymbol{\theta}_{\mathrm{ML}})]. \tag{6.35}$$

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

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

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

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) = \mathcal{I}^{-1}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}). \tag{6.37}$$

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 $\boldsymbol{\theta}$, that is, in the Cholesky decompositions of $\mathbf{V}_G$ and $\mathbf{V}_E$. Practically it may be more interesting to consider instead the covariance matrices $\mathbf{V}_G$ and $\mathbf{V}_E$, 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 $g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$ the function in the optimum $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is distributed as:

$$g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) \sim \mathcal{N}\left(g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}), \nabla g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})^{\top} \mathcal{I}^{-1}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) \nabla g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})\right), \tag{6.38}$$

where $\nabla g(\boldsymbol{\theta})$ is the gradient of $g()$ with respect to $\boldsymbol{\theta}$. Below, the functions $g(\boldsymbol{\theta})$ 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 $v_{ij}^{G}(\boldsymbol{\theta})$ be element ij of the genetic covariance matrix $\mathbf{V}_G$. Then:

$$v_{ij}^{G}(\boldsymbol{\theta}) = \boldsymbol{\theta}^{\top} \mathbf{A}_i^{G\top} \mathbf{A}_j^{G} \boldsymbol{\theta}, \tag{6.39}$$

where $\mathbf{A}_i^{G}$ is a matrix with zeros and ones such that $\mathbf{A}_i^{G}\boldsymbol{\theta}$ selects the row i elements of $\boldsymbol{\Gamma}_G$. The gradient of $v_{ij}^{G}(\boldsymbol{\theta})$ is given by:

$$\nabla v_{ij}^{G}(\boldsymbol{\theta}) = \left(\mathbf{A}_i^{G\top}\mathbf{A}_j^{G} + \mathbf{A}_j^{G\top}\mathbf{A}_i^{G}\right)\boldsymbol{\theta}.$$

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

The heritability of trait i can be written as:

$$h_i^2(\boldsymbol{\theta}) = \frac{v_{ii}^{G}(\boldsymbol{\theta})}{v_{ii}^{G}(\boldsymbol{\theta}) + v_{ii}^{E}(\boldsymbol{\theta})}. \tag{6.40}$$

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

$$\nabla h_i^2(\boldsymbol{\theta}) = \frac{2(1-h_i^2(\boldsymbol{\theta}))}{v_{ii}^{G}(\boldsymbol{\theta}) + v_{ii}^{E}(\boldsymbol{\theta})}\mathbf{A}_i^{G\top}\mathbf{A}_i^{G}\boldsymbol{\theta} - \frac{2h_i^2(\boldsymbol{\theta})}{v_{ii}^{G}(\boldsymbol{\theta}) + v_{ii}^{E}(\boldsymbol{\theta})}\mathbf{A}_i^{E\top}\mathbf{A}_i^{E}\boldsymbol{\theta}.$$

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

The genetic correlation matrix $\mathbf{R}_G$ has elements

$$r_{ij}(\boldsymbol{\theta}) = \frac{v_{ij}(\boldsymbol{\theta})}{v_{ii}^{1/2}(\boldsymbol{\theta}) v_{jj}^{1/2}(\boldsymbol{\theta})}, \tag{6.41}$$

where we have dropped the superscript G for notational simplicity. Then, its gradient becomes

$$\nabla r_{ij}(\boldsymbol{\theta}) = \frac{1}{v_{ii}^{1/2}(\boldsymbol{\theta})v_{jj}^{1/2}(\boldsymbol{\theta})}\left(\mathbf{A}_i{}^\top\mathbf{A}_j + \mathbf{A}_j{}^\top\mathbf{A}_i\right)\boldsymbol{\theta} - \frac{r_{ij}(\boldsymbol{\theta})}{v_{ii}^{1/2}(\boldsymbol{\theta})v_{jj}^{1/2}(\boldsymbol{\theta})}\left(\frac{v_{jj}^{1/2}(\boldsymbol{\theta})}{v_{ii}^{1/2}(\boldsymbol{\theta})}\mathbf{A}_i^\top\mathbf{A}_i\boldsymbol{\theta} + \frac{v_{ii}^{1/2}(\boldsymbol{\theta})}{v_{jj}^{1/2}(\boldsymbol{\theta})}\mathbf{A}_j^\top\mathbf{A}_j\boldsymbol{\theta}\right).$$

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

Trait	Description	Masurement units	UK Biobank data field
BMI	Logarithm average body mass index for all measurements	Kg/m^2	21001
Depression Score	Logarithm of first PC of depression intensity and frequency	NA	2050, 2060, 4609, 4620, 5375, 5386, 2090, 2100
Drinks consumed	Logarithm drinks per week	Number of units of alcohol per week	1558, 1568, 1578, 1588, 1598, 1608, 4407, 4418, 4429, 4440, 4451, 4462, 5364
Educational attainment	Highest self-reported schooling degree converted to US-schooling year equivalents using ISCED categories	years	6138
Grey matter in Amygdala	Volume of grey matter in Amygdala (left+right)	mm^3	25888, 25889
Grey matter in Angular Gyrus	Volume of grey matter in Angular Gyrus	mm^3	25822, 25823
Grey matter in Brain-Stem	Volume of grey matter in Brain-Stem	mm^3	25892
Grey matter in Caudate	Volume of grey matter in Caudate (left+right)	mm^3	25880, 25881
Grey matter in Central Opercular Cortex	Volume of grey matter in Central Opercular Cortex (left+right)	mm^3	25864, 25865

Trait	Description	Measurement units	UK Biobank data field
Grey matter in Cingulate Gyrus, (ad)	Volume of grey matter in Cingulate Gyrus, anterior division (left+right)	mm^3	25838, 25839
Grey matter in Cingulate Gyrus, (pd)	Volume of grey matter in Cingulate Gyrus, posterior division (left+right)	mm^3	25840, 25841
Grey matter in Crus I Cerebellum	Volume of grey matter in Crus I Cerebellum (left+right)	mm^3	25900, 25902
Grey matter in Crus I Cerebellum)	Volume of grey matter in Crus I Cerebellum (vermis)	mm^3	25901
Grey matter in Crus II Cerebellum	Volume of grey matter in Crus II Cerebellum (vermis)	mm^3	25904
Grey matter in Crus II Cerebellum	Volume of grey matter in Crus II Cerebellum (left+right)	mm^3	25903, 25905
Grey matter in Cuneal Cortex	Volume of grey matter in Cuneal Cortex (left+right)	mm^3	25844, 25845
Grey matter in Frontal Medial Cortex	Volume of grey matter in Frontal Medial Cortex (left+right)	mm^3	25830, 25831
Grey matter in Frontal Operculum Cortex	Volume of grey matter in Frontal Operculum Cortex (left+right)	mm^3	25862, 25863
Grey matter in Frontal Orbital Cortex	Volume of grey matter in Frontal Orbital Cortex (left+right)	mm^3	25846, 25847
Grey matter in Frontal Pole	Volume of grey matter in Frontal Pole (left+right)	mm^3	25782, 25783
Grey matter in Heschl's Gyrus	Volume of grey matter in Heschl's Gyrus (includes H1 and H2) (left+right)	mm^3	25870, 25871
Grey matter in Hippocampus	Volume of grey matter in Hippocampus (left+right)	mm^3	25886, 25887
Grey matter in I-IV Cerebellum	Volume of grey matter in I-IV Cerebellum (left+right)	mm^3	25893, 25894
Grey matter in Inferior Frontal Gyrus, po	Volume of grey matter in Inferior Frontal Gyrus, pars opercularis (left+right)	mm^3	25792, 25793
Grey matter in Inferior Frontal Gyrus, pt	Volume of grey matter in Inferior Frontal Gyrus, pars triangularis (left+right)	mm^3	25790, 25790

Trait	Description	Measurement units	UK Biobank data field
Grey matter in Inferior Temporal Gyrus, (tp)	Volume of grey matter in Inferior Temporal Gyrus, temporooccipital part (left+right)	mm^3	25812, 25813
Grey matter in Inferior Temporal Gyrus, (ad)	Volume of grey matter in Inferior Temporal Gyrus, anterior division (left+right)	mm^3	25808, 25808
Grey matter in Inferior Temporal Gyrus, (pd)	Volume of grey matter in Inferior Temporal Gyrus, posterior division (left+right)	mm^3	25810, 25811
Grey matter in Insular Cortex	Volume of grey matter in Insular Cortex (left+right)	mm^3	25784, 25785
Grey matter in Intracalcarine Cortex	Volume of grey matter in Intracalcarine Cortex (left+right)	mm^3	25828, 25829
Grey matter in IX Cerebellum	Volume of grey matter in IX Cerebellum (left+right)	mm^3	25915, 25917
Grey matter in Juxtapositional Lobule Cortex	Volume of grey matter in Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex) (left+right)	mm^3	25832, 25833
Grey matter in Lateral Occipital Cortex, (id)	Volume of grey matter in Lateral Occipital Cortex, inferior division (left+right)	mm^3	25826, 25827
Grey matter in Lateral Occipital Cortex, (sd)	Volume of grey matter in Lateral Occipital Cortex, superior division (left+right)	mm^3	25824, 25825
Grey matter in Lingual Gyrus	Volume of grey matter in Lingual Gyrus (left+right)	mm^3	25852, 25853
Grey matter in Middle Frontal Gyrus	Volume of grey matter in Middle Frontal Gyrus (left+right)	mm^3	25788, 25789
Grey matter in Middle Temporal Gyrus, (tp)	Volume of grey matter in Middle Temporal Gyrus, temporooccipital part (left+right)	mm^3	25806, 25807
Grey matter in Middle Temporal Gyrus, (ad)	Volume of grey matter in Middle Temporal Gyrus, anterior division (left+right)	mm^3	25802, 25803

Trait	Description	Measurement units	UK Biobank data field
Grey matter in Middle Temporal Gyrus, (pd)	Volume of grey matter in Middle Temporal Gyrus, posterior division (left+right)	mm^3	25804, 25805
Grey matter in Occipital Fusiform Gyrus	Volume of grey matter in Occipital Fusiform Gyrus (left+right)	mm^3	25860, 25861
Grey matter in Occipital Pole	Volume of grey matter in Occipital Pole (left+right)	mm^3	25876, 25877
Grey matter in Pallidum	Volume of grey matter in Pallidum (left+right)	mm^3	25884, 25884
Grey matter in Paracingulate Gyrus	Volume of grey matter in Paracingulate Gyrus (left+right)	mm^3	25836, 25837
Grey matter in Parahippocampal Gyrus, (ad)	Volume of grey matter in Parahippocampal Gyrus, anterior division (left+right)	mm^3	25848, 25849
Grey matter in Parahippocampal Gyrus, (pd)	Volume of grey matter in Parahippocampal Gyrus, posterior division (left+right)	mm^3	25850, 25851
Grey matter in Parietal Operculum Cortex	Volume of grey matter in Parietal Operculum Cortex (left+right)	mm^3	25866, 25867
Grey matter in Planum Polare	Volume of grey matter in Planum Polare (left+right)	mm^3	25868, 25869
Grey matter in Planum Temporale	Volume of grey matter in Planum Temporale (left+right)	mm^3	25872, 25783
Grey matter in Postcentral Gyrus	Volume of grey matter in Postcentral Gyrus (left+right)	mm^3	25814, 25815
Grey matter in Precentral Gyrus	Volume of grey matter in Precentral Gyrus (left+right)	mm^3	25794, 25795
Grey matter in Precuneous Cortex	Volume of grey matter in Precuneous Cortex (left+right)	mm^3	25842, 25843
Grey matter in Putamen	Volume of grey matter in Putamen (left+right)	mm^3	25882, 25883
Grey matter in Subcallosal Cortex	Volume of grey matter in Subcallosal Cortex (left+right)	mm^3	25834, 25835
Grey matter in Superior Frontal Gyrus	Volume of grey matter in Superior Frontal Gyrus (left)	mm^3	25786

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

Trait	Description	Measurement units	UK Biobank data field
Grey matter in VI Cerebellum)	Volume of grey matter in VI Cerebellum (vermis)	mm^3	25898
Grey matter in VIIb Cerebellum	Volume of grey matter in VIIb Cerebellum (left+right)	mm^3	25906, 25908
Grey matter in VIIb Cerebellum)	Volume of grey matter in VIIb Cerebellum (vermis)	mm^3	25907
Grey matter in VIIIa Cerebellum	Volume of grey matter in VIIIa Cerebellum (left+right)	mm^3	25909, 25911
Grey matter in VIIIa Cerebellum)	Volume of grey matter in VIIIa Cerebellum (vermis)	mm^3	25910
Grey matter in VIIIb Cerebellum	Volume of grey matter in VIIIb Cerebellum (left+right)	mm^3	25912, 25914
Grey matter in VIIIb Cerebellum)	Volume of grey matter in VIIIb Cerebellum (vermis)	mm^3	25913
Grey matter in X Cerebellum	Volume of grey matter in X Cerebellum (left+right)	mm^3	25918, 25920
Grey matter in X Cerebellum)	Volume of grey matter in X Cerebellum (vermis)	mm^3	25919
IQ	Standardized fluid intelligence score	correct-answers	20016, 20191
Neuroticism	Neuroticism standardized score	NA	1920, 1930, 1940, 1950, 1960, 1970, 1980, 1990, 2000, 2010, 2020, 2030
Reaction Time	Standardized reaction time	milliseconds	20023
Standing height	Standing height	cm	50
Subjective Well-being	Subjective well-being: In general how happy are you? (Average value over time)	NA	4526, 20458
Visual memory	Log standardized visual memory score	NA	399, 20132
Volume of brain	Volume of brain, grey+white matter	mm^3	25010

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 criteria.

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

Disease	UK Biobank data field	ICD10 code
Dementia or Alzheimer's disease	1263	F01, F02, G30
Parkinson's disease	1262	G20, G21
Chronic degenerative neurological	1258	G23, G31, G32
Guillain-Barré syndrome	1256	G610
Multiple Sclerosis	1261	G35
Other demyelinating disease	1397	G37
Stroke or ischaemic stroke	1081	G463, G464, I64, I694
Brain cancer	1031	C70, C71, D33
Brain haemorrhage	1491	I60, I61, I62, I691, I692, I693
Brain/intracranial abscess	1245	G060, G07
Cerebral aneurysm	1425	I671, Q282, Q283
Cerebral palsy	1433	G80, A521, A504, I64
Encephalitis	1246	A83, A86, B011, B020, B262, A85, B004, B582, A84, B050, B941, G04, A321, G05
Epilepsy	1264	G40, F803
Head injury	1266	S07, T040
Infections of the nervous system	1244	A80, A81, A82, A83, A84, A85, A86, A87, A88, A89
Ischaemic stroke	1583	G45
Meningeal cancer	1031	C70, C793
Meningioma (benign)	1659	D33, D32
Meningitis	1247	G03, A170, A171, A203, G01, G02, G00, G07
Motor Neuron Disease (ALS)	1259	G122
Neurological injury / trauma	1240	
Spina bifida	1524	Q05, Q760
Subdural haematoma	1083	P100
Subarachnoid haemorrhage	1086	I60, S066, P103
Transient ischaemic attack	1082	G45

6.D 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, age^2, age^3, sex x age, sex x age^2, sex x age^3, 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 $\mathbf{A}^* = \mathbf{MAM}$, where $\mathbf{M}$ is anti-projection matrix based on platform dummy and intercept, and recompute eigenvalue decomposition: $\mathbf{A}^* = \mathbf{[P1,P2]}diag(D1,D2)\mathbf{[P1,P2]}^\top$. 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 6.3 – *The estimated SNP-heritability for the different phenotypes in UK Biobank using M-GREML.*

Trait	h^2	Standard error
Height	0.561	0.046
log(BMI)	0.257	0.049
Years of Education	0.209	0.049
Visual memory	0.147	0.048
Reaction time	0.128	0.050
Neuroticism	0.159	0.049
Subjective well-being	0.085	0.049
Depressive symptoms	0.117	0.050
Log(drinks per week)	0.160	0.051
Absolute volume of grey matter	0.361	0.049
Absolute volume of grey and white matter	0.412	0.047
Relative volume of grey matter in Frontal Pole	0.255	0.048
Relative volume of grey matter in Insular Cortex	0.460	0.050
Relative volume of grey matter in Middle Frontal Gyrus	0.239	0.048
Relative volume of grey matter in Inferior Frontal Gyrus. pars triangularis	0.182	0.050
Relative volume of grey matter in Inferior Frontal Gyrus. pars opercularis	0.138	0.049
Relative volume of grey matter in Precentral Gyrus	0.229	0.049
Relative volume of grey matter in Temporal Pole	0.315	0.048
Relative volume of grey matter in Superior Temporal Gyrus. anterior division	0.248	0.047
Relative volume of grey matter in Superior Temporal Gyrus. posterior division	0.244	0.049
Relative volume of grey matter in Middle Temporal Gyrus. anterior division	0.205	0.049
Relative volume of grey matter in Middle Temporal Gyrus. posterior division	0.235	0.050
Relative volume of grey matter in Middle Temporal Gyrus. temporooccipital part	0.252	0.047
Relative volume of grey matter in Inferior Temporal Gyrus. anterior division	0.235	0.049
Relative volume of grey matter in Inferior Temporal Gyrus. posterior division	0.232	0.049
Relative volume of grey matter in Inferior Temporal Gyrus. temporooccipital part	0.224	0.048
Relative volume of grey matter in Postcentral Gyrus	0.245	0.048
Relative volume of grey matter in Superior Parietal Lobule	0.165	0.048

Trait	h^2	Standard error
Relative volume of grey matter in Supramarginal Gyrus. anterior division	0.170	0.049
Relative volume of grey matter in Supramarginal Gyrus. posterior division	0.158	0.048
Relative volume of grey matter in Angular Gyrus	0.178	0.047
Relative volume of grey matter in Lateral Occipital Cortex. superior division	0.167	0.048
Relative volume of grey matter in Lateral Occipital Cortex. inferior division	0.188	0.049
Relative volume of grey matter in Intracalcarine Cortex	0.446	0.049
Relative volume of grey matter in Frontal Medial Cortex	0.178	0.050
Relative volume of grey matter in Juxtapositional Lobule Cortex (formerly Supplementary M	0.175	0.047
Relative volume of grey matter in Subcallosal Cortex	0.391	0.048
Relative volume of grey matter in Paracingulate Gyrus	0.264	0.048
Relative volume of grey matter in Cingulate Gyrus. anterior division	0.216	0.046
Relative volume of grey matter in Cingulate Gyrus. posterior division	0.208	0.049
Relative volume of grey matter in Precuneous Cortex	0.287	0.049
Relative volume of grey matter in Cuneal Cortex	0.238	0.048
Relative volume of grey matter in Frontal Orbital Cortex	0.299	0.048
Relative volume of grey matter in Parahippocampal Gyrus. anterior division	0.300	0.048
Relative volume of grey matter in Parahippocampal Gyrus. posterior division	0.334	0.047
Relative volume of grey matter in Lingual Gyrus	0.275	0.048
Relative volume of grey matter in Temporal Fusiform Cortex. anterior division	0.262	0.050
Relative volume of grey matter in Temporal Fusiform Cortex. posterior division	0.314	0.048
Relative volume of grey matter in Temporal Occipital Fusiform Cortex	0.245	0.049
Relative volume of grey matter in Occipital Fusiform Gyrus	0.237	0.047
Relative volume of grey matter in Frontal Operculum Cortex	0.236	0.049
Relative volume of grey matter in Central Opercular Cortex	0.296	0.048
Relative volume of grey matter in Parietal Operculum Cortex	0.300	0.049
Relative volume of grey matter in Planum Polare	0.297	0.047
Relative volume of grey matter in Heschls Gyrus (includes H1 and H2)	0.353	0.048
Relative volume of grey matter in Planum Temporale	0.308	0.047
Relative volume of grey matter in Supracalcarine Cortex	0.277	0.047
Relative volume of grey matter in Occipital Pole	0.347	0.047
Relative volume of grey matter in Thalamus	0.346	0.047
Relative volume of grey matter in Caudate	0.239	0.047
Relative volume of grey matter in Putamen	0.355	0.047
Relative volume of grey matter in Pallidum	0.162	0.047
Relative volume of grey matter in Hippocampus	0.436	0.047

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

Bibliography

Acs, Z. and Lappi, E. *Entrepreneurship, culture, and the epigenetic revolution: a research note*. Small Business Economics, 2019.

Adda, J. and Cornaglia, F. Taxes, cigarette consumption, and smoking intensity. *American Economic Review*, 96(4):1013–1028, 2006.

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J.L., Bastiani, M., Miller, K.L., Nichols, T.E., and Smith, S.M. Confound modelling in uk biobank brain imaging. *NeuroImage, forthcoming*, 2020.

Allegrini, A., Selzam, S., Rimfeld, K., von, S.S., Pingault, J., and Plomin, R. Genomic prediction of cognitive traits in childhood and adolescence. *Molecular Psychiatry*, 24:819–827, 2019.

Anderson, L.R. and Mellor, J.M. Predicting health behaviors with an experimental measure of risk preference. *Journal of Health Economics*, 27(5):1260–1274, 2008.

Antshel, K.M. Attention deficit / hyperactivity disorder (ADHD) and entrepreneurship. *Academy of Management Perspectives*, 32(2):243–265, 2017.

Bansal, V., Mitjans, M., Burik, C.A., Linner, R.K., Okbay, A., Rietveld, C.A., Begemann, M., Bonn, S., Ripke, S., de Vlaming, R., et al. Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia. *Nature Communications*, 9(1):1–12, 2018.

Barsky, R.B., Juster, F.T., Kimball, M.S., and Shapiro, M.D. Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *The Quarterly Journal of Economics*, 112(2):537–579, 1997.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., et al. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, 2018.

Benjamin, D.J., Cesarini, D., Chabris, C.F., Glaeser, E.L., Laibson, D.I., Age, G., Guðnason, V., Harris, T.B., Launer, L.J., Purcell, S., et al. The promises and pitfalls of genoeconomics. *Annual Review of Economics*, 4(1), 2012a.

Benjamin, D.J., Cesarini, D., Van Der Loos, M.J., Dawes, C.T., Koellinger, P.D., Magnusson, P.K., Chabris, C.F., Conley, D., Laibson, D., Johannesson, M., et al. The genetic architecture of economic and political preferences. *Proceedings of the National Academy of Sciences of the United States of America*, 109(21):8026–8031, 2012b.

Benjamin, D.J., Chabris, C.F., Glaeser, E.L., Gudnason, V., Harris, T.B., Laibson, D.I., Launer, L.J., and Purcell, S. Genoeconomics. In: M. Weinstein, J. Vaupel, and K.W. Watcher, editors, *Biosocial surveys*, pp. 304–335. National Academies Press (US), Washington, DC, 2008.

Benowitz, N.L. Neurobiology of nicotine addiction: implications for smoking cessation treatment. *The American Journal of Medicine*, 121(4):S3–S10, 2008.

Benz, M. and Frey, B.S. The value of doing what you like: Evidence from the self-employed in 23 countries. *Journal of Economic Behavior & Organization*, 38(3-4):445–455, 2008.

Berkson, J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

Berzuini, C., Guo, H., Burgess, S., and Bernardinelli, L. A bayesian approach to mendelian randomization with multiple pleiotropic variants. *Biostatistics*, 21(1):86–101, 2020.

Bidwell, L., Palmer, R., Brick, L., McGeary, J., and Knopik, V. Genome-wide single nucleotide polymorphism heritability of nicotine dependence as a multidimensional phenotype. *Psychological Medicine*, 46(10):2059–2069, 2016.

Billingsley, P. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1995.

Boardman, J.D. State-level moderation of genetic tendencies to smoke. *American Journal of Public Health*, 99(3):480–486, 2009.

Böckerman, P., Viinikainen, J., Pulkki-Råback, L., Hakulinen, C., Pitkänen, N., Lehtimäki, T., Pehkonen, J., and Raitakari, O.T. Does higher education protect against obesity? evidence using mendelian randomization. *Preventive Medicine*, 101:195–198, 2017.

Bowden, J., Davey Smith, G., Haycock, P.C., and Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016.

Bowden, J., Del Greco, M., Minelli, C., Davey Smith, G., Sheehan, N., Thompson, J., et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine*, 36(11):1783–1802, 2017a.

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine*, 36(11):1783–1802, 2017b.

Bowden, J., Smith, G.D., and Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Internation Journal of Epidemiology*, 44(2):512–525, 2015.

Browning, S.R. and Browning, B.L. Population structure can inflate snp-based heritability estimates. *American Journal of Human Genetics*, 89(1):191–193, 2011.

Bulik-Sullivan, B.K., , Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, 2015.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 2019.

Burgess, S., Bowden, J., Dudbridge, F., and Thompson, S.G. Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. *arXiv:1606.03729*, 2016a.

Burgess, S., Butterworth, A.S., and Thompson, S.G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665, 2013.

Burgess, S., Dudbridge, F., and Thompson, S.G. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in Medicine*, 35(11):1880–1906, 2016b.

Burgess, S., Foley, C.N., Allara, E., Staley, J.R., and Howson, J.M. A robust and efficient method for mendelian randomization with hundreds of genetic variants: unravelling mechanisms linking hdl-cholesterol and coronary heart disease. *bioRxiv*, 2019.

Burgess, S. and Thompson, S.G. Interpreting findings from mendelian randomization using the mr-egger method. *European Journal of Epidemiology*, 32(5):377–389, 2017.

Burgess, S., Timpson, N.J., Ebrahim, S., and Smith, G.D. Mendelian randomization: where are we now and where are we going? *International Journal of Epidemiology*, 44(2):379–388, 2015.

Burgess, S., Zuber, V., Gkatzionis, A., and Foley, C.N. Modal-based estimation via heterogeneity-penalized weighting: model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid. *International Journal of Epidemiology*, 47(4):1242–1254, 2018.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., and Munafò, M.R. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Casella, G. and Searle, S.R. On a matrix identity useful in variance component estimation. Biometrics Unit Technical Reports; Number BU-875-M, 1985.

Chabris, C.F., Lee, J.J., Cesarini, D., Benjamin, D.J., and Laibson, D.I. The fourth law of behavior genetics. *Current Directions in Psychological Science*, 24(4), 2015.

Chaloupka, F. and Warner, K.E. The economics of smoking. In: A.J. Culyer and J.P. Newhouse, editors, *Handbook of Health Economics*, volume 1, chapter 29, pp. 1539–1627. Elsevier, 2000.

Chiou, L. and Muehlegger, E. Crossing the line: direct estimation of cross-border cigarette sales and the effect on tax revenue. *The BE Journal of Economic Analysis & Policy*, 8(1):1–41, 2008.

Clark, A. and Etilé, F. Do health changes affect smoking? evidence from british panel data. *Journal of Health Economics*, 21(4):533–562, 2002.

Davey Smith, G. and Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003.

Davies, N.M., Howe, L.J., Brumpton, B., Havdahl, A., Evans, D.M., and Davey Smith, G. Within family Mendelian randomization studies. *Human Molecular Genetics*, 28(R2):R170–R179, 2019.

Demmel, J.W. *Applied numerical linear algebra*. Siam, 1997.

Didelez, V. and Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330, 2007.

DiPrete, T.A., Burik, C.A.P., and Koellinger, P.D. Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. *Proceedings of the National Academy of Sciences*, 115(22):4970–4979, 2018a.

DiPrete, T.A., Burik, C.A.P., and Koellinger, P.D. Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. *Proceedings of the National Academy of Sciences*, 115(22):E4970–E4979, 2018b.

Dixon, P., Smith, G.D., von Hinke, S., Davies, N.M., and Hollingworth, W. Estimating marginal healthcare costs using genetic variants as instrumental variables: Mendelian randomization in economic evaluation. *Pharmacoeconomics*, 34(11):1075–1086, 2016.

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

Ebstein, R.P., Israel, S., Chew, S.H., Zhong, S., and Knafo, A. Genetics of human social behavior. *Neuron*, 65(6), 2010.

Ellis, P.D. *The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results*. United Kingdom: Cambridge University Press, 2010.

Erzurumluoglu, A.M., Liu, M., Jackson, V.E., Barnes, D.R., Datta, G., Melbourne, C.A., Young, R., Batini, C., Surendran, P., Jiang, T., et al. Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Molecular Psychiatry*, pp. 1–18, 2019.

Faraone, S.V. and Larsson, H. Genetics of attention deficit hyperactivity disorder. *Molecular Psychiatry*, 24:562–575, 2019.

Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

Fletcher, J.M. Why have tobacco control policies stalled? using genetic moderation to examine policy impacts. *PLOS ONE*, 7(12):e50576, 2012.

Fletcher, R. *Practical Methods of Optimization*. Wiley, New York, 2nd edition, 1991.

Fontana, M.A. *Mitigating Risk: Smartphone Notifications, Adaptive Surveying, and Genetics.* Ph.D. thesis, University of Michigan, 2015.

Gelernter, J., Kranzler, H.R., Sherva, R., Almasy, L., Herman, A.I., Koesterer, R., Zhao, H., and Farrer, L.A. Genome-wide association study of nicotine dependence in american populations: identification of novel risk loci in both african-americans and european-americans. *Biological Psychiatry*, 77(5):493–503, 2015.

Glymour, M.M., Tchetgen, E.J.T., and Robins, J.M. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339, 2012.

Goldberger, A.S. Heritability. *Economica*, 46(184):327–347, 1979.

Gorber, S.C., Schofield-Hurwitz, S., Hardt, J., Levasseur, G., and Tremblay, M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine & Tobacco Research*, 11(1):12–24, 2009.

Grasby, K.L., Jahanshad, N., Painter, J.N., Colodro-Conde, L., Bralten, J., Hibar, D.P., Lind, P.A., Pizzagalli, F., Ching, C.R., McMahon, M.A.B., et al. The genetic architecture of the human cerebral cortex. *Science*, 367(6484):eaay6690, 2020.

Grotzinger, A., Rhemtulla, M., de, V.R., Ritchie, S., Mallard, T., Hill, W., Ip, H., Marioni, R., McIntosh, A., Deary, I., Koellinger, P., Harden, K., Nivard, M., and Tucker-Drob, E. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour*, 3(5):513–525, 2019.

Guo, Z., Kang, H., Tony Cai, T., and Small, D.S. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815, 2018.

Hartwig, F.P., Davey Smith, G., and Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1998, 2017.

Harville, D.A. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

Haycock, P.C., Burgess, S., Wade, K.H., Bowden, J., Relton, C., and Davey Smith, G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American Journal of Clinical Nutrition*, 103(4):965–978, 2016.

Health and Study, R. *RAND HRS longitudinal file 2014 (V2) documentation [PDF file]*, 2018. https://www.rand.org/content/dam/rand/www/external/labor/aging/dataprod/randhrs1992_2014v2.pdf.

Health and Retirement Study. *Quality control report for genotypic data [PDF file]*, 2012. http://hrsonline.isr.umich.edu/sitedocs/genetics/HRS_QC_REPORT_MAR2012.pdf.

Health and Retirement Study. *HRS polygenic scores 2006-2012 genetic data - Release 3*, 2018. https://hrs.isr.umich.edu/news/hrs-polygenic-scores-2006-2012-genetic-data-release-3.

Hewitt, J. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior Genetics*, 42(1):1–2, 2012.

Hurd, M.D., Meijer, E., Moldoff, M., and Rohwedder, S. Improved wealth measures in the health and retirement study: asset reconciliation and cross-wave imputation. 2016.

Institute of Medicine. *Ending the tobacco problem: a blueprint for the nation*. National Academies Press, 2007.

International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109, 2011.

International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010.

International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

Ioannidis, J.P.A. Why most published research findings are false. *PLOS Medicine*, 2(8):124, 2005.

Jencks, C. Heredity, environment, and public policy reconsidered. *American Sociological Review*, 45(5):723–736, 1980.

Jiang, L., Oualkacha, K., Didelez, V., Ciampi, A., Rosa, P., Benedet, A.L., Mathotaarachchi, S.S., Richards, B., and Greenwood, C.M. Constrained instruments and their application to mendelian randomization with pleiotropy. *bioRxiv 22754. https://doi.org/10.1101/227454*, 2017.

Jones, A.M. Health, addiction, social interaction and the decision to quit smoking. *Journal of Health Economics*, 13(1):93–110, 1994.

Juster, F.T. and Suzman, R. An overview of the health and retirement study. *Journal of Human Resources*, 30:S7–S56, 1995.

Kandel, D.B., Kiros, G.E., Schaffran, C., and Hu, M.C. Racial/ethnic differences in cigarette smoking initiation and progression to daily smoking: a multilevel analysis. *American Journal of Public Health*, 94(1):128–135, 2004.

Kang, H., Zhang, A., Cai, T.T., and Small, D.S. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

Karp, F. *Growing older in America: The health & retirement study*. 7. National Institute on Aging, National Institutes of Health, US Department of Health and Human Services, 2007.

Kaufmann, L. and Rousseeuw, P.J. Finding groups in data: an introduction to cluster analysis. *New York: Jonh Wiley*, 1990.

Keller, M.C. Gene×environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biological Psychiatry*, 75(1):18–24, 2014.

Knopik, V.S., Neiderhiser, J.M., DeFries, J.C., and Plomin, R. *Behavioral Genetics (7th edition)*. New York: Worth Publishers, 2016.

Koellinger, P.D., van der Loos, M.J., Groenen, P.J., Thurik, A.R., Rivadeneira, F., van Rooij, F.J., Uitterlinden, A.G., and Hofman, A. Genome-wide association studies in economics and entrepreneurship research: promises and limitations. *Small Business Economics*, 35(1):1–18, 2010.

Kong, A., Thorleifsson, G., Frigge, M.L., Vilhjalmsson, B.J., Young, A.I., Thorgeirsson, T.E., Benonisdottir, S., Oddsson, A., Halldorsson, B.V., Masson, G., et al. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428, 2018.

Kosfeld, M., Heinrichs, M., Zak, P.J., Fischbacher, U., and Fehr, E. Oxytocin increases trust in humans. *Nature*, 435(7042):673–676, 2005.

Kristof-Brown, A.L., Zimmerman, R.D., and Johnson, E.C. Consequences of individual's fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. *Personnel Psychology*, 58(2):281–342, 2005.

Krugman, P. and Wells, R. *Economics: European Edition*. Macmillan, third edition, 2015.

Lahiri, K. and Song, J.G. The effect of smoking on health using a sequential self-selection model. *Health Economics*, 9(6):491–511, 2000.

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., Vaughan, B., et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics*, 47(7):692–695, 2015.

Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Linnér, R.K., et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50(8):1112–1121, 2018.

Lee, S., Yang, J., Goddard, M., Visscher, P., and Wray, N. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, 2012.

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. *International Journal of Epidemiology*, 44(4):1137–1147, 2015.

Lello, L., Avery, S.G., Tellier, L., Vazquez, A.I., de los Campos, G., and Hsu, S. Accurate genomic prediction of human height. *Genetics*, 210(2):477–497, 2018.

Lerner, D.A., Verheul, I., and Thurik, A.R. Entrepreneurship and attention deficit/hyperactivity disorder: a large-scale study involving the clinical condition of ADHD. *Small Business Economics*, 53(2), 2019.

Lindquist, M.J., Sol, J., Praag, V., and M. Why do entrepreneurial parents have entrepreneurial children? *Journal of Labor Economics*, 33(2), 2015.

Linnér, R.K., Biroli, P., Kong, E., Meddens, S.F.W., Wedow, R., Fontana, M.A., Lebreton, M., Tino, S.P., Abdellaoui, A., Hammerschlag, A.R., et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics*, 51(1):245–257, 2019.

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics*, 51(2):237–244, 2019.

Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

Luo, J. Effects of Ethanol on the Cerebellum: Advances and Prospects. *Cerebellum*, 14:383–385, 2015.

Lynch, M. and Walsh, B. *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA, 1998.

MacLean, J.C., Kessler, A.S., and Kenkel, D.S. Cigarette taxes and older adult smoking: evidence from the health and retirement study. *Health Economics*, 25(4):424–438, 2016.

Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., et al. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10):1181, 2007.

Manski, C.F. Genes, eyeglasses, and social policy. *Journal of Economic Perspectives*, 25(4):83–94, 2011.

Molinari, M., Petrosini, L., Misciagna, S., and Leggio, M. Visuospatial abilities in cerebellar disorders. *Journal of Neurology, Neurosurgery, and Psychiatry*, 75(2):235–40, 2004.

Monnereau, C., Vogelezang, S., Kruithof, C.J., Jaddoe, V.W., and Felix, J.F. Associations of genetic risk scores based on adult adiposity pathways with childhood growth and adiposity measures. *BMC Genetics*, 17(1):120, 2016.

Morris, T.T., Davies, N.M., and Smith, G.D. Can education be personalised using pupils' genetic data? *BioRxiv*, p. 645218, 2019.

Mosteller, F. and Tukey, J.W. *Data analysis and regression: a second course in statistics.* Addison-Wesley Series in Behavioral Science: Quantitative Methods, 1977.

National Human Genome Research Institute. *DNA sequencing costs: Data*, 2018a. https://www.genome.gov/27541954/dna-sequencing-costs-data/.

National Human Genome Research Institute. *Frequently asked questions about genetic and genomic science*, 2018b. https://www.genome.gov/19016904/faq-about-genetic-and-genomic-science/.

Nave, G., Jung, W., Karlsson, L.R., Kable, J., and Koellinger, P. Are Bigger Brains Smarter? Evidence From a Large-Scale Preregistered Study. *Psychological Science*, 30(1):43–54, 2019.

Nesson, E. Heterogeneity in smokers' responses to tobacco control policies. *Health Economics*, 26(2):206–225, 2017.

Ni, G., Moser, G., Wray, N., and Lee, S. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. *American Journal of Human Genetics*, 102(6):1185–1194, 2018.

Nicolaou, N. and Shane, S. Entrepreneurship and occupational choice: Genetic and environmental influences. *Journal of Economic Behavior and Organization*, 76(1):3–14, 2010.

Nicolaou, N., Shane, S., Adi, G., Mangino, M., and Harris, J. A polymorphism associated with entrepreneurship: Evidence from dopamine receptor candidate genes. *Small Business Economics*, 36(2):151–155, 2011.

Nicolaou, N., Shane, S., Cherkas, L., Hunkin, J., and Spector, T.D. Is the tendency to engage in entrepreneurship genetic? *Management Science*, 54(1):167–179, 2008a.

Nicolaou, N., Shane, S., Cherkas, L., and Spector, T.D. The influence of sensation seeking in the heritability of entrepreneurship. *Strategic Entrepreneurship Journal*, 2(1):7–21, 2008b.

Nicolaou, N., Shane, S., Cherkas, L., and Spector, T.D. Opportunity recognition and the tendency to be an entrepreneur: A bivariate genetics perspective. *Organizational Behavior and Human Decision Processes*, 110(2):108–117, 2009.

Nikpay, M., Goel, A., Won, H.H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P., Hopewell, J.C., et al. A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10):1121–1130, 2015.

Nocedal, J. and Wright, S. *Numerical Optimization*. Springer, 2006.

O'Connor, L.J. and Price, A.L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature Genetics*, 50(12):1728–1734, 2018.

Okbay, A., Baselmans, B.M., De Neve, J.E., Turley, P., Nivard, M.G., Fontana, M.A., Meddens, S.F.W., Linnér, R.K., Rietveld, C.A., Derringer, J., et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, 48(6):624–633, 2016a.

Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.B., Emilsson, V., Meddens, S.F.W., et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604):539–542, 2016b.

Okbay, A., Becker, J., Benjamin, D., Burik, C.A.P., Cesarini, and Turley, P. A repository of polygenic scores. *Behavior Genetics*, 49(6):507, 2018.

Orzechowski, W. and Walker, R.C. Tax burden on tobacco: The historical compilation, vol. 51. *Arlington, VA: Orzechowski and Walker*, 2016.

Palmer, T., Thompson, J., Tobin, M., Sheehan, N., and Burton, P. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *International Journal of Epidemiology*, 37(5):1161–1168, 2008.

Pare, G., Mao, S., and Deng, W.Q. A machine-learning heuristic to improve gene score prediction of polygenic traits. *Scientific Reports*, 7(1):1–11, 2017.

Parker, S.C. *The Economics of Entrepreneurship (2nd edition)*. Cambridge: Cambridge University Press, 2018.

Patel, P.C., Rietveld, C.A., and Verheul, I. Attention deficit hyperactivity disorder (ADHD) and earnings in later-life self-employment. *Entrepreneurship Theory and Practice. url = https://doi.org/10.1177/1042258719888641*, 2019.

Pearl, J. et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

Pierce, B.L. and Burgess, S. Efficient design for Mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184, 2013.

Polderman, T.J., Benyamin, B., Leeuw, D., A., C., Sullivan, P.F., Bochoven, V., A., V., M., P., and Posthuma, D. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7):702–709, 2015.

Power, R.A. and Pluess, M. Heritability estimates of the Big Five personality traits based on common genetic variants. *Translational Psychiatry*, 5(7):e604, 2015.

Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D., et al. Long-range LD can confound genome scans in admixed populations. *American Journal of Human Genetics*, 83:132–135, 2008.

Qi, G. and Chatterjee, N. Mendelian randomization analysis using mixture models (mrmix) for genetic effect-size-distribution leads to robust estimation of causal effects. *bioRxiv 367821. https://doi.org/10.1101/367821*, 2018.

Qi, G. and Chatterjee, N. A comprehensive evaluation of methods for mendelian randomization using realistic simulations of genome-wide association studies. *bioRxiv 702787. https://doi.org/10.1101/702787*, 2019.

Quaye, L., Nicolaou, N., Shane, S., and Mangino, M. A discovery genome-wide association study of entrepreneurship. *International Journal of Developmental Science*, 6(3-4):127–135, 2012.

Rakic, P. Evolution of the neocortex: a perspective from developmental biology. *Nature Reviews Neuroscience*, 10(10):724–35, 2009.

Ramsköld, D., Wang, E.T., Burge, C.B., and Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLOS Computational Biology*, 5(12):1000598, 2009.

Rietveld, C.A., Conley, D., Eriksson, N., Esko, T., Medland, S.E., Vinkhuyzen, A.A., Yang, J., Boardman, J.D., Chabris, C.F., Dawes, C.T., et al. Replicability and robustness of genome-wide-association studies for behavioral traits. *Psychological Science*, 25(11):1975–1986, 2014a.

Rietveld, C.A., Esko, T., Davies, G., Pers, T.H., Benyamin, B., Chabris, C.F., and Koellinger, P.D. Common genetic variants associated with cognitive performance identified using proxy-phenotype method. *Proceedings of the National Academy of Sciences*, 111(38):13790–13794, 2014b.

Rietveld, C.A., Hessels, J., and van der Zwan, P. The stature of the self-employed and its relation with earnings and satisfaction. *Economics and Human Biology*, 17(1):59–74, 2015.

Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340(6139):1467–1471, 2013.

Rietveld, C.A., Slob, E.A.W., and Thurik, A.R. A decade of research on the genetics of entrepreneurship: a review and view ahead. *Small Business Economics*, 2020.

Rietveld, C.A. and Webbink, D. On the genetic bias of the quarter of birth instrument. *Economics and Human Biology*, 21(1):137–146, 2016.

Schunkert, H., König, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., Preuss, M., Stewart, A.F., Barbalic, M., Gieger, C., et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, 43(4):333–338, 2011.

Searle, S.R., Casella, G., and McCulloch, C.E. *Variance Components*, chapter M.4, pp. 451–452. John Wiley and Sons, Hoboken, New Jersey, USA, 1992.

Shane, S. *Born entrepreneurs, born leaders: How your genes affect your work life*. Oxford University Press:New York, 2010.

Shane, S. and Nicolaou, N. Creative personality, opportunity recognition and the tendency to start businesses: A study of their genetic predispositions. *Journal of Business Venturing*, 30(3):407–419, 2015.

Shane, S. and Venkatamaran, S. The promise of entrepreneurship as a field of research. *Academy of Management Review*, 25(1):217–226, 2000.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J.G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J.F., and Campbell, H. Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89(5):607–618, 2011.

Slob, E.A.W. and Burgess, S. A comparison of robust mendelian randomization methods using summary data. *Genetic Epidemiology*, 44(4):313–329, 2020.

Slob, E.A.W., de Vlaming, R., Rietveld, C.A., and Groenen, P.J. Highly efficient multivariate greml. In: *Behavior Genetics*, volume 48, pp. 514–514. SPRINGER 233 SPRING ST, NEW YORK, NY 10013 USA, 2018.

Slob, E.A.W., Groenen, P.J., Thurik, A.R., and Rietveld, C.A. A note on the use of egger regression in mendelian randomization studies. *International Journal of Epidemiology*, 46(6):2094–2097, 2017.

Smith, G.D. and Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003.

Smith, G.D. and Ebrahim, S. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42, 2004.

So, H.C. and Sham, P.C. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. *Scientific Reports*, 7:41262, 2017.

Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013.

Song, L., Meng, J., Liu, Q., Huo, T., Zhu, X., Li, Y., Ren, Z., Wang, X., and Qiu, J. Polygenic Score of Subjective Well-Being Is Associated with the Brain Morphology in Superior Temporal Gyrus and Insula. *Neuroscience*, 414:210–218, 2019.

Standring, S. *Gray's anatomy e-book: the anatomical basis of clinical practice*. Elsevier Health Sciences, 2015.

Stephan, U. Entrepreneurs' mental health and well-being: A review and research agenda. *Academy of Management Perspectives*, 32(3):290–322, 2018.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12:e1001779, 2015.

Taylor, A.E., Davies, N.M., Ware, J.J., VanderWeele, T., Smith, G.D., and Munafò, M.R. Mendelian randomization in health research: using appropriate genetic variants and avoiding biased estimates. *Economics & Human Biology*, 13:99–106, 2014.

Tchetgen Tchetgen, E.J., Sun, B., and Walter, S. The GENIUS Approach to Robust Mendelian Randomization Inference. *arXiv:1709.07779 [stat.ME]*, 2017.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

Thomas, D., Lawlor, D., and Thompson, J. Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. *Annals of Epidemiology*, 17(7):511–513, 2007.

Thompson, O. Gene–environment interaction in the intergenerational transmission of asthma. *Health Economics*, 26(11):1337–1352, 2017.

Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*, 42(5):441–447, 2010.

Turkheimer, E. Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, 9(5):160–164, 2000.

Tyrrell, J., Jones, S.E., Beaumont, R., Astley, C.M., Lovell, R., Yaghootkar, H., Tuke, M., Ruth, K.S., Freathy, R.M., Hirschhorn, J.N., et al. Height, body mass index, and socioeconomic status: Mendelian randomisation study in UK biobank. *BMJ*, 352:i582, 2016.

US Department of Health and Human Services. The health consequences of smoking—50 years of progress: a report of the surgeon general. 2014.

Van der Loos, M.J., Koellinger, P.D., Groenen, P.J., Rietveld, C.A., Rivadeneira, F., van Rooij, F.J., Uitterlinden, A.G., Hofman, A., and Thurik, A.R. Candidate gene studies and the quest for the entrepreneurial gene. *Small Business Economics*, 37(3):269–275, 2011.

Van der Loos, M.J., Koellinger, P.D., Groenen, P.J., and Thurik, A.R. Genome-wide association studies and the genetics of entrepreneurship. *European Journal of Epidemiology*, 25(1):1–3, 2010.

Van der Loos, M.J., Rietveld, C.A., Eklund, N., Koellinger, P.D., Rivadeneira, F., Abecasis, G.R., Ankra-Badu, G.A., Baumeister, S.E., Benjamin, D.J., Biffar, R., et al. The molecular genetic architecture of self-employment. *PLOS ONE*, 8(4), 2013.

Van Kippersluis, H. and Rietveld, C.A. Pleiotropy-robust Mendelian randomization. *International Journal of Epidemiology*, 47(4):1279–1288, 2018.

VanderWeele, T.J., Tchetgen Tchetgen, E.J., Cornelis, M., and Kraft, P. Methodological challenges in Mendelian randomization. *Epidemiology*, 25(3):427–435, 2014.

VanRaden, P. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–23, 2008.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

Verbanck, M., Chen, C.Y., Neale, B., and Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, 50(5):693–698, 2018.

Verheul, I., Block, J., Burmeister-Lamp, K., Thurik, A.R., Tiemeier, H., and Turturea, R. Adhd-like behavior and entrepreneurial intentions. *Small Business Economics*, 45(1):85–101, 2015.

Verheul, I., Rietdijk, W., Block, J., Franken, I., Larsson, H., and Thurik, R. The association between attention-deficit/hyperactivity (ADHD) symptoms and self-employment. *European Journal of Epidemiology*, 31(8):793–801, 2016.

Verma, S.S., De Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou-Khales, B., Mukherjee, S., Jarvik, G.P., Kottyan, L.C., Burt, A., et al. Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in Genetics*, 5:370, 2014.

Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R., et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4):576–592, 2015.

Visscher, P., Wray, N., Zhang, Q., Sklar, P., McCarthy, M., Brown, M., and Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101(1):5–22, 2017a.

Visscher, P.M., Hill, W.G., and Wray, N.R. Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, 2008.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics*, 101(1):5–22, 2017b.

Von Hinke, S., Davey Smith, G., Lawlor, D.A., Propper, C., and Windmeijer, F. Genetic markers as instrumental variables. *Journal of Health Economics*, 45(1):131–148, 2016.

von Hinke, S., Smith, G.D., Lawlor, D.A., Propper, C., and Windmeijer, F. Genetic markers as instrumental variables. *Journal of Health Economics*, 45:131–148, 2016.

Wacholder, S., Chanock, S., Garcia-Closas, M., Ghormli, E., L., and Rothman, N. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96(6):434–442, 2004.

Ware, E., Schmitz, L., Gard, A., and Faul, J. Hrs polygenic scores—release 3: 2006–2012 genetic data. *Ann Arbor: Survey Research Center, University of Michigan*, 2018.

Weir, D. Quality control report for genotypic data. *University of Washington*, 2012.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, 2014.

Wernerfelt, N., Rand, D.G., Dreber, A., Montgomery, C., and Malhotra, D.K. *Arginine vasopressin 1a receptor (AVPR1a) RS3 repeat polymorphism associated with entrepreneurship*, 2012. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2141598.

White, J., Sofat, R., Hemani, G., Shah, T., Engmann, J., Dale, C., Shah, S., Kruger, F.A., Giambartolomei, C., Swerdlow, D.I., et al. Plasma urate concentration and risk of coronary heart disease: a Mendelian randomisation analysis. *Lancet Diabetes & Endocrinology*, 4(4):327–336, 2016.

Wiklund, J., Yu, W., Tucker, R., and Marino, L. Adhd, impulsivity, and entrepreneurship. *Journal of Business Venturing*, 32(6):627–656, 2017.

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114:1339–1350, 2019.

World Health Organization. *WHO report on the global tobacco epidemic, 2017: monitoring tobacco use and prevention policies*. World Health Organization, 2017.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565, 2010.

Yen, S.T. Zero observations and gender differences in cigarette consumption. *Applied Economics*, 37(16):1839–1849, 2005.

Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., Visscher, P.M., et al. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry. *Human Molecular Genetics*, 27(20):3641–3649, 2018.

Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

Zhang, Z., Zyphur, M.J., Narayanan, J., Arvey, R.D., Chaturvedi, S., Avolio, B.J., Lichtenstein, P., and Larsson, G. The genetic basis of entrepreneurship: Effects of gender and personality. *Organizational Behavior and Human Decision Processes*, 110(2):93–107, 2009.

Zhao, B., , Luo, T., Li, T., Li, Y., Zhang, J., Shan, Y., Wang, X., Yang, L., Zhou, F., Zhu, Z., and and, H.Z. Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nature Genetics*, 51(11):1637–1644, 2019.

Zhao, Q., Wang, J., Bowden, J., and Small, D.S. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *arXiv:1801.09652*, 2018.

Zheng, J., Erzurumluoglu, A.M., Elsworth, B.L., Kemp, J.P., Howe, L., Haycock, P.C., Hemani, G., Tansey, K., Laurin, C., Genetics, E., Consortium, L.E.E.E., Pourcain, B.S., Warrington, N.M., Finucane, H.K., Price, A.L., Bulik-Sullivan, B.K., Anttila, V., Paternoster, L., Gaunt, T.R., Evans, D.M., and Neale, B.M. LD Hub: a centralized database and web interface to perform ld score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, 33(2):272–279, 2017.

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 Mendeliaanse randomisatie studies helpen om te schatten of er een causaal verband is. Door ernstige zorgen over of bepaalde aannames die ten grondslag liggen aan Mendeliaanse 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 computationele 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 Publicatoins

- **A note on the use of Egger regression in Mendelian randomization studies.** E.A.W. Slob, P.J.F. Groenen, A.R. Thurik, and C.A. Rietveld, 2017, *International Journal of Epidemiology*, 46(6), 2094-2097.
- **A comparison of robust Mendelian randomization methods using summary data.** E.A.W. Slob, and S. Burgess, 2020, *Genetic Epidemiology*, 44(4),313-329.
- **A decade of research on the genetics of entrepreneurship: a review and view ahead.** C.A. Rietveld, E.A.W. Slob, and A.R. Thurik, 2020, *Small Business Economics*, https://doi.org/10.1007/s11187-020-00349-5.

Working Papers

- **Does the Genetic Predisposition to Smoking Moderate the Response to Tobacco Excise Taxes?** E.A.W. Slob, and C.A. Rietveld.
- **Multivariate analysis reveals the shared genetic architecture of brain regions and behavioral traits** R. de Vlaming, E.A.W. Slob, P.R. Jansen, P.D. Koellinger, P.J.F. Groenen, and C.A. Rietveld.

Grants and prizes

- 2020 Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), Call for Compute Time, EINF-403.
- 2017 Mendelian Randomization conference, Reserve poster prize.

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

Ahmadi, S., *A motivational perspective to decision-making and behavior in organizations*, Promotors: Prof. J.J.P. Jansen & Dr T.J.M. Mom, EPS-2019-477-S&E, https://repub.eur.nl/pub/116727.

Akemu, O., *Corporate Responses to Social Issues: Essays in Social Entrepreneurship and Corporate Social Responsibility*, Promotors: Prof. G.M. Whiteman & Dr S.P. Kennedy, EPS-2017-392-ORG, https://repub.eur.nl/pub/95768.

Albuquerque de Sousa, J.A., *International stock markets: Essays on the determinants and consequences of financial market development*, Promotors: Prof. M.A. van Dijk & Prof. P.A.G. van Bergeijk, EPS-2019-465-F&A, https://repub.eur.nl/pub/115988.

Alserda, G.A.G., *Choices in Pension Management*, Promotors: Prof. S.G. van der Lecq & Dr O.W. Steenbeek, EPS-2017-432-F&A, https://repub.eur.nl/pub/103496.

Arampatzi, E., *Subjective Well-Being in Times of Crises: Evidence on the Wider Impact of Economic Crises and Turmoil on Subjective Well-Being*, Promotors: Prof. H.R. Commandeur, Prof. F. van Oort & Dr. M.J. Burger, EPS-2018-459-S&E, https://repub.eur.nl/pub/111830.

Arslan, A.M., *Operational Strategies for On-demand Delivery Services*, Promotors: Prof. R.A. Zuidwijk & Dr N.A. H. Agatz, EPS-2019-481-LIS, https://repub.eur.nl/pub/126463.

Avci, E., *Surveillance of Complex Auction Markets: a Market Policy Analytics Approach*, Promotors: Prof. W. Ketter, Prof. H.W.G.M. van Heck & Prof. D.W. Bunn, EPS-2018-426-LIS.

https://repub.eur.nl/pub/106286

Balen, T.H. van, *Challenges of Early Stage Entrepreneurs: the Roles of Vision Communication and Team Membership Change*, Promotors: Prof. J.C.M. van den Ende & Dr M. Tarakci, EPS-2019-468-LIS, https://repub.eur.nl/pub/115654.

Bernoster, I., *Essays at the Intersection of Psychology, Biology, and Entrepreneurship*, Promotors: Prof. A.R. Thurik, Prof. I.H.A. Franken & Prof. P.J.F Groenen, EPS-2018-463-S&E, https://repub.eur.nl/pub/113907.

Blagoeva, R.R., *The Hard Power Of Soft Power: A behavioral strategy perspective on how power, reputation, and status affect firms*, Promotors: Prof. J.J.P. Jansen & Prof. T.J.M. Mom, EPS-2020-495-S&E, https://repub.eur.nl/pub/127681.

Bouman, P., *Passengers*, Crowding and Complexity: Models for Passenger Oriented Public Transport, Prof. L.G. Kroon, Prof.A.Schöbel&Prof.P.H.M.Vervest.

EPS-2017-420-LIS https://repub.eur.nl/pub/100767

Breugem, T., *'Crew Planning at Netherlands Railways: Improving Fairness, Attractiveness, and Efficiency'*, Promotors: Prof. D. Huisman & Dr T.A.B. Dollevoet, EPS-2020494-LIS, https://repub.eur.nl/pub/124016.

Bunderen, L. van, *Tug-of-War: Why and when teams get embroiled in power struggles*, Promotors: Prof. D.L. van Knippenberg & Dr. L. Greer, EPS-2018-446-ORG, https://repub.eur.nl/pub/105346.

Burg, G.J.J. van den, *Algorithms for Multiclass Classification and Regularized Regression*, Promotors: Prof. P.J.F. Groenen & Dr. A. Alfons, EPS-2018-442-MKT, https://repub.eur.nl/pub/103929.

Chammas, G., *Portfolio concentration*, Promotor: Prof. J. Spronk, EPS-2017-410-F&E, https://repub.eur.nl/pub/94975.

Chan, H.Y., *'Decoding the consumer's brain: Neural representations of consumer experience'*, Promotors: Prof. A. Smidts & Dr M. A.S. Boksem, EPS-2019-493-MKT, https://repub.eur.nl/pub/124931.

Dalmeijer, K., *Time Window Assignment in Distribution Networks*, Promotors: Prof A.P.M. Wagelmans & Dr R. Spliet, EPS-2019-486-LIS, https://repub.eur.nl/pub/120773.

Dennerlein, T., *Empowering Leadership and Employees' Achievement Motivations: the Role of Self-Efficacy and Goal Orientations in the Empowering Leadership Process*, Promotors: Prof. D.L. van Knippenberg & Dr J. Dietz, EPS-2017-414-ORG, https://repub.eur.nl/pub/98438.

Dolgova, E., *On Getting Along and Getting Ahead: How Personality Affects Social Network Dynamics*, Promotors: Prof. P.P.M.A.R Heugens & Prof. M.C. Schippers, EPS-2019-455-S& E, https://repub.eur.nl/pub/119150.

Duijzer, L.E., *Mathematical Optimization in Vaccine Allocation*, Promotors: Prof. R. Dekker & Dr W.L. van Jaarsveld, EPS-2017-430-LIS, https://repub.eur.nl/pub/101487.

Eijlers, E., *Emotional Experience and Advertising Effectiveness: on the use of EEG in marketing*, Prof. A. Smidts & Prof. M.A.S. Boksem, Eps-2019-487-MKT, https://repub.eur.nl/pub/124053.

El Nayal, O.S.A.N., *Firms and the State: An Examination of Corporate Political Activity and the Business-Government Interface*, Promotor: Prof. J. van Oosterhout & Dr. M. van Essen, EPS-2018-469-S&E, https://repub.eur.nl/pub/114683.

Feng, Y., *The Effectiveness of Corporate Governance Mechanisms and Leadership Structure: Impacts on strategic change and firm performance*, Promotors: Prof. F.A.J. van den Bosch, Prof. H.W. Volberda & Dr J.S. Sidhu, EPS-2017-389-S&E, https://repub.eur.nl/pub/98470.

Frick, T.W., *The Implications of Advertising Personalization for Firms, Consumer, and Ad Platfroms*, Promotors: Prof. T. Li & Prof. H.W.G.M. van Heck, EPS-2018-452-LIS, https://repub.eur.nl/pub/110314.

Fytraki, A.T., *Behavioral Effects in Consumer Evaluations of Recommendation Systems*, Promotors: Prof. B.G.C. Dellaert & Prof. T. Li, EPS-2018-427-MKT, https://repub.eur.nl/pub/110457.

Gai, J., *Contextualized Consumers: Theories and Evidence on Consumer Ethics, Product Recommendations, and Self-Control*, Promotors: Prof. S. Puntoni & Prof. S.T.L. Sweldens, EPS-2020-498-MKT, https://repub.eur.nl/pub/127680.

Ghazizadeh, P., *Empirical Studies on the Role of Financial Information in Asset and Capital Markets*, Promotors: Prof. A. de Jong & Prof. E. Peek, EPS-2019-470-F& A, https://repub.eur.nl/pub/114023.

Giurge, L., *A Test of Time; A temporal and dynamic approach to power and ethics*, Promotors: Prof. M.H. van Dijke & Prof. D. De Cremer, EPS-2017-412-ORG, https://repub.eur.nl/pub/98451.

Gobena, L., *Towards Integrating Antecedents of Voluntary Tax Compliance*, Promotors: Prof. M.H. van Dijke & Dr P. Verboon, EPS-2017-436-ORG, https://repub.eur.nl/pub/103276.

Groot, W.A., *Assessing Asset Pricing Anomalies*, Promotors: Prof. M.J.C.M. Verbeek & Prof. J.H. van Binsbergen, EPS-2017-437-F& A, https://repub.eur.nl/pub/103490.

Hanselaar, R.M., *Raising Capital: On pricing, liquidity and incentives*, Promotors: Prof. M.A. van Dijk & Prof. P.G.J. Roosenboom, EPS-2018-429-F& A, https://repub.eur.nl/pub/113274.

Harms, J. A., *Essays on the Behavioral Economics of Social Preferences and Bounded Rationality*, Prof. H.R. Commandeur & Dr K.E.H. Maas, EPS-2018-457-S& E, https://repub.eur.nl/pub/108831.

Hendriks, G., *Multinational Enterprises and Limits to International Growth: Links between Domestic and Foreign Activities in a Firm's Portfolio*, Promotors: Prof. P.P.M.A.R. Heugens & Dr. A.H.L Slangen, EPS-2019-464-S& E, https://repub.eur.nl/pub/114981.

Hengelaar, G.A., *The Proactive Incumbent: Holy grail or hidden gem? Investigating whether the Dutch electricity sector can overcome the incumbent's curse and lead the sustainability transition*, Promotors: Prof. R.J. M. van Tulder & Dr K. Dittrich, EPS-2018-438-ORG, https://repub.eur.nl/pub/102953.

Jacobs, B.J.D., *Marketing Analytics for High-Dimensional Assortments*, Promotors: Prof. A.C.D. Donkers & Prof. D. Fok, EPS-2017-445-MKT, https://repub.eur.nl/pub/103497.

Jia, F., *The Value of Happiness in Entrepreneurship*, Promotors: Prof. D.L. van Knippenberg & Dr Y. Zhang, EPS-2019-479-ORG, https://repub.eur.nl/pub/115990.

Kahlen, M. T., *Virtual Power Plants of Electric Vehicles in Sustainable Smart Electricity Markets*, Promotors: Prof. W. Ketter & Prof. A. Gupta, EPS-2017-431-LIS, https://repub.eur.nl/pub/100844.

Kampen, S. van, *The Cross-sectional and Time-series Dynamics of Corporate Finance: Empirical evidence from financially constrained firms*, Promotors: Prof. L. Norden & Prof. P.G.J. Roosenboom, EPS-2018-440-F& A, https://repub.eur.nl/pub/105245.

Karali, E., *Investigating Routines and Dynamic Capabilities for Change and Innovation*, Promotors: Prof. H.W. Volberda, Prof. H.R. Commandeur & Dr J.S. Sidhu, EPS-2018-454-S&E, https://repub.eur.nl/pub/106274.

Keko , E., *Essays on Innovation Generation in Incumbent Firms*, Promotors: Prof. S. Stremersch & Dr N.M.A. Camacho, EPS-2017-419-MKT, https://repub.eur.nl/pub/100841.

Kerkkamp, R.B.O., *Optimisation Models for Supply Chain Coordination under Information Asymmetry*, Promotors: Prof. A.P.M. Wagelmans & Dr. W. van den Heuvel, EPS-2018-462-LIS, https://repub.eur.nl/pub/109770.

Khattab, J., *Make Minorities Great Again: a contribution to workplace equity by identifying and addressing constraints and privileges*, Promotors: Prof. D.L. van Knippenberg & Dr A. Nederveen Pieterse, EPS-2017-421-ORG, https://repub.eur.nl/pub/99311.

Kim, T. Y., *Data-driven Warehouse Management in Global Supply Chains*, Promotors: Prof. R. Dekker & Dr C. Heij, EPS-2018-449-LIS, https://repub.eur.nl/pub/109103.

Klitsie, E.J., *Strategic Renewal in Institutional Contexts: The paradox of embedded agency*, Promotors: Prof. H.W. Volberda & Dr. S. Ansari, EPS-2018-444-S&E, https://repub.eur.nl/pub/106275.

Koolen, D., *Market Risks and Strategies in Power Systems Integrating Renewable Energy*, Promotors: Prof. W. Ketter & Prof. R. Huisman, EPS-2019-467-LIS, https://repub.eur.nl/pub/115655.

Kong, L., *Essays on Financial Coordination*, Promotors: Prof. M.J.C.M. Verbeek, Dr. D.G.J. Bongaerts & Dr. M.A. van Achter. EPS-2019-433-F&A, https://repub.eur.nl/pub/114516.

Kyosev, G.S., *Essays on Factor Investing*, Promotors: Prof. M.J.C.M. Verbeek & Dr J.J. Huij, EPS-2019-474-F&A, https://repub.eur.nl/pub/116463.

Lamballais Tessensohn, T., *Optimizing the Performance of Robotic Mobile Fulfillment Systems*, Promotors: Prof. M.B.M de Koster, Prof. R. Dekker & Dr D. Roy, EPS-2019-411-LIS, https://repub.eur.nl/pub/116477.

Leung, W.L., *How Technology Shapes Consumption: Implications for Identity and Judgement*, Promotors: Prof. S. Puntoni & Dr G Paolacci, EPS-2019-485-MKT, https://repub.eur.nl/pub/117432.

Li, X., *Dynamic Decision Making under Supply Chain Competition*, Promotors: Prof. M.B.M de Koster, Prof. R. Dekker & Prof. R. Zuidwijk, EPS-2018-466-LIS, https://repub.eur.nl/pub/114028.

Liu, N., *Behavioral Biases in Interpersonal Contexts*, Supervisors: Prof. A. Baillon & Prof. H. Bleichrodt, EPS-2017-408-MKT, https://repub.eur.nl/pub/95487.

Maas, A.J.J., *Organizations and their external context: Impressions across time and space*, Promotors: Prof. P.P.M.A.R Heugens & Prof. T.H. Reus, EPS-2019-478-S&E, https://repub.eur.nl/pub/116480.

Maira, E., *Consumers and Producers*, Promotors: Prof. S. Puntoni & Prof. C. Fuchs, EPS-2018-439-MKT, https://repub.eur.nl/pub/104387.

Mirzaei, M., *'Advanced Storage and Retrieval Policies in Automated Warehouses'*, Promotors: Prof. M.B.M. de Koster & Dr N. Zaerpour, EPS-2020-490-LIS, https://repub.eur.nl/pub/125975.

Nair, K.P., *Strengthening Corporate Leadership Research: The relevance of biological explanations*, Promotors: Prof. J. van Oosterhout & Prof. P.P.M.A.R Heugens, EPS-2019-480-S&E, https://repub.eur.nl/pub/120023.

Nullmeier, F.M.E., *Effective contracting of uncertain performance outcomes: Allocating responsibility for performance outcomes to align goals across supply chain actors*, Promotors: Prof. J.Y.F.Wynstra & Prof. E.M. van Raaij, EPS-2019-484-LIS, https://repub.eur.nl/pub/118723.

Okbay, A., *Essays on Genetics and the Social Sciences*, Promotors: Prof. A.R. Thurik, Prof. Ph.D. Koellinger & Prof. P.J.F. Groenen, EPS-2017-413-S&E, https://repub.eur.nl/pub/95489.

Peng, X., *Innovation, Member Sorting, and Evaluation of Agricultural Cooperatives*, Promotor: Prof. G.W.J. Hendriks, EPS-2017-409-ORG, https://repub.eur.nl/pub/94976.

Petruchenya, A., *Essays on Cooperatives: Emergence, Retained Earnings,and Market Shares*, Promotors: Prof. G.W.J. Hendriks & Dr Y. Zhang, EPS-2018-447-ORG, https://repub.eur.nl/pub/105243.

Plessis, C. du, *Influencers: The Role of Social Influence in Marketing*, Promotors: Prof. S. Puntoni & Prof. S.T.L.R. Sweldens, EPS-2017-425-MKT, https://repub.eur.nl/pub/103265.

Pocock, M., *Status Inequalities in Business Exchange Relations in Luxury Markets*, Promotors: Prof. C.B.M. van Riel & Dr G.A.J.M. Berens, EPS-2017-346-ORG, https://repub.eur.nl/pub/98647.

Pozharliev, R., *Social Neuromarketing: The role of social context in measuring advertising effectiveness*, Promotors: Prof. W.J.M.I. Verbeke & Prof. J.W. van Strien, EPS-2017-402-MKT, https://repub.eur.nl/pub/95528.

Qian, Z., *Time-Varying Integration and Portfolio Choices in the European Capital Markets*, Promotors: Prof. W.F.C. Verschoor, Prof. R.C.J. Zwinkels & Prof. M.A. Pieterse-Bloem, EPS-2020-488-F&A, https://repub.eur.nl/pub/124984.

Reh, S.G., *A Temporal Perspective on Social Comparisons in Organizations*, Promotors: Prof. S.R. Giessner, Prof. N. van Quaquebeke & Dr. C. Troster, EPS-2018-471-ORG, https://repub.eur.nl/pub/114522.

Riessen, B. van, *Optimal Transportation Plans and Portfolios for Synchromodal Container Networks*, Promotors: Prof. R. Dekker & Prof. R.R. Negenborn, EPS-2018-448-LIS, https://repub.eur.nl/pub/105248.

Romochkina, I.V., *When Interests Collide: Understanding and modeling interests alignment using fair pricing in the context of interorganizational information systems*, Promotors: Prof. R.A. Zuidwijk & Prof. P.J. van Baalen, EPS-2020-451-LIS, https://repub.eur.nl/pub/127244.

Schie, R. J. G. van, *Planning for Retirement: Save More or Retire Later?* , Promotors: Prof. B. G. C. Dellaert & Prof. A.C.D. Donkers, EOS-2017-415-MKT, https://repub.eur.nl/pub/100846.

Schouten, K.I.M., *Semantics-driven Aspect-based Sentiment Analysis*, Promotors: Prof. F.M.G. de Jong, Prof. R. Dekker & Dr. F. Frasincar, EPS-2018-453-LIS, https://repub.eur.nl/pub/112161.

Sihag, V., *The Effectiveness of Organizational Controls: A meta-analytic review and an investigation in NPD outsourcing*, Promotors: Prof. J.C.M. van den Ende & Dr S.A. Rijsdijk, EPS-2019-476-LIS, https://repub.eur.nl/pub/115931.

Smolka, K.M., *Essays on Entrepreneurial Cognition, Institution Building and Industry Emergence*, Promotors: P.P.M.A.R. Heugens, & Prof. J.P. Cornelissen, Eps-2019-483-S&E, https://repub.eur.nl/pub/118760.

Straeter, L.M., *Interpersonal Consumer Decision Making*, Promotors: Prof. S.M.J. van Osselaer & Dr I.E. de Hooge, EPS-2017-423-MKT, https://repub.eur.nl/pub/100819.

Stuppy, A., *Essays on Product Quality*, Promotors: Prof. S.M.J. van Osselaer & Dr N.L. Mead. , EPS-2018-461-MKT, https://repub.eur.nl/pub/111375.

Subaşi, B., *Demographic Dissimilarity, Information Access and Individual Performance*, Promotors: Prof. D.L. van Knippenberg & Dr W.P. van Ginkel, EPS-2017-422-ORG, https://repub.eur.nl/pub/103495.

Suurmond, R., *In Pursuit of Supplier Knowledge: Leveraging capabilities and dividing responsibilities in product and service contexts*, Promotors: Prof. J.Y.F Wynstra & Prof. J. Dul. , EPS-2018-475-LIS, https://repub.eur.nl/pub/115138.

Toxopeus, H.S. , *Financing sustainable innovation: From a principal-agent to a collective action perspective*, Promotors: Prof. H.R. Commandeur & Dr. K.E.H. Maas., EPS-2019-458-S&E, https://repub.eur.nl/pub/114018.

Turturea, R., *Overcoming Resource Constraints: The Role of Creative Resourcing and Equity Crowdfunding in Financing Entrepreneurial Ventures*, Promotors: Prof. P.P.M.A.R Heugens, Prof. J.J.P. Jansen & Dr. I. Verheuil, EPS-2019-472-S&E, https://repub.eur.nl/pub/112859.

Valboni, R., *'Building Organizational (Dis-)Abilities: The impact of learning on the performance of mergers and acquisitions'*, Promotors: Prof. T.H. Reus & Dr A.H.L. Slangen, EPS-2020-407-S&E, https://repub.eur.nl/pub/125226.

Vandic, D., *Intelligent Information Systems for Web Product Search*, Promotors: Prof. U. Kaymak & Dr Frasincar, EPS-2017-405-LIS, https://repub.eur.nl/pub/95490.

Verbeek, R.W.M., *Essays on Empirical Asset Pricing*, Promotors: Prof. M.A. van Dijk & Dr M. Szymanowska, EPS-2017-441-F&A, https://repub.eur.nl/pub/102977.

Visser, T.R. , *Vehicle Routing and Time Slot Management in Online Retailing*, Promotors: Prof. A.P.M. Wagelmans & Dr R. Spliet, EPS-2019-482-LIS, https://repub.eur.nl/pub/120772.

Vlaming, R. de., *Linear Mixed Models in Statistical Genetics*, Prof. A.R. Thurik, Prof. P.J.F. Groenen & Prof. Ph.D. Koellinger, EPS-2017-416-S&E, https://repub.eur.nl/pub/100428.

Vries, H. de, *Evidence-Based Optimization in Humanitarian Logistics*, Promotors: Prof. A.P.M. Wagelmans & Prof. J.J. van de Klundert, EPS-2017-435-LIS, https://repub.eur.nl/pub/102771.

Wang, R., *Corporate Environmentalism in China*, Promotors: Prof. P.P.M.A.R Heugens & Dr F. Wijen, EPS-2017-417-S&E, https://repub.eur.nl/pub/99987.

Wang, R., *Those Who Move Stock Prices*, Promotors: Prof. P. Verwijmeren & Prof. S. van Bekkum, EPS-2019-491-F&A, https://repub.eur.nl/pub/129057.

Wasesa, M., *Agent-based inter-organizational systems in advanced logistics operations*, Promotors: Prof. H.W.G.M van Heck, Prof. R.A. Zuidwijk & Dr A. W. Stam, EPS-2017-LIS-424, https://repub.eur.nl/pub/100527.

Wessels, C., *Flexible Working Practices: How Employees Can Reap the Benefits for Engagement and Performance*, Promotors: Prof. H.W.G.M. van Heck, Prof. P.J. van Baalen & Prof. M.C. Schippers, EPS-2017-418-LIS, https://repub.eur.nl/.

Wiegmann, P.M., *Setting the Stage for Innovation: Balancing Diverse Interests through Standardisation*, Promotors: Prof. H.J. de Vries & Prof. K. Blind, EPS-2019-473-LIS, https://repub.eur.nl/pub/114519.

Wijaya, H.R., *Praise the Lord!: Infusing Values and Emotions into Neo-Institutional Theory*, Promotors: Prof. P.P.M.A.R. Heugens & Prof. J.P. Cornelissen, EPS-2019-450-S&E, https://repub.eur.nl/pub/115973.

Williams, A.N., *Make Our Planet Great Again: A Systems Perspective of Corporate Sustainability*, Promotors: Prof. G.M. Whiteman & Dr. S. Kennedy, EPS-2018-456-ORG, https://repub.eur.nl/pub/111032.

Witte, C.T., *Bloody Business: Multinational investment in an increasingly conflict-afflicted world*, Promotors: Prof. H.P.G. Pennings, Prof. H.R. Commandeur & Dr M.J. Burger, EPS-2018-443-S&E, https://repub.eur.nl/pub/104027.

Ye, Q.C., *Multi-objective Optimization Methods for Allocation and Prediction*, Promotors: Prof. R. Dekker & Dr Y. Zhang, EPS-2019-460-LIS, https://repub.eur.nl/pub/116462.

Yuan, Y., *The Emergence of Team Creativity: a social network perspective*, Promotors: Prof. D. L. van Knippenberg & Dr D. A. Stam, EPS-2017-434-ORG, https://repub.eur.nl/pub/100847.

Zhang, Q., *Financing and Regulatory Frictions in Mergers and Acquisitions*, Promotors: Prof. P.G.J. Roosenboom & Prof. A. de Jong, EPS-2018-428-F&A, https://repub.eur.nl/pub/103871.