Advances and Current Methodological Problems in Understanding Depression: A Sociogenomic Approach

by

Evelina T. Akimova

St. Antony's College, University of Oxford

A thesis presented for the degree of
Doctor of Philosophy in Sociology

Michaelmas 2020
Word count: *47,388*

Evelina T. Akimova

Advances and Current Methodological Problems in Understanding Depression: A Sociogenomic Approach

Abstract

Depression is a global burden. It is one of the most common mental health disorders and top ten causes of sickness worldwide. The importance of obtaining a deeper understanding of depression and its development has been raised by various scholars, policy makers, and the media. All of these factors contribute to a burgeoning body of research into different aspects of depression. The lack of a multifaceted understanding of depression is one of the core obstacles for its treatment. Depression has both biological and non-biological risk factors driven by interconnected causes. The recognition of social and biological drivers and recent advances in molecular genetics permits an unprecedented and unique opportunity to use sociogenomic tools to deepen our knowledge of the multidimensional nature of the biological and social risks of developing depression.

This thesis bridges a gap in our knowledge on the historical, social, and biological predictors of depression by adopting a sociogenomic approach, focusing on the highly relevant context of the UK. The scope of the thesis is unique as I aim to not only expand on empirical evidence of the complex interplay between individual- and macro-level risk factors of depression, but also raise methodological concerns that arise due to such interplay. In four empirical chapters, I contribute to the existing literature on depression and social science genomics, in four different ways: (1) a methodological assessment of selection in genetic data and its role in the polygenic prediction of depression; (2) scrutiny of moderating patterns of birth cohorts and economic recessions associated with changing polygenic penetrance of depression; (3) exploration of the link between genetic predispositions to depression and instances of worklessness that position people into higher risks to experience depression; and, (4) investigation of the ways in which endogenous selection bias leads to spurious associations and biased variance statistics in the models with polygenic scores.

In memory of Irina Akimova and Dmitry Malyshev.

Acknowledgments

I would like to thank the following people without whom the journey of writing this dissertation would not have been possible. First and foremost, I would like to express my gratitude to my primary supervisor, Melinda C. Mills, who provided the opportunity to undertake the DPhil programme and guided me throughout the project. Her dedicated mentorship, depth of knowledge, commitment to scientific achievement, wisdom, and compassion have been a constant source of inspiration. I would also like to express sincere gratitude to my secondary supervisor, David M. Brazel. I am very grateful for his excellent advice, sharp feedback, numerous statistical suggestions, and encouragement for my research. I extend a special thank you to Richard Breen, who provided mentorship and guidance as an ESRC supervisor and research advisor. I am also eternally grateful to: David Kirk and Jennifer Beam Dowd for their generous time to provide me invaluable feedback during ToS and CoS examinations; Colin Mills, who was my MSc supervisor and who guided me into the field of social stratification; Christiaan Monden, Stephen D. Fisher, Jonathan Pointer, Natasha Cotton, and Khalid Omer for academic advice, assistance, and support; Patrick Präg, Nicola Barban, Felix C. Tropf, Charles Rahal, and Xuejie Ding for comments on my work and encouragements; and Jeremy Freese and Ben Domingue for the opportunity for a research stay at Stanford's Sociology Department, which benefited my research significantly. It would not have been possible to reach this stage without constant love and support from my family and friends, but I simply cannot express here the extent of my gratitude to them. I also want to thank my partner and friend Sébastien Barrey for his unending support, constant encouragement, and kindness.

The research leading to this dissertation project was funded by the European Research Council Consolidator Grant SOCIOGENOME, awarded to Melinda C. Mills (Grant no. 615603). I am grateful for the academic support provided by ESRC Doctoral Training Centre at Oxford.

Contents

List of Tables

List of Figures

List of Abbreviations

BDNF – Brain-Derived Neurotrophic Factor

BHPS – British Household Panel Survey

BMI – Body Mass Index

DNA – Deoxyribonucleic Acid

GxE – Gene-Environment Interaction

GHQ – General Health Questionnaire

GPS – General Population Sample

GWAS – Genome Wide Association Studies

H*n* – Hypothesis

HRS – Health and Retirement Study

HWE – Hardy-Weinberg Equilibrium

ICC – Inter-Class Correlation

ILO – International Labour Organization

LD – Linkage Disequilibrium

MAF – Minor Allele Frequency

MDD – Major Depressive Disorder

NHS – National Health System

NIH – National Institute of Health

OLS – Ordinary Least Squares (Regression)

PC – Principal Component

PGS – Polygenic Risk Score

rGE – Gene-Environment Correlation

SD – Standard Deviation

SE – Standard Error

SES – Socio-Economic Status

SF12 – Short Form (12) Health Survey

SNP – Single Nucleotide Polymorphism

UK – United Kingdom (England, Northern Ireland, Scotland, and Wales)

UKHLS – United Kingdom Household Longitudinal Study

US – United States

WWI – World War I

WWII – World War II

Chapter 1

Introduction

1.1 The problem of depression

Mental health is currently a global issue and a critical public health concern. It is high on the UK political agenda [Health & Social Care Department, 2019] and worldwide. According to data from the World Health Organisation [2020], around 20% of adults have a mental health condition and the prevalence of mental disorders increased by 13% in the last decade. Globally, mental health problems are one of the main causes of overall sickness [Vos et al., 2015] and disability [Lozano et al., 2012]. These findings align with the recent data showing that mental health conditions are associated with one in five years lived with a disability [World Health Organisation, 2020]. Mental health problems are also linked to suicides [Appleby et al., 2017].

Depression is one of the most common mental health disorders. Following the definition of the UK National Health System (NHS), depression is a mood disorder when an individual experiences feelings of unhappiness and hopelessness along with low self-esteem [NHS, 2019]. Gitterman [1991] calls depression the *'common cold'* of mental health to reflect the scope of its prevalence. Since 1991, the burden of depression has only increased. Depression is one of the top ten causes of sickness worldwide [World Health Organisation, 2018]. According to the Global Burden of Diseases, Injuries, and Risk Factors Study covering 195 countries, depressive disorders became the third leading cause for number of years lived with disability between 1990 and 2017 [James et al., 2018]. Moreover, there are more women suffering from depression than men [World Health Organisation, 2018]. The severity of depression varies, but the most severe is major depression. The frequency of major depression ranges from 8% to 12% in different countries [Flint and Kendler, 2014].

All of these factors have led to a burgeoning body of research into different aspects of depression. Although depression imposes a global burden, current understanding of its development is insufficient. The complexity of causes playing a role in the development of depression poses a core challenge to the task of tackling its different stages. Depression has both internal and external risk factors that represent a complex net of interconnected causes for its development. This has led to call for a more interdisciplinary focus, increasingly the most dominant approach that is adopted. Recent advances in molecular genetics have introduced a unique opportunity for the use of sociogenomic tools (i.e., inclusion of both social and genetic factors) to deepen our knowledge of the multidimensional nature of the biological and social risks of developing depression. For example, de Castro-Catala et al. [2020] show that while a childhood trauma is one of the most robust environmental trigger of depression, underlying genetic predispositions also play a role explaining inter-individual differences in responses to childhood adversities.

The motivation of this thesis is to bridge a gap in our knowledge of the historical, social, and biological predictors of depression. Accordingly, a complex interplay of individual- and macro-level risk factors of depression is the main research question I address over the course of the chapters through empirical and methodological lenses. Empirical lenses of my investigation on depression are covered in Chapters 3 and 4 where I investigate such substantive research topics as gene-by-cohort interactions and gene-environment correlations. Methodological aspects are addressed in Chapters 2 and 5 where the selection into genetic samples and the problem of endogeneity constitute two research topics.

My empirical focus is the context of the UK where the problem of depression is profound. The UK is on the list of the countries with high prevalence of depression [Ritchie and Roser, 2018]. Moreover, following the UK National Health System report [McManus et al., 2016], around 17% of the English population meet the criteria for common mental health disorders. The situation is further aggravated by the systemic underfunding in the UK mental health services [TUC, 2018]. The Austerity policy following the Great Recession further contributed to the budget cuts. These factors make the UK a highly relevant and essential context to study depression. Moreover, the relevance is further complimented by the availability of data, reliable population measures and samples.

Notably, there are various ways to measure and model depression. In broader terms, there are two main approaches: modelling diagnostic information or considering symptomatic data. I do not use depression diagnosis as my main phenotypic variable as symptomatic data on mental health is demonstrably more accurate with greater validity and reliability in the context of population-based surveys [Mandemakers, 2011]. While the incidence of depression and its diagnosis indicate the severity of someone's mental health, symptomatic data makes it possible to look at the issue across a broader spectrum and to model the risks for developing depression.

This introductory chapter presents the theoretical context for empirical inquiries into depression and details a comprehensive outline of the thesis. In the sections below, I discuss the three main scientific approaches to depression: psychological, biological, and sociological. Here, I highlight historical and field-specific differences in definitions and understanding of the causes for depression. Such differences have been the driver in the rise of interdisciplinary focus, which I also cover. I pay particular attention to the rise of social science genomics. Then, I provide an overview of the thesis comprised of short summaries of each chapter. I also describe methodological aspects of the work, where I cover information on data and analytical samples along with conceptualisation schemes and relevant terminology.

1.2 Multiple approaches to understanding depression

Despite the differences I explain in the following sections, scientific approaches to depression have some common assumptions I wish to mention here. Firstly, it is assumed that the mental illness can be meaningfully clustered and classified. Secondly, people in those clusters are similar to each other in ways other than their diagnosis (and it is not, therefore, randomly distributed). Thirdly, an understanding of the common causes of each cluster will point to treatments.

1.2.1 Psychological approaches

The importance of deeper knowledge on the complexity of depression has been raised not only in the current contexts driven by the growing prevalence of depression, but also dating back to the middle of the twentieth century. Back then, the most prevalent approach to mental health was based on psychological models. These aimed to understand the distinctive nature of feelings and behaviours observed among people with mental health problems. Importantly, the central analytical unit of the psychological approach is an individual and her/his peculiarities. While there are various psychological theories on mental health, these theories do not provide a unified profile of mental health problems [Peterson, 2009]. I wish to highlight three main approaches to understanding mental health that are directly linked to depression: psychoanalytic, family systems, and cognitive-behavioural theories. These theories play a key role in empirical studies on depression and its remedies. They also provide the theoretical foundations for an interdisciplinary approach.

Seeing depression through the lenses of the psychoanalytic model, our starting point becomes aspects of its development. For example, we would focus on how and what early childhood circumstances made us vulnerable to depression. On this assumption of the developmental nature of depression, more effective interventions could also be built. Psychoanalytic therapy aims to reveal insights from unpleasant experiences because another assumption of this theory sees uncovered insights as a source of healing power [Peterson, 2009]. The family systems model likewise highlights the importance of childhood in relation to depression. However, this notion is based on the assumption that mental health problems are driven by disturbances in the family [Jacobson and Addis, 1993]. Both of these theories provide the fundaments of modern talk therapies such as psychodynamic and interpersonal therapy (IPT), which has been shown to be effective in treating [de Mello et al., 2005, Shedler, 2010].

If we stand for cognitive-behavioural theory and its understanding of depression, we would emphasise the role of stressful situations along with the ways of thinking as driving forces. We would encourage therapies that teach people to develop adaptive habits and greater control over thinking processes. This approach provides the foundation for the big family of modern CBT techniques (Feldman [2007] provides a comprehensive review on this type of talking therapy).

One of the most important contributions of these psychological approaches are insights into the intraindividual mechanisms that contribute to the development of depression. Such insights are still playing critical roles in the development of individual therapies, further contributing to the clinical utility of psychotherapies [Nathan, 2007].

1.2.2 Biological approach

One of the most important shifts in the understanding of depression is rooted in a biological revolution in psychiatry that occurred during the last quarter of the twentieth century following an empirical success to demonstrate the role of biology in depression development [Schwartz and Corcoran, 2009]. What was most insightful and exciting for researchers is the ability to uncover the aetiology of mental diseases through the sequence of biological mechanisms [Andreasen, 1985]. Importantly, the biological perspective fundamentally sees depression as an illness of the brain. One of the main drivers of the revolution was the results from twin and adoptive studies revealing the importance of genetic grounds for mental health. Unsurprisingly, the issue of the heritability [1] of depression has been a major area of scientific interest. The earliest stages of research suggested that depression develops in families [Tsuang and Faraone, 1990]. Twin studies showed that if a person's parents suffer from depression, their risk of developing the illness ranges from 20% to 40% [Jansson et al., 2004, McGue and Christensen, 2003, Sullivan et al., 2000]. These findings resulted in the perception that genetics are a critical factor in understanding depression. Following recent advances in molecular genetics, we can now address heritability by looking directly at genetic patterns accompanying existing knowledge from twin studies. Although the role of genetics has not yet been exhaustively illustrated, there has been some progress in this direction.

There were several attempts to reveal a single gene responsible for depression (for a review, see Bosker et al. [2011]). It was known that some diseases, such as Huntington's disease, are caused by single genetic mutations [Chial, 2008] where depression is one element of its clinical representation [Epping and Paulsen, 2011]. Later, researchers demonstrated that depression does not have an inheritance pattern similar to single-genetic-mutation illnesses such as Huntington's disease. Further developments in the field have shown that no single gene is responsible for depression. But genes are important in another way: their influence is due to the small effects of many SNPs [2] across the genome - the observation that is called polygeneity. A widely used approach to discover the polygenic or multiple genetic variants associated with a trait are genome-wide-association studies (GWASs), where the rationale is to screen all of the SNPs in the human genome and to test their association with a certain phenotype. Multiple studies look at depression as a phenotype [Hek et al., 2013, Okbay et al., 2016, Terracciano et al., 2010, Wray et al., 2018]. The most recent and largest one

[1] Heritability is the amount of variation in a trait that is attributed to genetic differences between people[Meaney and Taylor, 2018]

[2] SNP is an abbreviation of single-nucleotide polymorphism. The human genome consists 3 billion pairs of nucleotide molecules which can be indexed by their location in the genome [Benjamin et al., 2012]. It is the most prevalent type of genetic variation among humans [National Institutes of Health, 2019]. These variations can be both common and unique (rare). SNPs act as biological markers that help us to find the genes associated with diseases or phenotypes [Carey, 2012]

was conducted by Howard et al. [2019], and I use this to construct polygenic scores [3]. Up until now, researchers have identified 102 independent variants, 269 genes, and 15 gene sets associated with depression; these numbers include both genes and gene pathways associated with synaptic structure and neurotransmission Howard et al. [2019]. Moreover, a portion of the identified genetic signals is linked to expressions of cortisol and serotonin [Bansal et al., 2016, McGowan et al., 2009]. Some of the SNPs associated with depression are also located in the coding region for the central nervous system, and exert their effects in transcriptions for the development of this system [Howard et al., 2019, Hyde et al., 2016].

Importantly, the genetics of depression is only one of the elements of a biological perspective. I deliberately focus on this element as it falls within the scope of my thesis. There are other angles within this paradigm as well, such as neuroanatomy, neurochemistry, and endocrinology. For a detailed discussion of these aspects, I refer the reader to Syvälahti [1994], Thase et al. [2002].

1.2.3 Sociological approach

While both psychological and biological perspectives mostly concentrate on depression with the individual at the centre of inquiry, sociological approaches are interested in social circumstances and bigger societal trends. Here, we move away from the individualistic paradigm to start wondering about the ways people coexist in the wider contexts of different cultures, countries, historical times, social structures, and institutions. It could be that these ways and differences play a role in the development and prevalence of depression. Horwitz [2009] admits that the sociological perspective is fundamentally distinct, but it nonetheless offers an additional range of insights that complement psychological and biological approaches to depression. Sociological analyses are less likely to conceptualise depression as its diagnosis, as symptomatic data and depression scores have greater validity and reliability in the context of population surveys [Mandemakers, 2011]. Such analyses are also

[3] Polygenic score is a single-value indicator of an individual's genetic predisposition to a certain trait. It is calculated as a sum of genome-wide SNPs weighted by a respective effect size that comes from the summary statistics of a GWAS [Choi et al., 2018, p. 2].

considerably less likely to study small groups of depression patients compared to psychological and biological approaches.

Traditionally, sociology has focused on uncovering and understanding the links between macro conditions and individual-level outcomes. The first sociological analysis of mental health is found in the work of Emile Durkheim (1858-1917). In his 1897 book on suicides [Durkheim, 1897], Durkheim analysed state-level statistics to show that suicides have social roots. He argued that this phenomenon should not be understood only on the basis of psychological mechanisms, proposing the importance of social factors.

If we study depression from a sociological approach, our fundamental assumption is that depression occurs and progresses due to individual susceptibilities such as different personalities and genetics – but only in part. Depression is also caused by different social conditions and the locations where individuals find themselves. Accordingly, depression can be triggered by factors such as educational attainment [Lee, 2011], marital status [Kessler and Essex, 1982, Pearlin and Johnson, 1977], and economic recessions [Frasquilho et al., 2015, Jahoda, 1988]. There is also a substantial literature on relationships between labour force status and depression (see reviews from Ezzy [1993], Bartley [1994], Fryers et al. [2003], McLean et al. [2005]). For example, periods of prolonged economic inactivity increase the risks of depression development across different age groups [Egan et al., 2015, Frese and Mohr, 1987, Strandh et al., 2014] in different countries [Butterworth et al., 2012, Jefferis et al., 2011, Winefield and Tiggemann, 1990].

Moreover, there is also a wider proposition that depression prevalence has a historical trend and occurs more frequently among recent birth cohorts [Bell, 2014, Marcus and Olfson, 2010]. Cohort trends in general are a major area of study in sociology and demography [Ryder, 1965]. This particular interest is linked to the notion that cohort effects indicate the importance of historical changes in potentially shaping people's experiences, at least to some extent. The first serious discussions of an increase in the prevalence of mental health problems emerged during the 1970s [Marcus and Olfson, 2010].

1.2.4 The rise of interdisciplinary approaches

Even though theoretical advances in understanding depression implied an interdisciplinary focus, empirical practices started from a single-predictor agendas and only after moved to modelling of complex interplay and interactions. Each of the scientific approaches described until now empirically showed that there are both internal and external factors for depression. This finding grounded the conceptualisation scheme of *endogenous depression* and *reactive depression* developed by Gillespie [1929]. The biological revolution in psychiatry resulted in the further rise of biomedical and psychosocial perspectives to depression, along with their split. While the biomedical approach locates causes for depression within disturbances in brain function, the psychosocial approach argues that life experiences cause depression [Garcia-Toro and Aguirre, 2007]. Notably, these two conflicting approaches have provided the two main fundamental aetiological grounds of depression.

Interdisciplinarity, which unfolded between psychology and sociology, was a relatively natural shift in focus. Childhood deprivation, family environments, and prenatal conditions, for example, were the main focus of psychoanalytic and family systems theories in psychology. Each of these were also shown to have larger contextual and cultural gradients that were revealed in sociological analyses. But it was more difficult to take biomedical aspects of depression into account and integrate them into psychosocial inquiry (and *vice versa*). It also did not help that the *biopsychosocial model* proposed by George and Engel [1980] intended to offer a transdisciplinary approach to depression without a clear methodological picture of this union [McLaren, 1998]. It is nonetheless largely acknowledged that biological and psychosocial aetiological aspects of depression are not only present in the developmental pattern, but also interact with each other [Gonda et al., 2019].

Social science genomics

In contemporary research, a transdisciplinary approach to depression, which takes into account a complex interplay between different by nature causes of depression, has an empirical manifestation.

Thanks to recent developments in technology, statistics, data science, and so on, it is now possible to advance our modelling of the complex network of different risk factors for depression. The rise of genotyping and large-scale samples that include both genetic and environmental/behavioural information motivated empirical inquiries into long-standing debates on *nature vs. nurture*. It further resulted in the emergence of the field of social science genomics [Conley and Fletcher, 2018, Freese, 2018, Mills and Tropf, 2020], which aims to study complex human traits and behaviours (such as depression) by applying sociogenomic tools. These tools include statistical and computational techniques used in population/quantitative genetics, epidemiology, sociology, demography, and econometrics.

One of the core notions of social science genomics is that estimates of the percentage of variation in complex outcomes explained by genetic and environmental differences are likely to be context-specific; this means they vary systematically across different social conditions, policy environments, or subgroups of the population [Boardman et al., 2011]. Such notions have led to a growing body of research on depression where childhood adversity [Brouillard et al., 2019], psychosocial stress [Arnau-Soler et al., 2019], partner loss [Domingue et al., 2017a], for instance, are all potential modifiers of genetic influences.

Thus, a novel sociogenomic approach acknowledges that the development of depression has a highly compound nature involving both genetic and environmental mechanisms. Complex interplay between genes and environments can take the form of gene × environment interactions and correlations. Conceptually, gene × environment interactions provide an opportunity to reveal environment-dependent variation in genetic associations. This theory states that environmental conditions are potential buffers and/or stressors of genetic predispositions, which are liable to cause certain health outcomes [Seabrook and Avison, 2010]. Another aspect of the interplay, gene-environment correlations, is less often studied. Gene-environment correlations take place when genes are associated with variations in exposure to adverse or protective environments. The presence of gene-environment correlations indicates that social environments may be a causal variable on the path from genes to the development of depression.

1.3 Areas for future research

1.3.1 Advancing empirical understanding of depression

Overall, a sociogenomic approach to understanding mental health offers a great potential to provide a new sets of insights. Such insights can be of particular importance for the relatively understudied but highly relevant contexts, such as the UK. For example, the focus on gene-by-cohort interactions has the potential to shed light on how historical contexts shape polygenic prediction across different generations. For social science, the particular insight is whether a rise in the prevalence of depression at certain historical points across the twentieth century is driven by those who have higher polygenic risks of depression or independent of genetic risks. While this question has been explored in the context of different countries, one of the frontiers would be to investigate the UK as it has not been done before.

Moreover, the concepts of gene × environment interactions and correlations not only address the questions on relative importance of external and internal risk factors but also raise the question of whether there are additional mechanisms in the etiological picture of depression. For instance, it is known that genetic predispositions to depression increase the probability of experiencing high-risk social environments. The presence of gene-environment correlations gives rise to the notion that social environments are potentially present on the causal path from genes to depression, contributing to our understanding of the complex nature of depression. Another frontier then would be to consider new aspects of social environments, such as worklessness that creates a toll of stressful life conditions and puts people at higher risk for experiencing depression.

1.3.2 Recognising methodological problems

It is important not only expand on empirical evidence of the complex interplay between individual- and macro-level risk factors of depression, but also reveal methodological concerns. Increased availability of data has further contributed to the integration of genetics with social science. The possibility of the inclusion of genetic risks as an additional variable into conventional empirical models is promising, but it also gives rise to substantial methodological considerations. One of the important considerations that potentially can bias results is the notion of sample selection. In order to deliver on the potential for genetic samples to comprehensively grasp the multidimensional nature of phenomena, it is essential to understand recruitment procedures and possible selection for genotyping based on various socio-demographic and health factors. A growing body of literature shows distinct features of genetic samples that have a potential to skew the picture of results (for example, see Domingue et al. [2017b], Fry et al. [2017], Manolio et al. [2012]). Thus, further research is needed to expand this issue on larger number of available samples.

Another important frontier considers the modelling of joint genetic and environmental variations. The rise of polygenic scores has resulted in a surge of studies investigating the mediating and moderating roles of environments along with genetic confounding. Yet disentangling the relative importance of polygenic scores and environmental covariates is difficult. Thus, in order to advance our understanding of such complex human traits and behaviours as depression, we should improve our modelling potential and use statistical techniques that provide results which are not methodologically flawed.

1.4 Thesis overview

My thesis seeks to address the research frontiers highlighted in the previous section. The title of this thesis, '*Advances and current methodological problems in understanding depression*', is directly expressing my goals: I aim to build knowledge of individual- and macro-level predictors of depression and to address methodological problems in the interdisciplinary field of social science genomics. My thesis presents a collection of one descriptive, two empirical, and one simulation-based chapters that altogether portray these objectives. The following paragraphs outline these chapters, while Table 1.4.1 further presents a summary.

Accordingly, I will contribute to existing literature on the sociogenomic approach to studying depression in four different ways: (1) through a methodological assessment of selection in genotyping and genetic data, including its role in the polygenic prediction of depression; (2) by looking at the changing polygenic penetrance of depression and associated moderating patterns of birth cohorts and economic recessions; (3) by exploring the link between genetic predispositions to depression and instances of worklessness that put people at higher risk of experiencing depression; and (4) by investigating how endogenous selection bias leads to spurious associations and biased variance statistics in models with polygenic scores.

In Chapter 2, I provide a detailed descriptive analysis of those who were genotyped in one of the largest UK datasets for social science genomics researchers (the UK Household Longitudinal Study, or UKHLS). The goal of this chapter is to answer two main questions: (1) does the UKHLS genetic sample suffer from socio-demographic and health selection? If so, (2) does selection into genotyping bias the polygenic prediction of depression? More broadly, this chapter focuses on one of the primary concerns regarding the rise of genetic data availability: the issue of selection in genetic samples. This fundamental problem tends to be neglected in empirical studies. To date, there is no comprehensive analysis of the representativeness of the UKHLS genetic sample even though published studies have used it. There are considerations, such as the *healthy volunteer effect*, which were empirically tested in other genetic samples. There are also observations that genotyped participants are likely from a distinct socio-demographic background. This led me to hypothesise that the UKHLS genetic sample has a portion of selection. I test my hypothesis with a regression analysis that estimates the differential probabilities of being genotyped attributed to socio-demographic and health characteristics of the general UKHLS survey. I further link the issue of sample selection to the polygenic prediction of depression, and test whether the weighted scenario performs differently or has a distinct predictive power. In doing so, I also extensively cover the topic of polygenic scores and the methods used to construct them.

Results from this chapter suggest that genotyped participants are a selective population, at least to a certain extent. I find small-to-moderate differences between genotyped and non-genotyped participants: on average, the former tend to have higher educational attainment, live in urban areas, and have better general self-reported health. Findings also suggest that these differences are likely to be germane to a gene-by-cohort interaction studies because genotyped and non-genotyped participants have distinctive mortality trends resulting in mortality selection in the UKHLS genetic sample. Mortality selection is a phenomenon whereby the group of respondents who died before a certain threshold form a non-random entity [Domingue et al., 2017b]. In my case, this threshold is genotyping that occurred in 2010–2012. Accordingly, genotyped participants born earlier in the century are likely survivors (85+) and a non-representative subset of their respective cohorts. However, the sample selection does not affect the polygenic prediction of depression. It is also important to note this chapter covers methodological considerations for my empirical investigations in Chapters 3 and 4, motivating the implementation of weights and describing the polygenic prediction approach.

In Chapter 3, I move towards examination of whether birth cohorts and recessions moderate genetic influence on depressive symptoms among adults in the UK. I aim to answer the question of whether the polygenic prediction of depression varies by birth cohorts of the twentieth century within the UK. To do so, I perform multilevel Poisson regression analyses. The chapter also takes into account important historical contexts, such as economic recessions, as a potential source of variation of polygenic prediction across birth cohorts. Consequently, this paper further contributes to existing knowledge by providing a gene × cohort interaction analysis of depression in the UK.

Another important contribution is demonstration that economic downturns contribute to the gene-by-cohort variations. Crucially, findings from this chapter indicate that the increase in depres-

Table 1.4.1: Thesis outline

Chapter	Research questions	Core variables studied	Data and methods
Ch 2. Selection into genotyping and the polygenic prediction of depression	1. What socio-demographic and health factors are associated with the probability of being genotyped? 2. Does selection into genotyping bias the polygenic prediction of depression?	Dependent variables: genotyping status, GHQ depressive symptoms. Independent variables: gender, age, marriage, children, area, neighbourhood, education, economic activity, smoking, drinking, physical health, depressive symptoms, polygenic score, PCs.	Data: 1991-2015 UK Household Longitudinal Study general and genetic samples. Methods: T-tests, logistic, Cox, and Poisson regressions, Kaplan-Meier curves, genotype imputation, genetic QC.
Contribution:	First comprehensive assessment of selection in the UKHLS genetic sample and its role in polygenic prediction		
Ch 3. Changing polygenic penetrance on depressive symptoms among adults in the United Kingdom	1. Does the polygenic prediction of depression vary across birth cohorts in the UK? 2. Do economic recessions moderate gene-by-cohort variations for depression in the UK?	Dependent variables: GHQ depressive symptoms. Independent variables: depression polygenic score, birth cohort, recession, gender, age, PCs.	Data: 1991-2015 UK Household Longitudinal Study genetic sample. Methods: Multilevel Poisson regressions, weights.
Contribution:	Exploration of the moderating patterns of birth cohorts and historical exposures to economic recessions on the polygenic prediction of depression in the UK		
Ch 4. Worklessness, genetic risks of depression, and gene-environment correlations	1. Does a genetic predisposition to depression increase the probability of worklessness? 2. If so, to what extent does this relationship depend on the respondent's sex?	Dependent variables: Worklessness status (employed, unemployed, economically inactive). Independent variables: depression polygenic score, gender, age, PCs.	Data: 1991-2015 UK Household Longitudinal Study genetic sample. Methods: Multilevel multinomial regressions, weights.
Contribution:	Investigation into whether labour market experiences are endogenous to genetic differences in depression predispositions, and the disentangling of differences across sexes		
Ch 5. Heritable environments: bias due to conditioning on a collider in models with polygenic scores	1. Does the correlation between polygenic scores and covariates bias the results of interaction, mediation, and genetic confounding analyses?	Dependent variables: Complex human traits and behaviours. Independent variables: Polygenic scores, environmental and phenotypic covariates.	Data: Simulations. Methods: Graphic methods, OLS regression simulations.
Contribution:	Methodological considerations and improvements to the modelling of joint genetic and environmental variations that are linked to complex traits		

sive symptoms is especially profound for the cohort of Baby Boomers who also display a significantly higher genetic penetrance of depressive symptoms. I also found a suggestive moderation for the Generation X cohort, but this aspect of the findings should be further replicated in future releases of the data to allow a wider age range for this cohort. More importantly, it appears periods of economic recession have the potential to shape the polygenic prediction of depression across generations in a different manner: my analysis reveals that economic recessions weaken the polygenic prediction of depressive symptoms across some cohorts, and strengthen the polygenic penetrance in others. Such findings highlight the importance of a cohort-specific approach in understanding phenotypic and genetic variations along with their interplay.

In Chapter 4, I explore whether genetic predispositions to depression are associated with higher chances of experiencing worklessness by using multilevel modelling techniques. To address this association, I conduct a multinomial regression analysis of working age participants in the UKHLS genetic sample. Conceptualising different reasons for worklessness, I investigate incidents of unemployment, economic inactivity due to family care, and economic inactivity due to sickness and disability. More broadly, this chapter contributes to theoretical knowledge of gene-environment correlations (rGE). The findings indicate that higher values in the polygenic score for depression are associated with a higher probability of unemployment. Moreover, such rGE patterns are observed for both females and males. The chapter also detects a significant positive association between polygenic scores for depression and the likelihood of economic inactivity. The analysis further reveals the non-linear nature of this relationship: as the number of risk alleles increases, the individual likelihood of becoming economically inactive becomes disproportionally higher. Importantly, this non-linearity is driven mainly by women. Polygenic prediction is stronger for absence from the labour force due to sickness or disability, as compared to absence due to family care. Additional findings in Chapter 4 are directly linked to debates about selection into worklessness, which complicate our understanding of the relationship between the absence of a job and depression. The presence of correlation between an underlying vulnerability factor, such as genetic risk and worklessness, is further evidence that job loss is not exogenous to depression susceptibilities. It implies that the relationship between mental health and unemployment is a complex one where the incidence of job loss increases the risk of developing depression. At the same time, job loss *per se* is a product of direct and indirect selection of prior-to-job- loss depression susceptibilities.

Chapter 5 makes an important contribution to the modelling of joint genetic and environmental variations linked to complex traits such as depression. We unpack methodological considerations that have not been previously addressed. The research question for this chapter is whether and how the presence of a correlation between polygenic scores and covariates in regression models biases the results of gene × environment interaction, mediation, and genetic confounding analyses. In broader terms, the rise of polygenic scores has resulted in a surge of studies investigating the mediating and moderating roles of environments along with genetic confounding. Yet disentangling the relative importance of polygenic scores and environmental covariates is difficult. To explore this question, we use graphical and simulation methods. We show that heritable covariates in regression models with polygenic scores as independent variables actually introduce an endogenous selection bias. This results in the problem of conditioning on a collider, which in turn leads to spurious associations and effect sizes. Collider bias is an important statistical problem that destabilises regression models and it can arise for a variety of reasons, including sample selection and attrition [Munafò et al., 2017].

Results from this chapter demonstrate that the degree of bias depends on the strength of not only the gene-covariate correlation, but also hidden heterogeneity linking environments with outcomes. This is true regardless of whether the main analytical focus is mediation, confounding, or gene × environment interactions. We offer potential solutions for obtaining unbiased estimates, highlighting the importance of causal inference methods and techniques. The issue itself was largely unrecognised before this study, which makes the focus of the chapter particularly important to the interdisciplinary field of sociogenomics. Taken together, we call attention to and urge further caution in fitting and interpreting models with polygenic scores and non-exogenous environments or covariates. We demonstrate how spurious associations and effect sizes are likely to arise, advancing our understanding of existing results and calling for caution in prospective studies.

1.5 Methodology

Each of four chapters in my thesis follow different analytical approaches and techniques in order to address the respective research questions. I thus describe the methods in each chapter separately; however, Table 1.4.1 summarises core variables studied and statistical methods used. The first empirical chapter (chapter 2 of this thesis) also covers the methodological agenda for Chapters 3 and 4 to motivate the inclusion of weights in the construction of polygenic scores. Thus, the sections below present a brief summary of the data and terminology used throughout the thesis.

1.5.1 Data

My thesis addresses a number of research questions with different scopes of analysis. However, the geographical scope of my investigations stays the same across chapters to cover the United Kingdom. I use one particular dataset, the UK Household Longitudinal Study (UKHLS). Though there are other genetic samples available in the UK, the UKHLS is the best source to address my research questions due to its rich information on socio-economic factors and representativeness. The UKHLS is a well-known and widely used longitudinal survey based on a national multi-stage sampling design [Benzeval et al., 2014, McFall et al., 2014]. It has the British Household Panel Survey (BHPS) subsample with 18 available waves of data spanning 1991 to 2008, and the General Population Sample (GPS) with available data starting in 2009. The UKHLS covers around 40,000 households in England, Scotland, Wales, and Northern Ireland. In 2010–2012 (Waves 2 or 3), the UKHLS invited adult respondents to take part in genotyping following data collection for the main annual survey. After the release of its genetics data, the UKHLS became a potential and valid data source for sociogenomics researchers. The genetic sample contains around 10,000 people, all adult members of households from both the BHPS and GPS subsamples of the UKHLS. The most recent and modified version of the genetic data release includes 9,944 individuals (5,568 women and 4,376 men).

Data protection notes

Since genetic data is rather sensitive, I provide some data protection notes. Procedures with data involve ensuring data protection and participant anonymity is maintained. For data sharing, I follow the US-based protocols from the NIH (National Institute of Health) on Genomic Data Sharing for which details can be found at: http://gds.nih.gov/08institutions.html. These guidelines ensure the anonymity of participants and represent higher standards in genetic research.

Considering participant anonymity, all data is anonymised and respondents are exclusively identified by an ID code. No researchers have access to respondent names or identifying information at any time. This means the following material is removed: names, all geographical subdivisions smaller than a region, all elements of dates (except years), telephone or fax numbers, email addresses, any identifying medical record numbers, IP address numbers, and biometric identifiers. Data is stored anonymously on a secure, password-protected institutional server at the University of Oxford.

1.5.2 Sample

In my UKHLS-based chapters, a common feature of analytical samples is the inclusion of the participants aged 16 and older. This decision is based on a specific feature of the depressive symptoms scale – namely, it is not valid measurement tool for children and teenagers younger than 16 years old. Only adult members of households (16+) were genotyped, which further contributed to this decision. But in the Chapter 4, where I focus on worklessness, I also exclude from my analysis students, those who are on maternity/paternity leaves, and retired respondents.

1.5.3 Terminology

Depression

Symptomatic data for depression is available in the UKHLS in the form of two scores, the GHQ and the SF-12. Both of these scores follow a traditional approach and have been assessed by measuring individual psychological state through item-based questionnaires with Likert scales. For the main body of my analysis, I use the GHQ score as it is one of the most widely used and consistently observed during the full survey period. The SF-12 score was included in questionnaires later, only becoming consistently present after 2009.

The GHQ was developed as a tool to screen for non-psychotic mental health problems [Goldberg et al., 1972]. Modelled symptoms primarily cover depression and anxiety disorders. The score is a constructed sum of 12 indicators such as usefulness, decision-making, unhappiness, confidence, self-worth, the ability to face problems, the ability to take joy in day-to-day activities, concentration, loss of sleep, the ability to overcome difficulties, and being under a strain. Each item in the GHQ asks the respondent to rate the degree of their symptoms from less than usual to more than usual. The rating scale ranges from 0 to 36 wherein higher scores indicate a higher severity of depression experiences. This score has been shown to be a valid and reliable instrument for detecting depression in the general population [Lundin et al., 2016].

Genetic predisposition to depression

In this thesis, the genetic predisposition to depression is conceptualised through polygenic scores for depression. A polygenic score is a tool for quantifying the genetic contribution to phenotypes [Wray et al., 2018]. In other words, it is a single-value indicator of the individual predisposition to a certain genetic trait. It is calculated as the sum of genome-wide SNPs weighted by a respective effect size that comes from summary statistics of a GWAS [Choi et al., 2018, p. 2]. For Chapters 2, 3, and 4, polygenic scores were constructed using the recent GWAS discovery of depression from Howard et al. [2019]. The construction of polygenic scores was performed using PRSice 2.0 software [Choi and O'Reilly, 2019]. Respondents with higher polygenic scores reported more depressive symptoms during follow-up.

Birth cohorts

In order to model broad historical contexts and shifts different UK birth cohorts were exposed to, I use the conceptualisation scheme described in Thomson and Katikireddi [2018]. There are five cohorts distinguished in this scheme and it also reflects conventional classification of demographic cohorts of 20th century. The first two cohorts are devoted to people who were exposed to the World Wars – a WWI cohort, born between 1916 and 1930, and a WWII cohort, with birth years in 1931–1945. To reflect the period of the baby boom, the demographic cohort of those born in 1946–1964 is distinguished as Boomers. Thereafter, there is Generation X – people born in 1965–1980 – followed by Millennials, or those who were born between 1981 and 1990. As genotyping was performed in 2010, it is not feasible to include Generation Z (or Zoomers) – those who were born in the 2000s. Additionally, due to the small number of survey participants from WWI cohort in the UKHLS genetic sample, I merge this with the WWII cohort, resulting in one World Wars birth cohort.

Economic recessions

I consider the periods of Early 1990s recession and Great Recession as economic recessions. During both recessions unemployment rates rose by minimum of 7% peaking at around 10% at the hardest-hit quarters [Jenkins, 2010]. Thus, 1990, 2008-2010 survey years are treated as recessive.

Worklessness

Chapter 4 of this thesis will focus on modelling the incidents of worklessness. In this thesis, worklessness is definied as instances of unemployment or economic inactivity (i.e. those not in labour force). Unemployment is conceptualised by the following ILO definition: the situation where an individual who has reached working age is unable to acquire a job and is actively in search of fulltime employment [Clegg, 2016, Hussmanns et al., 1990]. People who have not looked for a job in the past four weeks are treated as economically inactive, and there are two reasons for economic inactivity in the UKHLS sample - disability and family care which I also differentiate in my analyses.

Endogeneity

The main concept of Chapter 5 is endogeneity. Endogeneity is defined through situations in which an independent variable correlates with the error term of the regression model [Wooldridge, 2010]. In this thesis, I describe instances of endogenous variables in models with polygenic scores where endogeneity arises from the correlation of an explanatory variable with the model's error term through unmeasured confounders.

Chapter 2

Selection into genotyping and the polygenic prediction of depression

2.1 Abstract

This chapter covers methodological considerations for my empirical investigations in Chapters 3 and 4. My research aims to investigate whether there is socio-demographic and health selection in the UKHLS genetic sample and whether genotype selection leads to bias in the polygenic prediction of depression. Accordingly, this chapter focuses on one of the primary concerns about the rise of genetic data availability: the issue of selection in genetic samples. This fundamental problem tends to be neglected in empirical studies. No comprehensive analysis of the representativeness of the UKHLS genetic sample has occurred to date, despite the fact that published studies have used this sample. There are some considerations for representation, such as the healthy volunteer effect (which was empirically tested in other genetic samples). There are also

observations that genotyped participants are likely from a distinct socio-demographic background. This leads me to hypothesise that the UKHLS genetic sample has a portion of selection. I test this hypothesis by employing a regression analysis that estimates differential probabilities for genotyping attributed to the socio-demographic and health characteristics of the general UKHLS survey. I also show that the selection issue results in mortality selection in the UKHLS sample. This is a known source of potential bias in the cohort analysis (performed in Chapter 3). I further link the issue of sample selection to the polygenic prediction of depression, testing whether the corrected scenario performs differently and has a distinct predictive power. In doing so, I extensively cover the topic of polygenic scores and the methods used to construct them.

2.2 Introduction

One of the main difficulties in tackling different stages of depression is the complexity of causes that play a role in its development. Depression can be triggered by socio-economic factors, such as educational attainment [Lee, 2011], job loss [Drydakis, 2015, Paul and Moser, 2009], and recession [Frasquilho et al., 2015, Jahoda, 1988]. Depression has also been shown to have a genetic basis: twin studies suggest that the heritability of depression is 37% [Sullivan et al., 2000]. Quantitative genetics estimates, e.g. SNP heritability, vary from 2% to 9% [Howard et al., 2019, Okbay et al., 2016, Wray et al., 2018]. Following the GWAS approach, which is one of the most widely used tools in genetics, it has been shown that depression is a highly polygenic trait [Wray et al., 2018]. This has further resulted in the rising use of polygenic scores for depression across research fields. However, one issue that is usually overlooked is the selectiveness of genetic samples and whether selection affects the quality of polygenic prediction. In order to deliver on the potential for genetic samples to comprehensively grasp the multidimensional nature of phenomena, it is essential to understand recruitment procedures and possible selection for genotyping based on various socio-demographic and health factors. A growing body of literature shows distinct features of genetic samples that generally skew the picture of results.

Accordingly, this chapter investigates whether genotype selection leads to bias in the polygenic prediction of depression. The research question generates a set of goals to pursue. To begin, I show that the selection issue (caused by both recruitment criteria and respondents' consent) results in socio-demographic and health differences between genotyped and non-genotyped respondents. These differences are further linked to mortality selection. Even though the UKHLS genetic data claims to represent the white subpopulation with European ancestry based in the UK, possible selection mechanisms should be addressed as genotyping has not been performed randomly.

The second part of the chapter covers polygenic score analysis and assessment of whether correction for selection (addressed in the first part) changes the polygenic prediction of depression. Genetic data can be particularly challenging to work with as it requires special software skills along with significant computational resources. Thus along with the prediction assessment, I want to discuss decisions made about methodology within the process of polygenic score construction. Here, I also critically assess the quality of genetic markers and cover the imputation procedure. My main goal is to achieve higher precision in polygenic prediction, so detailed discussion of analysis is needed.

My motivation for performing such assessments is based on several points. First, there are no comprehensive data quality reports available for the UKHLS genetic dataset. Second, genetic quality controls and the procedures of DNA imputation are new to the sociological field. They should thus be described in detail, particularly in an interdisciplinary approach. Third, I wish to complement a common set of quality procedures with analysis of the representativeness of the sample. This is not yet perceived as a necessary step. However, I believe it is an important contribution that sociology can make to sociogenomics.

My analysis is structured as follows: I start by describing the procedure for genotyping and the range of eligibility factors that it implies. I further link data-generating decisions to possible selection issues. I assess the differential probabilities for genotyping across the socio-demographic and health characteristics of participants. I show that a lack of weighted correction leads to overrepresentation of educated individuals from urban areas. This overrepresentation likely introduces a *healthy volunteer* bias to analysis. I also link the discovered differences to a mortality bias, demonstrating that the genotyped participants are more likely to live longer than those who were not genotyped. I then turn to a discussion of the methodology and techniques behind polygenic score construction. I describe genetic quality controls to show the proportions of measured genetic markers, as well as the participants that did not pass control tests. After that, I detail the main concepts of genotype imputation. Finally, I show that constructed polygenic scores can significantly predict symptoms of depression in the UKHLS sample.

2.3 UKHLS data overview and genotyping

The UK Household Longitudinal Study (UKHLS) is a well-known and widely used longitudinal survey. Built from a national multi-stage sampling design (mostly through stratification and clustering), it covers around 40,000 households in England, Scotland, Wales, and Northern Ireland [Buck and McFall, 2011a]. Annual interviews cover a rich set of questions related to health, socio-economic conditions, and transitions along with family trajectories.

In 2010–2012 (Waves 2 or 3), following the annual main survey data collection, UKHLS invited adult respondents to take part in a nurse health assessment. The purpose of the nurse visit was to collect anthropometric information; to check blood pressure, grip strength, and lung function; and to take blood samples. The latter were used to produce a wide range of biomarkers that were included in the public access version of the data, and to build up the genotyped sample. Notably, respondents were asked to consent to the use of their blood samples for DNA extraction.

These nurse visits took place five months after the survey interviews. Nurse visitations of the British Household Panel Survey (BHPS) subsample of the UKHLS occurred in wave 3, while visitations of the main General Population Sample (GPS) occurred in wave 2 [McFall et al., 2014, p. 5]. A number of design decisions were made to implement nurse visits, and the full list of eligibility criteria can be found in the working paper by McFall et al. [2012]. Here, I list some eligibility criteria that are related to analytical decisions detailed in my analysis of representativeness. To be eligible for a nurse assessment, the respondent must have completed a full face-to-face interview in the corresponding wave, should be aged 16 or over, should live in England, Scotland, or Wales, and should speak English.

The study design further excluded people who were physically or mentally unable, pregnant women, and those who had moved to another location or could not be contacted for other reasons [McFall et al., 2014, pp. 8-13]. Further selection occurred in the blood sample routine, which did not sample respondents who had admitted they were HIV positive, those with hepatitis B or C, and those with a clotting or bleeding disorder (such as haemophilia or low platelets) [Benzeval et al., 2014, p. 9]. The response rate among eligible participants was 71.4% [Benzeval et al., 2014, p. 8]. More than half (58.6%) of eligible participants provided at least one measure during the health assessment. 38.2% of eligible participants provided blood samples [McFall et al., 2014, p. 8].

After the release of its genetics data, the UKHLS became a potential and valid data source for sociogenomics researchers. The genetic sample contains around 10,000 people, all of whom were adult members of households from the main Understanding Society survey. Originally, 10,484 adult members of households were selected for genome-wide array genotyping. Genotyping was performed with the Illumina Infinium HumanCoreExome BeadChip, which is considered an effective tool (for further details see `www.illumina.com`). In the initial quality control tests, 9,965 out of 10,484 individuals passed the checks [Prins, 2015]. However, the most recent and modified version of the data release includes 9,944 individuals (5,568 women and 4,376 men) [Hughes, 2018].

Genotyping is a procedure that aims to identify genetic variants across individuals. It is the most widely used method for decoding DNA. But it should not be confused with another method of decoding, DNA sequencing, as these two terms often appear in the literature. As distinction between these two concepts is an important issue that is further linked to imputation, a brief clarification is warranted. While genotyping implies data collection from specific locations in DNA (i.e. microarrays) known to be significant, sequencing presumes the detection of all SNPs in DNA. Even though the sequence approach produces more data without further need for imputation, it is an extremely time-consuming and resource-demanding process. This has led to the growing popularity of genotyping which, when combined with imputation techniques, is an effective tool for genetic analysis [Pistis et al., 2015].

2.4 UKHLS genetic sample representativeness

The rapid growth of genotyping and biomarkers has provided a wealth of new opportunities to the science community. However, excitement regarding new genetic insights and frontiers within an interdisciplinary focus of research has moved rapidly to discussion of generalisability.

The issue of generalisability is directly linked to the relevance of genetic discoveries. Genotyping in surveys is usually a non-random process. There are various selection mechanisms involved, starting from the eligibility of participants and recruitment criteria and ending with the personal decision to be genotyped – all these factors played a role in forming the UKHLS genetic data. Accordingly, these factors lead to the notion that genotyped samples are distinct groups of people that may or might not be similar to the general population. The assessment of selection importance is thus a vital step to understanding the relevance of any findings: are they indicative of broader trends or rather something specific, observed only in the genotyped group?

These points are also important for genomic discoveries and can be directly linked to both common and rare genetic variants. If the former are of critical importance to the development of a trait, the findings are likely to be valuable for a broader population. But if the latter make a profound contribution, the probability of observing rare variants in a homogenous sub-group is rather small. As a consequence, the findings will be non-generalisable.

The importance of sample selection in genomics has been widely addressed regarding ethnic diversity issues [Mills and Rahal, 2019], mortality selection [Domingue et al., 2017b], and socio-economic and health disparities [Fry et al., 2017]. I would like to contribute to this discussion by demonstrating that analysis of representativeness and possible selection of the genotyped samples should be included as an additional step in quality control procedures. Focus on selection when working with genetic samples should be further popularised and stressed: we should seek to achieve diversity not only based on ancestry, but also taking into account the socio-demographic and health profiles of population-based genetic samples.

In the following sections, I quantify the differences in socio-demographic and health factors associated with probabilities of genotyped status in the UKHLS survey. Because the main interest is to eliminate genotyping sampling trends that introduce biases to the associations (and particularly to gene × environment interactions), I further focus my investigation on the implementation of weights. I show that blood sampling weights provided by the UKHLS research team eliminate most of the disparities. Accordingly, they should not be ignored in any analysis based on the genetic sample. Notably, this stage of analysis was performed using publicly available data that can be downloaded from https://www.ukdataservice.ac.uk.

2.4.1 Factors associated with genotyping: socio-demographics, lifestyle, and health

I assessed individual differences between non-genotyped and genotyped participants. To do so, I provided t-statistics for mean differences. These statistics were accompanied with a regression analysis that investigated whether characteristics such as sex, age, marital status, economic activity, educational attainment, drinking, and smoking, along with self-reported general health, are significant predictors of the likelihood of being genotyped.

Measures

Following the literature on socio-economic selection and the *healthy volunteer effect* in genetic samples (for example, in Fry et al. [2017], Manolio et al. [2012]), I considered the following set of demographic factors and measures of socio-economic status (a comprehensive description of the variables can be found in Tables A.1.1, A.1.2, and A.1.3 in Annex A):

- sex (1=female; 0=male)
- age: age in years at a nurse assessment wave (wave 2 for GPS and wave 3 for BHPS participants; range is 16-99 y.o.);
- marital status: whether a respondent ever reported being married (1=yes; 0=no);
- having children: whether a respondent ever had a (biological) child (1=yes; 0=no);
- area: rural/urban area of living at a nurse assessment wave (1=rural; 0=urban);
- neighbourhood cohesion: the Buckner's Neighbourhood Cohesion Instrument (mean α=.87). Scores are provided by the UKHLS team in waves 1 and 3; they range from 1 to 5, with higher values indicating greater cohesion.
- educational attainment: highest qualification reported (from 1=no qualification to 5=higher tertiary);
- economic activity: labour force status at a nurse assessment wave;
- smoking: whether a respondent ever reported smoking (1=yes; 0=no);
- drinking: mean of how often respondent reports having an alcoholic drink for each survey year (from 1=almost every day to 8=not at all);
- general health: self-reported health (from 1=excellent to 5=poor) at a nurse assessment wave (wave 2 for GPS and wave 3 for BHPS participants);
- mental health: General Health Questionnaire (GHQ) 12-item score for depressive symptoms at a nurse assessment wave (ranges from 0 to 36 where higher values indicate greater depressive symptoms).

These measures allow me to not only assess potential selection due to various respondent characteristics, but also address the potential roots of the mortality selection that I discuss later in this chapter. Lifestyle, health, economic activity, educational attainment, characteristics of residential area where the respondent lives, and marital status are all linked to mortality disparities.

Table 2.4.1 provides the means and standard deviations of the variables by genotyping status along with t-statistics for differences. I considered the 26,382 respondents in the UKHLS study who provided the information and satisfied the eligibility criteria. To make the groups comparable, I restricted the non-genotyped group of participants from a white ethnic background born within the same birth window as those who were genotyped. I also focused my analysis on genotyped and non-genotyped participants for whom weighting information is available. I am particularly interested in the weighting scheme that corrects for selection in blood samples and adjustment for non-response [1]. As described earlier, there were a number of analytical decisions regarding eligibility criteria that made the genotyping non-random. Moreover, clustering and stratification are important elements of the UKHLS sample designs (both general survey and genotyped sub-sample) that affects the calculation of standard errors. I thus expect that the weights, provided by the UKHLS team, correct for disproportional probabilities to be genotyped, allowing for production of reliable standard errors. These weights are also publicly available. The weighting scheme further provides additional adjustment for non-response that takes into account drop-outs at each phase of obtaining a genotyped subsample, such as during the consent stage, blood taking *per se*, and analytes distraction. Therefore, weighted regression analysis adjusts for differential probabilities in selection for genotyping as well as for differences in response rates between subgroups of the sample [McFall et al., 2014].

As Tables 2.4.1 and 2.4.2 demonstrate, there are slightly more women in the non-genotyped and genotyped groups following both weighted and unweighted summary information. This difference

[1] The UKHLS team conducted the analysis of weights to enable estimations that would be representative of the general population. See the detailed discussion on weights derivation in [McFall et al., 2014, p. 51, 55]. I use cross-sectional weights that adjust for recruitment criteria and non-response.

Table 2.4.1: Unweighted descriptive statistics of analytic sample, by genotyping status with t-statistics of differences

	All participants		Genotyped		Non-genotyped			
	Mean	SD	Mean	SD	Mean	SD	*t*	*p*
Sex (female)	.561	.496	.562	.496	.561	496	-.10	.92
Age	48.43	17.71	51.87	16.57	47.25	17.93	-18.60	.00
Ever married	.772	.419	.836	.371	.750	.433	-14.42	.00
Ever had child	.758	.429	.804	.397	.741	.438	-10.41	.00
Rural area	.269	.444	.258	.437	.273	.446	2.47	.01
Neighbourhood	3.619	.645	3.673	.619	3.601	.652	-8.00	.00
Education								
No qualifications	.164	.370	.160	.367	.166	.372	1.11	.26
Lower secondary	.369	.483	.374	.484	.367	.482	-.99	.32
Higher secondary	.146	.353	.137	.344	.149	.356	2.43	.02
Lower tertiary	.103	.304	.115	.319	.099	.299	-3.68	.00
Higher tertiary	.218	.413	.214	.410	.219	.414	.78	.43
Economic activity								
Employed	.561	.496	.556	.497	.562	.496	.88	.38
Unemployed	.048	.213	.037	.188	.052	.221	4.00	.00
Retired	.241	.428	.288	.453	.224	.417	-10.63	.00
Student	.052	.222	.029	.167	.060	.237	9.88	.00
Inactive	.099	.299	.090	.286	.102	.303	2.84	.00
Lifestyle								
Smoking	.643	.479	.641	.480	.643	.479	0.28	.78
Drinking	4.446	1.833	4.302	1.832	4.494	1.831	7.46	.00
Health								
General health	2.609	1.036	2.585	1.016	2.617	1.043	2.216	.03
Depressive sympt	11.110	5.368	11.001	5.339	11.147	5.378	1.92	.05
No. of participants	*26,382*		*6,719*		*19,663*		*df* = *26,380*	

Table 2.4.2: Weighted descriptive statistics of analytic sample, by genotyping status with t-statistics of differences

	All participants		Genotyped		Non-genotyped			
	Mean	SD	Mean	SD	Mean	SD	*t*	*p*
Sex (female)	.552	.497	.557	.497	.534	.499	-1.62	.11
Age	48.14	18.19	48.19	17.10	47.95	18.94	-.39	.70
Ever married	.758	.429	.761	.427	.748	.434	-.86	.39
Ever had child	.748	.434	.761	.427	.707	.455	-3.52	.00
Rural area	.228	.419	.229	.420	.224	.418	-.25	.81
Neighbourhood	3.587	.646	3.597	.646	3.555	.648	-1.84	.07
Education								
No qualifications	.165	.371	.161	.367	.178	.382	1.50	.13
Lower secondary	.376	.484	.384	.486	.352	.478	-2.19	.03
Higher secondary	.146	.353	.147	.354	.143	.350	-.33	.74
Lower tertiary	.101	.302	.102	.303	.098	.298	-.44	.66
Higher tertiary	.212	.409	.207	.405	.229	.420	1.74	.08
Economic activity								
Employed	.565	.496	.569	.495	.550	.497	-1.18	.24
Unemployed	.051	.220	.050	.219	.052	.222	.21	.84
Retired	.240	.427	.238	.426	.248	.432	.80	.43
Student	.044	.205	.041	.198	.055	.228	1.71	.09
Inactive	.100	.300	.102	.303	.095	.293	-.76	.45
Lifestyle								
Smoking	.652	.476	.652	.476	.653	.476	0.09	.93
Drinking	4.433	1.820	4.416	1.810	4.491	1.850	1.28	.20
Health								
General health	2.595	1.038	2.600	1.036	2.577	1.043	-.73	.47
Depressive sympt	11.153	5.503	11.152	5.505	11.155	5.496	.02	.99
No. of participants	*26,382*		*6,719*		*19,663*		*df* = 1,261*	

**Design df which takes into account survey stratification*

is not statistically significant. In the unweighted scenario, genotyped participants are slightly older and on average 52 years old; non-genotyped individuals are on average 47 years old. The genetic sample includes more married people who ever had children. This trend is observed in the weighted means as well, but with a smaller difference. Differences in residential areas are more notable in the unweighted statistics. There are less individuals from rural areas with a smaller neighbourhood cohesion index in the genotyped group, but these differences are not distinguishable once weights are considered. Notably, differences between genotyped and non-genotyped participants regarding demographic characteristics are statistically significant (*.05* threshold) following an unweighted scenario. Weighted means are no longer statistically different with the exception of the having a child (where the difference is lower, but still shows significance).

There are some differences in educational attainment between genotyped and non-genotyped groups in the unweighted means. For example, there are more participants with higher secondary and lower tertiary qualifications in the genotyped group. However, these differences are not present once the weights are employed. Labour force participation status likewise displays differences that are further smoothed in the weighted analysis. Unweighted proportions indicate fewer unemployed and inactive people in the genetic sample (by around 1%-1.5%), fewer students (by 3%), and more retired people (by 6%) in comparison to non-genotyped participants.

In terms of lifestyle factors, both groups are quite similar in smoking experiences: around 64% of respondents in both groups reported they had smoked cigarettes at some point in their lives. The difference in the frequency of alcohol consumption is rather small, but statistically significant in unweighted comparison. The general health situation in the whole analytic sample is somewhat good, with a mean of 2.6 in the sample. Genotyped participants also reported better overall health and better scores on the depressive symptoms scale, but these differences are not distinguishable once the weights are applied. This trend also reflects the general picture where most potential differences between genotyped and non-genotyped groups are smoothed out with weights. Summary statistics for male and female respondents are provided separately in Tables A.2.1, A.2.2, A.2.3 and A.2.4 in

Annex A. Information on missing values can be found in Table A.1.4 in Annex A.

Methods

I further quantify the role of background factors as relates to the probability of being genotyped using logistic regressions:

$$Pr(respondent\ i\ genotyped|X_i) = \frac{exp(X_i'\beta)}{1 + exp(X_i'\beta)}$$

The choice for the predicted matrix X depends on the type of selection I am addressing, such as socio-economic, lifestyle, or health covariates. I am primarily interested in the probability of the inclusion of the genetic sample as a function of the factors described in Table 2.4.1. I also consider an alternative model that tests the sensitivity of the selection model by including blood sample weights. I further employ this selection model to conduct a sex-specific analysis for the weighted and non-weighted samples in order to address the issue of possible selective patterns across sex groups. This part of the analysis can be found in Annex A.

Results

Results from the regression analysis are displayed in Figure 2.4.1 and Table A.1.3 (in Annex A). No significant differences are revealed regarding sex, marital status, economic activity, neighbourhood cohesion, smoking, and depressive symptoms. However, genotyped participants are more likely to be older than those not genotyped (*β=.049, P<.001; β=-.000, P<.001* for squared term). In a regression with weights, this difference is around half the size and insignificant (*β=.021, P>.05; β=-.000, P>.05* for squared term). Also, people in the genetic sample people are more likely to have children than those in the non-genetic cohort (*β=.860, P<.05*). The weights inclusion does not correct for this difference, but makes it smaller (*β=.275, P<.001*). Accordingly, studies related to fertility and household composition including children should analyse the genetic sample more cautiously.

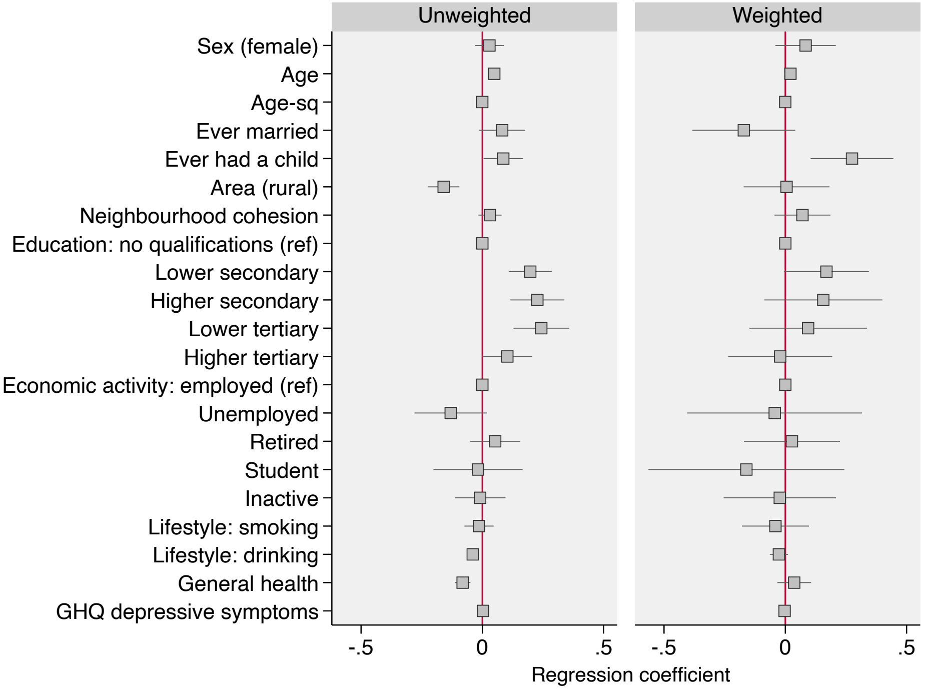


Figure 2.4.1: Coefficient plots for predictors of genotyping, for all respondents

Importantly, better education is a significant predictor of genotyping status. Higher (i.e. tertiary) education is associated with greater likelihood in appearing in the genotyped sample (*β=.243, P<.001* for lower tertiary and *β=.103, P<.05* for higher tertiary). Additionally, living in a rural area is associated with statistically significant lower chance of being genotyped (*β=-.159, P<.001*). The importance of education and area in relation to probability of genotyping indicates the socio-economic gradient of selection. However, these significant differences are not observed in the weighted regression model.

Important trends are also revealed in relation to lifestyle and health. Greater alcohol consumption is a significant predictor for genotyped status (*β=-.040, P<.001*). However, it is not observed in the model that corrects for the recruitment process. Better self-reported general health is likewise associated with greater likelihood in appearing in the genotyped sample (*β=-.081, P<.001* where higher values indicate poorer health). This factor is not significant once the weighting is applied.

The sex-specific analysis reveals further insights. In general, the trends are the same and the inclusion of weights smooths out differences in genotyping probabilities. Both sexes account for the observed differences described earlier, which include age, educational attainment, residential area type, alcohol intake, and self-reported general health (Table A.2.5 in Annex A). By contrast, women account for the greater probability in having children found among genotyped participants. As Table A.2.5 in Annex A displays, men in both genetic and non-genetic samples are equally likely to have children.

To conclude, there are a number of socio-demographic, lifestyle, and health differences between genotyped and non-genotyped UKHLS participants. A significant portion of these disparities disappears once correction for recruitment is employed. This thus indicates the importance of weights and their inclusion into models with polygenic scores as part of further empirical investigations.

Before I turn my discussion towards polygenic prediction, the following section conducts an additional analysis to investigate whether the UKHLS genetic sample suffers from a *healthy volunteer effect*. As the main focus of my thesis is a health-related phenotype and genotyped participants reported on average better general health, I conduct a more thorough analysis of the health histories of participants.

2.4.2 Healthy volunteer hypothesis

I expand the focus on UKHLS genetic sample characteristics by further quantifying whether genotyped participants can be distinguished by their health histories. This motivation is based on health-related criteria factors included on the eligibility list for blood sampling. Participants who passed the eligibility checks further volunteered for genotyping. It is known that such a set up can potentially lead to the *healthy volunteer effect*, where those who agreed to take part are likely in better health and more conscious of health-related behaviours than average [Delgado-Rodriguez and Llorca, 2004, Manolio et al., 2012].

Accordingly, I consider the following set of health determinants (a more detailed description of the variables can be found in Tables A.1.2 and A.1.3 in Annex A):

- body mass index (BMI) in kg/m2 measured at wave 1, which ranges from 10.1 to 74.4;
- blood pressure: whether a respondent was ever diagnosed as having a high blood pressure (1=yes; 0=no);
- diabetes: whether a respondent was ever diagnosed as having diabetes (1=yes; 0=no);
- heart: whether a respondent was ever diagnosed as having a heart condition, such as congestive heart failure, coronary heart disease, heart attack, myocardial infarction, or stroke (1=yes; 0=no);
- respiratory diseases: whether a respondent was ever diagnosed as having asthma, angina, or chronic bronchitis (1=yes; 0=no);
- cancer: whether a respondent was ever diagnosed as having cancer (1=yes; 0=no);
- depression: whether a respondent was ever diagnosed as having a clinical depression (1=yes; 0=no).

I considered 26,292 respondents who provided information and passed the eligibility criteria. Similar to the previous analysis, I constructed my sample by including participants from a white ethnic background who were born within the same birth window as those who were genotyped and for whom weighting information is available. Information on missing values can be found in Table A.1.4 in Annex A.

Table 2.4.3 below provides the means and standard deviations of the variables by genotyping status (summary statistics for male and female respondents are provided separately in Tables A.2.7 and A.2.8 in Annex A).

As Table 2.4.3 demonstrates, some health indicators vary notably across groups by status of genotyping. However, these differences are smoothed out once the weights are applied. For the whole

Table 2.4.3: Healthy volunteer hypothesis: descriptive statistics of analytic sample, by genotyping status with t-statistics of differences

	All participants		Genotyped		Non-genotyped			
	Mean	SD	Mean	SD	Mean	SD	*t*	*p*
Unweighted								
BMI	26.36	5.029	26.50	4.839	26.32	5.087	-2.49	.01
High blood pressure	.200	.400	.221	.415	.193	.395	-4.74	.00
Diabetes	.056	.230	.059	.235	.056	.229	-.92	.36
Heart condition	.056	.230	.051	.220	.057	.232	1.77	.08
Respiratory diseases	.174	.379	.166	.372	.177	.381	1.98	.04
Cancer	.041	.199	.047	.213	.039	.194	-2.81	.00
Clinical depression	.074	.262	.077	.267	.073	.261	-1.07	.28
Weighted								
BMI	26.27	4.968	26.33	4.920	26.06	5.111	-1.76	.08
High blood pressure	.195	.396	.199	.400	.182	.386	-1.58	.11
Diabetes	.056	.230	.054	.226	.062	.241	1.23	.22
Heart condition	.049	.217	.048	.214	.054	.226	.86	.39
Respiratory diseases	.173	.378	.172	.377	.176	.381	.31	.76
Cancer	.039	.193	.041	.198	.033	.178	-1.80	.07
Clinical depression	.077	.267	.078	.268	.076	.265	-.21	.83
No. of participants	*26,292*		*6,370*		*19,922*			

Note. Degrees of freedom for unweighted t-tests = 26,290; design df for weighted tests = 1,213

analytic sample, the mean BMI is 26.4. Approximately 20% of respondents reported a high blood pressure condition, 6% had diabetes, 5% reported heart conditions, 17% had respiratory diseases, and 4% had cancer. Around 8% of participants reported having clinical depression. There are slightly more cancer survivors in the genotyped group, with a statistically significant difference of .8%. On the contrary, less participants in the genetic sample reported respiratory diseases (by 1.1%, which is statistically significant at .05 threshold). It is also notable that genotyped participants, on average, a higher BMI and tend to report higher blood pressure.

The results from the regression analysis, where I assess whether health indicators are significant predictors of genotyping status, are displayed in Figure 2.4.2 and Table A.2.9 (in Annex A). Here, I include sex, age, age-sq., and educational attainment as covariates to control for basic socio-demographic characteristics. For most of the factors of interest, no significant differences are revealed. Furthermore, those differences that are significant become no longer distinguishable once the weighting is applied. In the unweighted scenario, only one factor displays a sign of health selection: having a heart condition (such as congestive failure, coronary disease, myocardial infarction, or the experience of a heart attack/stroke) is associated with a significantly lower probability of being genotyped (*β=-.283, P<.001*). Sex-specific analysis reveals that this trend is driven solely by males (Table A.2.9). Thus even though genotyped participants tend to report better general health, a more detailed analysis does not reveal specific health conditions that would drive it – with the exception of genotyped males, who have a significantly smaller prevalence of heart problems in comparison to non-genotyped participants.

All in all, there is a portion of certain differences regarding socio-demographic, health, and lifestyle characteristics between genotyped and non-genotyped UKHLS participants. I find small to moderate differences between genotyped and non-genotyped participants. For example, those who were genotyped are 16% less likely to live in rural areas and more likely to have higher educational qualifications in comparison to those not genotyped. Moreover, genotyped participants are 16% less likely

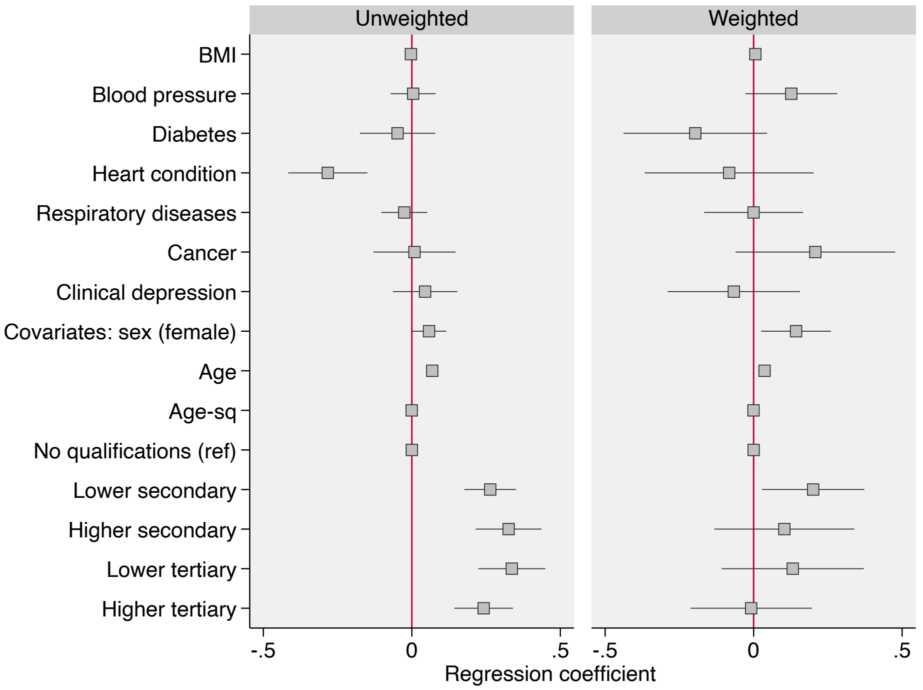


Figure 2.4.2: Coefficient plots for health profile models, for all respondents

to have bad general health and 28% less likely to have heart conditions. Since the disparities disappear once the eligibility correction is employed, the weights should also be used in further empirical studies based on the UKHLS genetic sample.

I also checked whether the probability of being genotyped differs among nurses by extracting inter-class correlations (ICC). Since at the last stage of recruitment it was nurses who played a role of interviewers, it is possible some of them recruited more than others. It would mean that the sample is generated not only by eligibility criteria and self-selection, but also by way of nurses themselves. The nurse ICC is quite modest, but statistically significant (*ICC=.05; 95CI=.04-.07*). It means that quantified differences among nurses are not likely a substantial factor in the UKHLS genetic sample and nurses did not play a role in respondents' decisions to participate in genotyping.

However, the samples will presumably always be different in some way from the general population. That is, with a reasonable N, the researcher will always find statistically significant differences. Hence the relevant question is not only how significant and large these differences are, but also whether these differences are germane to a particular analysis. To address this notion, I test whether UKHLS genetic sample has mortality selection (which is important for the gene-cohort study in Chapter 3) and whether the inclusion of weights re-shape polygenic prediction of depression (which is an important question for both Chapter 3 and 4). Two of these aspects are covered in the two following sections.

2.4.3 Mortality selection

Mortality selection is the phenomenon whereby a group of respondents who died before a certain threshold form a non-random entity. In my case, this threshold is genotyping that occurred in 2010–2012. Accordingly, those who were born earlier in the century and were genotyped are likely to be survivors (85+) and a non-representative subset of the respective cohorts. The issues of survivors and mortality selection are important in genetic epidemiology [Greenland, 2008] and were addressed in the Health and Retirement Study (HRS) [Domingue et al., 2017b, Liu and Guo, 2015, p. 726]. However, there is no such evaluation covering the UKHLS genetic sample. To fill the gap, I address the issue here. As the focus of Chapter 3 is the gene-by-cohort analysis of depression, it is also essential to address this possible source of bias.

Another motivation for performing such an analysis with respect to birth cohorts is the selection into genotyping by socio-demographic and health factors previously described. Educational attainment, marital status, general health, and heart conditions are all linked to mortality [Ben-Shlomo et al., 1993, DeSalvo et al., 2006, Franks et al., 2003, Martikainen and Valkonen, 1999, Thom, 1989]. Thus even when the focus is not a certain environmental exposure but a birth cohort in general, selection into the genetic sample imposes differences in survival chances. This means the genotyped cohorts have a distinct nature that might bias associations, in turn. This further leads to the inclusion of weights to correct these differences and obtain less biased estimates.

Table 2.4.4: Demographic characteristics of UKHLS sample by birth cohorts, genotyping status, and deaths

Cohort name	Birth years	N	N genotyped	% genotyped	N deaths	Deaths	
						% gen. sample	% non-gen. sample
All	1894-1995	77,706	9,944	12.8	4,494	3.3	6.1
WWI	1894-1930	6,081	431	7.1	2,449	24.8	41.5
WWII	1931-1945	11,432	2,106	18.4	1,329	7.1	12.7
Boomers	1946-1964	17,038	2,957	17.4	519	2.1	3.3
Generation X	1965-1980	20,604	2,740	13.3	146	0.4	0.8
Millennials	1981-1995	22,551	1,710	7.6	51	0.1	0.2

Survival models specification

I test mortality selection in the UKHLS sample by comparing survey participants who survived until DNA collection in 2010-2012 with those who did not. I compare survival differences between genotyped and non-genotyped participants using Kaplan-Meier survival curves [Kaplan and Meier, 1958] and Cox proportional hazard models [Cox, 1972]. Following the Kaplan-Meier method, I estimate cumulative time-to-event curves where the dependent variable is age at death and the main independent variable of interest is genotyping status. For Cox proportional hazards regression analysis, I model survival (i.e. the measurement of duration time) as the number of years between the first interview and the most recent available interview:

$$h(t|X) = h_0(t)exp(X'\beta)$$

where X is the covariates matrix - sex and age at first interview. $h(tX)$ is the hazard of mortality. The proportional hazards assumption is plausible once stratified by cohorts.

Results

Table 2.4.4 displays the demographic characteristics of UKHLS participants by incidents of death. Looking at its occurrence, one may notice a larger percentage of deaths among non-genotyped participants compared to the genotyped group.

Kaplan-Meier survival curves show that genotyped participants are more likely to live longer in comparison to the non-genotyped group (Figure 2.4.3). This trend is also observed for both sexes. Such increased longevity is not wholly unexpected, as genotyped respondents tend to be healthier than (and have different socio-demographic characteristics to) non-genotyped people.

Next, it is important to address whether survival curves are different across birth cohorts. Results from Cox regression models show that genotyping status is a significant predictor for surviving (Table A.3.1 in Annex A). Moreover, as Figure 2.4.4 below demonstrates, the differences are more profound across earlier birth cohorts. This is consistent with findings in the literature. Kaplan-Meier curves by sex and cohorts exploit the same trends (Figure A.3.1 in Annex A).

To sum up, I demonstrate that the UKHLS genetic sample is a distinct sub-group with a portion of socio-demographic- and health-related differences. The major methodological takeaway from this section is the critical importance of correction for recruitment, such as the blood sample weights in the UKHLS dataset. These weights should therefore be included in final models with polygenic scores. I showed that the selection issue in the genetic sample cannot be ignored and should be corrected in gene-cohort studies.

2.5 Polygenic score prediction

One of the main difficulties in tackling different stages of depression is the complexity of causes that play a role in its development. Genetics are believed to be a critical factor in understanding this mental health aspect. Though the role of genetics has not yet been exhaustively illustrated, some progress has been made. Through a consideration of GWAS, this section delves into the role of genetics in the development of depression.

The issue of the heritability of depression within behaviour studies has been a major area of sci-

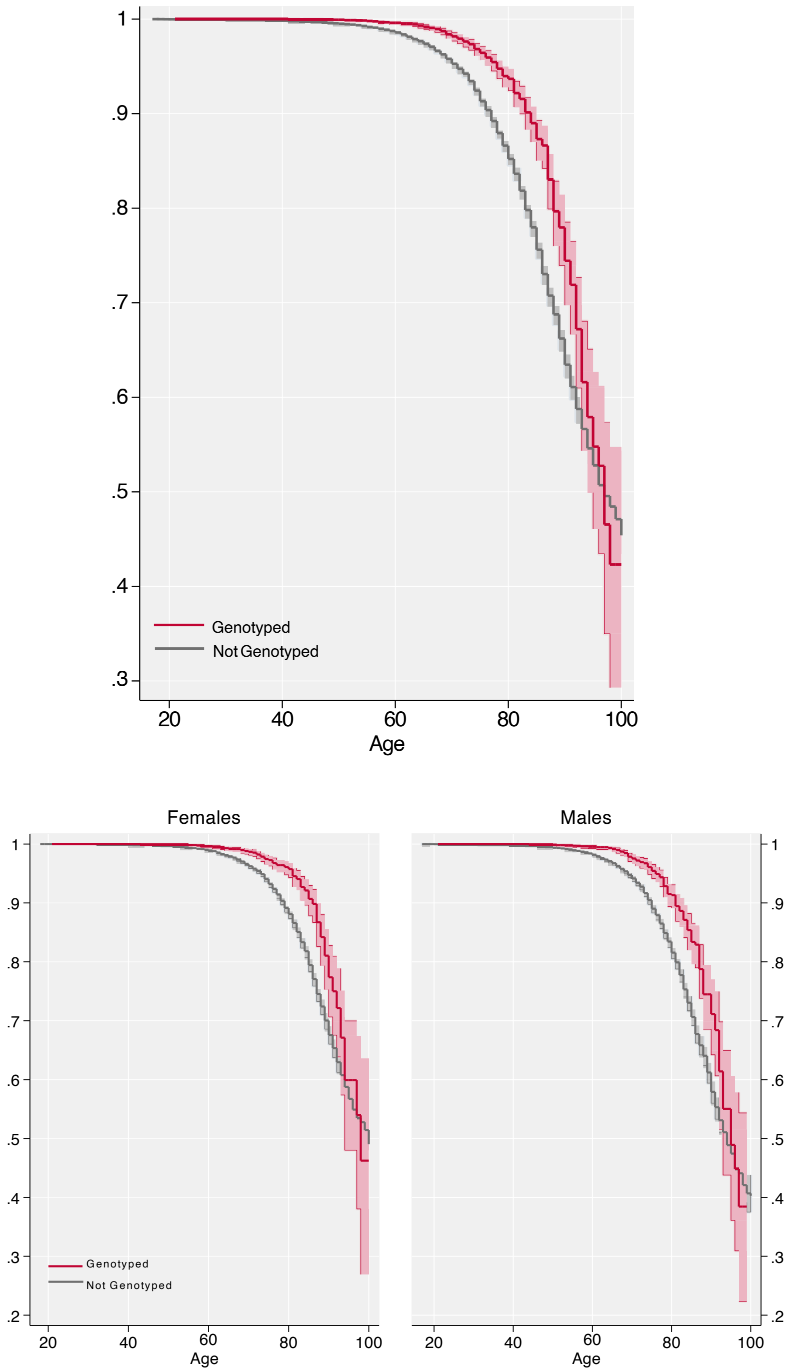


Figure 2.4.3: Kaplan-Meier survival curves for genotyped and non-genotyped UKHLS respondents, for all respondents and by sex, Y axis is a survival propensity

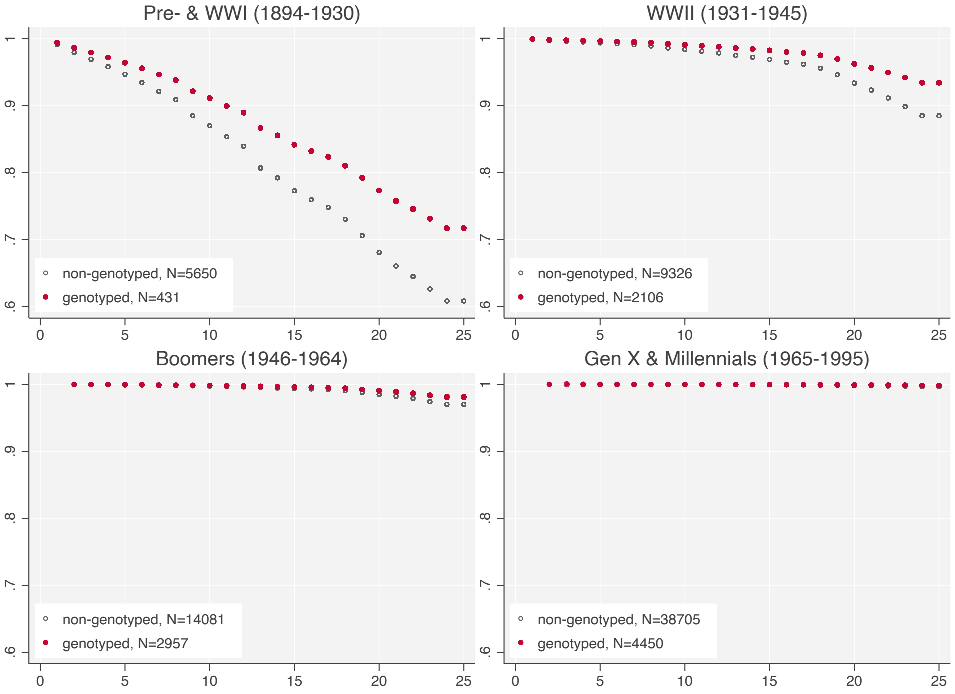


Figure 2.4.4: Cox survival curves for UKHLS respondents, by birth cohorts. Generation X and Millennials cohorts are merged due to small number of deaths in these cohorts, Y axis is a survival propensity and X axis is the number of years an individual is observed in the UKHL

entific interest for centuries. The earliest stages of research suggested that depression develops in families [Tsuang and Faraone, 1990]. Later, twin studies showed that if a person's parents suffer from depression, their risk of developing the illness ranges from 20% to 40% [Jansson et al., 2004, Keyes, 2005, McGue and Christensen, 2003, Sullivan et al., 2000]. But with the recent advances in molecular genetics, we are now addressing heritability by looking directly at the genetic structures accompanying existing knowledge from twin studies. One widely used approach is the so-called GWAS. The rationale behind this method is to screen all of the single nucleotide polymorphisms (SNPs) in the human genome and test their association with a certain phenotype. The Psychiatric GWAS Consortium (PGC), which takes a major interest in mental health disorders including depression, leads and coordinates the collaborative work in this field (and announces results). Multiple studies are looking at depression as a phenotype [Hek et al., 2013, Okbay et al., 2016, Terracciano et al., 2010], with the most recent ones conducted by Amare et al. [2020], Howard et al. [2019], Wray et al. [2018].

An important and critical insight that has changed conventional perceptions relates to polygenicity. It appears that for most common traits, no single gene is responsible for an outcome. Rather, the influence of genes is due to the small effects of many SNPs across the genome. This observation has even been proposed as a fourth law of behaviour genetics [Chabris et al., 2015] to accompany the three existing ones described by Eric Turkheimer [2000].

The notion of polygenicity also shifts the methodological prospects of research, particularly the conceptualisation of genes. In the 2000s, the dominant candidate-gene strategies started to seek stronger and stronger theoretical and empirical justifications [Moffitt et al., 2005]. The emerging polygenic score analysis appeared as a novel technique that made it possible to construct a single measure while taking into account the small effects of multiple SNPs [Plomin et al., 2009].

However, the first GWAS discoveries tended to be underpowered and were not able to find significant signals. An increase in the predictive power of the results from quantitative genetics was directly linked to methodological and technical advances in the field. In the example of depression, the earliest GWAS studies conducted in 2008 found only one SNP across the whole genome [Muglia et al., 2010]; in 2010/2011, the number of responsible variants increased to 10 [Rietschel et al., 2010, Shyn et al., 2011]. The most recent genetic discoveries from Howard et al. [2019] identified 102 independent variants, 269 genes, and 15 gene sets associated with depression, including both genes and gene pathways associated with synaptic structure and neurotransmission.

I include Figure 2.5.1 below, obtained from GWAS Catalog; it graphically demonstrates all known and discovered SNPs for depression located on respective chromosomes. Firstly, it further demonstrates the polygenicity of depression – all signals, which are indicated by dark blue points, are spread across the genome and located on different chromosomes. Secondly, it is hard to distinguish single locations even within chromosomes. Thirdly, there is limited evidence of sex-specific signals discovered up to date.

In terms of formal definition, the polygenic score is a single-value indicator of an individual's genetic predisposition to a certain trait. It is calculated as a sum of genome-wide SNPs weighted by a respective effect size that comes from the summary statistics of a GWAS [Choi et al., 2018, p. 2]. Sometimes polygenic scores are referred to as polygenic risk scores (PRS) or genetic risk scores (GRS). In my thesis, I use the term polygenic score (PGS).

Although polygenic scores have become quite popular, guidelines for performing this type of analysis are limited. I would argue that the absence of such guidelines leads to the heterogeneity of polygenic score performance observed by Erin Ware et al. [2017]. Such distinctive patterns in results indicate the necessity of presenting a clear picture of the research process – how scores were constructed and what decisions were made. Hence, this section of the chapter discusses polygenic score prediction and its construction.

Before polygenic scores are available for inclusion in the main models of interest, some initial steps are required to prepare and manage the data. There are generally three main preparatory stages: quality control, imputation, and polygenic score construction *per se*. The following sections are struc-

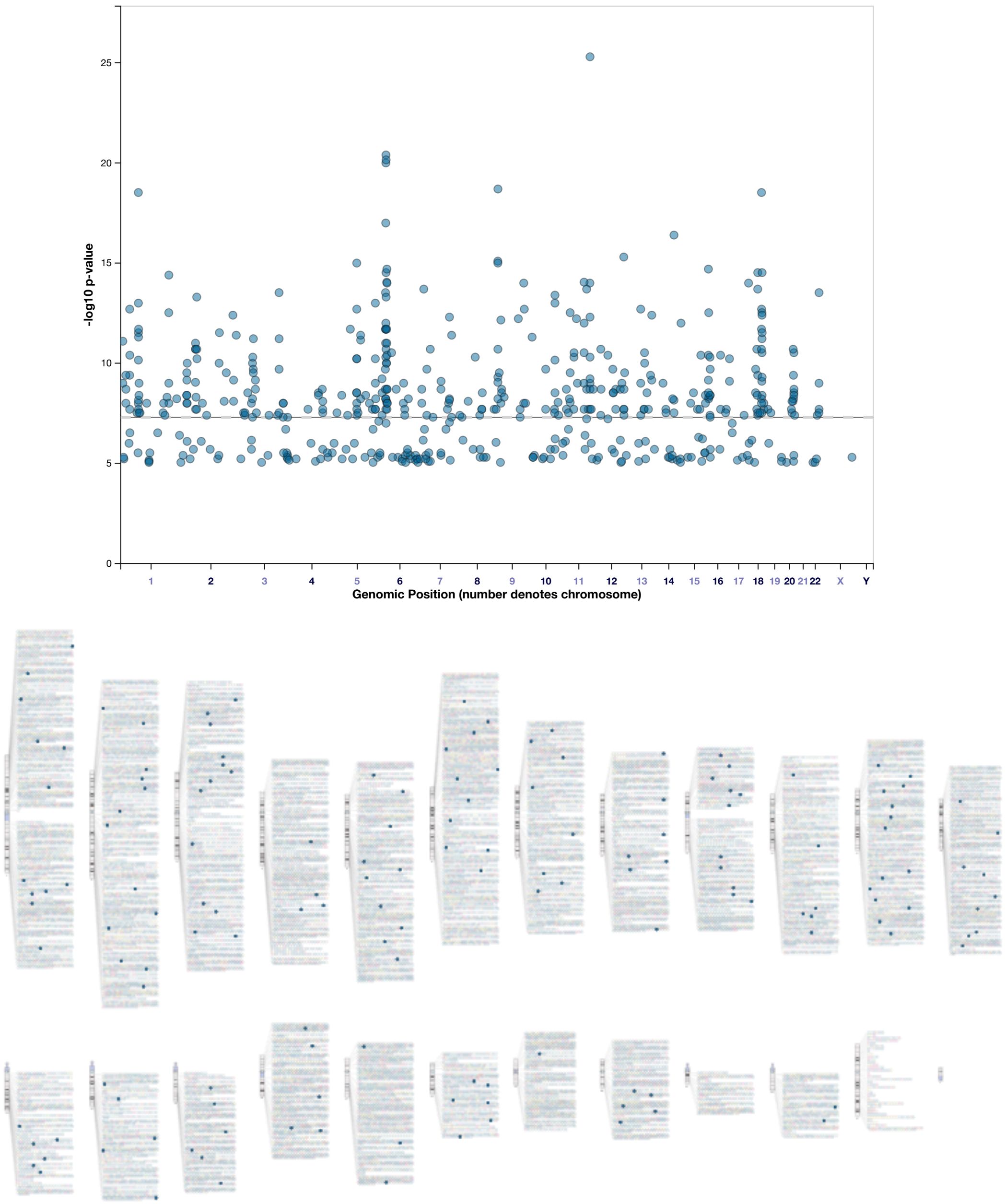


Figure 2.5.1: The NGRI-EBI Catalog (GWAS Catalog) graphics of published genome-wide associations studies for depression.
Source: https://www.ebi.ac.uk/gwas/
Date: 01-Oct-2019

tured with respect to these steps.

I start the discussion by describing the genetic quality control procedure for the UKHLS genetic sample – a critical step for preparing genetic data for analysis. I demonstrate that 0.7% of measured genetic markers and 8% of participants did not pass quality control tests. Then, I move to the next step for achieving a higher precision of polygenic prediction: the genotype imputation stage. I explain the concept of genotype imputation and demonstrate that the imputed UKHLS genetic sample can be categorised as having high imputation quality. Finally, I describe the methodology and technique for polygenic score construction. I show that constructed scores can significantly predict depression symptoms in the UKHLS sample. All stages were performed using the MacOS operating system. Importantly, this part of the analysis is based on a restricted version of the UKHLS data release.

2.5.1 Genetic quality controls

This section details the process of quality control that I performed. Quality control is an essential first step when working with genetic data. It aims not only to reveal patterns of possible biases and errors attributed to genotyping techniques, but also to prepare the data for imputation. The quality of the tests is directly linked to imputation performance [Li et al., 2009]. The conventional protocol is to generate quality checks per individual and per marker; participants and SNPs that fail the quality control are subsequently removed [Anderson et al., 2010, Meyer, 2019, Mills et al., 2020].

In the initial quality control tests, 9,965 out of 10,484 individuals passed the checks [Prins, 2015]. However, the most recent and modified version of the data release includes 9,944 individuals – 5,568 women and 4,376 men [Hughes, 2018]. This sample is a starting point for my analysis (the release version received from the UKHLS team). Because there is no comprehensive public report regarding initial quality checks and no such report was provided, my additional evaluation is essential to further analysis.

This stage of the analysis was performed using PLINK 2.0 software. [2] PLINK is a free command-line package widely used in genetic research [Chang et al., 2015, Purcell et al., 2007]. It is a comprehensive tool, allowing one to conduct all quality control procedures. Procedures can then be integrated using R software.

Per individual quality control

The main goal of per individual quality filters is to identify participants with low-quality genetic data. To achieve this aim, three conventional control tests are used [Meyer, 2019, Mills et al., 2020]:

- to detect participants with high rates of missing genotypes;
- to detect participants with high heterozygosity rates;
- to detect participants with discordant sex information;
- to detect genetically-related individuals;
- to detect participants with divergent ancestry.

Regarding the UKHLS data, no individuals were excluded due to missing genotypes. The missing rate cut-off was set at a conventional 0.05 level [Weir, 2012]. This means that for the genetic sample, each participant has at least 95% of the SNPs measured in the whole dataset. There were also no participants with discordant sex information or divergent ancestry. It should be noted that the UKHLS genotyped sample includes individuals with European ancestry only.

However, there were 41 participants with high and low heterozygosity rates. It is a common practice to drop participants with +/- 3 standard deviations from the sample heterozygosity mean; this step eliminates the possibility of inbreeding or other genotyping issues [Mills et al., 2020]. Figure

Table 2.5.1: Summary of per-individual quality control filters.

Filter	Excluded	Kept
Initial sample		9,944
Missing call rate >=5%	0	9,944
High heterozygosity rate ± 3 sd	41	9,903
Relatedness GRM >=5%	707	9,126
Discordant sex	0	9,126
Divergent ancestry	0	9,126
% sample lost due to quality filters	8%	
Total genotyping rate	99.92%	

A.4.1 in the Annex A shows the distribution of heterozygosity in the UKHLS sample, and the red dashed lines show quality control cut-offs.

To account for possible relatedness within the study, I adjusted the data by removing individuals who are more than 5% related. This is a critical step because the UKHLS is a household survey, and the genetic relatedness of participants will bias the results if it is not tackled appropriately. The threshold of 5% is commonly applied in genetic studies and also suggested in notes from the data release team. 707 individuals were excluded due to this issue.

Genotyping was performed on 9,944 individuals overall, and 9,126 individuals passed the quality control process. Below, Table 2.5.1 summarises statistics for the quality control filters. The total genotype rate is 99.92%, meaning that each participant has 99.92% of all measured SNPs.

Per marker quality control

In contrast to per individual quality filters, the per marker control procedure treats genetic variants as units of analysis. Accordingly, the main goal is not to exclude individual participants but to remove SNPs that do not pass certain quality thresholds. To achieve quality variants, the following stages are used as filters [Meyer, 2019, Mills et al., 2020]:

[2] Detailed information on PLINK and its properties can be found here: http://www.cog-genomics.org/plink/2.0/

Table 2.5.2: Summary of per-marker quality control filters.

Filter	SNPs excl.	SNPs kept
Initial #SNPs		248,604
Missing call rate >=5%	17	248,587
MAF <=1%	0	248,587
Hardy-Weinberg exact test	0	248,587
% of SNPs lost due to quality filters	0.7%	

- to detect variants with high rates of missingness (in terms of missing genotypes);
- to detect markers with low minor allele frequency (MAF);
- to detect variants with a significant deviation from the Hardy-Weinberg equilibrium (HWE).

Table 2.5.2 summarises the marker filters. No SNPs were excluded due to the low MAF issue (0.01 conventional threshold level). There were also no variants deviating from HWE, which indicates low likelihood in certain forms of genotyping error.

However, 17 variants were dropped due to high rates of omission. In other words, these SNPs were missing in a large proportion of the sample (the missing rate cut-off was set at a conventional 0.05 level). The rationale for excluding such markers is the observation that a missing call rate correlation with traits leads to spurious genetic associations because the omission is likely not random [Clayton et al., 2005, Weir, 2012].

At the final stage of the quality control procedure, a new sample is created from individuals and markers that passed through the filters. The new, clean version of the UKHLS includes 9,196 individuals and 248,587 SNPs. This is the version that I used for the imputation phase. No major or significant concerns emerged in relation to the quality of genetic data.

2.5.2 Genotype imputation

Imputation is an essential step in genetic research. Genotype imputation is the procedure of predicting unobserved genetic markers in a researcher's data based on haplotypes [3] from a more comprehensively genotyped or sequenced reference panel [Li et al., 2009]. This section describes the genotype imputation process and covers the output of imputation, including quality metrics. It also generally explains the methodological necessity of genotype imputation, highlighting primary differences from the imputation concept used in social surveys.

Although the concept is fundamentally the same, genotype imputation differs from imputation in social science. Thanks to evolutionary inheritance, certain chunks of chromosomes/DNA travel together across generations. This results in the existence of tag SNPs across genomes. With these tag SNPs, it is possible to impute nearby markers within regions one does not have. The genotype imputation procedure thus implies a process of guessing missing variants – one that is much more precise than guessing missing data in social surveys. Figure A.4.2 in the Annex graphically illustrates the process of imputation in general terms.

Imputation estimates unmeasured or missing genotypes based on measured SNPs and the haplotype structure of an external reference panel. The concept of haplotype blocks is central to genetic imputation, boosting the power of prediction. It leads to a notion that imputation capitalises on patterns of linkage disequilibrium (LD), or the correlations between alleles and pairs of SNPs. Perfect LD comes from no recombination, while no LD indicates perfect recombination. However, imputation does not ignore recombination and mutations; both are considered through iterations [Stephens and Scheet, 2005].

For this project, genetic data is imputed to achieve greater precision in polygenic prediction. The need for this is determined by increasing the number of SNPs overlapping the GWAS discovery genetic sample and the UKHLS dataset. The more markers are shared across groups, the better the

[3] Haplotypes (or haplotype structure) is an entity of genetic markers (SNPs or alleles) that are inherited together [Cox et al., 2016, p. 106].

polygenic scores perform. The main difficulty with imputation is a computationally demanding nature, which cannot be compromised without a loss of accuracy.

In chronological terms, a researcher begins with data from participants who were genotyped on an array with a certain number of variants. In my case, I have 9,196 individuals and 248,587 measured SNPs. In addition, there are also reference panels of sequenced genomes available from different projects.

Genotype imputation is a population-based concept and it is important to ensure use of an appropriate reference panel. Reference panels are usually sequenced, not genotyped (the distinction between these two concepts was clarified earlier in the text). The reference panel I am interested in is based on the 1000 Genomes Project (`http://www.internationalgenome.org/`). This interest is determined by the characteristics of the GWAS discoveries samples. The study that I use to construct polygenic scores refers to the 1000 Genomes Project as a panel for imputation. Imputation based on this reference sample allows me to combine UKHLS data with GWAS discovery cohorts for polygenic prediction. With these two pieces of information, we can start to find haplotype segments shared by UKHLS participants and the reference panel – in other words, finding imputation method.

I use the imputation technique developed by Das et al. [2016] at the University of Michigan, which is one of the most advanced among existing methods with regards to computational efficiency [Langmead and Nellore, 2018, Loh et al., 2016]. More information about the project can be found at the following website: `https://imputationserver.sph.umich.edu`. The main idea is as follows: the algorithm synthesises the haplotype-sharing analysis with state reduction of the hidden Markov-chain models. The former is a fundamental part of the improved computational performance achieved in recent years, while the latter is a novel step in genetic research. Please refer to Das et al. [2016] for a detailed description of the method and a comparative analysis of its performance.

Imputation is performed by chromosomes with MaCH pre-phasing [Li et al., 2010] and Eagle2 phasing [4] steps [Loh et al., 2016]. Imputation analysis *per se* was performed using minimac3 [Das et al., 2016]. Table A.4.1 in the Annex includes information on steps and tools used during the imputation phase.

There are also conventional standards regarding what is considered low- and high-quality imputation. Quality metrics depend on the software, but the idea is the same across tools: comparing the observed genotype distribution with the expected distribution. Minimac3 produces R-square statistics where:

$$R^2 = \frac{\textit{empirically observed variance}}{\textit{expected binomial variance } p(p\text{-}1)}$$

with assumptions that the expected variance follows the HWE and where p is the frequency of one allele [Browning and Browning, 2009].

As demonstrated in Figure A.4.3 in the Annex, imputed UKHLS has $R^2 = .991$, which is considered high imputation quality. Conventional thresholds are $R^2 \geq 0.8$ for high quality and ≤ 0.3 for poor quality [Browning and Browning, 2009, Li et al., 2010].

To conclude, imputation allocates genotypes at untyped variants with the highest degree of accuracy available. It thus boosts the genome coverage. As a result, the number of variants in the data increased along with a greater overlap in the number of SNPs with the GWAS discovery data. The precision of imputation is directly linked to the number of haplotypes in the reference panel, where the techniques developed by Das et al. [2016] are shown to be one of the most efficient tools for researchers. This was the final step in preparing the data for polygenic prediction.

2.5.3 Polygenic score construction

The polygenic score construction is based on recent guidelines presented in Mills et al. [2020] and in Choi et al. [2018]. Polygenic scores are computed from GWAS summary statistics where certain sets of genetic variants have proven valuable for predicting depression. I use one of the most recent

[4] Phasing is the process of the statistical estimation of haplotypes [Marchini and Howie, 2010]

studies on depression, which was conducted by Howard et al. [2019]. It is the most beneficial source for my research purposes because it is the largest GWAS on depression. The study took both self-reported and hospital admissions phenotypic data and identified 102 independent loci. For comparison, another recent GWAS from Wray et al. [2018] identified 44 independent variants. Since my goal is to operationalise polygenic grounds for depression, I do not consider GWASs on well-being spectrums as Baselmans et al. [2019] does or on multiple psychiatric phenotypes as Amare et al. [2020] does. There are also earlier studies by Hek et al. [2013] and Terracciano et al. [2010], but both of these have small sample sizes that limit claims to reliable results.

In line with the GWAS methodology, researchers are required to scan the whole human genome to identify SNPs associated with specific gene variants [Benjamin et al., 2012]. In the case of a given outcome of interest (such as depression) and a set of observed SNPs, a GWAS study is used to obtain estimates of separate regressions. This is presented in the equation below:

$$Y_i = \alpha X_i + \beta_j SNp_{ij} + \varepsilon_{ij}$$

where Y_i is a variable for depressive symptoms; SNP_{ij} is used to measure the number of reference allele copies possessed by person i for SNP_j. Likewise, X_i is referred to as a control for population genetic stratification, which comprises major components found in the genetic variables. Failure to control for these principal components could lead to biases in the data analysis [Price et al., 2006]. That is, instead of being focused on a specific genetic marker, the research is likely to be influenced by ethnic differences thereby altering the intended results. Accordingly, the estimated β_j would reflect a relationship that has genome-wide significance.

The second step is to follow a strategy whereby the estimated GWAS coefficients (β_j), and the observed SNPs are aggregated to form a polygenic score that can be used as an additional independent variable in the conventional regression models [Dudbridge, 2013]:

$$PGS_i = \sum_{j=1}^{J} \beta_j G_{ij}$$

where i is individual i (i=*1* to N), j is SNP j (j=*1* to J), β_j is the meta-analysis effect size for SNP j derived from the previous step and G is the genotype for individual i for SNP j. Consequently, each SNP is weighted by the magnitude of its effect, where for continuous outcome the estimate is a coefficient, while for dichotomous trait it is a log of odds ratios.

Once polygenic prediction is included in the research agenda, some important considerations must be taken into account. Firstly, the predictive power of the score depends on the nature of the trait in terms of whether the outcome is heritable and the extent to which it is polygenic. Depression is a heritable trait with a high extent of polygenicity. Secondly, the power of the GWAS discovery should be assessed. This notion mostly relates to the first genome discoveries; recent discoveries have over 1 million people in a sample (for example, see Lee et al. [2018]). Still, the notion of power is important for detecting common variants and their signals – and also rare ones. The history of the GWAS of psychiatric disorders demonstrates that rare variants can have moderate to strong effects, and they can only be detected with large sample sizes [Sullivan et al., 2018]. However, the role of rare variants in the development of depression is still unknown [Peterson et al., 2017]. Hence, current polygenic predictions of depression cover common SNPs and their signals.

Thirdly, it is important to consider what SNPs are used for an index – either all of them or only those whose p-value in GWAS is below a certain threshold. Here, the choice depends on the research question and the goal of the polygenic score. For mendelian-randomisation studies, the conventional preference is for a more biologically informed polygenic score. Thus, only a certain group of genetic markers is included. But if the goal is to address the heterogeneity of genetic prediction across environments, all SNPs (without any threshold) should be included in the score calculation (for example, see Belsky et al. [2018], Domingue et al. [2017a]). This decision is directly linked to the notion that both the strongest (in terms of prediction) and causal SNPs would survive in any environment. Hence, the score would not be sensitive enough to detect interactions with environments. Accordingly, my polygenic scores are constructed using all markers.

Fourthly, the LD structure of genetic inheritance is another valuable factor to consider when identifying sets of SNPs for inclusion in the polygenic score construction. As I discussed earlier, signals from genetic markers are grouped and concentrated in certain regions across the genome. One way to take them into account is to construct respective LD weights [Choi et al., 2018]. Since both of the GWAS studies that I use adjust for disequilibrium in the research design and provide summary statistics with appropriate weights, the issue of interconnectedness is proxied in my polygenic scores.

Last but not least, there is the notion of sensitivity to ancestry. Most GWAS studies have been done for cohorts with European ancestry, leading to the issue of reduced predictive power for non-European cohorts [Mills and Rahal, 2019]. While prediction in genetically diverse samples is not an appropriate procedure, polygenic scores should be appropriately calculated for ancestry groups independently. Since the UKHLS data and GWAS discoveries that I use cover populations with European ancestry, I do not have this source of a bias. However, I do admit that the focus on only one ancestry group is a limitation of the project. With future releases of the data, it will hopefully be possible to perform the analysis on other sub-populations living in the UK.

The polygenic scores were constructed after quality control of the summary statistics files. Since these files were provided by different research groups and downloadable online, I checked coding formats, whether file corruption occurred during transfer, and whether files have the same genome build as the UKHLS data. It is also critical to eliminate possible sample overlap between the GWAS discovery samples and the UKHLS [Choi et al., 2018]. Since the work by Howard et al. [2019] does not have Understanding Society in the list of their cohort, the potential threat is eliminated.

Accordingly, the construction of polygenic scores was performed using PRSice 2.0 software [Choi and O'Reilly, 2019]. There were 360,140 SNPs matched after clumping between reported results in a Howard et al.'s GWAS and the UKHLS dataset. Thereafter, the score was standardised with a 0 mean and standard deviation of 1, with further residualisation across the first 20 principal components (PCs) received from the UKHLS genetic team, to take into account population stratification. As a robustness check I compare the performance of polygenic scores based on imputed and non-imputed genetic data. Figure A.4.4 demonstrates normal distribution patterns for both scenarios of construction, which is consistent with the methodology of polygenic scores. Polygenic scores will always have a normal distribution according to the central limit theorem (Mills et al., 2020, p. 106, Plomin et al., 2009).

I assess the polygenic prediction by employing Poisson regression models with sex, age, age-sq., and the first 20 PCs as covariates. The choice of the model is determined by the nature and distribution of the GHQ depressive symptoms score in the UKHLS. This is a constructed sum of 12 indicators, such as usefulness, decision-making, unhappiness, confidence, and others. Full information on the GHQ questionnaire is provided in Table A.4.2. It is important to note that this is a count-based variable that ranges from 0 to 36 with a degree of skew, as demonstrated in Figure 2.5.2. Accordingly, Poisson models are an appropriate tool for taking the complexity of such skewed data into account [Wooldridge, 2010]. Moreover, it is a preferred method since the coefficients produced are more intuitive and interpretable than the ordinary least squares (OLS) regression estimates with logged dependent variable [Nichols et al., 2010]. The GHQ depressive symptoms are measured by averaging respondents' GHQ scores across the 25 waves of the UKHLS survey.

Table 2.5.3 shows the results from the regression models. As the reader will notice, imputed and non-imputed polygenic performances are very similar. This is something that is expected and it implies consistency. The aim of imputation is not to increase predictive power. Rather, it is to achieve a higher precision as indicated by a slightly higher coefficient for the score and slightly decreased standard errors. The incremental R-square is 0.4% in the imputed polygenic score model, which corresponds to the prediction from the GWAS discovery. As graphically presented in Figure 2.5.3, UKHLS respondents with higher polygenic scores (measured in standard deviation units) reported more depressive symptoms during follow-up. Additionally, Table 2.5.3 demonstrates that the implementation of weights to correct for selection into genotyping does not significantly change the

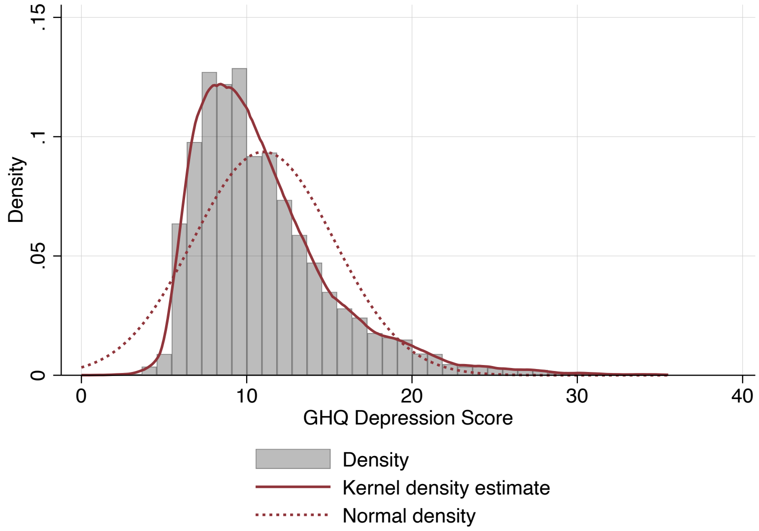


Figure 2.5.2: Distribution of GHQ depressive symptoms score in the UKHLS genetic sample.

polygenic prediction of depression. The PGS coefficient in the weighted regressions is slightly bigger along with standard errors.

Table A.4.5 in the Annex displays an additional scope of analysis wherein instead of averaging depressive symptoms across survey years, I consider respondent probability for crossing the symptomatic threshold at least once during the follow-up period. Following the UK's GL assessment guide (https://bit.ly/2FWql0E), this is conceptualised as a sign for minor psychiatric disorder. Accordingly, a 1 s.d. increase in PGS is associated with a 17% greater chance of crossing a threshold.

I also perform additional analysis. Table A.4.4 shows the results of Poisson models assessing nonlinearity in PGS prediction of GHQ depressive symptoms, where I include the squared term of the polygenic score in the set of predictors. This enquiry is based on the literature on liability threshold models using polygenic predictions (see review by Chatterjee et al. [2016]). Namely, the literature indicates that manifestation of outcomes occurs once the number of risk alleles exceeds a certain threshold. However, I find more support for a linear relationship between my genetic score and depressive symptoms.

Table 2.5.3: Coefficients (Robust SE) of Poisson Models Assessing Polygenic Score Prediction of GHQ Depressive Symptoms Score. (Weighted coefficients and robust SE)

	Baseline	Imputed PGS	Non-imp. PGS
PGS Depression		**.043** (.00)	**.038** (.00)
		.045 (.01)	**.041** (.00)
Sex (female)	**.118** (.01)	**.118** (.01)	**.118** (.01)
	.117 (.01)	**.117** (.01)	**.117** (.01)
Age	**.007** (.00)	**.007** (.00)	**.007** (.00)
	.008 (.00)	**.007** (.00)	**.007** (.00)
Age-sq.	**-.000** (.00)	**-.000** (.00)	**-.000** (.00)
	-.000 (.00)	**-.000** (.00)	**-.000** (.00)
Intercept	**2.100** (.04)	**2.099** (.04)	**2.097** (.04)
	2.099 (.05)	**2.099** (.05)	**2.109** (.05)
Pseudo R-sq.	*0.011*	*0.015*	*0.014*
No. of participants	*9,113*	*9,113*	*9,113*

Bold values are significant at 99.9% level; Genetic score is standardised;
All models include the largest 20 PCs

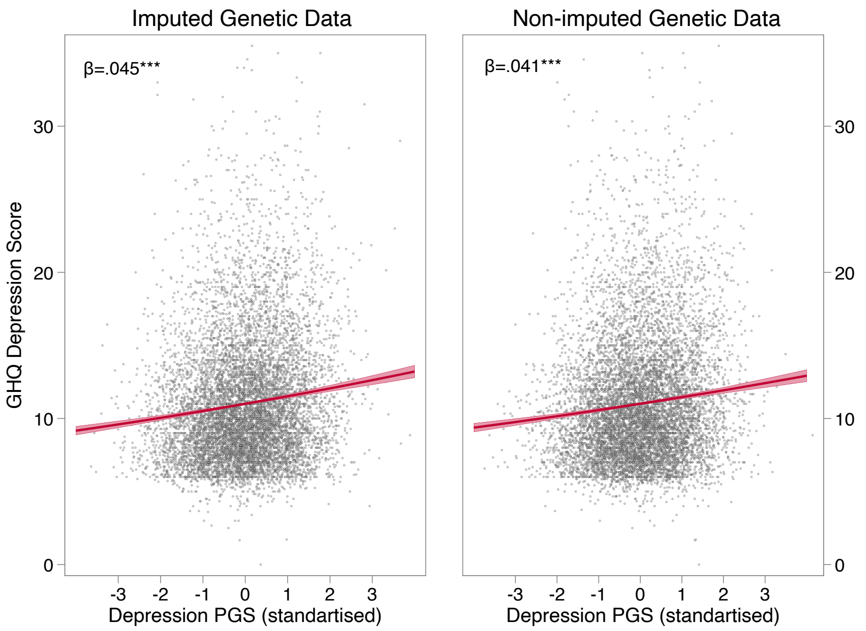


Figure 2.5.3: Association between polygenic score for depression and GHQ depressive symptoms in the UKHLS sample in weighted regressions.

Finally, I assess the predictive power of my polygenic score regarding the diagnosis of clinical depression. Table A.4.6 shows that 1 s.d. increase in polygenic score is associated with a 36% increase in the weighted probability of a diagnosis of clinical depression at least once during the 25-year period of the UKHLS survey. The incremental R-square of PGS is notably higher, equalling 1.5%. However, there are fewer people with diagnosis information. This is why the sample size for such models is smaller (*N=8,676*). Importantly, I do not use a diagnosis of depression as my main phenotypic variable since symptomatic data on mental health is shown to be more accurate. Such data has greater validity and reliability in the context of population-based surveys [Mandemakers, 2011].

2.6 Conclusion

The purpose of this chapter is to reveal methodological insights regarding the analysis of genetic samples and the construction of polygenic scores. I have provided an extensive overview of the UKHLS genetic data accompanied by analysis of sample selection. To my knowledge, this is the first sample selection analysis of the UKHLS dataset. I demonstrated the necessity of including weights in analytical agendas, as weighting eliminates sample selection differences associated with socio-demographic and health factors. I found small-to-moderate differences between genotyped and non-genotyped participants: those who were genotyped are 16% less likely to live in rural areas and more likely to have higher education; moreover, genotyped participants are 16% less likely to have a bad general health and 28% less likely to have heart conditions.

This further corresponds with existing literature as the unweighted scenario demonstrates the likelihood of socio-economic selection and a *healthy volunteer* effect in the UKHLS genetic sample. I also showed that these differences are likely germane to a gene-by-cohort studies because of mortality selection. I thus demonstrated that genetic samples should receive additional attention in terms of processes for generating data. In the case of the UKHLS sample, blood weights are sufficient for selection correction (except for studies aiming to reveal trends in reproductive behaviour). This means that weighting information should either be included on a mandatory list of variables provided by the UKHLS genetic team or acquired by researchers themselves. Notably, existing studies using the UKHLS genetic sample do not consider the consequences of selection (for an examples involving discoveries of gene × environment interactions and assortative mating, see Amin et al. [2017], Hugh-Jones et al. [2016]). This further illustrates the lack of such a focus within the literature.

I also performed a polygenic score analysis of depressive symptoms. Polygenic scores are based on information from bio-genetic research. They offer a demonstrably valuable method for capturing the notion of genetic predispositions and translating them into statistical usage [Vilhjalmsson et al., 2015]. In this project, the polygenic score is a genetic index of depression for each person based on the Howard et al. [2019] GWAS discovery, with further robustness checks based on the imputation status of the genetic data and UKHLS weights to correct for selection. The polygenic score identifies the level of genetic predisposition to depression. It is further used as an additional independent variable of interest in Chapters 3 and 4.

Chapter 3

Changing polygenic penetrance on depressive symptoms among adults in the United Kingdom

3.1 Abstract

Changes that occur over time across different birth cohorts is a major field of research in demography and sociology, as cohort effects reflect the importance of historical changes shaping people's lives. Ongoing discussion of depression also covers this aspect of research. The prevalence of depression is believed to have a historical trend and to occur more frequently among recent birth cohorts. Observed increases in the occurrence of depression could be due to various factors, including changes in policies, macro-economic conditions, and lifestyles. Genetic influences on depression may affect individual responses to contextual aspects, leading to variation in genetic penetrance on depression across birth cohorts (known as gene-by-cohort interactions). Accordingly, this chapter investigates whether polygenic prediction of depression varies by birth cohorts in *the UK. Through theoretical considerations of gene-environment interactions, I perform a regression analysis using the UKHLS genetic sample. I show some evidence supporting gene-by-cohort interactions in depression among adults in the UK, which I further link with exposures to economic recessions.*

3.2 Introduction

Mental health is a global health concern. Social media, newspapers, official reports, researchers, politicians, and others emphasise the importance of a deeper understanding of mental health disorders. Given the alarming statistics, this attention is unsurprising: mental disorder is the main cause of disability worldwide [Lozano et al., 2012] and one of the main causes of overall sickness [Vos et al., 2015].

Depression is one of the most common mental health disorders. The frequency of depression occurrence ranges from 8% to 12% in different countries [Flint and Kendler, 2014]. The severity of depression varies from mild symptoms to major depression. All of these factors contribute to growing research covering different aspects of depression. There are various causes of depression and social scientists refer to many different factors, including historical exposures. Depression is also heritable and there is an interdisciplinary field investigating how various environmental aspects interplay with genes.

Accordingly, depression can be triggered by socio-economic factors, such as educational attainment [Lee, 2011], job loss [Drydakis, 2015, Paul and Moser, 2009], and recessions [Frasquilho et al., 2015, Jahoda, 1988]. More broadly, the prevalence of depression is believed to have a historical trend and to occur more frequently among recent birth cohorts [Bell, 2014, Marcus and Olfson, 2010]. Observed increases in depression occurrence could be due to various factors, including environmental and lifestyle changes, policy contexts, and economic downturns. Genetic influences on depression may also affect individual responses to contextual components (for example, by shaping stress-internalisation processes). The latter would lead to variation in the genetic penetrance on depression across birth cohorts. As previously suggested, estimates of the percentage of variation in social outcomes explained by genetic and environmental differences are likely to be context specific, varying systematically across different social conditions, policy environments, or subgroups of the population [Boardman et al., 2011]. These notions have yielded a growing field of research wherein birth cohorts are potential modifiers of genetic influences.

This paper identifies changes in the polygenic penetrance on depression within the UK during the 20th century. The research investigates whether the polygenic prediction of depression varies by birth cohorts in the UK or, in other words, whether we observe gene-by-cohort interactions for this mental health trait. I also aim to answer the question of whether historical contexts (such as economic recessions) contribute to gene-by-cohort variations.

In a conventional demographic classification for the UK, there are six birth cohorts. Two cohorts are devoted to people exposed to the two World Wars: a WWI cohort born between 1916 and 1930; and a WWII cohort with birth years in 1931–1945. A demographic cohort of those born in 1946–1964 is distinguished as Boomers to reflect the period of the Baby Boom. After that, there is a Generation X cohort (people born in 1965–1980) followed by Millennials or Generation Y (those born between 1981 and 1995). People born at the very end of the century are referred to as Generation Z.

The focus on gene-by-cohort interactions has the potential to shed a light on how historical contexts shape polygenic prediction across different generations. The insight for social science, in particular, is whether the rise in the prevalence of depression at certain historical points in the 20th century is driven by those with a higher polygenic risk of depression; alternatively, prevalence could be independent of genetic risks. Within the literature, there is a notable gap in studies covering the UK context. Consequently, this paper contributes to existing knowledge by providing a gene-cohort interaction analysis of depression in the UK.

In this study, I use data from the Understanding Society genetic sample to investigate how associations between depressive symptoms and the polygenic risk score for depression differ across successive birth cohorts in a national sample of UK adults. I start my investigation by presenting the theoretical background for the issue. First, I discuss findings on the increasing prevalence of depression in the UK. Then, I demonstrate genetic factors related to the development of depression. Afterwards, I provide an overview of gene-by-cohort interaction studies where I also discuss the role of policy changes that could potentially shape heritability variations. In the sections thereafter, I explain my empirical strategy by describing measures and statistical methods. I discuss the results extensively and assess possible biases.

Next, I perform a robustness analyses that initially takes mortality selection into account mortality. As I showed in Chapter 2, selection into genotyping results in differential mortality curves among genotyped and non-genotyped participants in the UKHLS survey. This is known to be a source of potential bias in gene-cohort analysis [Domingue et al., 2017b]. Hence, I perform a weighted analysis to obtain estimates that are less biased by mortality selection. Finally, I check the robustness of my results through analysis of overlapping, age-comparable sub-samples of birth cohorts. Aging trends and cohort trajectories are inherently connected, and some age groups are omitted from some cohorts in the UKHLS dataset. Replication thus advances the understanding of observed results.

3.3 Background

3.3.1 Cohort trends in depression in the UK

Cohort trends are a major area of study in demography and sociology [Ryder, 1965]. This particular interest is linked to the notion that cohort effects indicate the importance of historical changes that potentially shape people's experiences at least to some extent. To reveal and to understand the links between macro conditions and individual-level outcomes is the traditional focus of sociology. In this section I provide a review of studies that address the differences in depression prevalence among UK birth cohorts born in the 20th century.

The first serious discussions of an increase in the prevalence of mental health problems occurred during the 1970s [Marcus and Olfson, 2010]. Bell [2014] found some support for this proposition in the UK context; through the use of BHPS, the study claims more recent cohorts have poorer mental health. In contradiction to Bell [2014], a study by Spiers et al. [2011] indicates no clear evidence of an upward linear trend in mental health problems across birth cohorts in the UK. However, some evidence suggests that a spike in depression occurred among people born in the middle of the 20th century, i.e. the cohort of Baby Boomers [Rice et al., 2010, Spiers et al., 2012]. Based on the National Psychiatric Morbidity Surveys, Spiers et al. [2012] show that the prevalence of depression is higher among men born between 1950 and 1956 than in the earlier cohort of those born between 1943 and 1949. Notably, trends among women showed less consistency. Researchers found some significant upward and downward trends among earlier cohorts with rates stabilising after 1963 as a birth year [Spiers et al., 2012, p. 2051]. To measure depression, researchers used a fully structured diagnostic instrument (the Clinical Interview Schedule-Revised, or CIS-R).

In the most recent cohort studies, the focus shifted to cohort-specific hypotheses. Thomson and Katikireddi [2018] investigated cohort variations in depression, for example, by paying particular attention to the so-called 'Jilted' generation (those born in the UK after the year 1979, or Millennials in conventional terms). The study involved repeat cross-sectional data obtained from the Health Survey for England for the period between 1991 and 2014. The researchers based their investigation on the GHQ depressive symptoms score that I apply in my analysis as well. Thomson and Katikireddi [2018] hypothesised that high material disadvantages caused by social and economic policies in the UK would lead to a higher prevalence of symptoms of depression (among others) in Millennials. But the researchers found no evidence to support this hypothetical claim. Moreover, there is limited insight into overall trends since the vast majority of associations are not significant and bounded around 0% changes [Thomson and Katikireddi, 2018, p. 137].

In terms of research trends for the latest cohorts, it is important to note that recent years have seen a growing volume of literature on the mental health problems of teenagers from different birth cohorts (those born at the end of the 20th century and those born in the 21st). Research suggests that more recent cohorts of teenagers suffer from a greater prevalence of mental health problems than teens born before the 2000s. Such trends are found both for the UK context [Collishaw et al., 2010, Fink et al., 2015] and worldwide [Hagquist, 2010, Twenge et al., 2010]. For example, Patalay and Gage [2019] identify an increase in depressive symptoms from 9% to 15% among UK teenagers born in 1991/92 (ALSPAC study) and those born in 2000/02 (MCS study) at the age of 15. Also, the prevalence of self-harm increased from 11.8% to 14.5%.

To date, there is little overall agreement on the shape of cross-cohort trends in depression in the UK. What seems to be robust across studies is a steady rise in the prevalence of mental health problems among those born prior to the 1960s. Also, there are consistent increases in depression observed among the most recent generations (those born after the 2000s). Regarding these trends, findings from the UK are consistent with those from the US. In a more detailed view, Twenge et al. [2019] show that the highest occurrence of depression and psychological distress in the US is likewise observed among the mid- and end-of-century cohorts. One plausible explanation for the similarity of trends in different contexts is the link to mechanisms that exist in both countries.

However, the literature on *explanans* linking birth cohorts and mental health is rather limited. This poses additional puzzles. Why do we observe some differences among UK cohorts across the 20th century? In the narrative for the UK context over the last century, there has been a wide range of historical and economic upturns and downturns that includes shifts in gender roles. Any of these could have impacted generational well-being in one way or another. Child labour was a common trend among people born at the beginning of the century, for instance, while female occupational employment was limited. A period of major recession and low growth occurred prior to World War II. In contrast, the post-war period (1950s–1960s) is known as the Golden Age – a time of economic boom and full employment. From the 1970s to the 1980s, a recessive period affected labour market participation. These periods of prosperity and challenge were experienced by people from different cohorts to varying extents. For women, the 20th century is of particular importance due to noticeable changes in labour participation and social norms in general. Increases in mental health awareness, coverage of regulations, and shifts in social norms regarding depression reporting during this time period are also important. All of these could conceivably contribute to cohort differences.

The work by Thomson and Katikireddi [2018] links cohort differences with the Great Recession of 2008. They observe an association between the global recession and austerity reforms in the UK with worsening mental health. This was especially evident among the young working generations, while gaps between earlier cohorts stayed constant. They do not give a definite answer on whether austerity measures broadened generational inequalities in mental health, however. Still, they offer an important insight: the contribution of an economic downturn that was experienced by different generations in different ways. It is known that prolonged economic recessions are associated with increased rates of depression [De Vogli, 2014, Stuckler et al., 2017]. Recessions do not affect everyone equally, but rather trigger social determinants of mental health (for instance, incidents of unemployment). These social determinants are partially cohort-specific due to age differences and status of labour market participation. Consequently, recessions widen the gaps between working-age generations and retired cohorts. Direct assessment of some of the factors, such as unemployment or economic inactivity, is notably problematic due to endogeneity and self-selection; however, birth cohorts are exogenous (this issue and its evaluation will be covered in Chapter 5).

To conclude, analysis of the literature shows that birth cohorts have different degrees of mental illness prevalence. There is no conclusive evidence showing this relationship follows one universal pattern. Moreover, much research into this topic is descriptive in nature and our knowledge of factors attributed to cohort differences is limited. This paper does not aim to provide a causal explanation. Instead, I offer insight into whether cohort differences in depression are associated with its changing polygenic penetrance. The next section thus discusses the genetic basis for depression and the phenomenon of gene-by-cohorts variation.

3.3.2 Genetic grounds for depression

To begin, the earliest stages of research noted that depression develops in families [Tsuang and Faraone, 1990]. Later, twin studies showed that if a person's parents suffer from depression, their risk of developing the illness ranges from 20% to 40% [Jansson et al., 2004, Keyes, 2005, McGue and Christensen, 2003, Sullivan et al., 2000]. This range is also known as the heritability estimate for depression. However, twin studies are also likely to overestimate the extent of heritability of traits and may not present the true genetic component or its size [Maher, 2008, Manolio et al., 2009].

New methods and technologies allow researchers to address the issue in a more straightforward manner as the screening of the whole genome is now possible. One widely used approach is the so-called GWAS method. The rationale behind this method is to screen all of the SNPs in the human genome and to test their association with a certain phenotype. Multiple studies are looking at depression as a phenotype [Hek et al., 2013, Okbay et al., 2016, Terracciano et al., 2010, Wray et al., 2018], with the most recent one conducted by Howard et al. [2019].

We have learned from GWAS that depression is a polygenic trait. The most recent genetic discoveries from Howard et al. [2019] identified 102 independent variants, 269 genes, and 15 gene sets associated with depression. This includes both genes and gene pathways associated with synaptic structure and neurotransmission.

It should be noted that researchers highlight the importance of certain SNPs for depression risk. In major depressive disorders (MDD), rs7647854 on chromosome 3 was found to play a significant role [Power et al., 2017]. Also, rs19323608 on chromosome 17 implicates the influence of genetics on the onset of depression [Okbay et al., 2016]. Such detailed analyses of SNPs and their locations is required for further investigation of the causal links because GWAS is a descriptive approach that only establishes associations.

Up to this point, I have intended to highlight an association between SNPs and depression. As the GWAS approach only provides a descriptive information, it cannot establish causality. In the following paragraphs, then, I would like to show that at least some of these associations (which we know from the GWAS studies) can be attributed as causal. This makes the genetic grounds for depression unignorable.

One of the biological mechanisms that links genes and depression is hormone regulation. I exemplify this with cortisol, but it should be noted cortisol is not the only hormone that could be listed here. [1] Cortisol is a hormone known to be produced in response to stress [Carey, 2012, p. 237]. A higher level of stress leads to a higher level of cortisol. There are biological and non-biological factors leading to the over-expression of cortisol, which makes an individual chronically stressed. The experience of traumatic childhood events, for example, can do so [Christine Heim et al., 2000, Yehuda et al., 2001]. DNA methylation at the cortisol receptor gene [McGowan et al., 2009] likewise contributes to the development of chronic stresses and further depressive symptoms. Cortisol also participates in long chains of switching other genes on and off, the cells of which contain cortisol receptors [Carey, 2012, p.150]. This suppression can even lower the immune system of the whole body.

Additionally, SNPs associated with depression are also located in the coding region for the central nervous system (CNS) and exert their effects in transcriptions for the development of this system [Howard et al., 2019, Hyde et al., 2016]. Under investigations for major depression, identified SNPs indicate that they are pleiotropic in other psychiatric conditions. Genes that play a significant role in the development of depression are brain-derived neurotrophic factors (BDNF) [Kaufman et al., 2006]. This implies the possibility of further investigation into the GWAS discoveries as genes play a significant role in the development of depression, influencing the function of the brain.

In sum, the genetic risk of experiencing depression is one factor in its occurrence. Growing research in this field reveals the associative nature of the genes-depression link, with further understanding of the biological causal mechanisms driving this correlation. These findings show that the genes-depression link is critical to a wider understanding of depression. It is also important to dis-

[1] Serotonin is another hormone that could be mentioned as the short allele of the serotonin transporter gene 5-HTTLPR leads to higher depression risk [Bansal et al., 2016, Lohoff, 2010, Pergamin-Hight et al., 2012].

tinguish causal SNPs from non-causal ones: while the former hits tend to survive in all environments, the latter are likely sensitive to environments. Thus, they are also included in analyses of gene-interaction studies. In the case of a phenotype such as depression, both groups of SNPs have been discovered.

3.3.3 Gene-by-cohort studies

So far, this background has focused on two different predictors of depression independently, i.e. birth cohorts and genetic associations. The following section introduces and explains gene-by-cohort variations. I begin by describing the conceptual basis of gene-cohort interactions, linking them to the gene-environment interplay theory. Next, I review the existing literature to show a profound gap in studies of gene-by-cohort variations in depression – particularly for the UK context, since the most significant portion of evidence comes from the US. Next, I motivate my investigation using findings from gene-by-cohort studies on smoking where I demonstrate how insightful such an interdisciplinary focus can be. I show that policies on smoking introduced during more recent US cohorts, which aimed to increase awareness of the harms of smoking, resulted in stronger genetic penetrance [Domingue et al., 2016]. This led to the notion that the decrease in smoking prevalence among more recent cohorts is driven by only those who are not genetically at risk of smoking. Therefore, gene-by-cohort variations can be particularly useful for extending our understanding of *explanans* in cohort differences.

In terms of the conceptual framework, gene-cohort interaction studies are based on the gene-environment (G×E) interaction theory. G×E theory states that environmental conditions are potential buffers and/or stressors of genetic predispositions, which are liable to cause certain health outcomes [Seabrook and Avison, 2010]. In gene-by-cohort interactions, an environmental variable is the incidence of birth during certain historical period. Accordingly, a birth cohort conceptualises time/period-specific peculiarities that are further represented as a sum of potential environmental stressors or controls. Since human genes respond to environmental variations differently [Courtiol

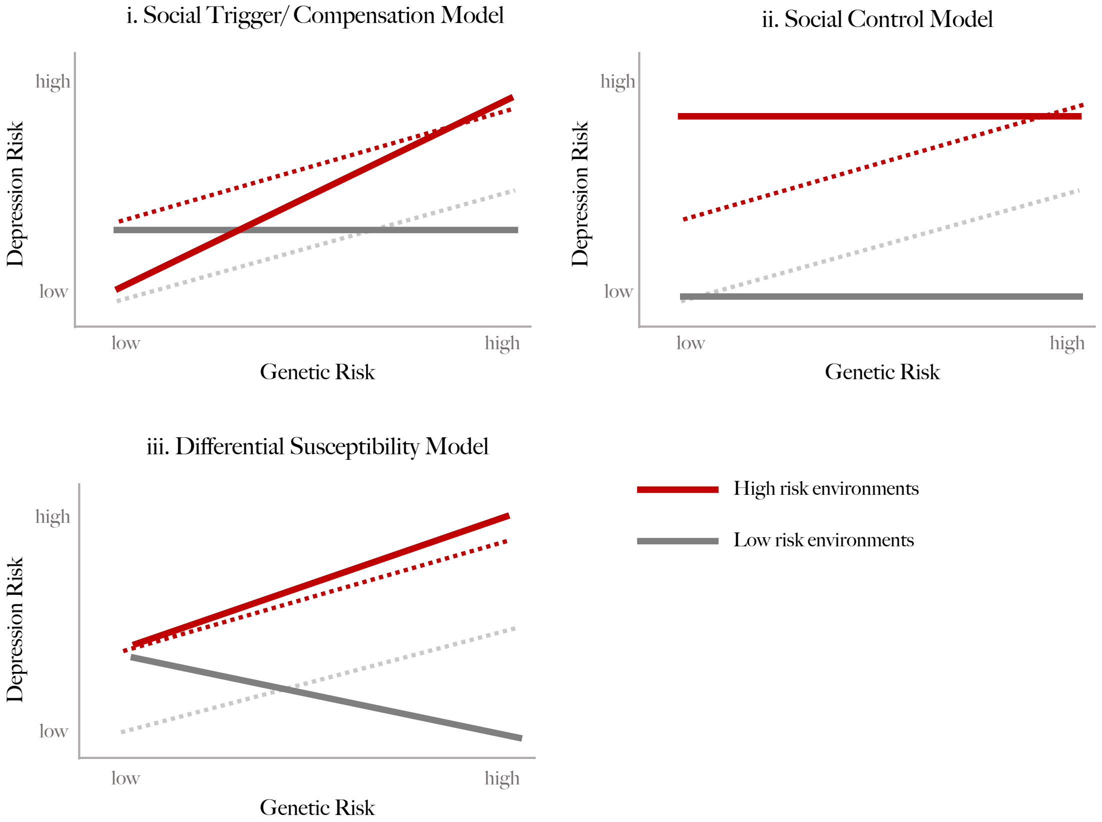


Figure 3.3.1: Conceptual gene × environment interaction models
Note: Red lines demonstrate genetic association with depression in groups exposed to stressful environment, compared to those who do not (gray lines). The amplitude of the genetic correlation is demonstrated by the slope: steeper indicates stronger genetic association. Square dot lines represent the absence of gene × environment interaction trends

et al., 2016], these variations can take different forms ranging from physical and social to contextual and historical. They are proxied in birth cohorts.

Three different models can be used to explain the relationship between contextual and behavioural factors that can moderate genetic predispositions. These models are *the social trigger/compensation model, the social control model,* and *the differential susceptibility model,* which are graphically illustrated in Figure 3.3.1. One example of a similar methodological agenda is the work of Liu and Guo [2015], which takes an analogous conceptual approach to demonstrate its promise for sociological analysis.

The first G×E model, *social compensation/trigger,* is divided into two distinct parts: diathesis-stress and the compensation component. Diathesis-stress associates a trait with stress brought about by extreme life experiences and vulnerability [Domingue et al., 2017a]. If a combination of stress factors is way above normal, then a person develops the associated trait [Domingue et al., 2017a]. For example, exposure to environmental risk factors would amplify the genetic association. This results in a steeper slope, as illustrated in Figure 3.3.1. For those who experience advantageous conditions, the occurrence of depression would not depend on genetic predisposition. This results in a flat slope that illustrates the compensation component.

According to Boardman et al. [2013], the *social control* model refers to situations when phenotypic risks are largely conditioned on environmental characteristics (extreme vs. typical) rather than stabled genetic variants. For example, people who lead a healthy and satisfactory life are less likely to experience depression irrespective of their genes. Exposing people to stressful environments puts them at greater risk of contracting mental health problems, regardless of whether or not they have such genetic predispositions. The latter can be referred to as the social push term described by Liu and Guo [2015], which implies that the most extreme and disadvantaged environments play a considerably larger role than genetic factors in boosting the chances of developing certain traits. Social push can thus be seen as a component of the social control model.

The *differential susceptibility* model considers cases where the same genetic variants are likely to have contrasting effects under adverse environmental conditions [Daw et al., 2013]. For instance, the *differential susceptibility* model predicts those who are more genetically prone to depressive symptoms have a higher chance of being depressed in extreme environmental conditions. Conversely, those who are genetically prone to depressive symptoms but not in extreme environments have reduced chances of depression.

A large portion of the literature on gene-by-cohort interactions focuses on smoking. In their twin study, for example, Boardman et al. [2010] shows that heritability estimates vary across different birth cohorts in the US. While those who were born in the 1920s, 1930s, and 1950s cohorts have strong genetic associations with smoking, cohorts of the 1940s and 1960s have considerably smaller influences. Researchers link these observations with changes in smoking policy in the US: following the passage of legislation making it illegal to smoke in public spaces, gene influences were reduced significantly. However, the first Surgeon General's Report is associated with an increase in genetic influences. This observation is further tested in Domingue et al. [2016]. Researchers show that despite people being aware of the harms associated with tobacco usage, a genetic influence on smoking continues to increase in more recent cohorts. Notably, the findings indicate that those who are genetically at risk of smoking are unlikely to respond to recent policy changes aimed at discouraging people from smoking. These studies of gene-by-cohort variations in smoking are good examples of how informative research in this field could help our understanding of policy changes.

There are also gene-by-cohort studies on alcohol consumption [Virtanen et al., 2019] and BMI [Walter et al., 2016]. Both studies found heritability differences among birth cohorts wherein those born earlier in the 20th century have lower levels of gene-phenotype correlations compared to those born in recent years. Additionally, Conley et al. [2016] conducted a study that aimed at evaluating how genetic penetration has changed in US society in relation to a broad spectrum of phenotypes across different birth cohorts. The research revealed that BMI and height continue having a higher genotypic penetration over the 20th century period, while heart diseases and education declined in genotypic effects. Notably, researchers did not find significant variation in genetic associations with depression across time. However, such a puzzling finding might be partly due to the use of earlier GWAS studies for the construction of polygenic scores – the case is particularly important for depression since, as I showed earlier, previous GWAS were able to identify less than 10 significant SNPs across human genome.

Moreover, findings of increased polygenic penetrance in later birth cohorts must be interpreted with caution as they could be biased from mortality selection in genetic samples. As demonstrated in Domingue et al. [2016], those who were born earlier in the century and genotyped are likely to be survivors (85+) and a non-representative subset of the respective cohorts. Accordingly, lack of correction for mortality selection might lead to the false impression of increased genetic penetrance under the condition of genetic homogeneity in earlier cohorts due to survival. Still, gene-by-cohort interactions offer potentially useful insights for understanding historical changes and cohort variations. There is a noticeable absence of studies covering the UK context in the literature of this field.

This literature review generates three hypotheses that aim to contribute to the existing knowledge. Firstly, I hypothesise that (1) *there are significant cohort variations in the prevalence of depressive symptoms, especially among mid-century cohorts* accompanied by (2) *changing genetic penetrance of depression*. Additionally, following the literature on the impact of economic downturns on mental health of different generations, I hypothesise that (3) *historical contexts (such as economic recessions) modify gene-by-cohort variations*.

3.4 Data

To investigate gene-by-birth cohort trends in depression in the United Kingdom, I use a well-known and widely used longitudinal survey called the UKHLS. Built from a national multi-stage sampling design, the survey covers approximately 40,000 households in England, Scotland, Wales, and Northern Ireland [Buck and McFall, 2011a]. Interviews are carried out every year covering a rich set of questions related to health, socio-economic conditions, and transitions along with family trajectories. The UKHLS has been collecting DNA data since 2010, and has introduced this genetic sample as an additional restricted data source.

After the release of its genetics data, the UKHLS became a unique data source for sociogenomics researchers. The genetic sample contains around 10,000 people, all of whom are adult members of households from the main Understanding Society survey. Originally, 10,484 adult members of households were selected for genome-wide array genotyping. Genotyping was performed with the Illumina Infinium HumanCoreExome BeadChip.

One of the main advantages of this dataset is that the UKHLS genetic sample includes people of all ages and is not restricted to certain birth cohorts. This is particularly important in light of my research question. For the UK context, such a data property offers a unique opportunity as it is common for genetic samples to cover specific age groups (for example, UK BioBank includes people who are more than 40 years old).

3.5 Measures

3.5.1 Depression score

To date, various methods have been developed and introduced to measure depression. Along with information on the diagnosis of this condition, depressive symptomatic data is another valuable source usually provided by surveys. While incidents of depression and its diagnosis indicate the severity of someone's mental health, symptomatic data makes it possible to look at the issue across a broader spectrum and to model the risk of developing depression. Importantly, I do not use depression diagnosis as my main phenotypic variable since it was shown that symptomatic data on mental health is more accurate and has greater validity and reliability in the context of population-based surveys [Mandemakers, 2011].

Depressive symptoms scores are available from multiple surveys and have been used in many empirical studies. The Understanding Society has two scores: GHQ and SF-12. Both of these scores follow a traditional way. Each is assessed by measuring individual psychological states through item-based questionnaires using Likert scales. I use the GHQ score for the main body of my analysis, as it is one of the most widely used and consistently observed during the full survey period. The SF-12 score was included in questionnaires later, consistently present only after 2009.

GHQ was developed as a tool to screen non-psychotic mental health problems [Goldberg et al., 1972]. The modelled symptoms primarily cover depression and anxiety disorders. The overall score is a constructed sum of 12 indicators (usefulness, decision-making, unhappiness, confidence, self-worth, the ability to face problems, joy in day-to-day activities, concentration, loss of sleep, overcoming difficulties, and being under strain). Each item in the GHQ asks respondents to rate the degree

Table 3.5.1: Descriptive statistics of UKHLS analytical sample, by birth cohorts

	All		World Wars		Boomers		GenX		Millennials		
	1919-95		1919-45		1946-64		1965-80		1981-95		Range
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
GHQ scale	10.99	5.26	10.32	4.60	11.32	5.49	11.18	5.36	10.95	5.73	0-36
PGS depression	-.00	1.00	-.05	1.02	-.00	.98	.02	1.00	.11	.99	-3.6-3.9
Female	.56	.50	.53	.50	.56	.50	.59	.49	.57	.50	0-1
Age	51.01	16.88	70.14	9.40	52.22	9.14	35.68	8.11	24.76	5.05	16-96
N participants	*9,113*		*2,458*		*3,607*		*2,372*		*769*		
N observations	*81,246*		*22,293*		*33,214*		*21,005*		*4,734*		

of symptoms from *less than usual* to *more than usual*. Ratings range from 0 to 36, with higher numbers indicating more severe experiences with depression. The GHQ score is a demonstrably valid and reliable instrument for detecting depression in the general population [Lundin et al., 2016].

Table 3.5.1 provides descriptive information for the analytical sample and indicates the mean value of 10.99 for the GHQ depression index. Figure 3.5.1 demonstrates the distribution patterns of GHQ score by birth cohorts. It is notable that the degree of skew differs across cohorts.

3.5.2 Polygenic risk score

Introduced in 2007, polygenic scores are conceptualised as a tool for quantifying the genetic contribution to phenotypes [Wray et al., 2018]. For this study, a polygenic score was constructed using the recent GWAS discovery of depression from Howard et al. [2019]. Additional details on genetic data preparation and construction are provided in the preceding chapter. Briefly, the construction of polygenic scores was performed using PRSice 2.0 software [Choi and O'Reilly, 2019]. There were 360,140 SNPs matched after clumping between reported results in a Howard's GWAS and the UKHLS dataset. The incremental R-square is 0.4% and corresponds to the prediction from the GWAS discovery. Respondents with higher polygenic scores (measured in standard deviation units) reported more depressive symptoms during follow-up (*β=.046, P<.001*).

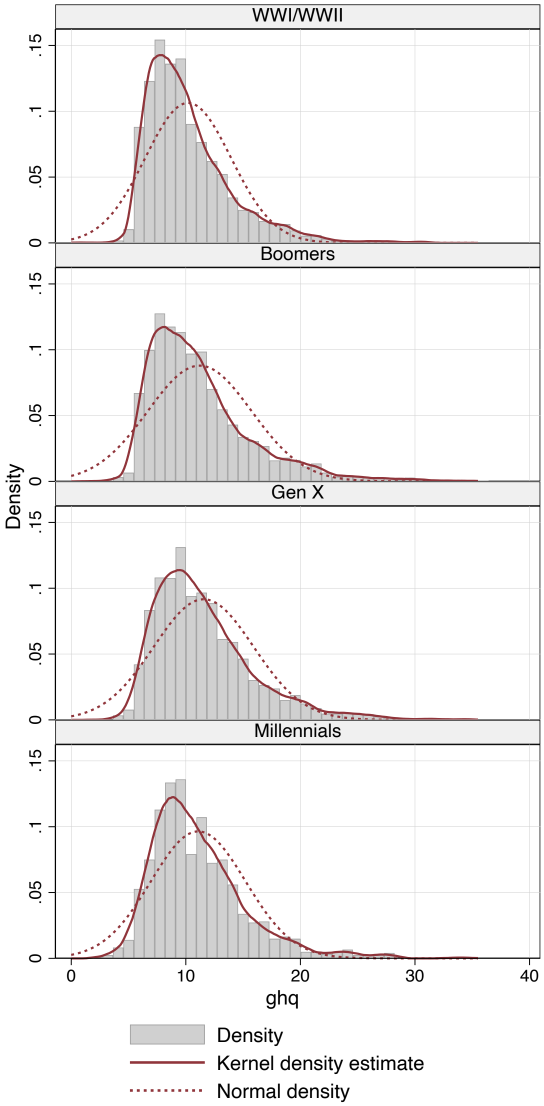


Figure 3.5.1: Distribution of GHQ depressive symptoms across birth cohorts in the UKHLS genetic sample.

3.5.3 Birth cohorts

To model broad historical contexts and exposure to various shifts across different UK birth cohorts, I use the conceptualisation scheme described in Thomson and Katikireddi [2018]. This scheme identifies five birth cohorts, which reflects the conventional classification of demographic cohorts for the 20th century. The first two cohorts are devoted to people exposed to the World Wars: a WWI cohort born between 1916 and 1930, and a WWII cohort with birth years in 1931–1945. The demographic cohort of those born in 1946–1964 is distinguished as Boomers to reflect the postwar period of the baby boom. Thereafter, there is Generation X (people born in 1965–1980) followed by Millennials. Thomson and Katikireddi [2018] use the term 'Jilted' to describe the Millennial cohort (those born between 1981 and 1990) whereas I use the conventional term instead. As genotyping was performed in 2010, it is not feasible to include Generation Z (or Zoomers, in reference to those born in the 2000s). Due to the small number of survey participants from the WWI cohort in the UKHLS genetic sample, I merged it with the WWII cohort to create one 'World Wars' birth cohort.

3.5.4 Recessions

To distinguish possible trends in recessive and non-recessive historical periods, I considered the timing of survey. Figure 3.5.2 shows the average GHQ scores over UKHLS survey years with a focus on periods of economic recession (i.e. the early 1990s and late 2000s). For both recessions, unemployment rates rose by minimum of 7% to peak around 10% during the hardest-hit quarters [Jenkins, 2010]. For this reason, I treat the 1990 and 2008-2010 survey years as recessive. Figure 3.5.2 shows no clear trend in the deterioration of mental health during economic downturns. At the same time, a visible portion of GHQ variation occurs across birth cohorts throughout the survey period.

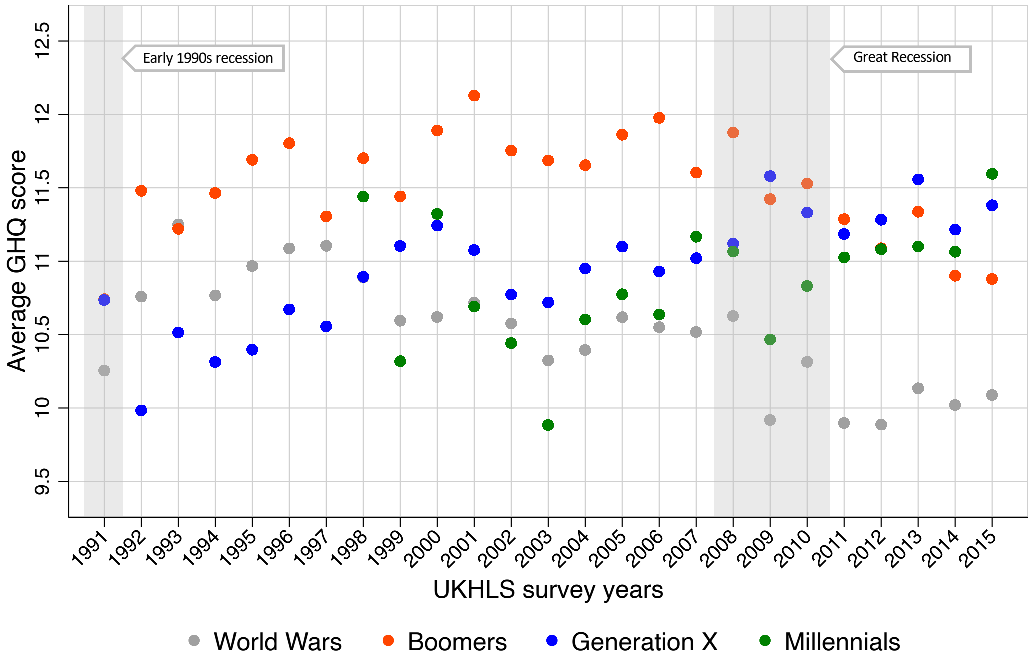


Figure 3.5.2: Average GHQ score by survey years, summarised by birth cohorts.

3.5.5 Covariates

Distinctive sex patterns exist for depressive symptoms, so it is important to control for sex. For instance, women are more likely to be depressed than men [Kuehner, 2017]. Men and women also have different ways of coping with stress [Matud, 2004].

Age is another important covariate, especially for studies of cohort trends. In Figure 3.5.3, I plot descriptive trends for the GHQ score averaged over age and cohorts. The shape is consistent with findings from Prior et al. [2020], wherein a general trend of mental health worsening with age reverses at around 50 years old before deteriorating again in older ages (after around 70). To reflect such a relationship, I include age along with its squared and cubic terms as covariates. This is also the basis for the additional scope of my sensitivity analysis. As I describe in the section to follow, the purpose of this analysis is to further disengage age and cohort trends by analysing age-comparable sub-samples. Lastly, phenotypes have an age-related genetic basis [Kulminski et al., 2016]. It is thus necessary to consider age as an additional covariate in gene-by-cohorts variations.

Although I focus on respondents with European ancestry, the interaction estimates could be confounded by population stratification [Price et al., 2006]. To rule out the possibility of this confounding, 20 first PCs were included as additional covariates for all models (which were provided in the release version of the data, and thus calculated by the UKHLS team).

3.6 Empirical strategy

Gene-by-birth-cohort interactions are examined using multilevel Poisson models. The choice of model is initially determined by the nature of UKHLS data, which contains multiple observations over time. The strategy permits consideration for the correlation of repeated measurements [Hox et al., 2017, Raudenbush and Bryk, 2002]. This correlation is particularly important for the depressive symptoms data.

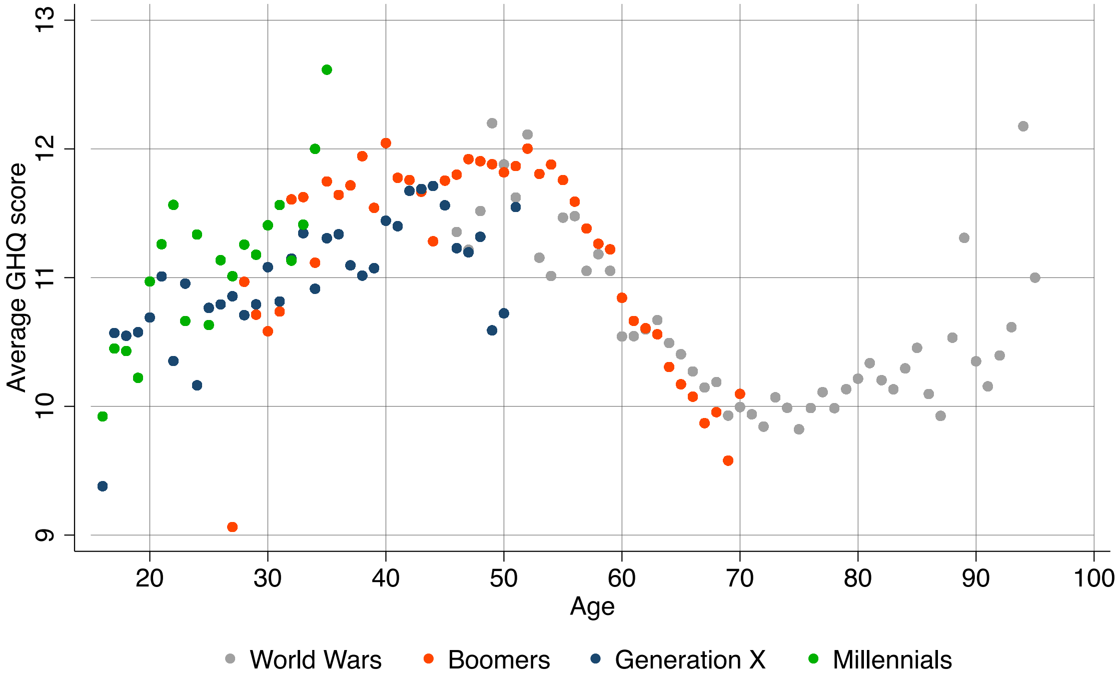


Figure 3.5.3: Average GHQ score by age, summarised by birth cohorts.

Due to the count-based and skewed nature of the depressive symptoms scores (Figure 2.5.2), I employ a Poisson family of regression models. This type of model is an appropriate tool for taking the complexity of skewed data into account [Wooldridge, 2010]. Moreover, it is a preferred method since the produced coefficients are more intuitive and interpretable than the OLS regression estimates with a logged dependent variable [Nichols et al., 2010].

I constructed a two-level Poisson model with the following definitions for hierarchy:

- Level 2: waves, denoted by j;
- Level 1: individuals, denoted by i.

The model is introduced as follows:

$$GHQ_{ij} \sim poisson(\mu_{ij})$$

$$log(\mu_{ij}) = \beta_0 + \beta_1 PGS_i + \beta_2 BirthCohort_i + \beta_3 PGS_i \times BirthCohort_i + \sum_p \gamma_p C_{pi} + \sum_q \gamma_q C_{qij} + u_j$$

$$var(GHQ_{ij}|\mu_{ij}) = \alpha\mu_{ij}$$

where GHQ_{ij} denotes the depression symptom score GHQ of respondent i in a wave j; μ_{ij} denotes the expected score based on a Poisson distribution; C_{pi} represents time-invariant covariates, such as sex and genetic principal components (PCs); and C_{qij} represents time-varying controls, including age and age-sq. The key parameter of interest is β_3, which indicates the marginal association of polygenic scores for different birth cohorts. To test the modifying potential of economic downturns, I perform an analysis of three-way interactions - $PGS_i \times BirthCohort_i \times Recession_{ij}$ in Poisson models.

3.6.1 Gene-by-cohort analysis adjusting for age

The UKHLS is a rich source for the research questions stated earlier. As a multi-cohort longitudinal survey with an analytic period of observation spanning a quarter century, it allows me to disentangle age from cohort trends. However, one limitation is that not all ages are observed for all birth cohorts. For instance, the World Wars cohort was not observed before 45 years of age. Meanwhile, Millennials did not reach this age during data collection. Consequently, differential age distributions within each cohort have a potential to affect the accuracy of cohort estimates [Yang and Land, 2013]. To address this possibility, I perform a sensitivity test where I consider age-comparable sub-samples from overlapping age groups. I am able to conduct pair-wise comparison where I first consider World Wars vs. Boomers for the ages of 46-71, Boomers vs. Generation X for 27-51 years of age, and Generation X vs. Millennials between 16-34 years old. I replicate cohort analysis following a similar statistical approach, then compare the results obtained from the entire analytic sample.

3.7 Results

3.7.1 Cohort variations in the prevalence of depressive symptoms (H1)

In terms of the research hypotheses stated earlier, Table 3.7.1 presents the results obtained from the Poisson regression models assessing cohort and gene-cohort variations in the Understanding Society genetic sample. Following the results from Model 1 in Table 3.7.1, I find evidence of an increase in depressive symptoms occurring among people born in the second half of the 20th century. If the World Wars cohort is used as the reference category, the greatest increases occur in the cohorts of Boomers and Millennials. These results are significant at the *p=.01* level. Generation X experiences a smaller but nonetheless statistically significant increase, as well. These findings are in line with observations from Spiers et al. [2011], Spiers et al. [2012], and Rice et al. [2010]. They do not provide definite support for the proposition of linear trends in the prevalence of depression across birth cohorts proposed by Bell [2014]. There is no evidence of a Jilted generation hypothesis consistent with the study by Thomson and Katikireddi [2018], since the increase is observed not only for Millennials. Age and its quadratic and cubic terms display the trends described in Prior et al. [2020], which further link my findings to the existing literature.

3.7.2 Moderating trends of birth cohorts on the genetic association with depressive symptoms (H2)

In terms of gene-cohort moderation, Model 2 in Table 3.7.1 evidences significant positive interaction in the cohort of Baby Boomers when the earlier World Wars cohort becomes the baseline group. It is likely that the positive interactional trends are present for Generation X as well, but these results are significant at the *p=.10* level. Notably, the size of the interaction coefficient for Baby Boomers is almost the same as the coefficient for PGS. Figure 3.7.1 further illustrates this observation: the strength of the gene-phenotype correlation is greater among Boomers (i.e. participants born be-

Table 3.7.1: Coefficients and standard errors of Poisson multilevel models assessing the moderation of birth cohorts and recessions on the genetic association with depressive symptoms

	Model 1		Model 2		Model 3	
	Beta	Std. Err.	Beta	Std. Err.	Beta	Std. Err.
Cohorts (World Wars - ref.)						
Boomers	.062***	.011	.060***	.011	.059***	.012
Generation X	.039*	.016	.037*	.016	.039*	.016
Millennials	.060**	.023	.055*	.023	.059*	.023
PGS depression			.028***	.006	.032***	.007
PGS × Cohort (World Wars - ref.)						
Boomers			.025**	.009	.020*	.009
Generation X			.010	.009	.008	.010
Millennials			.019+	.012	.015	.015
Recession					.002	.006
Recession × Cohort (World Wars - ref.)						
Boomers					.011	.008
Generation X					.008	.009
Millennials					-.003	.016
Recession × PGS					-.014***	.006
Recession × PGS × Cohort (World Wars - ref.)						
Boomers					.016*	.007
Generation X					.005	.009
Millennials					.015	.017
Female	.114***	.007	.114***	.007	.114***	.007
Age	.032***	.004	.032***	.004	.032***	.004
Age^{2}	-.000***	.000	-.001***	.000	-.001***	.000
Age^{3}	.000***	.000	.000***	.000	.000***	.000
Random-Effect Variance						
$\sigma^{2}u$	.093	.002	.092	.002	.092	.002
AIC	453138.9		452997.4		452991.7	
BIC	453408.8		453304.5		453373.2	
Sample Size						
No. of participants	9,113		9,113		9,113	
No. of observations	81,246		81,246		81,246	

+p<0.1, *p<0.05, **p<0.01, ***p<0.001

Genetic score is standardised; Models include first 20 PCs as covariates

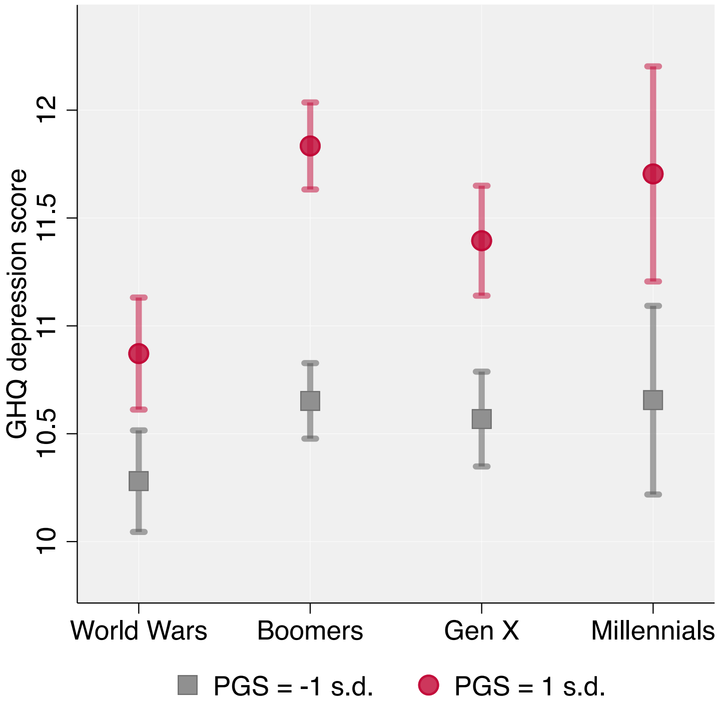


Figure 3.7.1: Birth cohorts as moderators for the association between depression polygenic risk score and depressive symptoms. Marginal probabilities from Poisson multilevel models.

tween 1946-1964) as the gap widens between those with a -1 and +1 standard deviation (s.d.) in genetic risk score.

The results thus show a significant and positive correlation between the polygenic score for depression and depressive symptoms in all birth cohorts. However, the strength of this correlation varies with steeper slopes occuring among Baby Boomers and Generation X. It is apparent that even though these results indicate variation in genetic penetrance among birth cohorts, they do not point to a specific conceptual model of gene × environment interactions in a definite manner. Still, the positive and significant signs of the PGS and PGS × cohort interaction estimates for Baby Boomers suggest that this birth window is likely to be a trigger component reflected in the *social trigger/compensation* G × E model.

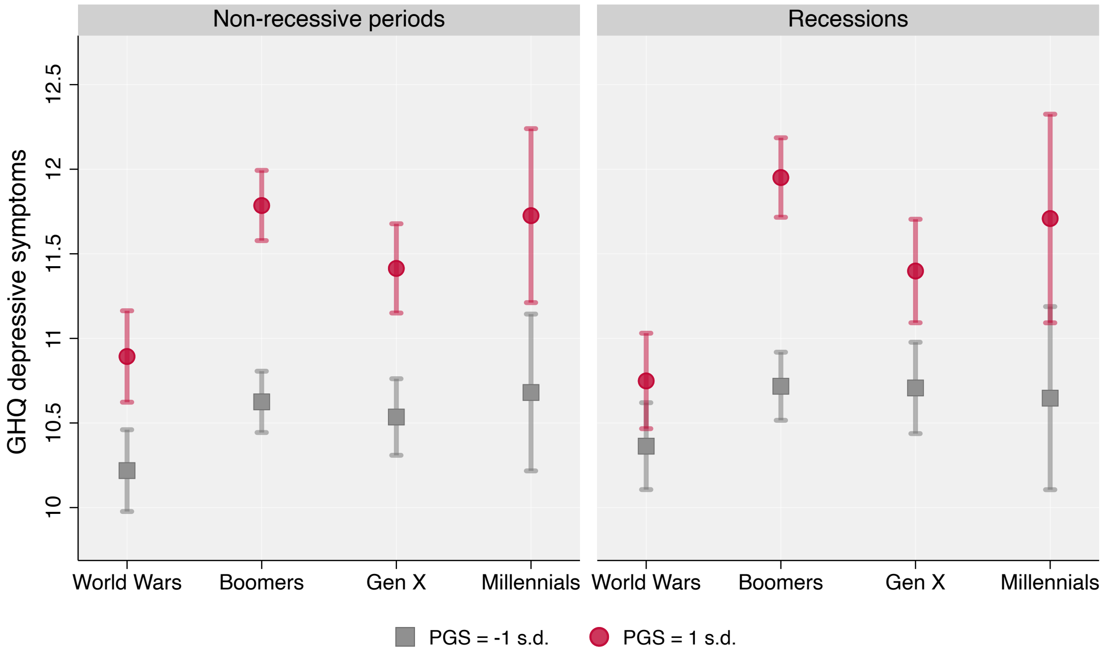


Figure 3.7.2: Birth cohorts as moderators for the association between depression polygenic risk score and depressive symptoms. Marginal probabilities from Poisson multilevel models.

3.7.3 Gene-by-cohort interactions during recessive and non-recessive periods (H3)

Model 3 in Table 3.7.1 and Figure 3.7.2 demonstrate the differences in the moderating patterns of birth cohorts on the genetic penetrance of depressive symptoms during recessive and non-recessive periods. Firstly, I find the polygenic prediction of depressive symptoms weakens during periods of recession; this result is statistically significant (p<.001 for *PGS* × *recessions* interaction term). It implies that recessions, while being exogenous economic shocks, also play the role of an environmental control. This weakens the importance of polygenic signals for a trait, which is consistent with the *social control* model from the conceptual gene × environment framework.

Moreover, I find positive three-way *PGS* × *cohort* × *recession* interaction for Baby Boomers. The result is significant at the p=.05 level. Consistent with H3, I thus find some evidence that gene-by-cohort trends are different during times of economic downturn. In the World Wars, Generation X, and Millennials cohorts, genetic associations with depression tend to be weaker during recessions. For Baby Boomers, the gap widens between those who predisposed to depression and those who are not. This is further evidence of how historical times have the potential to shape polygenic predictions within populations and across generations. It is also an important finding towards the notion that the same environmental condition, such as economic recession in my case, can show potential as both a social trigger and a social control for genetic penetrance (depending on the times people are born into or living in).

3.7.4 Sensitivity analysis (1): Correction for mortality selection

Mortality selection has a potential to bias gene-by-cohort interactions [Domingue et al., 2017b], so an investigation would be incomplete without correction for such potential bias. As I showed in Chapter 2, the issue is especially profound in earlier cohorts. I thus expect estimates for earlier generations to be sensitive to the inclusion of weights. Below, I demonstrate the results of a Poisson regression model with the inclusion of blood weights to reveal gene-by-cohort estimates that are less biased due to mortality selection. As blood weights provided in the UKHLS study correct for the greatest portion of health and socio-demographic selection in the genetic sample, which I address in Chapter 2, these weights are expected to redress the issue sufficiently.

Table B.0.2 displays the results obtained from the weighted Poisson regression models. Looking at columns for Models 1 and 2, it is apparent that the weighted results do not differ significantly from the main analysis. Thus, cohort variations in depressive symptoms (along with increased polygenic penetrance of depression in Boomers) are not sensitive towards the differential probability of selection for genotyping. However, weighted results on differences in the moderating patterns of birth cohorts in terms of the genetic penetrance of depression during recessive and non-recessive periods do not display statistical significance indicating quite marginal trends (observed in the main analysis) sensitive to the implementation of weights.

3.7.5 Sensitivity analysis (2): Age-comparable cohort analysis

So far, the current study has found some evidence supporting gene-by-cohort variations in depression among adults in the UK. While mortality selection is one of the concerns covering a broad aspect of the UKHLS genetic sample, the findings are also likely sensitive to other plausible errors. The list of potential parameters that the findings might be sensitive to includes differential age distributions within each birth cohort, which is not taken into account (even after the inclusion of age along with its squared and cubic terms in the net of covariates). Accordingly, the following section evaluates the consistency of my findings once age-comparable cohort comparisons are applied. I show that the implication of age-comparable cohort analyses indeed changes some of my results.

Tables 3.7.2, 3.7.2, and 3.7.3 below illustrate the results of separate Poisson regressions for age-comparable sub-samples from overlapping age groups (World Wars vs. Boomers aged between 46-71, Boomers vs. Generation X aged between 27-51, and Generation X vs. Millennials aged between 16-34). What stands out in these results is the consistency of trend observed earlier, wherein Baby Boomers have higher depressive symptoms compared to the World Wars cohort. They also likely have higher scores than Generation X as well. However, it is not possible to compare Boomers with Millennials in an age-overlapping manner.

In terms of gene-cohort variations, Figure 3.7.3 graphically displays the observed trend to accompany the results tables. Following analysis of age-comparable sub-samples, the notion that genetic penetrance is greater among Baby Boomers is suggestive. But this result is significant at .10 level only; it is also unsurprising, as the estimate for the interaction term is around 30% smaller than in the main analysis and the sample size shrank as well. In line with previous results, no significant variation was observed for Generation X and Millennials.

Table 3.7.2: Coefficients and standard errors of Poisson multilevel models assessing the moderation of birth cohorts and recessions on the genetic association with depressive symptoms for age-comparable cohorts **[World Wars vs. Boomers, 46-71 y.o.]**

	Model 1		Model 2	
	Beta	Std. Err.	Beta	Std. Err.
Cohorts (World Wars - ref.)				
Boomers	.035**	.012	.029*	.012
PGS depression	.034***	.008	.041***	.009
PGS × Cohort (World Wars - ref.)				
Boomers	.018+	.009	.010	.010
Recession			-.010	.008
Recession × Cohort (World Wars - ref.)				
Boomers			.025*	.010
Recession × PGS			-.020**	.008
Recession × PGS × Cohort (World Wars - ref.)				
Boomers			.025**	.009
Female	.121***	.009	.121***	.009
Age	.364***	.078	.354***	.079
Age^2	-.006***	.001	-.006***	.001
Age^3	.000***	.000	.000***	.000
Random-Effect Variance				
$\sigma^2 u$	.098	.002	.098	.002
AIC	216914.6		216900.7	
BIC	217163.6		217184.1	
Sample Size				
No. of participants	5,205		5,205	
No. of observations	39,562		39,562	

*+p<0.1, * p<0.05, ** p<0.01, *** p<0.001*

Genetic score is standardised; Models include first 20 PCs as covariates

Table 3.7.3: Coefficients and standard errors of Poisson multilevel models assessing the moderation of birth cohorts and recessions on the genetic association with depressive symptoms for age-comparable cohorts **[Boomers vs. Generation X, 27-51 y.o.]**

	Model 1		Model 2	
	Beta	Std. Err.	Beta	Std. Err.
Cohorts (Boomers - ref.)				
Generation X	-.025*	.012	-.028*	.012
PGS depression	.041***	.008	.046***	.008
PGS × Cohort (Boomers - ref.)				
Generation X	-.004	.010	-.008	.011
Recession			-.000	.009
Recession × Cohort (Boomers - ref.)				
Generation X			.013	.011
Recession × PGS			-.017+	.008
Recession × PGS × Cohort (Boomers - ref.)				
Generation X			.009	.011
Female	.111***	.010	.111***	.010
Age	.010	.050	.005	.050
Age^2	-.000	.001	.000	.001
Age^3	-.000	.000	-.000	.000
Random-Effect Variance				
$\sigma^2 u$	.088	.002	.088	.002
AIC	190336.1		190331.9	
BIC	190579.7		190609.1	
Sample Size				
No. of participants	4,178		4,178	
No. of observations	32,890		32,890	

*+p<0.1, * p<0.05, ** p<0.01, *** p<0.001*

Genetic score is standardised; Models include first 20 PCs as covariates

Table 3.7.4: Coefficients and standard errors of Poisson multilevel models assessing the moderation of birth cohorts and recessions on the genetic association with depressive symptoms for age-comparable cohorts **[Generation X vs. Millennials, 16-34 y.o.]**

	Model 1		Model 2	
	Beta	Std. Err.	Beta	Std. Err.
Cohorts (Generation X - ref.)				
Millennials	.011	.017	.012	.018
PGS depression	.033**	.010	.038**	.011
PGS × Cohort (Generation X - ref.)				
Millennials	.014	.016	.010	.017
Recession			-.003	.014
Recession × Cohort (Generation X - ref.)				
Millennials		-.006	.020	
Recession × PGS			-.015	.014
Recession × PGS × Cohort (Generation X - ref.)				
Millennials			.015	.021
Female	.111***	.016	.111***	.016
Age	.191**	.073	.189*	.074
Age^2	-.007*	.003	-.007*	.003
Age^3	.000*	.000	.000*	.000
Random-Effect Variance				
$\sigma^2 u$	.086	.004	.086	.004
AIC	70242.2		70247.5	
BIC	70456.6		70491.5	
Sample Size				
No. of participants	1,847		1,847	
No. of observations	11,997		11,997	

*+p<0.1, * p<0.05, ** p<0.01, *** p<0.001*

Genetic score is standardised; Models include first 20 PCs as covariates

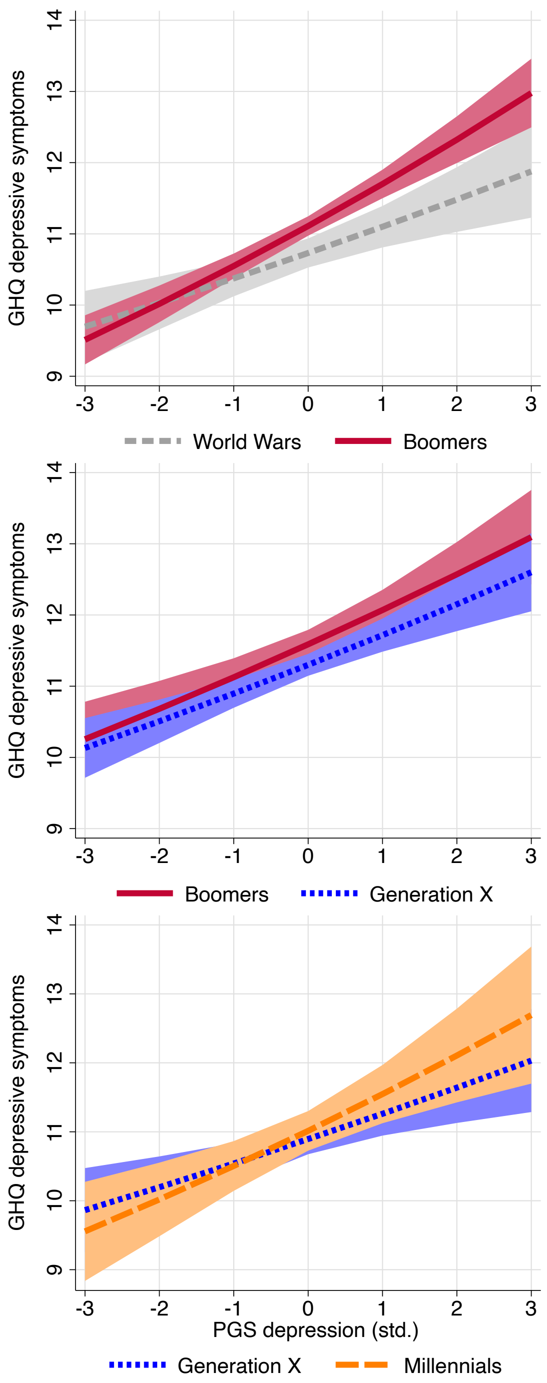


Figure 3.7.3: Birth cohorts as moderators for the association between depression polygenic risk score and depressive symptoms. Marginal probabilities from Poisson multilevel models.

Turning to the discussion of the role of economic downturns on genetic penetrance of depression among birth cohorts, I include Figure 3.7.4 to graphically represent my results. I find that the result of weakening polygenic prediction of depressive symptoms (during periods of recession) is also suggested in age-comparable robustness analysis. Yet the likelihood of this trend is rather marginal for Millennials, which is directly linked to the considerably smaller sample size for comparison of Generation X - Millennials. I also find the observation that genetic penetrance of depressive symptoms in Baby Boomers is stronger during recessions is robust and present in age-comparable analysis.

3.8 Results and discussion

This study examines how birth cohorts and recessions moderate genetic influence on depressive symptoms among adults in the UK. First, the paper contributes to the G×E literature in mental health research and represents an interdisciplinary research agenda. This approach to understanding mental health is necessary to provide new sets of insights. One of the challenges for the field is the absence of consistent results; effect sizes are usually small, necessitating increased power and rich data sources. This results in conceptual difficulties along with the challenge of operationalising the 'environment' in a meaningful way (see Boardman et al. [2014], Pehkonen et al. [2017] for more detailed discussion). This paper examines multiple individual observations in time to grasp trends more accurately than other G×E studies, which typically consider one environmental factor at a time. The rich set of controls and sensitivity tests aims to obtain more robust inferences. The results for the covariates are consistent with those in the existing literature. I have critically assessed G×E models and built my argument on distinctive components of the *social trigger/compensation*, *social control*, and *differential susceptibility* theories.

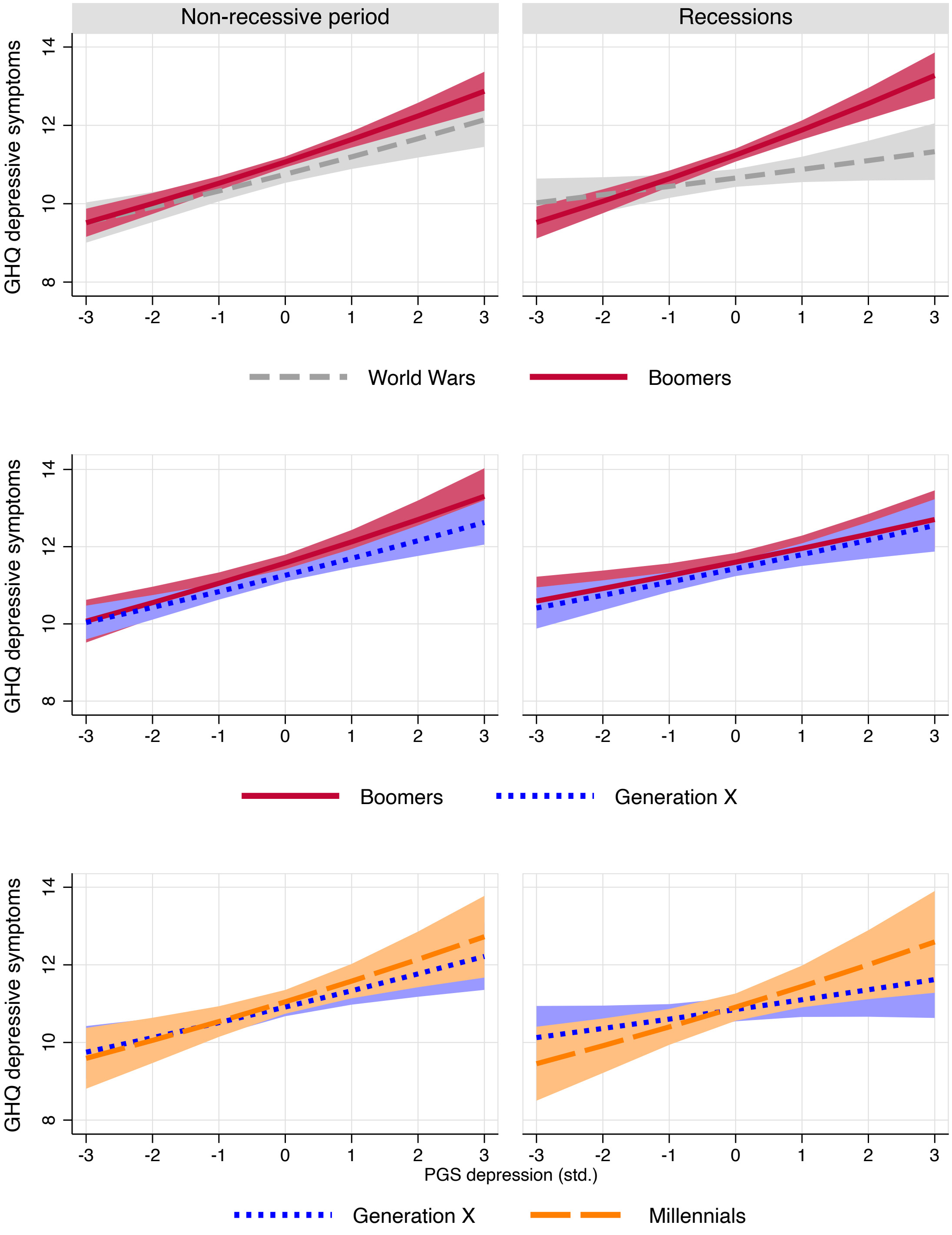


Figure 3.7.4: Birth cohorts as moderators for the association between depression polygenic risk score and depressive symptoms. Marginal probabilities from Poisson multilevel models.

Second, I find evidence that an increase in depressive symptoms occurred among people born in the second half of 20th century. Age-comparable analysis indicates that the increase is especially profound for Baby Boomers. Third, I do not find support for the proposition of increased genetic penetrance on depressive symptoms across all cohorts. I find evidence of increased depression prevalence in two birth cohorts, as well as significant gene-cohort moderation patterns for one of the cohorts (Baby Boomers) and suggestive moderation for the Generation X cohort. These findings are robust towards mortality selection, and marginal in the range of age-comparable robustness checks. These findings further contribute to the notion that the cohort of Baby Boomers is different from others: they achieved higher educational attainments, experienced more marital disruptions and changes in family structures that constitute their distinctive life histories [Dennis and Migliaccio, 1997] and results in greater polygenic penetrance of depression. Lastly, I find that the polygenic prediction of depressive symptoms weakens across all birth cohorts during periods of recession – except for Baby Boomers. Thus, historical times have the potential to shape polygenic predictions within populations and across generations differently. Moreover, my findings on recessions also indicate that not only cohort-specific historical exposures can potentially shape genetic penetrance, but also historical exposures experienced by everyone have differentiating trends.

However, there are limitations to this study. The most significant is the issue of depressive symptoms measurement. The notion of measurement challenges arises from psychology and psychiatry. There is an ongoing debate about what harmonised and meaningful procedures must be developed to grasp the non-diagnosed dimensions of depression. These dimensions affect how we understand mental health problems and the true scale of these problems in society (a more detailed discussion of this issue can be found in Kinderman [2014]). There are no systematic patterns in how the response bias of mental health affects estimated correlations [Gove and Geerken, 1977]. Especially taking into account that symptomatic data of depression compared to diagnostic is more accurate and has greater validity and reliability in the context of population-based surveys [Mandemakers, 2011]. In using the widely documented and analysed GHQ depressive symptoms index, the threat of measurement challenges might not be critical but it is necessary to acknowledge possible silencing. Furthermore, I did not consider qualitative differences between birth cohorts and their experiences of different economic periods. These can be explored in further research on this topic. I also did not analyse minority samples as there is no genetic sample in the UKHLS presenting different ancestries, which is one of the most critical issues in the field of sociogenomics (see Mills and Rahal [2019] for a more detailed discussion). Future research should extend the analysis to other racial populations once data become available.

Despite these limitations, this study demonstrates that the occurrence of depressive symptoms among adults in the UK is a consequence of complex interplay among individuals' genes, incidence of birth, and the historical context of economic recession. The results illustrate the importance of applying molecular genetic data to advance our understanding of well-established links in public health and social science enquiry.

Chapter 4

Worklessness, genetic risks of depression, and gene-environment correlations

4.1 *Abstract*

The development of depression has a highly compound nature involving both genetic and environmental mechanisms. Complex interplay between genes and environments can take the form of gene × environment interactions and gene-environment correlations. As discussed in the previous chapter, the concept of gene × environment interactions provides an opportunity to reveal environment-dependent variation in genetic associations. Another aspect of the interplay, though less often studied, is gene-environment correlations. Such correlations take place when genes are associated with variations in exposure to adverse or protective environments. The presence of a gene-environment correlation implicates social environments as a potential causal pathway between genes and the onset or development of depression. This paper focuses *on worklessness as an aspect of the socio-economic environment because it creates stressful life conditions and puts people at higher risk of experiencing depression. Accordingly, this chapter investigates whether genetic predispositions to depression are associated with a higher probability of experiencing worklessness. The presence of rGE trends suggests that, as one of the risk factors of depression, the absence of a job is a new element to consider in understanding genetic influences on depression. Results on the importance of baseline mental health as a selection factor for worklessness have been mixed. The presence of correlation between an underlying vulnerability factor (such as genetic risk) is further evidence that job loss is not exogenous to susceptibility to depression. Using data from the Understanding Society genetic sample, I conduct a multinomial regression analysis with additional sensitivity tests that take selection into genotyping into account.*

4.2 Introduction

The core of the difficulty in tackling different stages of depression lies in the complexity of the causes that play a role in the development of depression. Depression has been shown to have a genetic basis: twin studies suggest that the heritability of depression is 37% [Sullivan et al., 2000] while GWAS heritability estimates vary from 2 to 9% [Howard et al., 2019, Okbay et al., 2016, Wray et al., 2018]. Depression can be also triggered by socio-economic factors. Among these, worklessness is one of the most serious concerns.

Worklessness creates serious problems for both policymakers and society in ways beyond individual mental health. Worklessness can be caused by an individual becoming economically inactive (i.e. no longer searching for a job) or unemployed. Extensive reviews of current evidence show that, when combined with economic inactivity, unemployment increases the risks of developing depression [Bartley, 1994, Ezzy, 1993, Fryers et al., 2003, McLean et al., 2005]. Unemployment spells are further associated with general health deterioration [Heggebø and Elstad, 2017] and even a higher risk of suicide [McDaid, 2017]. People without jobs may experience psychological tensions (especially anxiety and depression) that negatively impact their health, the security of their family, and the stability of the society in general. Worklessness is mainly understood as an economic problem and enquiries into its nature tend to stem from such perspectives [Drydakis, 2015]. Yet worklessness often creates a toll that ranges from economic to psychological concerns. This is because as a condition, it does not merely impose financial constraints. It also creates an absence of connection to other individuals, routine, and security.

Researches have extensively addressed the link between worklessness and depression. On the one hand, evidence shows that worklessness increases the risks of developing this mental health problem (for example, see Drydakis [2015], Paul and Moser [2009], Murphy and Athanasou [1999]). Moreover, this link is demonstrably causal. On the other hand, there is discussion that depression leads to worklessness. While depression is one of the causes of disability that might determine economic inactivity, evidence for a baseline status of depression as a significant predictor of unemployment status has been mixed [Butterworth et al., 2012, Dooley et al., 1994, 2000, Egan et al., 2015, Jefferis et al., 2011].

We thus know that the onset of depression is highly complex in nature, wherein both genetic and non-genetic mechanisms are involved [Kendler et al., 2006]. From the perspective of social science, we know that worklessness is one of the socio-economic factors causing depression. However, our knowledge is limited and rooted in the notion that genes and worklessness play independent roles. To date, no study has investigated the interplay between the genetic risks of depression and worklessness.

Worklessness creates a toll of concerns that determines one's life experience. Following formal definitions for sociogenomics, it can be understood as socio-economic environmental factor. In the literature on sociogenomics, complex interplay between genes and environments can take the forms of gene × environment interactions and correlations [Plomin et al., 1977]. Conceptually, gene × environment interactions provide an opportunity to reveal environment-dependent variation in genetic associations. Empirically, they are challenged on various bases – especially when environmental exposures are of a non-exogenous nature, which is implied in situations of economic inactivity and unemployment. Moreover, the presence of a correlation between genetic risks and environments further biases interaction analysis [Elbaz and Alpérovitch, 2002]. Another aspect of the interplay takes place when genes are associated with variations in exposures towards adverse or protective environments (namely, gene-environment correlations or rGE) [Jaffee and Price, 2007]. In terms of empirical study, gene-environment correlations receive less attention compared to gene × environment interactions.

However, several studies on rGE have shown that genetic predispositions towards depression increase the probability of experiencing high-risk social environments. These environments include bullying [Veldkamp et al., 2019], the absence of emotional parental support [Wilkinson et al., 2013], and adverse life events [Lau and Eley, 2008]. The presence of such gene-environment correlations gives rise to the notion that social environments may function as a causal pathway linking genes to depression. The novelty of the current study is its focus on worklessness as another aspect of environmental exposure putting people at a higher risk of experiencing depression.

Accordingly, this paper analyses whether genetic predispositions towards depression are associated with a higher chance of experiencing worklessness which constitutes the main research question of this chapter. I conceptualise worklessness as incidents of unemployment or economic inactivity. I further distinguish economically inactive individuals by reason for inactivity, whether disability or family care. Genetic predispositions are conceptualised as polygenic scores for depression, which were constructed using the recent GWAS discovery from Howard et al. [2019]. In particular, following the conceptualisation of worklessness, this paper examines the following sub-questions: (1) *are polygenic scores for depression associated with increased likelihood in unemployment (for those who are actively seeking a job)?* (2) *Are polygenic scores for depression associated with increased likelihood in economic inactivity?*

Following the literature on sex differences for depression, I also conduct sex-specific analyses in order to investigate potentially different trends across the sexes. The presence of rGE trends would suggest that, as one of the social-environmental risk factors for depression, worklessness is a new element to consider when understanding genetic influences on depression. Since results regarding the importance of baseline mental health have been mixed, the presence of a correlation between an underlying vulnerability factor (such as genetic risk) and unemployment is further evidence that job loss is not exogenous to depression susceptibility.

This study uses data from the Understanding Society genetic sample. I conduct a multinomial regression analysis and perform a sensitivity test to correct for selection into genotyping, both for all respondents and by sex. The paper begins by presenting the theoretical background for the issue, discussing worklessness an environmental risk factor for depression. Next, I demonstrate the relationship between genetic influence and the development of depression. I then turn to a conceptual framework of gene-environment correlations, wherein I cover different mechanisms of rGE. I detail my empirical strategy in an overview of measures and statistical methods. Finally, I provide extensive discussion of the results along with an assessment of non-linearities and selection biases.

4.3 Background

4.3.1 Worklessness increases risks of depression

There is a substantial body of literature on relationships between labour force status and depression (see reviews from Ezzy [1993], Bartley [1994], Fryers et al. [2003], McLean et al. [2005]). Along with short-term job loss, periods of prolonged economic inactivity increase the risks for depression development across different age groups [Egan et al., 2015, Frese and Mohr, 1987, Strandh et al., 2014, Winefield and Tiggemann, 1990] in different countries [Butterworth et al., 2012, Jefferis et al., 2011, Winefield and Tiggemann, 1990]. In the UK context, macro-level factors (such as welfare benefits and payments) along with individual-level characteristics (sex, relationship status, familial and social supports) have demonstrably confounded the link between worklessness and mental health deterioration [McLean et al., 2005].

Worklessness typically increases during economic recessions, but it can also occur as a result of disease or physical injury [Brydsten et al., 2015]. It is thus common practice to distinguish between those who are seeking a job while unemployed and those who are not. The 'economically active' group of unemployed people refers to situations where an individual has reached the working age and actively searches for full-time employment [Clegg, 2016, Hussmanns et al., 1990]. The 'economically inactive' (i.e. not in the labour force) group of unemployed people refers to those who have stopped looking for a job for certain amount of time. This situation can occur due to care-giving or family care needs, prolonged diseases, injuries, etc. Notably, the absence of a job is one of the most significant predictors of mental health deterioration – regardless of whether or not an individual is seeking re-employment [Owen and Watson, 1995, Rai et al., 2013]. Moreover, long term unemployment is shown to be more harmful than short-term instances [Herbig et al., 2013]. Stankunas et al. [2006] compared the group of long-term unemployed with those unemployed for a short time, for example, to find that long-term unemployment results in more frequent occurrences of depressive mood.

People who do not have jobs tend to think about future ways of survival and how well they can take care of their family. This leads to higher stress levels and depression because individuals have to meet their daily needs and without a job, they may lack the income to do so [Huffman et al., 2015]. A job routine can make people feel as though they are contributing to the community, which may provide a sense of purpose [Schwartz, 2015]. The unexpected loss of a job can trigger even greater insecurity and fear, which can lead to depression especially in nations with lower levels of pre-crisis unemployment. According to data from World Health Organisation [2020], every year over 800,000 people around the world commit suicide due to depression particularly in light of job deficiency.

From a psychological perspective, the absence of a job leads to lower self-esteem and stronger external locus control; the risk of developing depressive symptoms rises in turn [McLean et al., 2005, Waters and Moore, 2002]. According to Statt [1994], the initial stages of unemployment are usually marked with optimism as people think they will find other opportunities. Individuals believe they are not the only ones who are not working and their economic status is a consequence of external factors.

As their financial situation worsens, people's hopes about the future deteriorate. They develop a sense of helplessness that can give rise to anxiety, insomnia, reduced self-esteem, drug abuse, and other psychosocial problems (e.g. domestic violence). As a result, depression sets in and mental health is adversely affected. The overall effect on well-being is a loss of optimism about finding a new job, with individuals blaming themselves for the inability to retain a position [Daly and Delaney, 2013].

One of the most robust observations in the epidemiology of mental health is the existence of a sex gap in depression. Accordingly, there are also distinctive sex patterns linking depression and joblessness. First, women are more likely to be depressed [Kuehner, 2017] and more likely to be unemployed [Albanesi and Şahin, 2018, Antecol, 2000]. Second, there are sex differences in the psychological distress caused by unemployment [Ensminger and Celentano, 1990]. Third, men and women have different stress coping styles [Matud, 2004]. This suggests that the harmful consequences of worklessness likely operate through different mechanisms. For instance, marriage is shown to be a protective factor against developing depression during spells of unemployment mostly among female employees [McLean et al., 2005].

4.3.2 Depression as a predictor of worklessness

While studies from the list of reviewed literature have addressed the question of mental health and worklessness, most efforts aimed at tracking the relationships between unemployment as a cause and depression as an effect. Few studies suggest that the correlation may operate in both ways: with depression triggering the unemployment, and unemployment causing depression [Jefferis et al., 2011]. Therefore, evaluation of the effects of unemployment on levels of depression should be considered alongside the possibility of reverse causality. It further suggests a need to address how different states of mental health might determine employment status before investigating any related consequences for unemployment.

A relatively smaller number of studies track the role of depression as a determinant of worklessness. Hence, there is some evidence to suggest that being depressed is associated with higher chances of worklessness (though the results are mixed for the unemployment component of worklessness) [Butterworth et al., 2012, Egan et al., 2015, Frese and Mohr, 1987, Winefield and Tiggemann, 1990]. In various ways, mental illness has always had an adverse effect on labour market outcomes. This is observable in possibilities for employment. In a study conducted by Nelson and Kim [2011], the authors examine the relationship between the diagnosis of mental disorders and the risk of employment termination. They find that depression increases the chances of unemployment in the future. There is a large employment gap between depressed individuals and those with depressive symptoms when compared to those without mental disorders. This can be linked to the harmful consequences of depression. For example, it is known that depression can severely influence individual performance and productivity in the workplace [Keyes, 2005]. Depression can also lead to lack of sleep, which can interfere negatively with daytime performance in turn [Keyes, 2005].

Moreover, some alarming results show that a history of mental disease can make someone vulnerable to work-related discrimination. For example, people with mental disorders are more likely to be absent from work than those with a chronic physical disability [Tefft, 2011]. Employers can tend to ignore mental issues [Kessler et al., 1999]. This may result not only in unemployment, but also in the *discouraged worker* phenomenon – when an individual has given up looking for a job and becomes economically inactive, even though they would prefer to be employed [Akyeampong, 1989].

All in all, worklessness is a prominent risk factor for both major and minor depression [Rosholm and Andersen, 2010]. However, current literature indicates that the association between unemployment and depression likely operates both ways: depression is associated with a higher chance of being unemployed, while the job loss experience worsens one's state of mental health in turn. The possibility of reverse causality is challenging for drawing a conclusive picture since it implies the existence of selection bias. Notably, findings on the harmful consequences of worklessness are not necessarily in question. This is because studies take into account both cross-sectional differences and longitudinal aspects of depression development along with the baseline depressive symptoms of respondents. Hence, unemployment and economic inactivity are risk factors for depression.

Selection bias poses a challenge to understanding the mechanisms driving these associations. For example, depressive symptoms might lead to job loss [Egan et al., 2015] that increases the risk of developing clinical depression. Under this scenario, worklessness would amplify depressive symptoms. It is also possible that several depressive episodes are not necessarily causally associated with each other. Instead, they mark the existence of depression vulnerability factors such as genetic predispositions [Wilkinson and Goodyer, 2011]. In this case, the absence of a job is likely an environmental risk factor presenting in a developmental pattern of depression through the mechanism of gene-environment correlation (rGE), which I cover in the sections below. As studies on the importance of baseline mental health have diverse results (for example, Dooley et al. [1994], Dooley et al. [2000], Butterworth et al. [2012], Jefferis et al. [2011]), the existence of a correlation between an underlying vulnerability factor (such as genetic risk) and worklessness would be further evidence on the importance of baseline depression factors.

4.3.3 Genetic grounds for depression

Most people experiencing unemployment or economic inactivity (even for prolonged spells) do not develop serious mental health problems or exhibit significant increases on a depressive symptoms scale. But it is crucial to note that worklessness is only one of the predictors that can be included on the list of so-called risk factors of depression. In quantitative forms of research, the puzzle of depression exists without a strong singular predictor. We know that a variety of pressures boost the probability of depression, including those from childhood trauma, substance use, negative family relationships, social disadvantages, and so on [Brown and Harris, 2012, Lopizzo et al., 2015, Peters et al., 2015]. But still, the predictive power of statistical models is relatively low. We thus cannot conclude that trauma and risk factors (in their broad conceptualisation) are the only relevant pieces of the puzzle. What is missing is our biology.

There have been several attempts to reveal a single gene responsible for depression (for a review, see Bosker et al. [2011]). Such a research focus aligns with the past trend and prevalence of a candidate gene approach to genetic studies. Moreover, it was known that some diseases are caused by single mutations. For example, depression is an important factor in the clinical representation [Epping and Paulsen, 2011] of Huntington's disease [Chial, 2008]. Research hypotheses are also based on observations that depression develops in families [Tsuang and Faraone, 1990]. Twin studies show that if a person's parents suffer from depression, their risk of developing the illness ranges from 20 to 40% [Jansson et al., 2004, McGue and Christensen, 2003, Sullivan et al., 2000]. However, such a concordance rate indicates that most people with depression do not necessarily have a depressed family member. In that regard, depression does not have the inheritance pattern similar to single-genetic-mutation illnesses such as Huntington's disease. With further developments in the field and poor replication of rates of candidate gene results for depression [Bosker et al., 2011], it is now clear that no single gene is responsible for depression. While genes are important, their influence is due to the small effects of many SNPs across the genome. This indicates that depression is a polygenic trait.

The approach widely used to discover polygeneity is called GWAS. As I noted in Chapter 2, the rationale behind this method is to screen all of the SNPs in the human genome and to test their association with a certain phenotype. Multiple studies have looked at depression as a phenotype [Hek et al., 2013, Okbay et al., 2016, Terracciano et al., 2010, Wray et al., 2018]. The most recent and largest one, which I use to construct polygenic scores, was conducted by Howard et al. [2019]. 102 independent variants, 269 genes, and 15 gene sets have been associated with depression to date. This includes both genes and gene pathways associated with synaptic structure and neurotransmission [Howard et al., 2019].

Researchers have also highlighted the importance of certain SNPs for depression risk. In major depressive disorders (MDD), rs7647854 on chromosome 3 was found to play a significant role [Power et al., 2017]. Rs19323608 on chromosome 17 further points towards the influence of genetics in the onset of depression [Okbay et al., 2016]. Such detailed analyses of SNPs and their locations are required for further investigation of causal links because GWAS is a descriptive approach that only establishes associations. But as I show in Chapter 2, some of the identified SNPs and regions likely contribute to the known biological causes responsible for the development of depression. The portion of identified genetic signals are linked to expressions of cortisol and serotonin [Bansal et al., 2016, McGowan et al., 2009]. Additionally, some of the SNPs associated with depression are also located in the coding region for the central nervous system and exert their effects in transcriptions for the development of this system [Howard et al., 2019, Hyde et al., 2016].

Importantly, the up-to-date, SNP-based heritability of depression is around 9% [Howard et al., 2019]. We thus cannot treat genes in deterministic terms for such a complex trait. Genetic predisposition is yet another risk (or protective) factor. It is an important element to take into account along with environmental pressures.

4.3.4 Gene-environment correlations (rGE)

Depression has a genetic and environmental net of causes. With the growing prevalence of interdisciplinary approaches, we know that an important and complex interplay exists between genes and other risk factors for depression. One under-examined but promising area of research into health inequalities investigates how genes and different forms of economic activity (or inactivity) are jointly responsible for observed mental health outcomes. Following the literature on sociogenomics, depression can result from gene × environment interactions (G×E), gene-environment correlations (rGE), and epigenetic mechanisms of gene expressions. In the paragraphs below, I mainly focus on the first two scenarios for interplay and set epigenetics aside. [1]

In the G×E approach, environmental conditions are associated with buffers and/or stressors of genetic predispositions that are liable to cause certain health outcomes [Seabrook and Avison, 2010]. Human genes respond to environmental variations differently [Courtiol et al., 2016]. Environmental variations can take different forms ranging from physical to social. Having covered gene × envi-

[1] While the first two types of interplay treat genetic predispositions as constant over the course of a lifetime, epigenetics aims to address the dynamic structure of gene expressions. Since this paper is not concerned whether and how environments modify genetic expressions that were not necessarily inherited, I instead refer readers to Shanahan and Hofer [2011] for a more detailed discussion on this matter.

ronment interactions in the previous chapter of my thesis, I will only mention here that G×E interactions remain a central aspect of sociogenomics. For my purposes here, it should be stressed that robust statistical identification of G×E patterns requires the absence of a correlation between genes and environments. This is because the observed G×E might be a direct product of the correlation, which would generate a misleading picture of development.

Gene-environment correlations are another aspect of the interplay between genes and environmental factors that can be present in the developmental pattern of traits. Compared to G×E interactions, gene-environment correlations are often less studied. rGE occurs in situations when individual exposure to an environmental context depends on heritable inclinations [Jaffee and Price, 2007]. But strictly speaking, rGE is a statistical correlation between genetic predisposition and exposure to an environment. Consequently, rGEs could be perceived as statistical abstracts but they have conceptual basis coming from behavioural genetics. Behavioural geneticists characterise three mechanisms through which genetic and environmental factors correlate with one another: passive, evocative, and active [Plomin et al., 1977].

The first type is passive rGE, which arises because parents do not merely transmit genes that might promote the development of a particular phenotype or trait. They may also create a rearing environment that promotes development of the same trait. Consequently, it induces a correlation between the genes parents transmit and the environment they create for their children (Panel A of the Figure 4.3.1 illustrates this notion).

We might consider a depressed parent, for instance, where we know depression is a heritable trait. As shown in Wilkinson et al. [2013], a depressed parent passes genes on to his/her children that increase the probability of developing depression. However, the parent is also less likely to provide high levels of emotional support and warmth due to their own depression. In turn, this creates an environment that fosters the development of depression in his/her children. Transmitted genes thus become correlated with the environment the parent provides for a child. In behavioural genetics, this situation is referred to as passive rGE because a child does not play an active role in it. It happens by the will of his/her parents, who essentially enforce the situation. Methodologically, this means the correlation between genes and environment is not causal. It is driven by parental characteristics, as demonstrated on Panel A Figure 4.3.1 by the dotted arrow between G and E. Once parental characteristics are omitted from a regression analysis (as a confounder that is not controlled for), we observe a correlation between G and E. In the other types of gene-environment correlations, the child/person involved plays a much more prominent role.

Let's consider passive rGE in relation to the genetic risk of depression and labour force participation. Labour force status is an aspect of adulthood. It could be argued that parents with a greater genetic risk of depression pass on to their children more than just associated risk alleles. They also contribute to their child's experiences in the labour market and career trajectories. This can happen unintentionally through indirect risk factors that increase the likelihood of their children experiencing job loss or economic inactivity. Such risk factors can include (but are not limited to) neighbourhood and school characteristics. Following this scenario, labour force status is one of the products of passive rGE.

The second type of gene-environment correlation is evocative. Evocative rGE describes a situation where environments, including social environments we all experience, are partly a function of our own behaviour. This is genetically influenced to some extent, and certain illicit reactions from others creates a correlation between our environments and our genes. In other words, individual behaviour or characteristics that are (even partially) genetically driven provoke a certain response from the environment or causes for the environment. In terms of genetic predispositions and labour force status, this type of correlation could occur in the form of a recruitment agent who is more likely to make a positive hiring decision about candidates without symptoms of depression and anxiety. The extent to which such symptoms are heritable will induce a correlation between genes and joblessness mediated by the recruitment process. Methodologically, evocative rGE is illustrated in Panel B of Figure 4.3.1. In contrast to passive rGE, the evocative correlation refers to a causal association

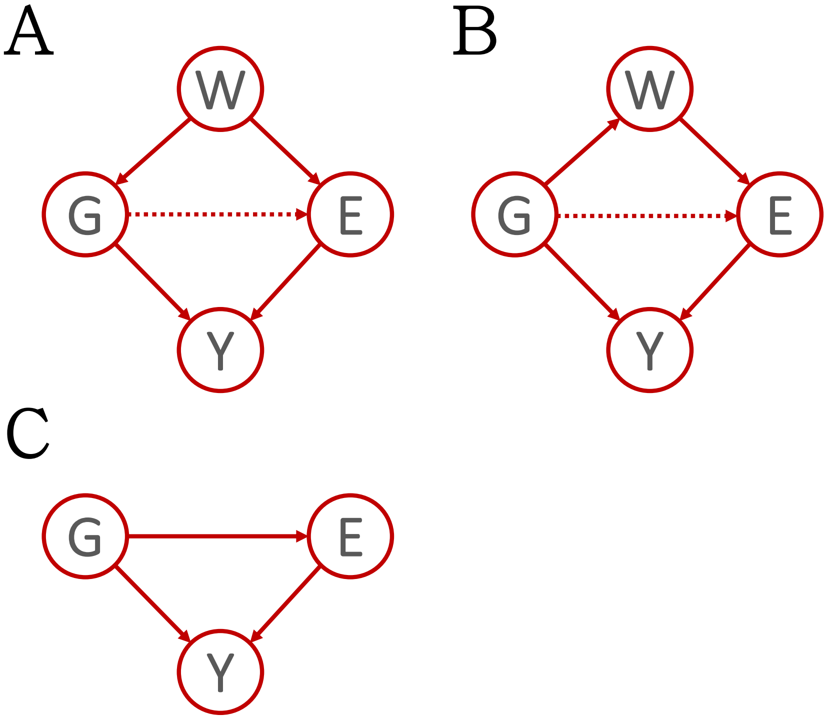


Figure 4.3.1: Conceptualisation of gene-environment correlations (rGE) following causal graphic methods. G represents genetic predisposition, E is an environmental factor, Y is an outcome, W represents parental characteristics and choices (for more straightforward illustration I restrict W to parents only).

Panel A. Passive rGE. Correlation between G and E is not causal, which reflected by a dashed line, and observed once parental characteristic, W, is not controlled for in a regression analysis. W here is a confounder.

Panel B. Evocative rGE. Correlation between G and E is not causal, reflected by a dashed line, and is driven by a mechanism, W, which includes modified behaviours of parents in relation to values of G.

Panel C. Active rGE. Correlation between G and E is causal, reflected by a solid line, where G causes E.

between G and E that can be explained by mechanism W. This mechanism includes the modified behaviours of parents, teachers, peers, colleagues, employers, etc. Empirically, this type of correlation is shown for educational achievements (where parents are more likely to adjust education decisions about their children in regards to their child's genotype) [Avinun and Knafo, 2014, Klahr and Burt, 2014].

The third type of rGE is active. This occurs under conditions wherein our experiences result from personal choices we make based on our abilities and interests. To the extent that our abilities and interests are genetically influenced, it will induce gene-environment correlation. It thus describes situations where selection into environmental exposure occurs depending on heritable inclinations; 'active' is used here to indicate an individual actively shaping the environment he/she experiences. This correlation is empirically illustrated in twin studies on lifestyles and choices. For example, McGue et al. [2014] shows that monozygotic twins are more similar than dizygotic twins in terms of lifestyle preferences (regarding physical, intellectual, and social activities along with diet). To some extent, these lifestyle choices are products of personal decisions. Methodologically, we would call such choices *self-selection*. Greater similarity in monozygotic twins is explained as the result of genetic differences to a certain degree. We know that periods of depression can lead to prolonged economic inactivity; hence, while depression has genetic grounds, we might observe gene-environment correlation that would reflect direct selection into environment (as demonstrated in Panel C of Figure 4.3.1. However, the presence of rGE is not a proof in itself of a causal relationship between genes and environments. Rather, it is a potentially important element in understanding how life choices and circumstances are generated.

The conceptual framework of gene-environment correlation is of great importance to both behaviour geneticists and social scientists. Firstly, rGE is an important element in understanding parent-child resemblance (primarily of concern to social mobility research). Secondly, most substantial topics in sociology cover complex human behaviours and actions that have both social and biological grounds (i.e. educational attainment, fertility, mental health, and so on). For such outcomes, rGE is one of the mechanisms of behavioural development.

I have introduced the conceptual framework of rGE to demonstrate the theoretical motivation for my research enquiry. But as I explicitly discuss in the following sections, this paper does not aim to test the sub-types and mechanisms of rGE. I intend to conduct an exploratory and descriptive analysis as the first step in developing an empirical understanding of the role of worklessness in the complex link between genes and depression.

4.4 Research hypotheses

The purpose of this study is to shine a new light on interdisciplinary research and to identify whether genetic risks for depression are associated with worklessness – a question that has not yet been answered. The presence of such rGE trends would suggest the absence of job is one of the environmental risk factors for depression. This would make worklessness a new element in our consideration of genetic influences on depression. More specifically, this paper aims to examine whether a polygenic score for depression is associated with increased likelihood in unemployment (for those who are actively seeking a job) or in economic inactivity.

Following the literature revealing sex differences, I also conduct sex-specific analyses in order to reveal potentially different trends across sex groups. Importantly, I test a general presence of rGE. This is an essential first step for further research into specific subtypes of rGE, which would constitute a separate set of research questions. Hence the exact mechanisms of selection into labour force status, which is due to genetic differences in depression risks, are yet to be established.

4.5 Data and measures

4.5.1 UK Household Longitudinal Study (UKHLS) – BHPS genetic sample

The UK Household Longitudinal Study (Understanding Society) is a well-known and widely used data source. It is built on a national multi-stage sampling design and covers approximately 40,000 households in England, Scotland, Wales, and Northern Ireland [Buck and McFall, 2011b]. Interviews are carried out every year covering a rich set of questions related to health, socio-economic conditions, and transitions along with family trajectories. Since 2010, the UKHLS has been collecting DNA data and has introduced its genetic sample as an additional restricted data source. The genetic sample contains adult members of households from the main Understanding Society and BHPS surveys. DNA genotyping was performed in waves 2 and 3 after follow-up health assessment visits from registered nurses. Genotyping was performed with the Illumina Infinium HumanCore-Exome BeadChip.

After the release of genetics data, this study became an efficient instrument for sociogenomics researchers. One of the main properties of this data is its representativeness of the white subpopulation with European ancestry in the UK. Additionally, the UKHLS genetic sample includes people of all ages. Such range is particularly important for the research questions addressed here. For the UK context, this data property offers a unique opportunity as it is common for genetic samples to cover specific age groups (for example, UK BioBank only includes people older than 40 years old). This leads to age-specific findings as labour force trajectories vary across age groups.

But as I discuss in Chapter 2, genotyping was not performed randomly. Consequently, the issue of genetic sample selection should be addressed responsibly. This necessity is further justified by the range of health criteria included in a recruitment process, which might lead to a lack of generalisability for the findings. As a part of the robustness analysis, I focus my investigation on the implementation of weights and check whether the gene-environment correlations are consistent.

The initial genotyped sample contains 69,608 observations on 9,944 individuals, 5,568 women and 4,376 men. This is the data release version. After applying the genetic quality control filters covered in Chapter 2, the sample size decreased to 9,196 participants. Considering age bounds for labour force participation eligibility and missing responses on the main variables of interest and covariates included in the models, the final analytical sample consists of 58,891 observations on 7,211 individuals. This amounts to 4,070 women and 3,141 men of European ancestry. For the majority of individuals, there are multiple observations over time.

4.5.2 Worklessness

I conceptualise worklessness as unemployment or economic inactivity. The unemployed are those who do not have a job but are actively seeking one. Here, unemployment is conceptualised by the following ILO definition: the situation where an individual who has reached working age is unable to acquire a job and is actively in search of fulltime employment [Clegg, 2016, Hussmanns et al., 1990]. People who have not looked for a job in the past four weeks are treated as economically inactive (i.e. not in the labour force). In the UKHLS sample, there are two reasons for economic inactivity: disability and family care (as differentiated in my analysis). The baseline group for comparison is those who are employed, whether in terms of a job or self-employment. My analysis excludes respondents who are students, on maternity/paternity leave, or retired.

My study compares three groups: the employed, the unemployed, and those not in the labour force. Unemployed and economic inactivity present worklessness. Tables 4.5.1 and 4.5.2 provide descriptive information for the whole analytical sample and by sex. In general, 16.3% of respondents reported unemployment and 24.1% were economically inactive at least once throughout the survey.

4.5.3 Polygenic risk score

Introduced in 2007, polygenic scores are conceptualised as a tool to quantify the genetic contribution to phenotypes [Wray et al., 2018]. For this study, a polygenic score was constructed using the recent GWAS discovery of depression from Howard et al. [2019]; Chapter 2 provides additional details on genetic data preparation and construction. Briefly, the construction of polygenic scores was performed using PRSice 2.0 software [Choi and O'Reilly, 2019]. After clumping between results reported in Howard's GWAS and the UKHLS dataset, 360,140 SNPs were matched The incremental R-square is 0.4% and corresponds to prediction from the GWAS discovery. Respondents with higher

Table 4.5.1: Descriptive statistics of UKHLS analytical sample, by sex and worklessness status

Whole analytic sample								
	All		Employed		Unemployed		Not in lab. force	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Depression PGS	0.01	0.99	-0.02	0.99	0.19	0.99	0.15	1.03
Sex	0.56	0.50	0.53	0.50	0.49	0.50	0.81	0.41
Age	45.26	12.61	45.12	12.16	42.46	13.48	47.18	13.95
N participants	*7211*		*6370*		*1172*		*1740*	
N observations	*58891*		*49793*		*2201*		*6897*	
Men								
	All men		Employed		Unemployed		Not in lab. force	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Depression PGS	0.02	0.98	-0.00	0.98	0.17	0.98	0.25	0.98
Age	45.88	12.70	45.62	12.53	43.56	13.58	52.40	13.00
N participants	*3141*		*2892*		*553*		*364*	
N observations	*26031*		*23561*		*1132*		*1338*	
Women								
	All women		Employed		Unemployed		Not in lab. force	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Depression PGS	0.00	1.01	-0.03	1.00	0.22	1.00	0.12	1.04
Age	44.77	12.54	44.67	11.81	41.30	13.31	45.93	14.02
N participants	*4070*		*3478*		*619*		*1376*	
N observations	*32860*		*26232*		*1069*		*5559*	

Table 4.5.2: Detailed descriptive statistics for worklessness group, by sex

Whole analytic sample of economically inactive				
	Family care		Sickness/disability	
	Mean	SD	Mean	SD
Depression PGS	0.05	1.03	0.31	0.99
Sex	0.94	0.22	0.58	0.49
Age	43.94	13.17	51.31	7.87
N participants	*1195*		*695*	
N observations	*4310*		*2587*	
Men				
	Family care		Sickness/disability	
	Mean	SD	Mean	SD
Depression PGS	0.20	0.99	0.27	0.94
Age	45.63	11.88	52.78	7.79
N participants	*94*		*287*	
N observations	*231*		*1107*	
Women				
	Family care		Sickness/disability	
	Mean	SD	Mean	SD
Depression PGS	0.04	1.03	0.35	1.02
Age	43.85	13.23	50.20	7.78
N participants	*1101*		*408*	
N observations	*4079*		*1480*	

polygenic scores (measured in standard deviation units) reported more depressive symptoms during follow-up (*β=.046, P<.001*). Tables 4.5.1 and 4.5.2 demonstrate that the unemployed and those who are not in a labour force have higher means of polygenic scores in comparison to those who are employed. Figures C.0.1 and C.0.2 in Annex C show the distribution of standardised values for polygenic scores by employment status.

4.5.4 Covariates

Gene-environment correlation analysis includes sex and age (and its squared term) as a set of controls along with the first 20 genetic principal components (PCs). In the sex-specific analysis, I include age, age-sq., and PCs as covariates. As mentioned earlier in the paper, women and men differ in labour force participation trends [Albanesi and Şahin, 2018, Antecol, 2000], stress coping styles [Matud, 2004], and experiences with depression [Kuehner, 2017]. All these factors imply the likelihood that rGE varies across sexes, which motivates its inclusion as one of the controls and expansion to a sex-specific hypothesis and analysis. Age is another important covariate, since phenotypes have an age-related genetic basis [Kulminski et al., 2016]. Labour force participation also differs with age: there is a U-shaped relationship between age and unemployment chances [Hughes and Hutchinson, 1986], for example, indicating that younger and older workers are at higher risks for job loss.

Although I focus on those of European ancestry, the estimates might be confounded by population stratification [Price et al., 2006]. To rule out the possibility of this confounding, 20 first PCs were included as additional covariates to all models (which were provided with the release version of data, and thus calculated by the UKHLS team).

Importantly, the set of controls does not include socio-economic factors, such as educational attainment, marital status, and first-employment occupation characteristics. These variables are conventionally controlled for when modelling worklessness. This analytical decision extends from the notion that, rather than confounding regression estimates, any phenotypic variations in adulthood are actually present on the mediation path from polygenic score to worklessness (being divorced, for example, does not change your genetic predisposition to depression). Therefore, the gene- environment correlation analysis includes sex, age, and PCs as covariates.

4.6 Empirical strategy

Gene-environment correlations are examined using multilevel multinomial logistic regressions. Multilevel models were chosen due to the nature of the UKHLS genetic sample, which contains multiple observations over time; hence, labour force status varies across waves depending on participant. As noted in the previous chapter, this strategy accounts for the correlation of repeated observations [Hox et al., 2010, Raudenbush and Bryk, 2002] – an important aspect in modelling labour force participation. For example, Heckman and Borjas [1980] shows that the probability of future unemployment varies by the number of unemployment spells experienced in the past; this implies that repeated observations of labour force status for the same individuals are indeed likely to be correlated.

Since labour force participation is a discrete response variable with three unordered categories (i.e. *employed*, *unemployed*, and *not in labour force*), I use multinomial logistic regressions. This is one of the most commonly used statistical methods for modelling labour force status (for example, see De Jong and Madamba [2001], Woodland [1987]), and it complements the multilevel structure of the data (for instance, see Ward and Dale [1992]).

Two-level multinomial models were constructed with the following definitions for hierarchy:

- Level 2: waves, denoted by j;
- Level 1: individuals, denoted by i.

With the *employed* category as the reference, I am interested in estimating the relative probability of being either unemployed or not in the labour force. Consequently, the model introduces the following:

$$P(Y_{ij} = k|PGS_i, x_{ij}) = \frac{exp[\alpha_k + \beta_{k1}PGS_i + \beta_{k2}PGS_i \times PGS_i + \beta_{k3}x_{ij}]}{\sum_{h=1}^{K} exp[\alpha_h + \beta_{h1}PGS_i + \beta_{h2}PGS_i \times PGS_i + \beta_{h3}x_{ij}]}$$

where *k* = *1, 2,* and 3 representing categories of the response variable for respondent *i* in a wave *j*. The sum in the numerator is provided for the probabilities over the categories to be 1. The numerical values of *k* are used as labels, where *K* represents a baseline group and *h* represents the parameters of baseline category (i.e. *employed*). x_{ij} represents a vector of control variables that includes both time-invariant covariates (e.g. polygenic score, sex, and genetic principal components) and time-varying controls (e.g. age). The key parameter of interest is β_{k1}, which indicates the marginal association of polygenic score with labour force status. I am also interested in β_{k2} to address possible non-linearity in the associations of interest.

4.7 Results

4.7.1 Depression polygenic score and the probability of unemployment (H1)

The first set of research questions aimed to address whether the genetic risks of depression are associated with a higher probability of unemployment (following the ILO definition of this term). Table 4.7.1 and Figure 4.7.1 below present the results obtained from multinomial multilevel regression models using the Understanding Society genetic sample. Firstly, I find evidence that higher values of polygenic scores for depression are indeed associated with a higher probability of unemployment. Following the parameters reported in the top panel of Table 4.7.1, one standard deviation (+1 s.d.) in polygenic scores increases the odds of unemployment versus employment by 1.57 times (*β=.45; p<.001*) while holding all other variables in the model constant. To further illustrate the relationship between polygenic scores for depression and unemployment, Figure 4.7.1 below plots the estimated marginal probabilities against standardised values for polygenic score. There are separate colour bars for employed, unemployed, and economically inactive groups, respectively.

The results of the sex-specific analysis are displayed in Table 4.7.2 and Figure 4.7.2 below. Gene-environment correlations between depression polygenic scores and unemployment can be observed for both females and males. For women, 1 s.d. increase in polygenic scores is associated with 1.65 higher odds of unemployment compared to employment (*β=.50; p<.001*) while holding all other variables in the model constant. For men, the same 1 s.d. increase in depression genetic score is associated with 1.43 higher odds of unemployment (*β=.36; p<.001*) compared to employment.

In line with existing literature, my analysis also shows that women are more likely to experience worklessness compared to men [Albanesi and Şahin, 2018, Antecol, 2000]. I find further support for the existence of a U-shaped relationship between age and unemployment as indicated in Hughes and Hutchinson [1986], demonstrating that younger and older workers are at higher risks for job loss compared to the middle-aged group.

4.7.2 Depression polygenic score and the probability of economic inactivity (H2)

The second research question addressed whether the genetic risks of depression are associated with a higher probability of economic inactivity. Accordingly, the bottom panel of Table 4.7.1 reveals a significant positive association between depression polygenic scores and the probability of becoming economically inactive (or not in the labour force). For 1 s.d. increase in polygenic scores, the odds of economic inactivity are 1.79 times larger compared to those who are employed (*β=.58; p<.001*) while holding all other variables in the model constant. These findings indicate the presence of rGE wherein the genetic risks of depression are associated with a higher probability for both types of worklessness.

Table 4.7.1: Coefficients and standard errors of multinomial multilevel model assessing the association between depression polygenic score and worklessness

	Parameters	β	Std. Err.
$\frac{P(Y_i=unemployed)}{P(Y_i=employed)}$	PGS depression	0.45***	0.06
	PGS depression^2	0.05	0.04
	Female	0.51***	0.12
	Age	-0.26***	0.02
	Age^2	0.00***	0.00
	Intercept	-0.30	0.39
$\frac{P(Y_i=not\ in\ lab.\ frc.)}{P(Y_i=employed)}$	PGS depression	0.58***	0.08
	PGS depression^2	0.12*	0.06
	Female	3.64***	0.18
	Age	-0.25***	0.02
	Age^2	0.00***	0.00
	Intercept	-4.57***	0.44
Random-Effect Variance			
	σ^2_{u1}	19.84***	1.25
	σ^2_{u2}	39.43***	2.18
	AIC	35,584.07	
	BIC	36,078.16	
Sample Size			
No. of participants		7,211	
No. of observations		58,891	

*+p<0.1, *p<0.05, **p<0.01, ***p<0.001*

Genetic score is standardised; Model includes first 20PCs as additional covariates.

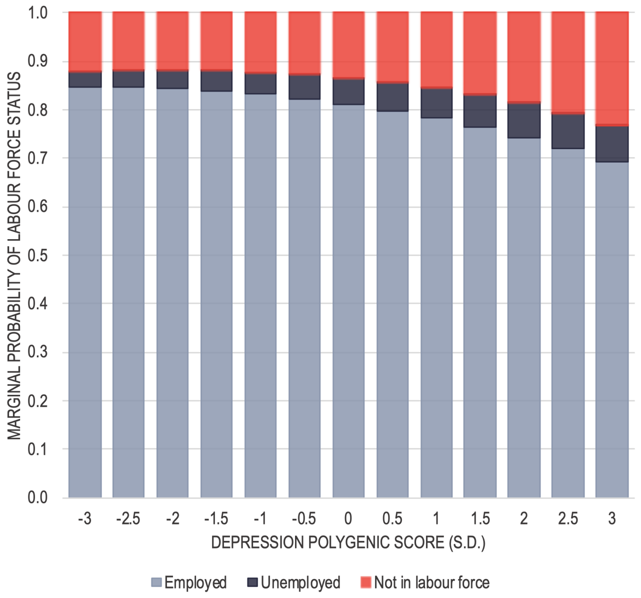


Figure 4.7.1: Marginal probabilities of worklessness in relation to polygenic score of depression

Figure 4.7.1 also illustrates the relationship between polygenic scores for depression and economic inactivity. One of the interesting aspects is the visualisation of a non-linear relationship between genetic score and the probability of economic inactivity. The presence of a significant and positive squared-term polygenic score in the model of economic inactivity (*β=.12; p<.05*) indicates that if a person has greater numbers of risk alleles associated with depression, his/her chances of becoming economically inactive are disproportionally higher. Such a finding further corresponds to the literature on liability threshold models using polygenic predictions (see review by Chatterjee et al. [2016]). Namely, it indicates that the manifestation of outcomes occurs once the number of risk alleles exceeds a certain threshold. This is not a surprising observation in the context of this paper as one of the reasons of economic inactivity is mental disability, which has a polygenic basis in turn. All in all, this result suggests that rGE patterns likely arise not only in the form of general linearities, but also as subjects for further investigation in line with the liability threshold approach for stratified disease preventions. Notably, the squared-term polygenic score in the model for ILO unemployment is positive but this parameter is not significant (*β=.05; p>.1*).

Table 4.7.2: Coefficients and standard errors of multinomial multilevel model assessing the association between depression polygenic score and worklessness, by sex

		Men		Women	
	Parameters	β	Std. Err.	β	Std. Err.
$\frac{P(Y_i\text{=}unemployed)}{P(Y_i\text{=}employed)}$	PGS depression	0.36^{***}	0.09	0.50^{***}	0.09
	PGS depression^2	0.04	0.06	0.04	0.06
	Age	-0.26^{***}	0.03	-0.26^{***}	0.02
	Age^2	0.00^{***}	0.00	0.00^{***}	0.00
	Intercept	-0.74	0.60	0.36	0.50
$\frac{P(Y_i\text{=}not\ in\ lab.\ frc.)}{P(Y_i\text{=}employed)}$	PGS depression	0.95^{***}	0.18	0.50^{***}	0.10
	PGS depression^2	0.01	0.12	0.15^{*}	0.07
	Age	-0.29^{***}	0.05	-0.23^{***}	0.02
	Age^2	0.00^{***}	0.00	0.00^{***}	0.00
	Intercept	-8.07^{***}	1.16	-0.34	0.45
Random-Effect Variance					
	σ^2_{u1}	24.22^{***}	2.36	17.51^{***}	1.40
	σ^2_{u2}	76.26^{***}	7.52	29.64^{***}	1.93
	AIC	11,246.46		24,248.00	
	BIC	11,679.31		24,693.20	
Sample Size					
No. of participants		3,141		4,070	
No. of observations		26,031		32,860	

$+p<0.1$, $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

Genetic score is standardised; Model includes first 20PCs as additional covariates.

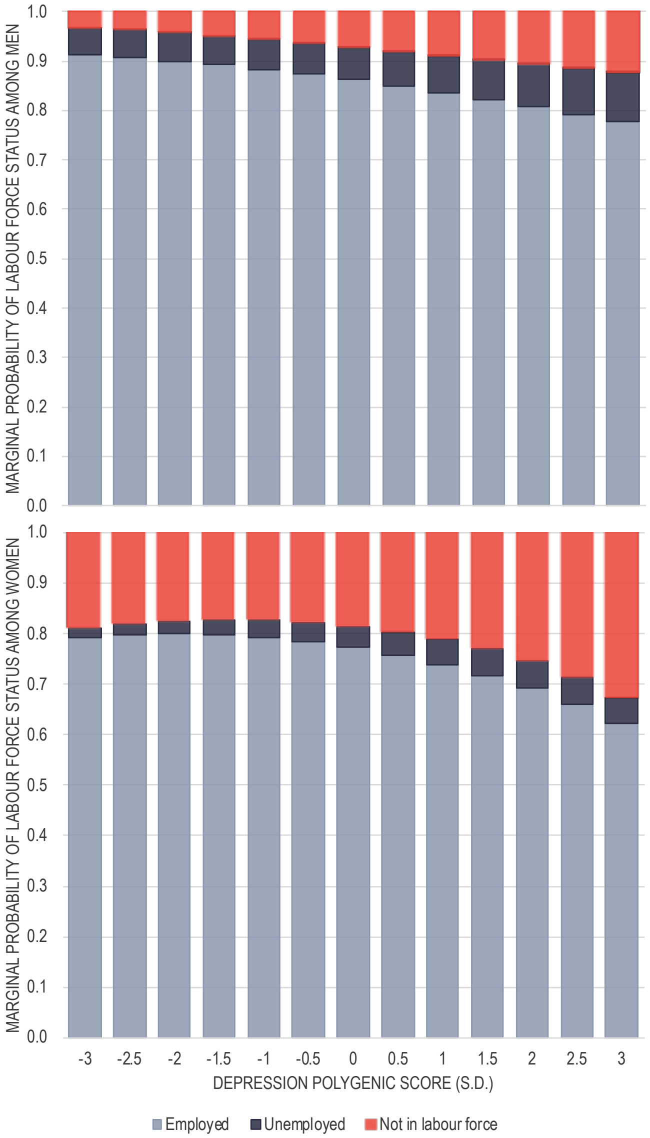


Figure 4.7.2: Marginal probabilities of worklessness in relation to polygenic score of depression, by sex

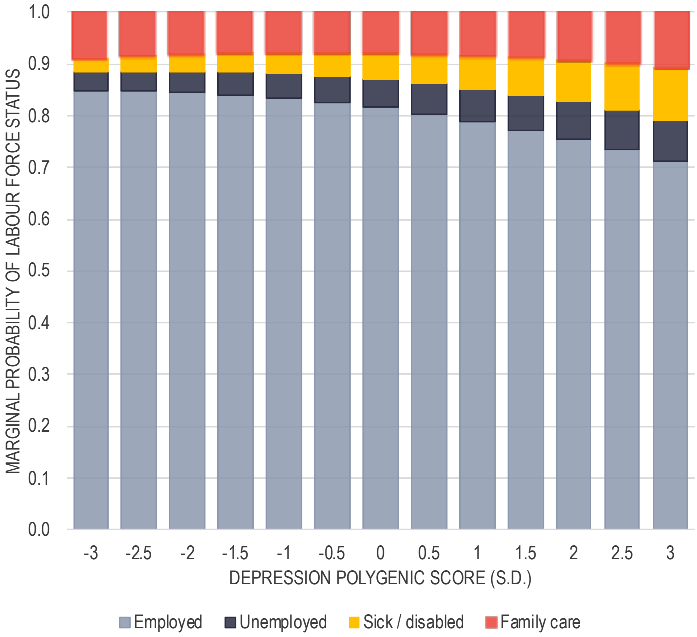


Figure 4.7.3: Marginal probabilities of worklessness with detailed profile of economic inactivity in relation to polygenic score of depression

The correlation between depression polygenic score and economic inactivity was also tested using a more detailed profile of economic inactivity. Here, I distinguish the reasons for not being in the labour force (as subtracted from the UKHLS questionnaire): inactivity due to sickness or disability, and inactivity due to family care needs. Table 4.7.3 and Figure 4.7.3 display the results obtained from the regression analysis following the same empirical strategy as before. Unsurprisingly, once the reasons for economic inactivity are taken into account, we see that polygenic prediction is the strongest for absence from the labour force due to sickness or disability (*β=1.07; p<.001*) in comparison to absence due to family care (*β=.30; p<.001*). Notably, the results for both outcomes are statistically significant.

The correlation between polygenic score and the sickness/disability reason for inactivity further reflects the importance of underlying health. A higher genetic predisposition for depression is also associated with a higher genetic predisposition for other traits, such as schizophrenia [Grotzinger et al., 2019]. Such a correlation can reflect direct selection, which would correspond to active rGE.

Table 4.7.3: Coefficients and standard errors of multinomial multilevel model assessing the association between depression polygenic score and worklessness with detailed profile of economic inactivity

	Parameters	β	Std. Err.
$\frac{P(Y_i\text{=}unemployed)}{P(Y_i\text{=}employed)}$	PGS depression	0.39***	0.06
	PGS depression^2	0.04	0.04
	Female	0.44***	0.12
	Age	-0.25***	0.02
	Age^2	0.00***	0.00
	Intercept	-0.58	0.40
$\frac{P(Y_i\text{=}family\ care)}{P(Y_i\text{=}employed)}$	PGS depression	0.30***	0.08
	PGS depression^2	0.10+	0.06
	Female	5.45***	0.23
	Age	-0.26***	0.02
	Age^2	0.00***	0.00
	Intercept	-5.32***	0.50
$\frac{P(Y_i\text{=}sick\ /\ disable)}{P(Y_i\text{=}employed)}$	PGS depression	1.07***	0.14
	PGS depression^2	0.08	0.09
	Female	1.33***	0.25
	Age	0.05	0.04
	Age^2	0.00**	0.00
	Intercept	-19.60***	1.19
Random-Effect Variance			
	σ^2_{u1}	19.02***	1.24
	σ^2_{u2}	30.52***	1.86
	σ^2_{u3}	82.64***	6.50
	AIC	38,680.8	
	BIC	39,435.40	
Sample Size			
No. of participants		7,211	
No. of observations		58,891	

*+p<0.1, * p<0.05, ** p<0.01, *** p<0.001*

Genetic score is standardised; Model includes first 20PCs as additional covariates.

Insights into the drivers of the correlation between polygenic scores and inactivity due to family care can be taken from the economics literature on *discouraged workers*. In the Background section, I detail how it is harder for people with depression to find a job due to social and professional stigmas. Prolonged lack of success can result in taking turns to cover family care responsibilities. However, the mechanisms should be empirically tested; here, I refer only to possible explanations from existing theoretical knowledge.

Correlation between polygenic score and economic inactivity is observed for both females and males, as Table 4.7.2 and Figure 4.7.2 show. For women, 1 s.d. increase in polygenic score is associated with 1.65 higher probability of becoming economically inactive compared to those who are employed (*β=.50; p<.001*) while holding covariates in the model constant. For men, 1 s.d. in depression genetic score raises the likelihood of dropping out of the labour force by 2.59 times (*β=.95; p<.001*) compared to those who have a job. A striking feature of these sex-specific trends is the considerably stronger marginal association between polygenic score and economic inactivity among males, which mainly takes a linear form. For females, this same association is not linear because of the presence of a significant and positive squared-term polygenic score (*β=.15; p<.05*). Accordingly, the non-linearity observed for economic inactivity within the whole genetic sample is mainly driven by women.

Detailed analysis of the reasons for economic inactivity further explains sex-specific trends. Firstly, the greater association of polygenic score with economic inactivity among men is mainly driven by the sickness and disability reason for worklessness. Inactivity due to family care does not have significant results, which can be at least partly attributed to the observation that the vast majority of men are not inactive for this reason. For women, this is different: polygenic score is a significant predictor for both causes for absence from the labour force. Moreover, there is marginal significance for the quadratic term of polygenic score (*p=.10*). Thus, while the presence of rGE appears for both sexes, rGE patterns take different forms for men and women. This suggests that the mechanisms driving rGEs are likely unique to each group, which is a subject for future investigation.

Table 4.7.4: Coefficients and standard errors of multinomial multilevel model assessing the association between depression polygenic score and worklessness with detailed profile of economic inactivity, by sex

		Men		Women	
	Parameters	β	Std. Err.	β	Std. Err.
$\frac{P(Y_i=unemployed)}{P(Y_i=employed)}$	PGS depression	0.36^{***}	0.09	0.46^{***}	0.09
	PGS depression^2	0.04	0.06	0.04	0.06
	Age	-0.26^{***}	0.02	-0.25^{***}	0.04
	Age^2	0.00^{***}	0.00	0.00^{***}	0.00
	Intercept	-0.65	0.52	-0.11	0.52
$\frac{P(Y_i=family\ care)}{P(Y_i=employed)}$	PGS depression	0.23	0.13	0.31^{**}	0.10
	PGS depression^2	0.04	0.09	0.11+	0.07
	Age	-0.29^{***}	0.05	-0.25^{***}	0.02
	Age^2	0.01^{***}	0.00	0.00^{***}	0.00
	Intercept	-18.15^{***}	4.45	0.07	0.47
$\frac{P(Y_i=sick\ /\ disable)}{P(Y_i=employed)}$	PGS depression	1.93^{***}	0.33	1.04^{***}	0.17
	PGS depression^2	-0.19	0.22	0.15+	0.12
	Age	-0.30^{***}	0.12	0.12^{*}	0.05
	Age^2	0.00^{***}	0.00	0.00	0.00
	Intercept	-16.76^{***}	2.94	-18.59^{***}	1.45
Random-Effect Variance					
	σ^2_{u1}	24.15^{***}	2.34	17.03^{***}	1.34
	σ^2_{u2}	32.04^{***}	3.52	27.38^{***}	1.86
	σ^2_{u3}	70.15^{***}	9.56	64.51^{***}	6.19
	AIC	10,569.21		26,974.61	
	BIC	11,679.31		27,687.78	
Sample Size					
No. of participants		3,141		4,070	
No. of observations		26,031		32,860	

*+p<0.1, * p<0.05, ** p<0.01, *** p<0.001*

Genetic score is standardised; Model includes first 20PCs as additional covariates.

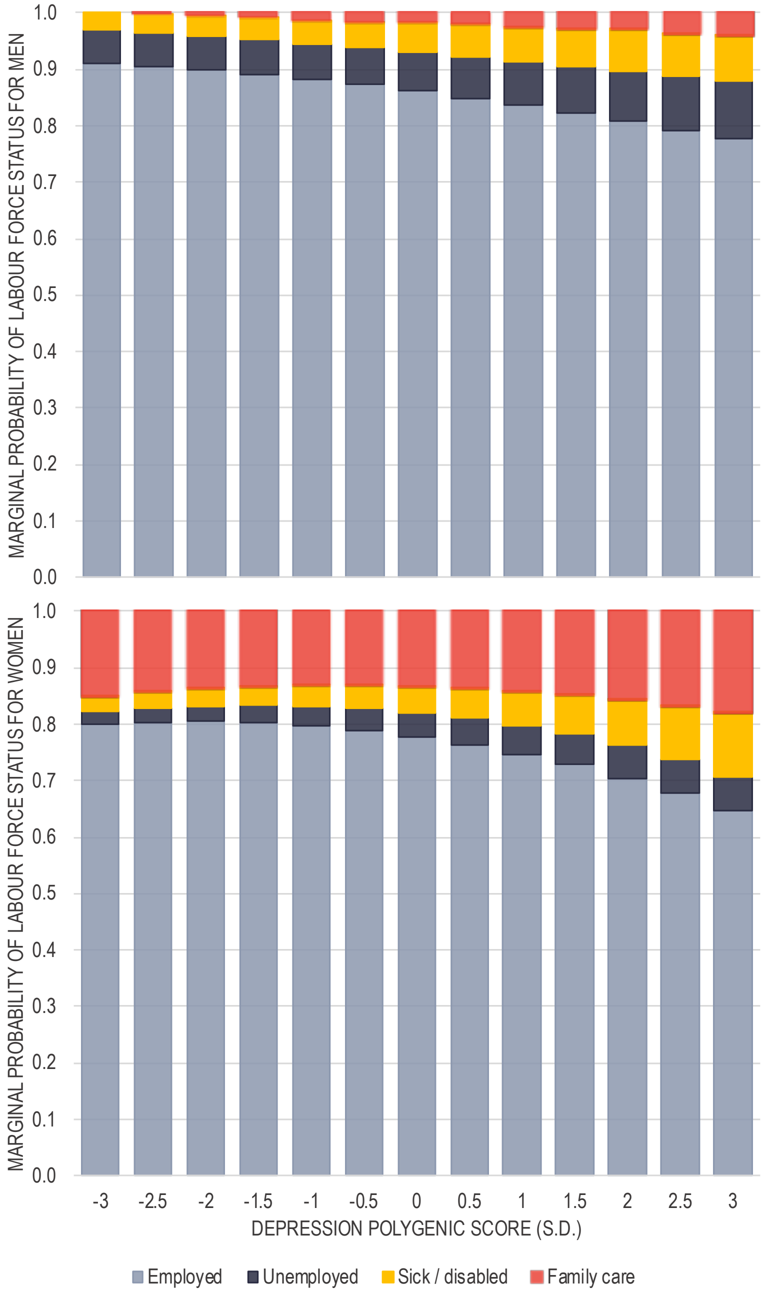


Figure 4.7.4: Marginal probabilities of worklessness with detailed profile of economic inactivity in relation to polygenic score of depression, by sex

4.7.3 Sensitivity analysis

For the sensitivity analysis, I examine gene-environment correlations with the inclusion of weights correcting for differential probabilities in selection for genotyping. This is because genotyping in the Understanding Society survey was not a random process. Various selection mechanisms were involved starting from the eligibility of participants and recruitment criteria and ending with the personal decision to be genotyped (as discussed in greater detail in Chapter 2). For example, those who were genotyped are less likely to live in rural areas and more likely to have higher educational qualifications compared to those who were not genotyped. Moreover, genotyped participants are less likely to have bad general health. Since these factors also contribute to economic activity, gene-environment correlation analysis can be potentially biased. As blood weights provided in the UKHLS study correct for the greatest portion of health and socio-demographic selection in the genetic sample, these weights should be sufficient for tackling the sample selection issue. Accordingly, I provide the results of weighted multinomial multilevel regression models in the Annex: Table C.0.2 and Figure C.0.1 present the results for the whole sample, while Table C.0.3 and Figure C.0.2 cover sex-specific analysis. Overall, no differences are revealed. This means that the rGE trends described here are not sensitive to the sampling procedure and can be considered representative of the UK adult population. These findings indicate the presence of rGE where the genetic risks of depression are associated with a higher probability for both types of worklessness.

4.8 Results and discussion

Several studies have shown that genetic predispositions to depression increase the probability of experiencing high-risk social environments, which include bullying [Veldkamp et al., 2019], the absence of emotional parental support [Wilkinson et al., 2013], and adverse life events [Lau and Eley, 2008]. The presence of such gene-environment correlations suggests social environments could function as a causal pathway from genes to depression. This contributes to our understanding of the complex nature of depression and potentially to more efficient remedies. The current study thus has a novel focus on worklessness as another aspect of the social environment that creates a toll through stressful life conditions, putting people at higher risk for experiencing depression. Using the Understanding Society genetic sample, I show that the polygenic risks of depression are associated with a higher probability for worklessness that includes being unemployed and economically inactive.

By conceptualising different reasons of worklessness, my aim was to investigate incidents of unemployment, economic inactivity due to family care, and economic inactivity due to sickness and disability. Firstly, I find evidence that higher values of polygenic score of depression are associated with higher probability to be unemployed, and such rGE patterns are observed for both females and males. Secondly, I detected a significant positive association between depression polygenic score and the probability to be economically inactive; moreover, my analysis revealed that this relationship also has a non-linear nature which indicates that if a person has greater numbers of risk alleles, his/her chances to be economically inactive are disproportionally higher. This non-linearity is mainly driven by women. Another important observation is that once the reasons for economic inactivity are taken into account, polygenic prediction becomes the strongest for the absence in the labour force that is due to sickness or disability compared to the absence due to family care. Importantly, while the presence of rGE appears for both sexes, rGE patterns are of different strengths between men and women. It suggests that it is likely that the forms and the mechanisms driving rGEs are likely to be unique for men and women which is an important highlight. More broadly, my results suggest that genetic predispositions to depression increase the probability to experience high-risk socio-economic environments.

The current study also contributes to debates about selection into worklessness [Bartley, 1994], complicating our understanding of the relationship between the absence of a job and depression. Consequently, the correlation between an underlying vulnerability factor (such as genetic risk) and worklessness is further evidence that job loss is not exogenous when it comes to depression susceptibilities. It also implies that the relationship between mental health and unemployment is a complex one. Incidences of worklessness increase the risks of developing depression even as job loss *per se* is a product of both direct and indirect selection prior to unemployment, in terms of depression susceptibilities and experiences. Here, gene-environment correlations play an important role as an additional aspect of selection into unemployment and its harmful consequences for health.

There are a number of important considerations and frontiers for prospective studies. To begin, I have introduced the conceptual framework of rGE as a methodologically and empirically novel and important element in the research on worklessness and mental health. The current study is the first step, primarily aiming to explore the descriptive aspects of the research interest. Although genes are randomly assigned at birth, specification of the link between depression polygenic score and worklessness cannot be characterised as causal and instead refers to a specific type of rGE. Future research is thus needed for a deeper understanding of existing trends; one frontier might be to test the specific types of rGE that drive observed correlations. Another frontier for further research is the longitudinal aspect of rGE, which is implicated in the cross-sectional differences shown here. More research is needed to reveal possible mechanisms driving the association between depression polygenic score and worklessness. In terms of socio-demographic determinants for worklessness, such mechanisms could be educational attainment and neighbourhood characteristics; for such a mediation analysis, researchers will need to take possible endogeneity into account (an issue that I cover in the next chapter of my thesis). Lastly, the presence of rGE amplifies the need to re-consider the harmful consequences of worklessness in regards to genetic confounding issue. However, researchers should be careful in employing statistical models on mediation and genetic confounding with additional covariates such as genetic predisposition. This is because the presence of rGE would bias the estimates – a notion that is fully explained in the following chapter of my thesis.

Chapter 5

Heritable environments: bias due to conditioning on a collider in models with polygenic scores

Author's Note: In the previous chapters, my focus was restricted to depression as the main phenotype of my interest and trends in the UK. The following chapter expands the lens of my focus and addresses methodological concerns that are relevant not only to depression but also to all complex human traits and behaviours. This chapter [1] was co-authored with Professor Richard Breen [2], Professor Melinda Mills [3], and Dr. David Brazel [4].

[1] A revised version of this chapter is published as Evelina T. Akimova, Richard Breen, David M. Brazel, Melinda C. Mills. (2021). Gene-environment dependencies lead to collider bias in models with polygenic scores. *Scientific Reports*, 11: 9457. doi: https://doi.org/10.1038/s41598-021-89020-x

[2] Nuffield Professor of Sociology, Department of Sociology, University of Oxford

[3] Nuffield Professor of Sociology, Department of Sociology and Leverhulme Centre for Demographic Science, University of Oxford

[4] Senior Research Officer, Department of Sociology and Leverhulme Centre for Demographic Science, University of Oxford

5.1 Abstract

The application of polygenic scores has transformed our ability to investigate whether and how genetic and environmental factors jointly contribute to the variation of complex traits. Modelling the complex interplay between genes and environment, however, raises serious methodological challenges. Here we illustrate the largely unrecognised impact of gene-environment dependencies on the identification of the effects of genes and their variation across environments. We show that heritable covariates in regression models that include polygenic scores as independent variables introduce endogenous selection bias when one of the covariates depends on unmeasured factors that also affect the outcome. This results in the problem of conditioning on a collider, which in turn leads to spurious associations and effect sizes. Using graphical and simulation methods we demonstrate that the degree of bias depends on the strength of the gene-covariate correlation and of hidden heterogeneity linking environments with outcomes, regardless of whether the main analytic focus is mediation, confounding, or gene × environment interactions. We offer potential solutions, highlighting the importance of causal inference. We also urge further caution when fitting and interpreting models with polygenic scores and non-exogenous environments or phenotypes and demonstrate how spurious associations are likely to arise, advancing our understanding of such results.

5.2 Introduction

The importance of understanding the joint contributions of genetic and environmental variation underlying complex traits is widely recognised. The rise of polygenic scores has resulted in a surge of studies investigating the mediating and moderating roles of environments along with genetic confounding (i.e., whether genes confound associations between environments or phenotypes) [Barbaro et al., 2017]. Yet, disentangling the relative importance of polygenic scores and environmental covariates is difficult; various methodological concerns have been raised, including but not limited to the power and predictive accuracy of polygenic scores [Morris et al., 2020, Mostafavi et al., 2020, Ware et al., 2017] and the non-exogenous nature of environmental exposures and their consequences

[Conley, 2009, Dudbridge and Fletcher, 2014, Fletcher and Conley, 2013]. Moreover, genes and environments do not operate independently, necessitating greater scrutiny of both conventional models and new methods addressing gene-environment-trait correlations [Avinun, 2019, Ni et al., 2019]. Here we address further methodological problems arising from gene-environment correlations that have gone largely unrecognised yet make identification of causal effects and the accurate estimation of associations even more challenging.

We illustrate how the presence of genetic predispositions associated with exposures to an environment (or phenotype in cases of genetic confounding) introduces *endogenous selection bias* in a regression analysis. Heritable covariates in regression models with polygenic scores are endogenous variables, and this can give rise to the problem of conditioning on a collider. Collider bias is an important statistical problem that destabilises regression models and it can arise for a variety of reasons, including sample selection and attrition [Munafò et al., 2017]. We demonstrate that collider bias is likely to occur not only in genetic association studies but also in other analyses where polygenic scores are included, regardless of whether the main focus is mediation, confounding, or gene × environment (G×E) interaction.

Moreover, the issue we describe here is linked to a growing body of literature showing the heritability of environments known as gene-environment correlation (rGE). To date, discussion of the methodological implications of these findings has focussed on the implications for G×E interaction studies (e.g. Dudbridge and Fletcher [2014]). However, if both genetic and environmental covariates are included in a statistical model, gene-environment correlations may lead to spurious estimates and effect sizes. Understanding the mechanisms that generate these dependences is crucial for how we interpret such results. We emphasise the conceptual differences between passive, evocative, and active gene-environment correlations [Plomin et al., 1977] and potential sources of endogeneity of environmental covariates.

We use a graphical approach to demonstrate these methodological problems, illustrated by simulations. We then discuss the consequences of the bias in linear models and offer some potential solutions. The problems outlined here are relevant for making both causal and non-causal claims, with serious implications for the interpretation of results.

5.3 Endogenous selection bias

The notion of *endogenous selection bias* arises from the broader concept of *selection bias*. While the term *selection bias* is very widely used [Infante-Rivard and Cusson, 2018], endogenous selection bias commonly arises in analyses in which we adjust for an endogenous variable – that is, a variable caused by other, unmeasured variables which also affect the outcome. In this case, bias arises through the adjusting variable operating as a collider. Collider bias was demonstrated in Day et al. [2016] in the context of genetic association studies where such bias led to false-positive and biologically implausible associations between GWAS significant SNPs for height and sex once the respondent's height was adjusted for. The bias arose because the respondent's height is a collider variable since it is a direct product of another covariate (SNPs of height) and an outcome (sex).

In what follows we consider a situation in which genes and environment are correlated (for reasons discussed below) and the environmental variable(s) is affected by variables not measured in the study and which also affect the outcome. We discuss the consequences of the resulting collider bias for both additive and G×E interaction models.

5.3.1 Additive models with polygenic scores

The first type of model we consider is the rather straightforward design where polygenic scores and environmental covariates (or phenotypes if they are used as covariates) are jointly included as a set of predictors for an outcome of interest. Such models are intended to reveal whether genetic influences confound associations between environments and outcomes or whether environments are mediators of the link between genetic variants and phenotypes. Examples include linking health disparities with socio-economic outcomes such as the relationship between attention-deficit hyperactivity disorder (ADHD), its polygenic prediction, and IQ on educational outcomes among teenagers (e.g. Stergiakouli et al. [2016]). Another example is studies on labour market outcomes predicted by educational measures (e.g., grades, years of education) along with an educational attainment polygenic score (e.g. Ayorech et al. [2019], Papageorge and Thom [2019]). The variation of exam scores in relation to school types and polygenic prediction of education is another instance (e.g. Smith-Woolley et al. [2018]).

All of these studies are similar regarding the nature of environmental variables – they are, firstly, not exogenous and, secondly, are direct or indirect products of polygenic scores which are also included into the models. Dependence of these covariates could arise through the inclusion of a phenotypic variable – the scenario that is prevalent in genetic confounding studies. For example, the polygenic score for ADHD is directly linked to the diagnosis of ADHD and is a phenotype in the associated GWAS studies. Dependences could also arise due to active and evocative selection in environments. Applying the polygenic prediction of educational attainment as an example, we see that it contributes to the variation of school grades [Smith-Woolley et al., 2018] which likely reflects active rGE (i.e., children selecting their environments for genetically influenced reasons). It could also be linked to school type since parents adjust their educational choices for children depending on their child's characteristics which are partially due to genetic differences, reflecting evocative rGE (i.e., child indirectly shapes the environment via the reaction of parent's to child's behaviour, traits) [Avinun and Knafo, 2014, Klahr and Burt, 2014]. Therefore, non-exogenous environments of this type vary depending on the values of polygenic scores. Such dependencies accompanied by the presence of hidden heterogeneity linking environments with outcomes will result in endogenous selection bias, which we describe now.

Whether the aim is to address genetic confounding or to reveal mediation, models of this kind include polygenic scores and environments as predictors for an outcome of interest. The simple case of such a model is illustrated in Figure 5.3.1 Panel A. The polygenic score, G, has an independent association with the outcome, Y, along with an indirect path through the environment, E. The exclusion from the model of the environmental covariate, E, results in the estimation of the total effect of G on Y, while the exclusion of G and the inclusion of E produces the association between E and Y, confounded by unobserved G.

The challenges for the model in Figure 5.3.1 Panel A are to produce reliable estimates of the direct effects from G to Y and from E to Y in the face of confounding by the unmeasured U. Since the focus of our paper is not related to the issues of polygenic prediction *per se*, we do not include a discussion on the sources of bias between G and Y caused by confounders that are likely to arise due to assortative mating, population stratification, which have been amply explored elsewhere Kerminen et al. [2019]. Here, we focus on the role of confounders of the link between environment and outcome since this is directly linked to our argument.

The first and most important problem of the presence of unmeasured factors causing E and Y is non-exogenous environments. Socio-economic conditions, parental characteristics, health policies, and cultural norms, along with neighbourhoods and other factors, could all be confounders in specific cases. Unless included in the model, these factors cause E and Y to be correlated, as they are jointly present in the error structure of both variables. The issue is further problematic because the confounding can be driven by both observed and unobserved factors. Hence, even an extensive set of controls would not necessarily yield unbiased estimates if substantial confounding on observables remained unaddressed.

Moreover, unobserved confounder(s), U, linking E and Y biases not only their association, but also the estimate from G to Y. This is driven by the fact that E is now a collider since it is a product of both G and U, as indicated by the arrows from U to E and G in Figure 5.3.1 Panel A. It is known that if we do not control for a collider variable, the path between its sources is blocked; however, once a collider is included in the set of covariates, the associated path is now open [Elwert and Winship, 2014]. Accordingly, conditioning on E introduces a new path from G to E to Y through U: this is the green path denoted in Figure 5.3.1 Panel A. This path is the source of the collider bias in these

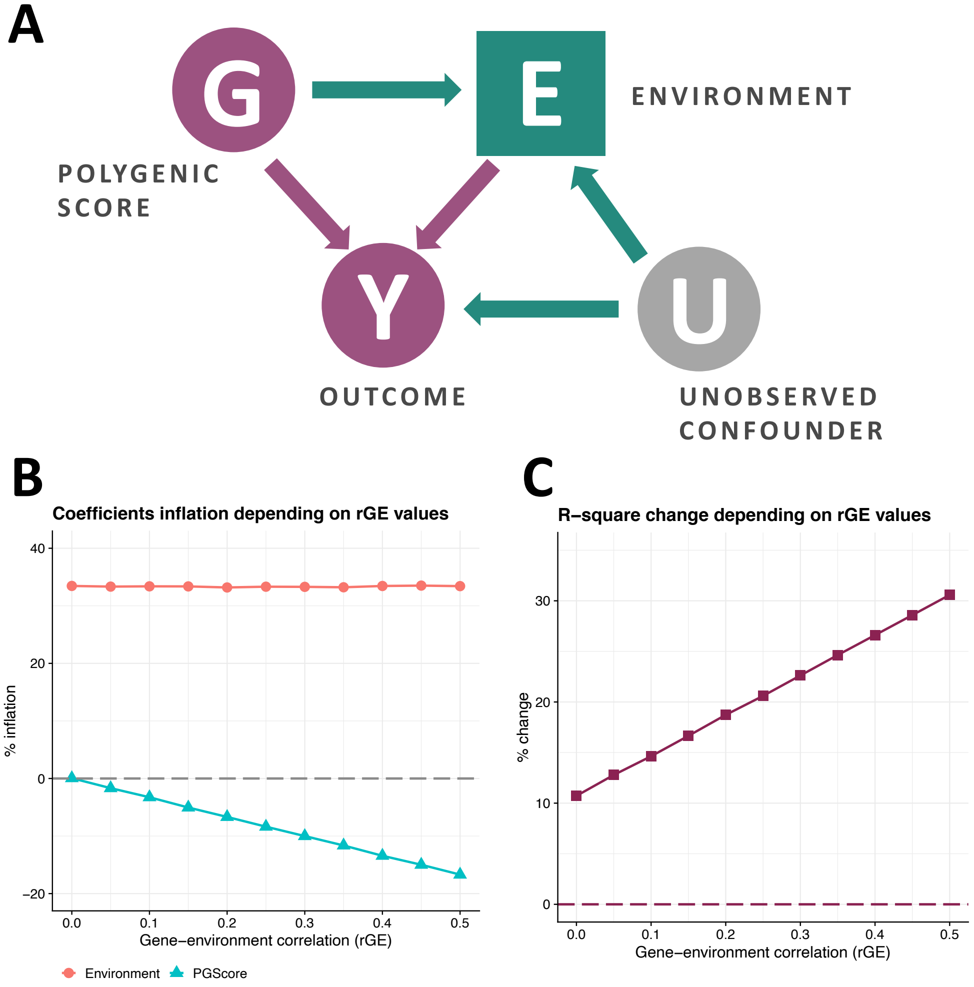


Figure 5.3.1: Collider bias in polygenic gene-environment models.
Note: Panel A. Schematic diagram of the collider bias which occurs between polygenic score, environment, and outcome in cases of gene-environment interdependence. Dark purple circles represent variables, unobserved confounders are shown in grey circles, collider variables are indicated by squares. By adding E into the model with the polygenic score G, we make E a collider. A collider that is not conditioned on, blocks the path between its sources (G and U); once a collider is controlled for, the path is opened as indicated by green nodes. Panel B. Spurious regression estimates for polygenic score and environment from the series of OLS simulations reflecting the range of gene-environment interdependence. Collider bias due to positive values of gene-environment correlation and the presence of uncontrolled confounder, which is positively correlated with covariate and outcome, results in deflation of polygenic score estimates. Estimates of the environmental effect are upwardly biased but are not affected by the gene-environment correlation. Panel C. R-squared inflation plot from the series of OLS simulations; collider bias results in inflated values of explained variance statistics. R-squared statistics for the model with endogenous covariate and polygenic score includes not only the true share of the variance in Y explained by G and E (baseline estimate indicated by 0), but also the elements of variance that are due to gene-environment correlation and confounder(s), U.

models.

To further illustrate this bias, we conducted a series of simulations of the simple linear model from Figure 5.3.1 Panel A. We considered the presence of direct effects from G to Y and E to Y, allowed the G-E correlation to vary from 0 indicating no heritability to 0.5 indicating a highly heritable covariate E, and included an uncontrolled confounder, U, which is positively correlated with both E and Y at a fixed value. Figure 5.3.1 Panel B illustrates the deviations of coefficients from the true simulated values. Notably, the presence of G-E correlation and an omitted confounder, U, where both U-E and U-Y associations are positive, results in the deflation of polygenic score estimates and inflation of environmental coefficients. Deflation of the G-Y association is greater with higher values of G-E correlation, while the models without this association produced results free of collider bias. The path coefficient from E to Y is biased regardless of the strength of the G-E correlation reflecting the importance of the omitted confounder, U, as a source of bias of this path.

It is also possible to exemplify the source of inflation of the G-Y association by considering examples from the existing literature. For instance, Papageorge and Thom [2019] regress the educational attainment polygenic score and years of schooling on standardised job tasks. If we take models on nonroutine analytic and interactive tasks (where the association between polygenic score and outcome is positive), we see that the inclusion of educational controls results in about 70% smaller polygenic score coefficients. It is likely that such a change is largely attributable to the extended set of educational controls, which includes both parental and respondent educational attainment. However, if we consider a moderate strength of association between the genetic score and respondent's years of schooling along with additional assumptions about unobserved confounders linking educational attainment and the type of job tasks, we can show that around 15-20% of the polygenic score coefficient decrease is plausibly due to collider bias, following the derivations we include in the section below.

We also show in Figure 5.3.1 Panel C that the described bias results in greater values of explained variance statistics, which are R-squared estimates in the case of our simulations. This is because statistical models suffering from this bias explain both true and artificial (due to collider) variation in a dependent variable. It further complicates the assessment of the relative predictive power of polygenic scores and environments. As demonstrated in Figure 5.3.1 Panel C, R-squared statistics for the model with an endogenous covariate and a polygenic score would include not only the true share of the variance in Y explained by G and E, but also the elements of variance that are due to rGE and confounder(s), U.

To conclude, the inclusion of associated polygenic scores and covariates in regression models results in spurious estimates and greater explained variance statistics. The direction and strength of coefficient bias depends on the strength of gene-covariate correlation and on the underlying structure of endogeneity which links the covariate to the outcome variable.

5.3.2 Gene × environment interaction models

A growing literature estimates the moderating patterns of environmental risks in the associations between polygenic scores and phenotypes. Here, in the same fashion as in additive models, environmental exposures of interest are not usually exogenous. For example, recent studies on gene × environment interaction analysis consider such environments as relationship status [Barr et al., 2019], educational attainment [Amin et al., 2017], occupational exposure [Zeng et al., 2019], neighbourhood characteristics [Robinette et al., 2019] and others. There is ongoing discussion on the implications of non-exogeneity of environments [Dudbridge and Fletcher, 2014, Keller, 2014, Schmitz and Conley, 2017]. Also, the issue of collider bias has been demonstrated in the context of case-only gene × environment interaction studies [Balazard et al., 2017]. We expand on these concerns by showing that moderation models also suffer from collider bias.

Firstly, the problem outlined until this point is also relevant for gene × environment interaction studies. One difference, however, between additive and moderation models is the presence of G×E interaction in a set of covariates. As indicated in Figure 5.3.2 Panel A by green nodes, the bias path

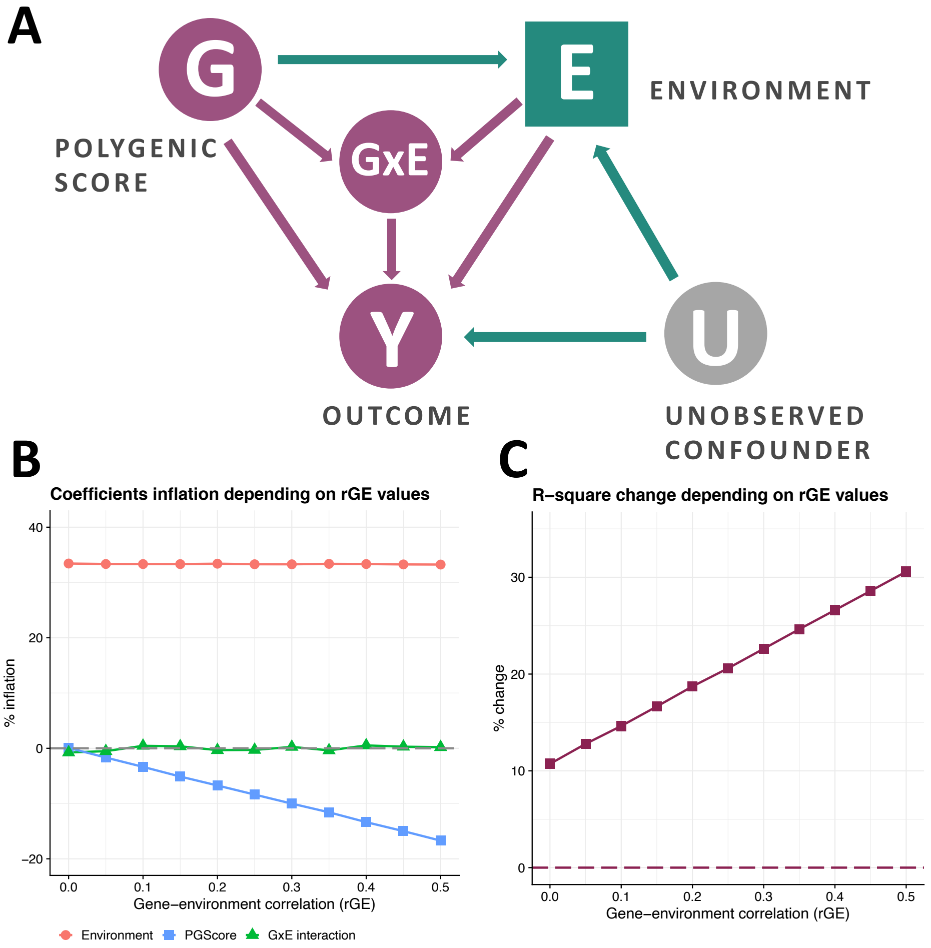


Figure 5.3.2: Collider bias in polygenic gene × environment interaction models.
Note: Panel A. Schematic diagram of the collider bias which occurs between polygenic score, environment, and outcome in cases of gene-environment interdependence. Dark purple circles represent variables, unobserved confounders are shown in grey circles, collider variables are indicated in squares. By adding E into the model with the polygenic score G, we make E a collider. A collider that is not conditioned on, blocks the path between its sources (G and U); once a collider is controlled for, the path is opened as indicated by green nodes. Panel B. Spurious regression estimates for the polygenic score and environment along with non-inflated interaction terms from the series of OLS simulations reflecting a range of gene-environment interdependence. Collider bias due to positive values of gene-environment correlation and the presence of an uncontrolled confounder, which is positively correlated with covariate and outcome, results in deflation of polygenic score estimates. Deflation is greater the higher the gene-environment correlation; greater confounding also results in greater bias. The interaction term is not affected but results for moderation analysis are biased as long as direct effects are spurious. Panel C. R-squared inflation plot from the series of OLS simulations; collider bias results in inflated values of explained variance statistics. R-squared statistics for the model with endogenous covariate and polygenic score includes not only the true share of the variance in Y explained by G and E (baseline estimate indicated by 0), but also the elements of variance that are due to gene-environment correlation and confounder(s), U.

from G to Y through E and U would still lead to spurious results. Considering the examples of G×E studies mentioned earlier, the environments may, to some degree, be products of self-selection, which leads to a greater likelihood of G and E interdependence along with the presence of unobserved confounder(s), U. Secondly, since the overall G×E interaction pattern depends on the direct estimates of G on Y and E on Y, results for moderation analyses are biased when direct effects are spurious. However, the G×E coefficient *per se* is not inflated due to collider bias. This can be seen in Figure 5.3.2 Panel B, along with the inflation of R-square statistics in Figure 5.3.2 Panel C which were obtained from similar simulations as earlier but with the inclusion of interaction terms. Consequently, endogeneity of environmental covariates biases both additive and moderation models.

5.3.3 Mathematical expression of bias

Previous sections provided a general description of endogenous selection bias when polygenic scores and environments are not independent. Here we further illustrate this issue by deriving exact expressions for the bias under the assumption of linear relationships that can be modelled using regression analysis.

We assume that the data have been generated by the DAG shown in Figure 5.3.1 Panel A. Here U is an unobserved variable or set of variables that confounds the E – Y relationship (this is equivalent to, but, in our view more transparent than, a depiction that would include correlated error terms for E and Y). We further assume that all the variables have unit standard deviation and that G is exogenous.

Additive model

When the effects of G and E on Y are additive the true linear models given the data-generating process are:

$$E(E|G,U) = \alpha_0 + \alpha_1 G + \alpha_2 U \tag{5.1}$$

$$E(Y|G, E, U) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 U \tag{5.2}$$

We assume that the parameters α_1, β_1, and β_2 are all positive.

If estimate the models:

$$E(E|G) = a_0 + a_1 G \tag{5.3}$$

$$E(Y|G, E) = b_0 + b_1 G + b_2 E \tag{5.4}$$

the relationships between the true and estimated parameters are:

$$b_1 = \beta_1 - \frac{\alpha_1 \alpha_2 \beta_3}{1 - \alpha_1^2} \tag{5.5}$$

$$b_2 = \beta_2 + \frac{\alpha_2 \beta_3}{1 - \alpha_1^2} \tag{5.6}$$

The proof is as follows. Let β_{YG} denote the coefficient of the unconditional regression of Y on G and likewise for β_{YE}. Then, tracing the paths linking G and Y in the Figure 5.3.1 Panel A we have:

$$\beta_{YG} = \beta_1 + \beta_2 \alpha_1 \tag{5.7}$$

and

$$\beta_{YE} = \beta_2 + \beta_1 \alpha_1 + \beta_3 \alpha_2 \tag{5.8}$$

Given that $\beta_{EG} = \alpha_1$ we then apply the standard formula to derive conditional regression coefficients from unconditional:

$$b_1 = \frac{\beta_{YG} - \beta_{YE}\alpha_1}{1 - \alpha_1^2} = \frac{\beta_1 + \beta_2 \alpha_1 - (\beta_2 + \beta_1 \alpha_1 + \beta_3 \alpha_2)\alpha_1}{1 - \alpha_1^2} \tag{5.9}$$

Straightforward algebra yields (Equation 5.5). b_2 is derived similarly. Notice, however, that $a_1 = \alpha_1$ because G and U are unconditionally independent.

The bias in both estimates depends on the sign of $\alpha_2\beta_3$: if this is positive the estimate of the partial effect of G on Y, given E, will be downwardly biased and the estimate of the effect of E on Y, given G, will be upwardly biased. If there is no correlation between G and E ($\alpha_1 = 0$) then b_1 will be unbiased. If there is no effect of an unmeasured confounder (either $\alpha_2 = 0$ and/or $\beta_3 = 0$) both b_1 and b_2 will be unbiased. The bias in the effect of G on Y has a different sign than the bias in the effect of E on Y: if the bias in the latter is positive, the size of the genetic effect will be underestimated relative to the environmental effect.

The example of coefficient deflation from Papageorge and Thom [2019][Table 6: 41] can be demonstrated following Equation 5.5. For instance, considering the case on nonroutine interactive job tasks as the dependent variable, we see that the baseline coefficient of the educational attainment polygenic score is 0.185, which reflects a model without any environmental and phenotypic covariates. In the model with educational controls (respondent's years of schooling and parental education), the polygenic score coefficient drops to 0.055 reflecting a 70% negative change. Since the dependent variable is standardised, we can assess the relative importance of collider bias which is $\frac{\alpha_1\alpha_2\beta_3}{1-\alpha_1^2}$ from Equation 5.5 under additional assumptions. If we allow the coefficient of the correlation between educational attainment polygenic score and respondents years of schooling $\alpha_1 = 0.300$, and the presence of unobserved confounder U, positively correlated with both years of schooling and job task (for example, living in advantaged neighbourhood as a child), we have $\alpha_2 = 0.250$ and $\beta_3 = 0.250$. These are all plausible and rather modest suggestions following correlation matrix from Papageorge and Thom [2019][Table 6: 41], leading the inflation bias to be:

$$\frac{\alpha_1\alpha_2\beta_3}{1-\alpha_1^2} = \frac{0.300 \times 0.250 \times 0.250}{1-0.300^2} = 0.021 \tag{5.10}$$

which explains 16% downward change of polygenic score coefficient.

G×E interaction model

The DAG in Figure 5.3.2 Panel A shows the case in which the effect of G on Y varies with E. In this case, the true linear models given the data-generating process are:

$$E(E|G,U) = \alpha_0 + \alpha_1 G + \alpha_2 U \tag{5.11}$$

$$E(Y|G,E,U) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 U + \beta_4 (GE) \tag{5.12}$$

We estimate:

$$E(E|G) = a_0 + a_1 G \tag{5.13}$$

$$E(Y|G,E) = b_0 + b_1 G + b_2 E + b_4 GE \tag{5.14}$$

In this case, b_4 is an unbiased estimate of β_4 because the backdoor path from G-E to Y is blocked by E. The bias in b_1 and b_2 will be the same as above. In the case in which E is a binary variable, coded 0 and 1, b_4 will be an unbiased estimate of the difference in the effect of G at $E = 1$ and $E = 0$, but the estimate of the baseline effect of G on Y when $E = 0$ will be biased.

Bias in R^2

The R^2 for models in Equations 5.4 and 5.10 will be biased. In the additive case, for example, the true R^2 attributable to G and E is:

$$\frac{\beta_1^2 var(G) + \beta_2^2 var(E) + 2\beta_1\beta_2 cov(G,E)}{var(Y)} = \beta_1^2 + \beta_2^2 + 2\beta_1\beta_2\alpha_1 \tag{5.15}$$

(using the assumption that all the variables have unit standard deviation). But the reported R^2

from model in Equation 5.4 is:

$$b_1^2 + b_2^2 + 2b_1 b_2 a_1 \tag{5.16}$$

Substituting Equations 5.5 and 5.6 into Equation 5.12 we calculate the inflation of R^2 due to confounding and collider bias. This is:

$$R^2 bias = \alpha_2\beta_3[\frac{\alpha_2\beta_3}{1-\alpha_1^2} + 2\beta_2] \tag{5.17}$$

Confounder bias arises from $\alpha_2\beta_3$. The derivative of Equation 5.13 with respect to this is positive provided that $1 - \alpha_1^2 > 0$. The derivative of Equation 5.13 with respect to α_1 (which captures the association between G and E) is:

$$\frac{2\alpha_1\alpha_2^2\beta_3^2}{(1-\alpha_1^2)^2} \tag{5.18}$$

The sign of this depends on the sign of the numerator. When it is positive both confounding and collider bias will inflate the reported R^2. As an example, consider a case in which $\beta_1 = 0.465$, $\beta_2 = 0.505$, $\beta_3 = 0.231$, $\alpha_1 = 0.209$, $\alpha_2 = 0.693$. Then the observed $R^2 = 0.758$, whereas the true share of the variance in Y explained by G and E is 0.569. The inflation bias here is:

$$0.693 \times 0.231[\frac{0.693 \times 0.231}{1 - 0.209^2} + 2 \times 0.505] = 0.188 \tag{5.19}$$

If the correlation between G and E had been larger and/or if the confounding of E had been greater, the reported R^2 would have been larger because of the greater bias.

5.4 Solutions

To resolve this issue, it is important to understand the nature of the correlation between genes and environment (or phenotype if it is used as a covariate) – whether a correlation is conditional and observed because of omitted confounders between G and E and/or a correlation reflects causal interdependency. The former would necessitate controlling for the confounding factor: this could be parental characteristics (passive rGE) [Kong et al., 2018, Trejo and Domingue, 2019], ancestry [Mostafavi et al., 2020], and so forth. If a correlation arises as a result of active or evocative selection, the assumption of non-causal association would be violated and require another set of solutions to avoid collider bias. The latter is relevant when a phenotypic variable is used as a covariate along with its polygenic prediction, since the association would be at least partially causal.

To obtain unbiased estimates, we need to apply causal inference methods that seek to exploit the exogenous variation in an environmental covariate. A comprehensive discussion of methods available for researchers and applicable to the context of this paper is provided by Fletcher and Conley [2013]. Briefly, techniques such as regression discontinuity and difference-in-difference designs, instrumental variables and quasi-natural experiments will produce unbiased results for both additive and gene × environment interaction models, conditional on certain assumptions being met.

There are also existing ways to assess the magnitude of the bias for the general collider cases [Greenland, 2003, VanderWeele, 2010]. Since the type of collider we described here is the product of both observed and unobserved factors, calculation of bias magnitude would rely on additional assumptions about the structure of the error correlation between the environmental covariate and the outcome. The presence of unobserved variables, however, makes it impossible to provide a definitive estimate of the bias. However, following the mathematical expression of the bias we show the sources of bias in coefficient estimates and error in R-squared. We demonstrate that the strength of collider bias is positively associated with the strength of rGE and unobserved confounders. The use of sensitivity analyses is a valuable tool in showing how robust conclusions are to different degrees of unobserved confounding and thus of collider bias [Ding and VanderWeele, 2016].

5.5 Conclusion

We have discussed methodological considerations arising due to heritable environments (or phenotypes that are included into models as covariates) that have not yet been previously recognised. We demonstrated that the inclusion of environments that are products of polygenic scores may introduce endogenous selection bias through conditioning on a collider, leading to spurious associations. Particularly, we illustrated that the degree of bias depends on the strength of gene-covariate correlation and of the omitted variable(s) linking the covariate and outcome. We also showed that the portion of explained variance is overestimated proportionally. We proposed some solutions that exploit the strength of causal inference methods: these are likely to be important not only for obtaining reliable results but also in the interpretation of existing studies.

Chapter 6

Conclusion

6.1 Main contributions

Depression is one of the most common mental health disorders and one of the top ten causes for sickness worldwide [Vos et al., 2015]. Various scholars, policy makers, and the media have all raised the importance of obtaining a deeper understanding of depression and its development. This has resulted in a growing body of research into different aspects of depression. My aim was to further contribute to our understanding of depression. Driven by both biological and socio-economic factors, depression is a demonstrably complex phenomenon in our society. Nowadays, mental health researchers often take an interdisciplinary focus to study. Moreover, recent advances in molecular genetics have introduced a unique opportunity for use of sociogenomic tools to deepen our knowledge of the multidimensional nature of the biological and social risks of developing depression.

Increased availability of data has further contributed to the integration of genetics with social science. The possibility of the inclusion of genetic risks as an additional variable into conventional empirical models is promising, but it also gives rise to substantial methodological considerations. This thesis partly aims to address ongoing methodological problems with the study of depression: the issue of selection in genetic samples and biases associated with causal interconnections between genes and complex environments. Accordingly, Chapter 2 contributes to the discussion on sample selection in social science genomics by demonstrating that analysis of representativeness and possible selection should be included as an additional step for quality control procedures. Such a focus should be further popularised and stressed in work with genetic samples. This is because it is important to achieve diversity not only in terms of ancestry, but also in accounting for the socio-demographic and health profiles of population-based genetic samples.

Continuing the discussion on methodological insights, Chapter 5 makes an important contribution to the modelling of joint genetic and environmental variations linked to complex traits such as depression. The rise of polygenic scores has resulted in a surge of studies investigating the mediating and moderating roles of environments along with genetic confounding. Yet disentangling the relative importance of polygenic scores and environmental covariates is difficult. I unpacked methodological considerations arising from heritable environments (or phenotypes included in models as covariates) that have not been previously recognised. I demonstrated how the inclusion of environments that are products of polygenic scores can introduce endogenous selection bias through conditioning on a collider, leading to spurious associations and biased variance statistics. In particular, I illustrated how the degree of bias depends on the strength of gene-covariate correlation and of the omitted variable(s) linking covariate with outcome. These notions are important not only for obtaining reliable results, but also in the interpretation of existing studies.

Alongside these methodological contributions, the thesis makes several empirical contributions to the existing body of knowledge on depression. I base my analysis on the context of the United Kingdom; my motivation is driven by the fact that mental health is high on the UK political agenda.

I investigated the role of unemployment and cohort trends to show changing genetic penetrance among adults in the UK across the twentieth century. Chapter 3 examined how birth cohorts and recessions moderate genetic influence on depressive symptoms among adults in the UK. The focus on gene-by-cohort interactions can potentially shed light on how historical contexts shape polygenic prediction across different generations. For social science, the particular insight is whether a rise in the prevalence of depression at certain historical points across the twentieth century is driven by those who have higher polygenic risks of depression or independent of genetic risks. Moreover, there is a gap in the literature in terms of studies covering the UK context. Consequently, this paper also contributes to existing knowledge by providing gene × cohort interaction analysis of depression in the UK. For example, in other contexts, such as US, Conley et al. [2016] did not find significant variation in genetic associations with depression across birth cohorts. Another important contribution is the demonstration that economic downturns contribute to gene-by-cohort variations.

Several studies have shown that genetic predispositions to depression increase the probability of experiencing high-risk social environments. The presence of gene-environment correlations gives rise to the notion that social environments are potentially likewise present on the causal path from genes to depression, contributing to our understanding of the complex nature of depression. The contribution of Chapter 4 is its focus on worklessness as another aspect of the social environment that creates a toll of stressful life conditions, putting people at higher risk for experiencing depression. I showed that polygenic risks for depression are associated with a higher probability of worklessness, which includes unemployment and economic inactivity. An additional contribution of Chapter 4 is directly linked to debates about selection into worklessness that complicate our understanding of the relationship between the absence of a job and depression. The presence of correlation between an underlying vulnerability factor (such as genetic risk) and worklessness is further evidence that job loss is not exogenous to depression susceptibilities. It further implies that the relationship between mental health and unemployment is a complex one wherein the incidence of job loss increases the risks of developing depression. At the same time, job loss *per se* is a product of direct and indirect selection of prior-to-job-loss depression susceptibilities.

6.2 Summary of main findings

This thesis has aimed to answer five main research questions. The first question addressed who was genotyped in the UKHLS survey, which methodologically translates into determining whether the UKHLS genetic sample suffers from selection. To address this question, I assessed differences between genotyped and non-genotyped respondents through a regression analyses that estimated differential probabilities in being genotyped attributed to socio-demographic and health characteristics. Consistent with findings from other genetic samples, the results suggest that genotyped participants are a selective population (at least to a certain extent). I revealed small-to-moderate differences between genotyped and non-genotyped participants: on average, the former tend to have higher educational attainment, live in urban areas, and have better general self-reported health. Findings also suggest that these differences are likely to be germane to a gene-by-cohort interaction studies because genotyped and non-genotyped participants have distinctive mortality trends resulting in mortality selection in the UKHLS genetic sample. All in all, these patterns indicate that analyses based on these genetic samples should take into account the notion of selection and use tools for selection correction (i.e. weights).

The second research question was whether selection into genotyping biases the polygenic prediction of depression. Here, my findings suggest the implementation of weights to correct for genotyping selection does not significantly change the polygenic prediction to depression. These results mirror those of the previous studies that have examined the performance of polygenic prediction in other genetic samples.

The third question this thesis aimed to answer was whether the polygenic prediction of depression varies by birth cohorts in the UK during the twentieth century. While this question has been explored in the context of different countries, a major goal of this thesis was to investigate the UK context as it has not been done before. Additionally, I aimed to take into account important historical contexts (e.g., economic recessions) as a potential source of variation in polygenic prediction across birth cohorts. My results indicate that an increase in depressive symptoms is especially profound for the cohort of Baby Boomers who also display a significantly higher genetic penetrance of depressive symptoms. I also found a suggestive moderation for the Generation X cohort, but this aspect of the findings should be further replicated in future releases of the data to allow a wider age range for this cohort. More importantly, it appears periods of economic recession have the potential to shape the polygenic prediction of depression across generations in a different manner: my analysis reveals that economic recessions weaken the polygenic prediction of depressive symptoms across some cohorts, and strengthen the polygenic penetrance in others. Such findings highlight the importance of a cohort-specific approach in understanding phenotypic and genetic variations along with their interplay. It is important to take into account not only cohort-specific historical exposures (currently the prevailing approach to study), but also wider contexts experienced by everyone as the patterns of their consequences can be also cohort-specific.

The fourth research question my thesis explored was whether genetic predispositions to depression are associated with higher chances of experiencing worklessness. To answer this question, I conducted a multinomial regression analysis of working age participants in the UKHLS genetic sample. By conceptualising different reasons for worklessness, my aim was to investigate incidents of unemployment, economic inactivity due to family care, and economic inactivity due to sickness and disability. The findings indicate that higher values in the polygenic score for depression are associated with a higher probability of unemployment. Moreover, such rGE patterns are observed for both females and males. The analysis further reveals the non-linear nature of this relationship: as the number of risk alleles increases, the individual likelihood of becoming economically inactive becomes disproportionally higher. Importantly, this non-linearity is driven mainly by women. Another important observation is that once the reasons for economic inactivity are taken into account, polygenic prediction becomes strongest for absence from the labour force due to sickness or disability compared to absence due to family care. While the presence of rGE appears for both sexes, rGE patterns are of different strengths between men and women. This suggests that the forms and mechanisms driving rGEs are likely unique for men and women, which is an important highlight. More broadly, my results suggest that a genetic predisposition to depression increases the probability of experiencing high-risk socio-economic environments.

The fifth research question my thesis sought to address was whether and how the presence of a correlation between polygenic scores and covariates in regression models biases the results of gene × environment interaction, mediation, and genetic confounding analyses. In broader terms, this chapter aimed to uncover methodological challenges arising on our way to model genetic and non-genetic factors that jointly contribute to the development of complex behaviours and traits, including depression. To explore this question, I used graphical and simulation methods. I showed that heritable covariates in regression models with polygenic scores as independent variables actually introduce endogenous selection bias. This results in the problem of conditioning on a collider, which in turn leads to spurious associations and effect sizes. My results also demonstrate that the degree of bias depends on the strength of gene-covariate correlation and of hidden heterogeneity linking environments with outcomes, regardless of whether the main analytical focus is mediation, confounding, or gene × environment interactions. I offer potential solutions for obtaining unbiased estimates where I highlight the importance of causal inference methods and techniques. This issue has gone largely unrecognised until now, which makes the focus of the chapter particularly important for the interdisciplinary field of sociogenomics. Taken together, I call attention to and urge further caution in fitting and interpreting models with polygenic scores and non-exogenous environments or covariates. I demonstrate how spurious associations and effect sizes are likely to arise, advancing our understanding of existing results and calling for caution in prospective studies.

6.3 Limitations

In regards to my analyses and research methods, some limitations should be acknowledged. Along with chapter-specific limitations covered in the discussion sections of my empirical investigations, I would like to draw attention to a broader set of concerns related to my analysis in general. The first issue touches on the complexity of measuring depression. The notion of measurement challenges arises from psychology and psychiatry, where some practitioners and researchers have raised the problem of a diagnostic crisis [Paris, 2008]. There is an ongoing debate about what harmonised and meaningful procedures must be developed to grasp the non-diagnosed dimensions of depression that affect how we understand mental health problems, as well as the true scale of these problems in society (a more detailed discussion of this issue can be found in Kinderman [2014]).

In light of that, it is not surprising when researchers raise concerns about the validity and reliability of mental health measurements in surveys including depression. Gove and Geerken [1977] found three forms of response bias for depression that included naysaying, perceived trait desirability, and a need for social approval. The paper also argues there is random noise. Along with these findings, more recent evidence shows self-reported mental health indicators are subject to recall bias [Patten, 2003]. Using the widely documented and analysed GHQ depressive symptoms index, the threat of measurement challenges might not be critical – especially after taking into account that symptomatic data for depression is more accurate is with greater validity and reliability in the context of population-based surveys when compared to diagnostic data [Mandemakers, 2011]. However, acknowledgement of possible silencing is necessary.

Since there is no genetic sample in the UKHLS presenting different ancestries, I did not analyse non-European ancestry samples. This is another limitation on the study, as my results do not cover diverse ancestral populations (one of the most critical issues in the field of sociogenomics; for a more detailed discussion, see Mills and Rahal [2019]). As demonstrated in Figure 6.3.1, the vast majority of participants in genetic discoveries are of European ancestry. Future research should extend the

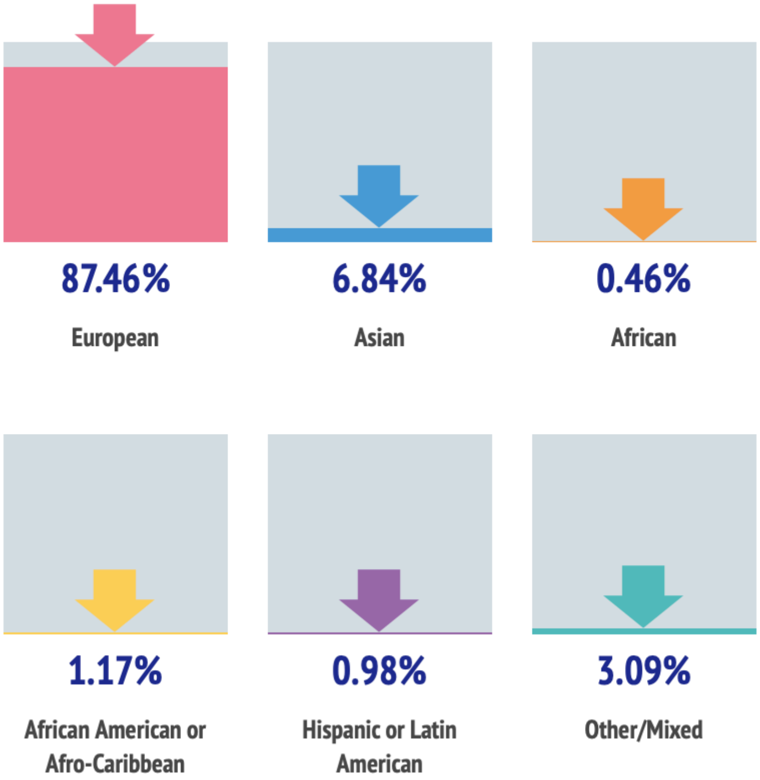


Figure 6.3.1: Total GWAS participants diversity over 2000-2020 period downloaded from the diversity monitor developed by Mills and Rahal [2019].
Source: `https://gwasdiversitymonitor.com`
Date: 01-Dec-2020

analysis to other populations once data become available. Since most GWAS studies were carried out on those of European ancestry [Mills and Rahal, 2019], the polygenic prediction provided in this thesis cannot be generalised to non-European ancestry groups living in the UK. While prediction in genetically diverse samples is not an appropriate procedure [Martin et al., 2017], polygenic scores should be appropriately calculated for ancestry groups independently. Thus, this project is limited in its focus on only one ancestry group. Hopefully, with future data releases, it will be possible to perform the analysis on other sub-populations living in the UK.

A third important limitation that I would like to stress is the issue of genetic principal components (PCs). PCs aim to control for population stratification, since it is one of the strongest sources of confounding for genetic associations. Interaction estimates can also be confounded by population stratification [Price et al., 2006]. To rule out the possibility of this confounding, twenty first PCs were included as additional covariates for all of the statistical models of my thesis involving provision of PCs in the release version of the data (and thus calculated by the UKHLS team). But it would be ambiguous to claim to have fully controlled for the possibility of population stratification. The most important limitation is that the current PCs – not only those calculated in the UKHL data, but also in other genetic samples – are calculated using common genetic variants. There have been recent inquiries, such as Zaidi and Mathieson [2020], showing the population stratifications that arose recently should be accounted for by calculating PCs with rare genetic variants, as well. Such an approach requires sequenced, not imputed, genetic data. This would make it possible to complement the current set of PCs.

6.4 Frontiers for future research

In terms of directions for future research, I point towards the following areas. To begin, important areas for further research arise from consideration of the limitations addressed in the previous section. For one, analysis should be expanded to a wider range of non-European ancestry groups. It should also take into account the notion of rare variants in the calculation of both principal components and polygenic scores, which constitutes a frontier for future empirical enquiries.

Another important avenue would be to unpack the notion of selection in genetic samples by revealing its fundamental causes. Methodologically, it is important to understand whether respondents self-select themselves into genetic survey data or selection is happening due to eligibility criteria and availability. While correction can be later applied with weighting tools (as in my thesis), the issue of self-selection has more considerable implications for bias in estimates and effect sizes. In these contexts, weighting alone is insufficient for correction. Self-selection can be an important aspect of genetic samples, since the decision to be genotyped can be deeply rooted in the personal beliefs of an individual in terms of family history, religiosity, and all other factors. We thus should understand whether the *healthy volunteer effect* and socio-economic selection has a self-selection aspect to improve the reliability of our results. This, too, is an area for future research that should aim to deepen our knowledge on the data-generating process of genetic samples.

Variation in the polygenic penetrance of depression across birth cohorts can be explored by taking genetic susceptibility to the effectiveness of antidepressants into account. As Wigmore et al. [2020] shows, there is a genetic component that responds to the medical treatment of depression. It is thus important to expand my analysis towards the changes in the genetic penetrance of the efficacy of antidepressants across birth cohorts. This would provide a bridge to more informed interventions and a deeper understanding of observed trends. Another important area would be an empirical investigation into how and why birth cohort variations arise or, in other words, finding possible mechanisms that would explain the descriptive trends I observed. This could be done by considering qualitative differences between birth cohorts and their experiences in the broader spectrum of historical and political changes.

Further research questions should consider why and how rGE associations in worklessness and depression PGS have emerged. This implies testing what specific type of rGE is the main source for the observed correlations. Also, it is important to implement a life-course prospective for the development of depression and the importance of rGE on such development patterns. Moreover, it is important to consider further differences in labour force experiences. This could be accomplished by, for example, distinguishing types of work and job security for those employed. In the context of the UK, this can be particularly important following the good vs. bad jobs discussions. It would also be interesting to address different types of unemployment experiences, e.g., the recipients of unemployment benefits or voluntary vs. involuntary joblessness. These can be explored in further research on the topic. Additionally, rGE trends can be analysed in cross-national comparisons to countries with different welfare policies. This would further explore the links between macro- and individual-level factors.

A final area for potential research would be to re-assess existing results with endogenous environments in order to understand the role of this bias in our current empirical insights into gene-environment interplay. It would also be important to consider calculating the effect sizes of polygenic score predictions wherein environments play the role of a covariate. Lastly, the application of causal inference methods is further advised for future inquiries in order to enable empirical investigation that does not suffer from endogeneity within statistical analyses.

References

Akyeampong, E. (1989). Discouraged workers. *Perspectives on Labour and Income*, 2:64–69.

Albanesi, S. and Şahin, A. (2018). The gender unemployment gap. *Review of Economic Dynamics*, 30:47–67.

Amare, A. T., Vaez, A., Hsu, Y.-H., Direk, N., Kamali, Z., Howard, D. M., McIntosh, A. M., Tiemeier, H., Bültmann, U., Snieder, H., and Hartman, C. A. (2020). Bivariate genome-wide association analyses of the broad depression phenotype combined with major depressive disorder, bipolar disorder or schizophrenia reveal eight novel genetic loci for depression. *Molecular Psychiatry*, 25(7):1420–1429. doi: https://doi.org/10.1038/s41380-018-0336-6.

Amin, V., Böckerman, P., Viinikainen, J., Smart, M. C., Bao, Y., Kumari, M., ..., and Pehkonen, J. (2017). Gene-environment interactions between education and body mass: Evidence from the UK and Finland. *Social Science and Medicine*, 195:12 – 16. doi: https://doi.org/10.1016/j.socscimed.2017.10.027.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5:1564 EP –. doi: https://doi.org/10.1038/nprot.2010.116.

Andreasen, N. C. (1985). *The broken brain: The biological revolution in psychiatry*, volume 1272. Harper Collins.

Antecol, H. (2000). An examination of cross-country differences in the gender gap in labor force participation rates. *Labour Economics*, 7(4):409–426. doi: https://doi.org/10.1016/S0927-5371(00)00007-5.

Appleby, L., Hunt, I. M., and Kapur, N. (2017). New policy and evidence on suicide prevention. *The Lancet Psychiatry*, 4(9):658–660.

Arnau-Soler, A., Macdonald-Dunlop, E., Adams, M. J., Clarke, T.-K., MacIntyre, D. J., Milburn, K., Navrady, L., Hayward, C., McIntosh, A. M., and Thomson, P. A. (2019). Genome-wide by environment interaction studies of depressive symptoms and psychosocial stress in uk biobank and generation scotland. *Translational psychiatry*, 9(1):1–13.

Avinun, R. (2019). The e is in the g: Gene–environment–trait correlations and findings from genome-wide association studies. *Perspectives on Psychological Science*, 15(1):81–89. doi: https://doi.org/10.1177/1745691619867107.

Avinun, R. and Knafo, A. (2014). Parenting as a reaction evoked by children's genotype: a meta-analysis of children-as-twins studies. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology*, 18(1):87–102. doi: https://doi.org/10.1177/1088868313498308.

Ayorech, Z., Plomin, R., and von Stumm, S. (2019). Using dna to predict educational trajectories in early adulthood. *Developmental Psychology*, 55(5):1088–1095. doi: https://doi.org/10.1037/dev0000682.

Balazard, F., Le Fur, S., Bougnères, P., and Valleron, A.-J. (2017). Interactions and collider bias in case-only gene-environment data. *bioRxiv*, page 124560. doi: https://doi.org/10.1101/124560.

Bansal, R., Peterson, B. S., Gingrich, J., Hao, X., Odgerel, Z., Warner, V., ..., and Weissman, M. M. (2016). Serotonin Signaling Modulates the Effects of Familial Risk for Depression on Cortical Thickness. *Psychiatry Research - Neuroimaging*, 248:83–93.

Barbaro, N., Boutwell, B. B., Barnes, J. C., and Shackelford, T. K. (2017). Genetic confounding of the relationship between father absence and age at menarche. *Evolution and Human Behavior*, 38(3):357–365. doi: https://doi.org/10.1016/j.evolhumbehav.2016.11.007.

Barr, P. B., Kuo, S. I.-C., Aliev, F., Latvala, A., Viken, R., Rose, R. J., Kaprio, J., Salvatore, J. E., and Dick, D. M. (2019). Polygenic risk for alcohol misuse is moderated by romantic partnerships. *Addiction*, 114(10):1753–1762. doi: https://doi.org/10.1111/add.14712.

Bartley, M. (1994). Unemployment and ill health: understanding the relationship. *Journal of Epidemiology & Community Health*, 48(4):333–337. doi: https://jech.bmj.com/content/48/4/333.

Baselmans, B. M. L., van de Weijer, M. P., Abdellaoui, A., Vink, J. M., Hottenga, J. J., Willemsen, G., ..., and Bartels, M. (2019). A genetic investigation of the well-being spectrum. *Behavior Genetics*, 49(3):286–297. doi: https://doi.org/10.1007/s10519-019-09951-0.

Bell, A. (2014). Life-course and cohort trajectories of mental health in the uk, 1991–2008 – a multilevel age–period–cohort analysis. *Social Science and Medicine*, 120:21 – 30. doi: https://doi.org/10.1001/archgenpsychiatry.2010.151.

Belsky, D. W., Domingue, B. W., Wedow, R., Arseneault, L., Boardman, J. D., Caspi, A., ..., and Harris, K. M. (2018). Genetic analysis of social-class mobility in five longitudinal studies. *Proceedings of the National Academy of Sciences*, 115(31):E7275–E7284. doi: https://doi.org/10.1073/pnas.1801238115.

Ben-Shlomo, Y., Smith, G. D., Shipley, M., and Marmot, M. G. (1993). Magnitude and causes of mortality differences between married and unmarried men. *Journal of Epidemiology & Community Health*, 47(3):200–205.

Benjamin, D. J., Cesarini, D., Chabris, C. F., Glaeser, E. L., Laibson, D. I., Guðnason, V., ..., and Lichtenstein, P. (2012). The promises and pitfalls of genoeconomics. *Annual Review of Economics*, 4(1):627–662. doi: https://doi.org/10.1146/annurev-economics-080511-110939.

Benzeval, M., Davillas, A., Kumari, M., and Lynn, P. (2014). Understanding society: the uk household longitudinal study biomarker user guide and glossary. *Institute for Social and Economic Research, University of Essex*. http://doc.ukdataservice.ac.uk/doc/7251/mrdoc/pdf/7251_understandingsociety-biomarker-userguide-2014.pdf. Online; accessed 15-June-2019.

Boardman, J. D., Blalock, C. L., and Pampel, F. C. (2010). Trends in the genetic influences on smoking. *Journal of health and social behavior*, 51(1):108–123.

Boardman, J. D., Blalock, C. L., Pampel, F. C., Hatemi, P. K., Heath, A. C., and Eaves, L. J. (2011). Population composition, public policy, and the genetics of smoking. *Demography*, 48(4):1517–1533.

Boardman, J. D., Daw, J., and Freese, J. (2013). Defining the Environment in Gene-Environment Research: Lessons from Social Epidemiology. *American journal of public health*, 103 Suppl(Suppl 1):S64–S72.

Boardman, J. D., Domingue, B. W., Blalock, C. L., Haberstick, B. C., Harris, K. M., and McQueen, M. B. (2014). Is the gene-environment interaction paradigm relevant to genome-wide studies? the case of education and body mass index. *Demography*, 51(1):119–139. doi: https://doi.org/10.1007/s13524-013-0259-4.

Bosker, F. J., Hartman, C. A., Nolte, I. M., Prins, B. P., Terpstra, P., Posthuma, D., van Veen, T., Willemsen, G., DeRijk, R. H., de Geus, E. J., Hoogendijk, W. J., Sullivan, P. F., Penninx, B. W., Boomsma, D. I., Snieder, H., and Nolen, W. A. (2011). Poor replication of candidate genes for major depressive disorder using genome-wide association data. *Molecular Psychiatry*, 16(5):516–32.

Brouillard, C., Brendgen, M., Vitaro, F., Dionne, G., and Boivin, M. (2019). Predictive links between genetic vulnerability to depression and trajectories of warmth and conflict in the mother–adolescent and father–adolescent relationships. *Developmental psychology*.

Brown, G. W. and Harris, T. (2012). *Social origins of depression: A study of psychiatric disorder in women*, volume 2. Routledge.

Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics*, 84(2):210–223. doi: https://doi.org/10.1016/j.ajhg.2009.01.005.

Brydsten, A., Hammarström, A., Strandh, M., and Johansson, K. (2015). Youth unemployment and functional somatic symptoms in adulthood: results from the Northern Swedish cohort. *European Journal of Public Health*, 25(5):796–800. doi: https://doi.org/10.1093/eurpub/ckv038.

Buck, N. and McFall, S. (2011a). Understanding society: design overview. *Longitudinal and Life Course Studies*, 3(1). doi: https://doi.org/10.14301/llcs.v3i1.159.

Buck, N. and McFall, S. (2011b). Understanding Society: design overview. *Longitudinal and Life Course Studies*, 3(1):5–17. doi: http://doi.org/10.14301/llcs.v3i1.159.

Butterworth, P., Leach, L. S., Pirkis, J., and Kelaher, M. (2012). Poor mental health influences risk and duration of unemployment: a prospective study. *Social psychiatry and psychiatric epidemiology*, 47(6):1013–21. doi: https://doi.org/10.1007/s00127-011-0409-1.

Carey, N. (2012). *The epigenetics revolution: How modern biology is rewriting our understanding of genetics, disease, and inheritance*. Columbia University Press.

Chabris, C. F., Lee, J. J., Cesarini, D., Benjamin, D. J., and Laibson, D. I. (2015). The fourth law of behavior genetics. *Current Directions in Psychological Science*, 24(4):304–312. doi: https://doi.org/10.1177/0963721415580430.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1). doi: https://doi.org/10.1186/s13742-015-0047-8.

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17(7):392–406. doi: https://doi.org/10.1038/nrg.2016.27.

Chial, H. (2008). Huntington's disease: The discovery of the huntingtin gene. *Nature Education*, 1(1):71.

Choi, S. W., Heng Mak, T. S., and O'Reilly, P. F. (2018). A guide to performing polygenic risk score analyses. *bioRxiv*. doi: https://doi.org/10.1101/416545.

Choi, S. W. S. and O'Reilly, P. (2019). Prsice-2: Polygenic risk score software for biobank-scale data. *GigaScience*, 8. doi: https://doi.org/10.1093/gigascience/giz082.

Christine Heim, Newport, D. J., Heit, S., Graham, Y. P., Wilcox, M., Bonsall, R., Miller, A. H., and Nemeroff, C. B. (2000). Pituitary-Adrenal and Autonomic Responses to Stress in Women After Sexual. *The Journal of American Medical Association*, 284(5):592–597.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., ..., and Todd, J. A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, 37(11):1243–1246. doi: https://doi.org/10.1038/ng1653.

Clegg, R. (2016). A guide to labour market statistics: Explanation of the major concepts that exist within the labour market and their relationship to each other. *Office for National Statistics*.

Collishaw, S., Maughan, B., Natarajan, L., and Pickles, A. (2010). Trends in adolescent emotional problems in england: a comparison of two national cohorts twenty years apart. *Journal of Child Psychology and Psychiatry*, 51(8):885–894.

Conley, D. (2009). The promise and challenges of incorporating genetic data into longitudinal social science surveys and research. *Biodemography and Social Biology*, 55(2):238–251. doi: https://doi.org/10.1080/19485560903415807.

Conley, D. and Fletcher, J. (2018). *The Genome Factor: What the social genomics revolution reveals about ourselves, our history, and the future*. Princeton University Press.

Conley, D., Laidley, T. M., Boardman, J. D., and Domingue, B. W. (2016). Changing polygenic penetrance on phenotypes in the 20 th century among adults in the us population. *Scientific reports*, 6:30348.

Courtiol, A., Tropf, F. C., and Mills, M. C. (2016). When Genes and Environment Disagree: Making Sense of Trends in Recent Human Evolution. *Proceedings of the National Academy of Sciences*, 113(28):7693 LP – 7695.

Cox, C., Moore, P., and Ladle, R. (2016). *Biogeography: An Ecological and Evolutionary Approach*. Wiley. ISBN: 9781118968604.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Daly, M. and Delaney, L. (2013). The scarring effect of unemployment throughout adulthood on psychological distress at age 50: estimates controlling for early adulthood distress and childhood psychological factors. *Social Science and Medicine*, 80:19–23. doi: https://doi.org/10.1016/j.socscimed.2012.12.008.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., ..., and Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, 48:1284 EP –. doi: https://doi.org/10.1038/ng.3656.

Daw, J., Shanahan, M., Harris, K. M., Smolen, A., Haberstick, B., and Boardman, J. D. (2013). Genetic Sensitivity to Peer Behaviors: 5HTTLPR, Smoking, and Alcohol Consumption. *Journal of health and social behavior*, 54(1):92–108.

Day, F., Loh, P.-R., Scott, R., Ong, K., and Perry, J. B. (2016). A robust example of collider bias in a genetic association study. *The American Journal of Human Genetics*, 98(2):392–393. doi: https://doi.org/10.1016/j.ajhg.2015.12.019.

de Castro-Catala, M., Papiol, S., Barrantes-Vidal, N., and Rosa, A. (2020). *Interaction Between Genes and Childhood Trauma on the Outcome of Psychiatric Disorders*, pages 105–124. Springer International Publishing, Cham. doi: https://doi.org/10.1007/978-3-030-49414-8_6.

De Jong, G. F. and Madamba, A. B. (2001). A double disadvantage? minority group, immigrant status, and underemployment in the united states. *Social Science Quarterly*, 82(1):117–130. doi: https://doi.org/10.1111/0038-4941.00011.

de Mello, M. F., de Jesus Mari, J., Bacaltchuk, J., Verdeli, H., and Neugebauer, R. (2005). A systematic review of research findings on the efficacy of interpersonal therapy for depressive disorders. *European archives of psychiatry and clinical neuroscience*, 255(2):75–82.

De Vogli, R. (2014). The financial crisis, health and health inequities in europe: the need for regulations, redistribution and social protection. *International journal for equity in health*, 13(1):58.

Delgado-Rodriguez, M. and Llorca, J. (2004). Bias. *Journal of Epidemiology & Community Health*, 58(8):635–641.

Dennis, H. and Migliaccio, J. (1997). Redefining retirement: The baby boomer challenge. *Generations: Journal of the American Society on Aging*, 21(2):45–50.

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., and Muntner, P. (2006). Mortality prediction with a single general self-rated health question: A meta-analysis. *Journal of general internal medicine*, 21(3):267–275.

Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368–377. doi: https://doi.org/10.1097/EDE.0000000000000457.

Domingue, B., Liu, H., Okbay, A., and Belsky, D. (2017a). Genetic heterogeneity in depressive symptoms following the death of a spouse: Polygenic score analysis of the u.s. health and retirement study. *The American journal of psychiatry*, 174:appiajp201716111209. doi: https://doi.org/10.1176/appi.ajp.2017.16111209.

Domingue, B. W., Belsky, D. W., Harrati, A., Conley, D., Weir, D. R., and Boardman, J. D. (2017b). Mortality selection in a genetic sample and implications for association studies. *International Journal of Epidemiology*, 46(4):1285–1294.

Domingue, B. W., Conley, D., Fletcher, J., and Boardman, J. D. (2016). Cohort effects in the genetic influence on smoking. *Behavior genetics*, 46(1):31–42.

Dooley, D., Catalano, R., and Wilson, G. (1994). Depression and unemployment: panel findings from the epidemiologic catchment area study. *American journal of community psychology*, 22(6):745–765.

Dooley, D., Prause, J., and Ham-Rowbottom, K. A. (2000). Underemployment and depression: longitudinal relationships. *Journal of Health and Social Behavior*, pages 421–436.

Drydakis, N. (2015). The effect of unemployment on self-reported health and mental health in greece from 2008 to 2013: A longitudinal study before and during the financial crisis. *Social Science and Medicine*, 128:43 – 51. doi: https://doi.org/10.1016/j.socscimed.2014.12.025.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLOS Genetics*, 9(3):1–17. doi: https://doi.org/10.1371/journal.pgen.1003348.

Dudbridge, F. and Fletcher, O. (2014). Gene-environment dependence creates spurious gene-environment interaction. *The American Journal of Human Genetics*, 95(3):301–307. doi: https://doi.org/10.1016/j.ajhg.2014.07.014.

Durkheim, E. (1897). *Le suicide: étude de sociologie*. Alcan.

Egan, M., Daly, M., and Delaney, L. (2015). Childhood psychological distress and youth unemployment: Evidence from two british cohort studies. *Social Science and Medicine*, 124:11–17. doi: https://doi.org/10.1016/j.socscimed.2014.11.023.

Elbaz, A. and Alpérovitch, A. (2002). Bias in association studies resulting from gene-environment interactions and competing risks. *American Journal of Epidemiology*, 155(3):265–272. doi: https://doi.org/10.1093/esr/jcx080.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology*, 40:31–53. doi: https://doi.org/10.1146/annurev-soc-071913-043455.

Ensminger, M. E. and Celentano, D. D. (1990). Gender differences in the effect of unemployment on psychological distress. *Social Science and Medicine*, 30(4):469–477. doi: https://doi.org/10.1016/0277-9536(90)90349-W.

Epping, E. A. and Paulsen, J. S. (2011). Depression in the early stages of huntington disease. *Neurodegenerative Disease Management*, 1(5):407–414. doi: https://doi.org/10.2217/nmt.11.45.

Ezzy, D. (1993). Unemployment and mental health: A critical review. *Social Science and Medicine*, 37(1):41 – 52. doi: https://doi.org/10.1016/0277-9536(93)90316-V.

Feldman, G. (2007). Cognitive and behavioral therapies for depression: overview, new directions, and practical recommendations for dissemination. *Psychiatric Clinics of North America*, 30(1):39–50.

Fink, E., Patalay, P., Sharpe, H., Holley, S., Deighton, J., and Wolpert, M. (2015). Mental health difficulties in early adolescence: a comparison of two cross-sectional studies in england from 2009 to 2014. *Journal of Adolescent Health*, 56(5):502–507.

Fletcher, J. M. and Conley, D. (2013). The challenge of causal inference in gene-environment interaction research: leveraging research designs from the social sciences. *American journal of public health*, 103 Suppl 1(Suppl 1):S42–S45. doi: https://doi.org/10.2105/AJPH.2013.301290.

Flint, J. and Kendler, K. S. (2014). The genetics of major depression. *Neuron*, 81(3):484–503. doi: https://doi.org/10.1016/j.neuron.2014.01.027.

Franks, P., Gold, M. R., and Fiscella, K. (2003). Sociodemographics, self-rated health, and mortality in the us. *Social Science and Medicine*, 56(12):2505–2514.

Frasquilho, D., Matos, M. G., Salonna, F., Guerreiro, D., Storti, C. C., Gaspar, T., and Caldas-de Almeida, J. M. (2015). Mental health outcomes in times of economic recession: a systematic literature review. *BMC public health*, 16(1):115.

Freese, J. (2018). The arrival of social science genomics. *Contemporary Sociology*, 47(5):524–536.

Frese, M. and Mohr, G. (1987). Prolonged unemployment and depression in older workers: A longitudinal study of intervening variables. *Social Science and Medicine*, 25(2):173–178. doi: https://doi.org/10.1016/0277-9536(87)90385-6.

Fry, A., Littlejohns, T., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N. (2017). Comparison of sociodemographic and health-related characteristics of uk biobank participants with the general population. *American Journal of Epidemiology*, 186. doi: https://doi.org/10.1093/aje/kwx246.

Fryers, T., Melzer, D., and Jenkins, R. (2003). Social inequalities and the common mental disorders. *Social Psychiatry and Psychiatric Epidemiology*, 38(5):229–237. doi: https://doi.org/10.1007/s00127-003-0627-2.

Garcia-Toro, M. and Aguirre, I. (2007). Biopsychosocial model in depression revisited. *Medical hypotheses*, 68(3):683–691. doi: https://doi.org/10.1016/j.mehy.2006.02.049.

George, E. and Engel, L. (1980). The clinical application of the biopsychosocial model. *American journal of Psychiatry*, 137(5):535–544.

Gillespie, R. (1929). The clinical differentiation of types of depression. *Guy's Hospital Reports*, 79:306–344.

Gitterman, A. (1991). *Handbook of social work practice with vulnerable populations*. Columbia University Press.

Goldberg, D., McDowell, I., and Newell, C. (1972). General health questionnaire (ghq), 12 item version, 20 item version, 30 item version, 60 item version [ghq12, ghq20, ghq30, ghq60]. *Measuring health: A guide to rating scales and questionnaire*, pages 225–36.

Gonda, X., Petschner, P., Eszlari, N., Baksa, D., Edes, A., Antal, P., Juhasz, G., and Bagdy, G. (2019). Genetic variants in major depressive disorder: From pathophysiology to therapy. *Pharmacology & therapeutics*, 194:22–43. doi: https://doi.org/10.1016/j.pharmthera.2018.09.002.

Gove, W. R. and Geerken, M. R. (1977). Response bias in surveys of mental health: An empirical investigation. *American journal of Sociology*, 82(6):1289–1317.

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306. doi: https://doi.org/10.1097/01.Ede.0000042804.12056.6c.

Greenland, S. (2008). Multiple comparisons and association selection in general epidemiology. *International journal of epidemiology*, 37(3):430–434.

Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., Ip, H. F., Marioni, R. E., McIntosh, A. M., Deary, I. J., et al. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature human behaviour*, 3(5):513–525.

Hagquist, C. (2010). Discrepant trends in mental health complaints among younger and older adolescents in sweden: an analysis of who data 1985–2005. *Journal of Adolescent Health*, 46(3):258–264.

Health & Social Care Department (2019). The government's revised mandate to NHS England for 2018-19. *National Health Service Act*, 13A(1):1–27.

Heckman, J. J. and Borjas, G. J. (1980). Does unemployment cause future unemployment? definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica*, 47(187):247–283. doi: https://doi.org/10.2307/2553150.

Heggebø, K. and Elstad, J. I. (2017). Is it easier to be unemployed when the experience is more widely shared? effects of unemployment on self-rated health in 25 european countries with diverging macroeconomic conditions. *European Sociological Review*, 34(1):22–39. doi: https://doi.org/10.1093/esr/jcx080.

Hek, K., Demirkan, A., Lahti, J., Terracciano, A., Teumer, A., Cornelis, M. C., ..., and Murabito, J. (2013). A Genome-Wide Association Study of Depressive Symptoms. *Biological Psychiatry*, 73(7):667–678. doi: https://doi.org/10.1016/j.biopsych.2012.09.033.

Herbig, B., Dragano, N., and Angerer, P. (2013). Health in the long-term unemployed. *Deutsches Arzteblatt international*, 110(23-24):413–419. doi: https://doi.org/10.3238/arztebl.2013.0413.

Horwitz, A. V. (2009). *An Overview of Sociological Perspectives on the Definitions, Causes, and Responses to Mental Health and Illness*, page 6–19. In T. L. Scheid and T. N. Brown (Ed.), *A Handbook for the Study of Mental Health: Social Contexts, Theories, and Systems*. Cambridge University Press, 2 edition. doi: https://doi.org/10.1017/CBO9780511984945.004.

Howard, D. M., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., Shirali, M., ..., and McIntosh, A. M. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature Neuroscience*, 22(3):343–352. doi: https://doi.org/10.1038/s41593-018-0326-7.

Hox, J. J., Moerbeek, M., and Van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

Hox, J. J., Moerbeek, M., and Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Huffman, A. H., Culbertson, S. S., Wayment, H. A., and Irving, L. H. (2015). Resource replacement and psychological well-being during unemployment: The role of family support. *Journal of Vocational Behavior*, 89:74–82. doi: https://doi.org/10.1016/j.jvb.2015.04.011.

Hugh-Jones, D., Verweij, K. J., Pourcain, B. S., and Abdellaoui, A. (2016). Assortative mating on educational attainment leads to genetic spousal resemblance for polygenic scores. *Intelligence*, 59:103 – 108. doi: https://doi.org/10.1016/j.intell.2016.08.005.

Hughes, A. (2018). The UK Household Longitudinal Study (UKHLS): overview for health researchers. Health Studies User Conference 2018.

Hughes, P. R. and Hutchinson, G. (1986). *Unemployment, Irreversibility and the Long Term Unemployed*. Queen Mary College Department of Economics.

Hussmanns, R., Mehran, F., and Varmā, V. (1990). *Surveys of economically active population, employment, unemployment, and underemployment: an ILO manual on concepts and methods*. International Labour Organization.

Hyde, C. L., Nagle, M. W., Tian, C., Chen, X., Paciga, S. A., Wendland, J. R., ..., and Winslow, A. R. (2016). Identification of 15 Genetic loci Associated with Risk of Major Depression in Individuals of European Descent. *Nature Genetics*, 48(9):1031–1036.

Infante-Rivard, C. and Cusson, A. (2018). Reflection on modern methods: selection bias—a review of recent developments. *International Journal of Epidemiology*, 47(5):1714–1722. doi: https://doi.org/10.1093/ije/dyy138.

Jacobson, N. S. and Addis, M. E. (1993). Research on couples and couple therapy: What do we know? where are we going? *Journal of consulting and clinical psychology*, 61(1):85.

Jaffee, S. R. and Price, T. S. (2007). Gene–environment correlations: A review of the evidence and implications for prevention of mental illness. *Molecular psychiatry*, 12(5):432. doi: https://doi.org/10.1093/esr/jcx080.

Jahoda, M. (1988). Economic recession and mental health: Some conceptual issues. *Journal of social Issues*, 44(4):13–23.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., et al. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1789–1858.

Jansson, M., Gatz, M., Berg, S., Johansson, B., Malmberg, B., McClearn, G. E., Schalling, M., and Pedersen, N. L. (2004). Gender differences in heritability of depressive symptoms in the elderly. *Psychological Medicine*, 34:471–479. doi: https://doi.org/10.1017/S0033291703001375.

Jefferis, B. J., Nazareth, I., Marston, L., Moreno-Kustner, B., Bellón, J., Svab, I., Rotar, D., Geerlings, M. I., Xavier, M., Goncalves-Pereira, M., Vicente, B., Saldivia, S., Aluoja, A., Kalda, R., and King, M. (2011). Associations between unemployment and major depressive disorder: evidence from an international, prospective study (the predict cohort). *Social Science and Medicine*, 73(11):1627–34. doi: https://doi.org/10.1016/j.socscimed.2011.09.029.

Jenkins, J. (2010). The labour market in the 1980s, 1990s and 2008/09 recessions. *Economic & Labour Market Review*, 4(8):29–36. doi: https://doi.org/10.1057/elmr.2010.110.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kaufman, J., Yang, B. Z., Douglas-Palumberi, H., Grasso, D., Lipschitz, D., Houshyar, S., Krystal, J. H., and Gelernter, J. (2006). Brain-derived Neurotrophic factor-5-HTTLPR Gene Interactions and Environmental Modifiers of Depression in Children. *Biological Psychiatry*, 59(8):673–680.

Keller, M. C. (2014). Gene × environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biological psychiatry*, 75(1):18–24. doi: https://doi.org/10.1016/j.biopsych.2013.09.006.

Kendler, K. S., Gardner, C. O., and Prescott, C. A. (2006). Toward a comprehensive developmental model for major depression in men. *The American Journal of Psychiatry*, 163(1):115–24. doi: 10.1176/appi.ajp.163.1.115.

Kerminen, S., Martin, A. R., Koskela, J., Ruotsalainen, S. E., Havulinna, A. S., Surakka, I., Palotie, A., Perola, M., Salomaa, V., Daly, M. J., Ripatti, S., and Pirinen, M. (2019). Geographic variation and bias in the polygenic scores of complex diseases and traits in finland. *The American Journal of Human Genetics*, 104(6):1169–1181. doi: https://doi.org/10.1016/j.ajhg.2019.05.001.

Kessler, R. C., Barber, C., Birnbaum, H. G., Frank, R. G., Greenberg, P. E., Rose, R. M., Simon, G. E., and Wang, P. (1999). Depression in the workplace: Effects on short-term disability: Could treating workers' depression help employers to save money on disability? these results are encouraging. *Health affairs*, 18(5):163–171.

Kessler, R. C. and Essex, M. (1982). Marital status and depression: The importance of coping resources. *Social forces*, 61(2):484–507.

Keyes, C. L. M. (2005). Mental illness and/or mental health? Investigating axioms of the complete state model of health. *Journal of Consulting and Clinical Psychology*, 73(3):539–548. doi: https://doi.org/10.1037/0022--006X.73.3.539.

Kinderman, P. (2014). *A prescription for psychiatry: Why we need a whole new approach to mental health and wellbeing*. Palgrave Macmillan UK. doi: https://doi.org/10.1057/9781137408716.

Klahr, A. M. and Burt, S. A. (2014). Elucidating the etiology of individual differences in parenting: A meta-analysis of behavioral genetic research. *Psychological Bulletin*, 140(2):544–586. doi: https://doi.org/10.1037/a0034205.

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., Gudbjartsson, D. F., Helgason, A., Bjornsdottir, G., Thorsteinsdottir, U., and Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428. doi: https://doi.org/10.1126/science.aan6877.

Kuehner, C. (2017). Why is Depression More Common among Women than among Men? *The Lancet Psychiatry*, 4(2):146–158. doi: http://dx.doi.org/10.1016/S2215-0366(16)30263-2.

Kulminski, A., Loika, Y., Culminskaya, I., Arbeev, K., Arbeeva, L., Christensen, K., Stallard, P., and Yashin, A. (2016). Genetic Predisposition to Age-Related Phenotypes in the Light of Evolution. *Gerontologist*, 56(3):49–49.

Langmead, B. and Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, 19(4):208. doi: https://doi.org/10.1038/nrg.2017.113.

Lau, J. Y. and Eley, T. C. (2008). Disentangling gene-environment correlations and interactions on adolescent depressive symptoms. *Journal of child psychology and psychiatry, and allied disciplines*, 49(2):142–50. doi: https://doi.org/10.1111/j.1469-7610.2007.01803.x.

Lee, J. (2011). Pathways from education to depression. *Journal of cross-cultural gerontology*, 26(2):121–135.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., ..., and Consortium, S. S. G. A. (2018). 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. doi: https://doi.org/10.1038/s41588-018-0147-3.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10(1):387–406. PMID: 19715440.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834. doi: https://doi.org/10.1002/gepi.20533.

Liu, H. and Guo, G. (2015). Lifetime Socioeconomic Status, Historical Context, and Genetic Inheritance in Shaping Body Mass in Middle and Late Adulthood. *American sociological review*, 80(4):705–737.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., ..., and Abecasis, G. R. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443. doi: https://doi.org/10.1038/ng.3679.

Lohoff, F. W. (2010). Overview of the Genetics of Major Depressive Disorder. *Current Psychiatry Reports*, 12(6):539–546.

Lopizzo, N., Bocchio Chiavetto, L., Cattane, N., Plazzotta, G., Tarazi, F. I., Pariante, C. M., Riva, M. A., and Cattaneo, A. (2015). Gene-environment interaction in major depression: focus on experience-dependent biological systems. *Frontiers Psychiatry*, 6:68. doi: https://doi.org/10.3389/fpsyt.2015.00068.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., ..., and Murray, C. J. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2095–2128. doi: https://doi.org/10.1016/S0140-6736(12)61728-0.

Lundin, A., Hallgren, M., Theobald, H., Hellgren, C., and En, M. (2016). Validity of the 12-item version of the general health questionnaire in detecting depression in the general population. *Public Health*, 136. doi: https://doi.org/10.1016/j.puhe.2016.03.005.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21. doi: https://doi.org/10.1038/456018a.

Mandemakers, J. (2011). *Socio-economic differentials in the impact of life course transitions on well-being*. PhD thesis. Pagination: 181.

Manolio, T., Weis, B., Cowie, C., Hoover, R., Hudson, K., Kramer, B., ..., and Collins, F. (2012). New models for large prospective studies: Is there a better way? *American journal of epidemiology*, 175:859–66. doi: https://doi.org/10.1093/aje/kwr453.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ..., and Chakravarti, A. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753. doi: https://doi.org/10.1038/nature08494.

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

Marcus, S. C. and Olfson, M. (2010). National Trends in the Treatment for Depression From 1998 to 2007. *JAMA Psychiatry*, 67(12):1265–1273. doi: https://doi.org/10.1001/archgenpsychiatry.2010.151.

Martikainen, P. and Valkonen, T. (1999). Bias related to the exclusion of the economically inactive in studies on social class differences in mortality. *International journal of epidemiology*, 28(5):899–904.

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

Matud, M. P. (2004). Gender Differences in Stress and Coping Styles. *Personality and Individual Differences*, 37(7):1401–1415.

McDaid, D. (2017). Socioeconomic disadvantage and suicidal behaviour during times of economic recession and recovery. http://eprints.lse.ac.uk/69795/1/McDaid_Socioeconomic%20disadvantage%20and%20suicidal%20behaviour_published_2017%20LSERO%20edit.pdf. Online; accessed 20-May-2020.

McFall, S., Petersen, J., Kaminska, O., and Lynn, P. (2014). Understanding society waves 2 and 3 nurse health assessment, 2010-2012. *Guide to Nurse Health Assessment. ISER, University of Essex*. https://www.understandingsociety.ac.uk/documentation/health-assessment. Online; accessed 15-June-2019.

McFall, S. L., Booker, C., Burton, J., Conolly, A., et al. (2012). Implementing the biosocial component of understanding society - nurse collection of biomeasures. *Institute for Social and Economic Research, University of Essex*. https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2012-04. Online; accessed 15-June-2019.

McGowan, P. O., Sasaki, A., D'Alessio, A. C., Dymov, S., Labonté, B., Szyf, M., Turecki, G., and Meaney, M. J. (2009). Epigenetic Regulation of the Glucocorticoid Receptor in Human Brain Associates with Childhood Abuse. *Nature Neuroscience*, 12(3):342–348.

McGue, M. and Christensen, K. (2003). The heritability of depression symptoms in elderly Danish twins: Ocassion-specific versus general effects. *Behavior Genetics*, 33(2):83–93. doi: https://doi.org/10.1023/A:1022545600034.

McGue, M., Skytthe, A., and Christensen, K. (2014). The nature of behavioural correlates of healthy ageing: a twin study of lifestyle in mid to late life. *International Journal of Epidemiology*, 43(3):775–782. doi: https://doi.org/10.1093/ije/dyt210.

McLaren, N. (1998). A critical review of the biopsychosocial model. *Australian & New Zealand Journal of Psychiatry*, 32(1):86–92. doi: https://doi.org/10.3109/00048679809062712.

McLean, C., Carmona, C., Francis, S., Wohlgemuth, C., and Mulvihill, C. (2005). Worklessness and health: What do we know about the causal relationship. *Evidence review. London: Health Development Agency*.

McManus, S., Bebbington, P., Jenkins, R., and Brugha, T. (2016). Adult psychiatric morbidity survey: survey of mental health and wellbeing, England, 2014. *NHS Digital Leeds, UK*.

Meaney, F. J. and Taylor, C. (2018). Heritability. https://www.britannica.com/science/heritability. Online; accessed 01-November-2019.

Meyer, H. (2019). Genotype quality control with plinkQC. https://cran.r-project.org/web/packages/plinkQC/vignettes/plinkQC.pdf. Online; accessed 22-June-2018.

Mills, M., Tropf, F., and Barban, N. (2020). *An Introduction to Statistical Genetic Data Analysis*. Cambridge, MA: MIT Press. (Forthcoming).

Mills, M. C. and Rahal, C. (2019). A scientometric review of genome-wide association studies. *Communications Biology*, 2(1):9. doi: https://doi.org/10.1038/s42003-018-0261-x.

Mills, M. C. and Tropf, F. C. (2020). Sociology, genetics, and the coming of age of sociogenomics. *Annual Review of Sociology*, 46.

Moffitt, T. E., Caspi, A., and Rutter, M. (2005). Strategy for investigating interactions between measured genes and measured environments. *Archives of general psychiatry*, 62(5):473–481.

Morris, T. T., Davies, N. M., Hemani, G., and Smith, G. D. (2020). Population phenomena inflate genetic associations of complex social traits. *Science Advances*, 6(16):eaay0328. doi: https://doi.org/10.1126/sciadv.aay0328.

Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J. K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. *eLife*, 9:e48376. doi: https://doi.org/10.7554/eLife.48376.

Muglia, P., Tozzi, F., Galwey, N., Francks, C., Upmanyu, R., Kong, X., Antoniades, A., Domenici, E., Perry, J., Rothen, S., et al. (2010). Genome-wide association study of recurrent major depressive disorder in two european case–control cohorts. *Molecular psychiatry*, 15(6):589.

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., and Davey Smith, G. (2017). Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology*, 47(1):226–235. doi: https://doi.org/10.1093/ije/dyx206.

Murphy, G. C. and Athanasou, J. A. (1999). The effect of unemployment on mental health. *Journal of Occupational and Organizational Psychology*, 72(1):83–99. doi: https://onlinelibrary.wiley.com/doi/abs/10.1348/096317999166518.

Nathan, P. E. (2007). Efficacy, effectiveness, and the clinical utility of psychotherapy research. *The art and science of psychotherapy*, pages 69–83.

National Institutes of Health (2019). Genetics home reference - genomic research. https://ghr.nlm.nih.gov. Online; accessed 01-November-2019.

Nelson, R. E. and Kim, J. (2011). The impact of mental illness on the risk of employment termination. *The journal of mental health policy and economics*, 14(1):39–52.

NHS (2019). Overview: Clinical depression. https://www.nhs.uk/conditions/clinical-depression/#overview. Online; accessed 13-December-2019.

Ni, G., van der Werf, J., Zhou, X., Hyppönen, E., Wray, N. R., and Lee, S. H. (2019). Genotype–covariate correlation and interaction disentangled by a whole-genome multivariate reaction norm model. *Nature Communications*, 10(1):2239. doi: https://doi.org/10.1177/1745691619867107.

Nichols, A. et al. (2010). Regression for nonnegative skewed dependent variables. In *BOS10 Stata Conference*, volume 2, pages 15–16. Stata Users Group.

Okbay, A., Baselmans, B. M., De Neve, J. E., Turley, P., Nivard, M. G., Fontana, M. A., ..., and Gratten, J. (2016). Genetic Variants Associated with Subjective Well-Being, Depressive Symptoms, and Neuroticism Identified through Genome-Wide Analyses. *Nature Genetics*, 48(6):624–633. doi: https://doi.org/10.1038/ng.3552.

Owen, K. u. and Watson, N. (1995). Unemployment and mental health. *Journal of psychiatric and mental health nursing*, 2(2):63–71.

Papageorge, N. W. and Thom, K. (2019). Genes, education, and labor market outcomes: Evidence from the health and retirement study. *Journal of the European Economic Association*, jvz072. doi: https://doi.org/10.1093/jeea/jvz072.

Paris, J. (2008). *Prescriptions for the mind: A critical view of contemporary psychiatry*. Oxford University Press.

Patalay, P. and Gage, S. H. (2019). Changes in millennial adolescent mental health and health-related behaviours over 10 years: a population cohort comparison study. *International journal of epidemiology*.

Patten, S. B. (2003). Recall bias and major depression lifetime prevalence. *Social Psychiatry and Psychiatric Epidemiology*, 38(6):290–296.

Paul, K. I. and Moser, K. (2009). Unemployment impairs mental health: Meta-analyses. *Journal of Vocational Behavior*, 74(3):264 – 282. doi: https://doi.org/10.1016/j.jvb.2009.01.001.

Pearlin, L. I. and Johnson, J. S. (1977). Marital status, life-strains and depression. *American sociological review*, pages 704–715.

Pehkonen, J., Viinikainen, J., Böckerman, P., Lehtimäki, T., Pitkänen, N., and Raitakari, O. (2017). The challenges of gxe research: A rejoinder. *Soc Sci Med*, 188:204–205. doi: https://doi.org/10.1016/j.socscimed.2017.07.010.

Pergamin-Hight, L., Bakermans-Kranenburg, M. J., Van Ijzendoorn, M. H., and Bar-Haim, Y. (2012). Variations in the Promoter Region of the Serotonin Transporter Gene and Biased Attention for Emotional information: A Meta-Analysis. *Biological Psychiatry*, 71(4):373–379.

Peters, A. T., Shankman, S. A., Deckersbach, T., and West, A. E. (2015). Predictors of first-episode unipolar major depression in individuals with and without sub-threshold depressive symptoms: A prospective, population-based study. *Psychiatry Research*, 230(2):150–6. doi: https://doi.org/10.1016/j.psychres.2015.08.030.

Peterson, C. (2009). *Psychological Approaches to Mental Illness*, page 89–105. In T. L. Scheid and T. N. Brown (Ed.), *A Handbook for the Study of Mental Health: Social Contexts, Theories, and Systems*. Cambridge University Press, 2 edition. doi: https://doi.org/10.1017/CBO9780511984945.008.

Peterson, R. E., Cai, N., Bigdeli, T. B., Li, Y., Reimers, M., Nikulova, A., Webb, B. T., Bacanu, S.-A., Riley, B. P., Flint, J., and Kendler, K. S. (2017). The genetic architecture of major depressive disorder in han chinese women. *JAMA psychiatry*, 74(2):162–168. doi: https://doi.org/10.1001/jamapsychiatry.2016.3578.

Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., ..., and Busonero, F. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *European journal of human genetics : EJHG*, 23(7):975–983. doi: http://doi.org/10.1038/ejhg.2014.216.

Plomin, R., DeFries, J. C., and Loehlin, J. C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin*, 84(2):309–322. doi: https://doi.org/10.2217/nmt.11.45.

Plomin, R., Haworth, C. M. A., and Davis, O. S. P. (2009). Common disorders are quantitative traits. *Nature Reviews Genetics*, 10(12):872–878. doi: https://doi.org/10.1038/nrg2670.

Power, R. A., Tansey, K. E., Buttenschøn, H. N., Cohen-Woods, S., Bigdeli, T., Hall, L. S., ..., and Steinberg, S. (2017). Genome-wide association for major depression through age at onset stratification: major depressive disorder working group of the psychiatric genomics consortium. *Biological psychiatry*, 81(4):325–335.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909. doi: https://doi.org/10.1038/ng1847.

Prins, B. (2015). Genome-wide genotyping in the UKHLS. Understanding Society Scientific Conference 2015. https://www.understandingsociety.ac.uk/scientific-conference-2015/papers/78.html. Online; accessed 3-March-2019.

Prior, L., Jones, K., and Manley, D. (2020). Ageing and cohort trajectories in mental ill-health: An exploration using multilevel models. *PLOS ONE*, 15(7):1–14. doi: https://doi.org/10.1371/journal.pone.0235594.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ..., and Sham, P. C. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses.

The American Journal of Human Genetics, 81(3):559–575. doi: https://doi.org/10.1086/519795.

Rai, D., Zitko, P., Jones, K., Lynch, J., and Araya, R. (2013). Country- and individual-level socioeconomic determinants of depression: multilevel cross-national comparison. *British Journal of Psychiatry*, 202(3):195–203. doi: https://doi.org/10.1192/bjp.bp.112.112482.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage.

Rice, N. E., Lang, I. A., Henley, W., and Melzer, D. (2010). Baby boomers nearing retirement: the healthiest generation? *Rejuvenation research*, 13(1):105–114.

Rietschel, M., Mattheisen, M., Frank, J., Treutlein, J., Degenhardt, F., Breuer, R., Steffens, M., Mier, D., Esslinger, C., Walter, H., et al. (2010). Genome-wide association-, replication-, and neuroimaging study implicates homer1 in the etiology of major depression. *Biological psychiatry*, 68(6):578–585.

Ritchie, H. and Roser, M. (2018). Mental health. *Our World in Data*. https://ourworldindata.org/mental-health.

Robinette, J. W., Boardman, J. D., and Crimmins, E. M. (2019). Differential vulnerability to neighbourhood disorder: a gene×environment interaction study. *Journal of Epidemiology and Community Health*, 73(5):388–392. doi: https://doi.org/10.1136/jech-2018-211373.

Rosholm, M. and Andersen, H. L. (2010). The effect of changing mental health on unemployment duration and destination states after unemployment. *SSRN Electronic Journal*, 21(1):23–26. doi: http://dx.doi.org/10.2139/ssrn.1672026.

Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review*, 30(6):843–861. doi: http://www.jstor.org/stable/2090964.

Schmitz, L. and Conley, D. (2017). Modeling gene-environment interactions with quasi-natural experiments. *Journal of Personality*, 85(1):10–21. doi: https://doi.org/10.1111/jopy.12227.

Schwartz, B. (2015). *Why we work*. Simon and Schuster.

Schwartz, S. and Corcoran, C. (2009). *Biological Theories of Psychiatric Disorders: A Sociological Approach*, page 64–88. In T. L. Scheid and T. N. Brown (Ed.), *A Handbook for the Study of Mental Health: Social Contexts, Theories, and Systems*. Cambridge University Press, 2 edition. doi: https://doi.org/10.1017/CBO9780511984945.007.

Seabrook, J. A. and Avison, W. R. (2010). Genotype-environment interaction and sociology: Contributions and complexities. *Social Science and Medicine*, 70:1277–1284.

Shanahan, M. J. and Hofer, S. M. (2011). Molecular genetics, aging, and well-being: Sensitive period, accumulation, and pathway models. In *Handbook of aging and the social sciences*, pages 135–147. Elsevier.

Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. *American psychologist*, 65(2):98.

Shyn, S. I., Shi, J., Kraft, J., Potash, J. B., Knowles, J., Weissman, M., Garriock, H., Yokoyama, J., McGrath, P., Peters, E., et al. (2011). Novel loci for major depression identified by genome-wide association study of sequenced treatment alternatives to relieve depression and meta-analysis of three studies. *Molecular psychiatry*, 16(2):202.

Smith-Woolley, E., Pingault, J.-B., Selzam, S., Rimfeld, K., Krapohl, E., von Stumm, S., Asbury, K., Dale, P. S., Young, T., Allen, R., Kovas, Y., and Plomin, R. (2018). Differences in exam performance between pupils attending selective and non-selective schools mirror the genetic differences between them. *npj Science of Learning*, 3(1):3. doi: https://doi.org/10.1038/s41539-018-0019-8.

Spiers, N., Bebbington, P., McManus, S., Brugha, T. S., Jenkins, R., and Meltzer, H. (2011). Age and birth cohort differences in the prevalence of common mental disorder in england: National psychiatric morbidity surveys 1993–2007. *The British Journal of Psychiatry*, 198(6):479–484.

Spiers, N., Brugha, T., Bebbington, P., McManus, S., Jenkins, R., and Meltzer, H. (2012). Age and birth cohort differences in depression in repeated cross-sectional surveys in england: the national psychiatric morbidity surveys, 1993 to 2007. *Psychological Medicine*, 42(10):2047–2055.

Stankunas, M., Kalediene, R., Starkuviene, S., and Kapustinskiene, V. (2006). Duration of unemployment and depression: a cross-sectional survey in lithuania. *BMC Public Health*, 6(1):174. doi: https://doi.org/10.1186/1471-2458-6-174.

Statt, D. A. (1994). *Psychology and the World of Work*. NYU Press.

Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics*, 76(3):449–462. doi: https://doi.org/10.1086/428594.

Stergiakouli, E., Martin, J., Hamshere, M. L., Heron, J., St Pourcain, B., Timpson, N. J., Thapar, A., and Davey Smith, G. (2016). Association between polygenic risk scores for attention-deficit hyperactivity disorder and educational and cognitive outcomes in the general population. *International Journal of Epidemiology*, 46(2):421–428. doi: https://doi.org/10.1093/ije/dyw216.

Strandh, M., Winefield, A., Nilsson, K., and Hammarström, A. (2014). Unemployment and mental health scarring during the life course. *European Journal of Public Health*, 24(3):440–5. doi: https://doi.org/10.1093/eurpub/cku005.

Stuckler, D., Reeves, A., Loopstra, R., Karanikolos, M., and McKee, M. (2017). Austerity and health: the impact in the uk and europe. *European journal of public health*, 27(suppl_4):18–21.

Sullivan, P., Neale, M., and Kendler, K. (2000). Genetic epidemiology of major depression: Review and meta-analysis. *American Journal of Psychiatry*, 157(10):1552–1562. doi: https://doi.org/10.1176/appi.ajp.157.10.1552.

Sullivan, P. F., Agrawal, A., Bulik, C. M., Andreassen, O. A., Børglum, A. D., Breen, G., Cichon, S., Edenberg, H. J., Faraone, S. V., Gelernter, J., Mathews, C. A., Nievergelt, C. M., Smoller, J. W., and O'Donovan, M. C. (2018). Psychiatric genomics: An update and an agenda. *Am J Psychiatry*, 175(1):15–27. doi: https://doi.org/10.1176/appi.ajp.2017.17030283.

Syvälahti, E. K. (1994). Biological aspects of depression. *Acta Psychiatrica Scandinavica*, 89:11–15. doi: https://doi.org/10.1111/j.1600-0447.1994.tb05795.x.

Tefft, N. (2011). Insights on unemployment, unemployment insurance, and mental health. *Journal of Health Economics*, 30(2):258–264.

Terracciano, A., Tanaka, T., Sutin, A. R., Sanna, S., Deiana, B., Lai, S., Uda, M., Schlessinger, D., Abecasis, G. R., Ferrucci, L., and Costa, Paul T., J. (2010). Genome-wide association scan of trait depression. *Biological Psychiatry*, 68(9):811–817. doi: https://doi.org/10.1016/j.biopsych.2010.06.030.

Thase, M. E., Jindal, R., and Howland, R. H. (2002). Biological aspects of depression.

Thom, T. J. (1989). International mortality from heart disease: rates and trends. *International journal of epidemiology*, 18(3_Supplement_1):S20–S28.

Thomson, R. M. and Katikireddi, S. V. (2018). Mental health and the jilted generation: Using age-period-cohort analysis to assess differential trends in young people's mental health following the great recession and austerity in england. *Social Science and Medicine*, 214:133–143.

Trejo, S. and Domingue, B. W. (2019). Genetic nature or genetic nurture? quantifying bias in analyses using polygenic scores. *bioRxiv, 524850*.

Tsuang, M. and Faraone, S. (1990). *The Genetics of Mood Disorders*. The Johns Hopkins University Press, Baltimore, US.

TUC (2018). Breaking point: the crisis in mental health funding. https://www.tuc.org.uk/sites/default/files/Mentalhealthfundingreport2_0.pdf. London: Trades Union Congress. Online; accessed 02-December-2020.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, 9(5):160–164. doi: https://doi.org/10.1111/1467-8721.00084.

Twenge, J. M., Cooper, A. B., Joiner, T. E., Duffy, M. E., and Binau, S. G. (2019). Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017. *Journal of abnormal psychology*.

Twenge, J. M., Gentile, B., DeWall, C. N., Ma, D., Lacefield, K., and Schurtz, D. R. (2010). Birth cohort increases in psychopathology among young americans, 1938–2007: A cross-temporal meta-analysis of the mmpi. *Clinical psychology review*, 30(2):145–154.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass.)*, 21(4):540–551. doi: https://doi.org/10.1097/EDE.0b013e3181df191c.

Veldkamp, S. A. M., Boomsma, D. I., de Zeeuw, E. L., van Beijsterveldt, C. E. M., Bartels, M., Dolan, C. V., and van Bergen, E. (2019). Genetic and environmental influences on different forms of bullying perpetration, bullying victimization, and their co-occurrence. *Behavior Genetics*, 49(5):432–443. doi: https://doi.org/10.1007/s10519-019-09968-5.

Vilhjalmsson, B., Yang, J., Finucane, H. K., Gusev, A., Lindstrom, S., Ripke, S., ..., and Price, A. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97:576–592. doi: https://doi.org/10.1016/j.ajhg.2015.09.001.

Virtanen, S., Kaprio, J., Viken, R., Rose, R. J., and Latvala, A. (2019). Birth cohort effects on the quantity and heritability of alcohol consumption in adulthood: A finnish longitudinal twin study. *Addiction*, 114(5):836–846.

Vos, T., Barber, R. M., Bell, B., Bertozzi-Villa, A., Biryukov, S., Bolliger, I., ..., and Murray, C. J. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries: a systematic analysis for the global burden of disease study 2013. *The Lancet*, 386(9995):743–800. doi: https://doi.org/10.1016/S0140-6736(15)60692-4.

Walter, S., Mejía-Guevara, I., Estrada, K., Liu, S. Y., and Glymour, M. M. (2016). Association of a genetic risk score with body mass index across different birth cohorts. *Jama*, 316(1):63–69.

Ward, C. and Dale, A. (1992). Geographical variation in female labour force participation: An application of multilevel modelling. *Regional Studies*, 26(3):243–255. doi: https://doi.org/10.1080/00343409212331346941.

Ware, E. B., Schmitz, L. L., Faul, J., Gard, A., Mitchell, C., Smith, J. A., Zhao, W., Weir, D., and Kardia, S. L. (2017). Heterogeneity in polygenic scores for common human traits. *bioRxiv*. doi: https://doi.org/10.1101/106062.

Waters, L. E. and Moore, K. A. (2002). Predicting self-esteem during unemployment: The effect of gender, financial deprivation, alternate roles, and social support. *Journal of Employment Counseling*, 39(4):171–189. doi: https://doi.org/10.1002/j.2161-1920.2002.tb00848.x.

Weir, D. R. (2012). Quality control report for genotypic data. http://hrsonline.isr.umich.edu/sitedocs/genetics/HRS_QC_REPORT_MAR2012.pdf. Online; accessed 22-December-2018.

Wigmore, E. M., Hafferty, J. D., Hall, L. S., Howard, D. M., Clarke, T.-K., Fabbri, C., Lewis, C. M., Uher, R., Navrady, L. B., Adams, M. J., et al. (2020). Genome-wide association study of antidepressant treatment resistance in a population-based cohort using health service prescription data and meta-analysis with gendep. *The pharmacogenomics journal*, 20(2):329–341.

Wilkinson, P. O. and Goodyer, I. M. (2011). Childhood adversity and allostatic overload of the hypothalamic-pituitary-adrenal axis: a vulnerability model for depressive disorders. *Development and Psychopathology*, 23(4):1017–37. doi: http://doi.org/10.1017/S0954579411000472.

Wilkinson, P. O., Trzaskowski, M., Haworth, C. M., and Eley, T. C. (2013). The role of gene-environment correlations and interactions in middle childhood depressive symptoms. *Dev Psychopathol*, 25(1):93–104. doi: https://doi.org/10.1017/S0954579412000922.

Winefield, A. H. and Tiggemann, M. (1990). Length of unemployment and psychological distress: Longitudinal and cross-sectional data. *Social Science and Medicine*, 31(4):461–465. doi: https://doi.org/10.1016/0277-9536(90)90041-P.

Woodland, A. D. (1987). Determinants of the Labour Force Status of the Aged. *The Economic Record*, 63(181):97–114.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

World Health Organisation (2018). Depression. http://www.who.int/mediacentre/factsheets/fs369/en/. Online; accessed 12-October-2018.

World Health Organisation (2020). Mental health prevention of suicidal behaviours: A task for all. http://www.who.int/mental_health/prevention/suicide/background. Online; accessed 20-May-2020.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., ..., and Adams, M. J. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, 50(5):668–681. doi: https://doi.org/10.1038/s41588--018--0090--3.

Yang, Y. and Land, K. C. (2013). *Age-period-cohort analysis: New models, methods, and empirical applications*. CRC press.

Yehuda, R., Halligan, S., and Grossman, R. (2001). Childhood Trauma and Risk for PTSD: Relationship to Intergenerational Effects of Trauma, Parental PTSD, and Cortisol Excretion. *Development and Psychopathology*, 13(3):733–753.

Zaidi, A. A. and Mathieson, I. (2020). Demographic history mediates the effect of stratification on polygenic scores. *eLife*, 9:e61548.

Zeng, X., Vonk, J. M., van der Plaat, D. A., Faiz, A., Paré, P. D., Joubert, P., Nickle, D., Brandsma, C.-A., Kromhout, H., Vermeulen, R., Xu, X., Huo, X., de Jong, K., and Boezen, H. M. (2019). Genome-wide interaction study of gene-by-occupational exposures on respiratory symptoms. *Environment International*, 122:263–269. doi: https://doi.org/10.1016/j.envint.2018.11.017.

Supplementary material - Chapter 2

A.1 Variables and sample

Table A.1.1: Variable definitions

Dependent variable	
Genotyping status	1=genotyped; 0=not genotyped
Independent variables	
Socio-economic characteristics	
sex	1=female; 0=male
Age (Age-sq.)	Age in years at nurse assessment wave (wave 2 for GPS and wave 3 for BHPS participants)
Married	Binary (0/1) indicator whether a respondent ever reported being married

Table A.1.2: Continuation of variable definitions

Children	Binary (0/1) indicator whether a respondent ever had a (biological) child
Education	Highest educational attainment reported over the survey (1=no qualifications; 2=lower secondary; 3=higher secondary; 4=lower tertiary; 5=higher tertiary)
Economic activity	Labour force status at nurse assessment wave (wave 2 for GPS and wave 3 for BHPS participants) 1=employed (including self-employment); 2=unemployed (ILO definition); 3=retired; 4=full time student; 5=economically inactive
Neighbourhood characteristics	
Type of area	1=urban; 0=rural at nurse assessment wave (wave 2 for GPS and wave 3 for BHPS participants)
Neighbourhood	The mean of the Buckner's Neighbourhood Cohesion Instrument (mean a=.87) over available waves prior to genotyping (waves 1, 3)
Life-style	
Smoking	Binary (0/1) indicator whether a respondent ever reported smoking
Drinking	Mean of how often a respondent had an alcoholic drink during the each survey year (from 1=almost every day to 8=not at all)
Physical health indicators	
BMI	Body mass index in kg/m^2
High blood pressure	Binary (0/1) indicator whether a respondent ever diagnosed with high blood pressure
Diabetes	Binary (0/1) indicator whether a respondent ever diagnosed having diabetes
Heart condition	Binary (0/1) indicator whether a respondent ever diagnosed having such a heart condition as congestive heart failure, coronary heart disease, heart attack, myocardial infarction or stroke
Respiratory diseases	Binary (0/1) indicator whether a respondent ever diagnosed having asthma, angina, chronic bronchitis
Cancer	Binary (0/1) indicator whether a respondent ever diagnosed having cancer
General health	Self-reported health (1=excellent; 2=very good; 3=good; 4=fair; 5=poor)

Table A.1.3: Continuation of variable definitions

Mental health indicators	
Clinical depression	Binary (0/1) indicator whether a respondent ever diagnosed having clinical depression
Depressive symptoms	Composite GHQ depressive symptoms score based on such indicators as concentration, loss of sleep, usefulness, decision making, being constantly under strain, ability to overcome difficulties, to enjoy day-to-day activities, ability to face problems, feelings of unhappiness and depression, loss of confidence, worthlessness, general happiness. Varies from 0 to 36, higher values indicating greater depressive symptoms

Table A.1.4: Non-response statistics of UKHLS analytical sample, by genotyping status

	N genotyped	N non-genotyped	% reduction	
			genotyped	non-genotyped
Weights availability	9,530	31,026		
sex and age	9,530	31,026	0.0%	0.0%
Rural area	9,529	29,783	0.0%	4.0%
Ever married	9,530	31,026	0.0%	0.0%
Ever had child	9,530	31,026	0.0%	0.0%
Education	6,970	24,168	26.9%	22.1%
Economic activity	9,530	29,787	0.0%	4.0%
Neigh. cohesion	9,425	29,318	1.1%	5.5%
Smoking	9,497	29,617	34.6%	4.5%
Drinking	9,327	27,697	2.1%	10.7%
BMI	6,370	20,692	33.2%	33.3%
High blood pressure	6,573	22,135	31.0%	28.7%
Diabetes	6,573	22,135	31.0%	28.7%
Heart condition	6,573	22,135	31.0%	28.7%
Respiratory diseases	6,573	22,135	31.0%	28.7%
Cancer	6,573	22,135	31.0%	28.7%
General health	9,482	28,676	0.5%	7.6%
Depression	6,573	22,135	31.0%	28.7%
Depressive symptoms	9,342	27,377	2.0%	11.8%

Table A.2.1: Unweighted descriptive statistics for **males**, by genotyping status with t-statistics of differences

	All males		Genotyped		Non-genotyped			
	Mean	SD	Mean	SD	Mean	SD	*t*	*p*
Age	48.99	17.82	52.39	16.88	47.84	17.98	-12.04	.00
Ever married	.758	.428	.828	.378	.734	.442	-10.22	.00
Ever had child	.719	.450	.770	.421	.701	.458	-7.15	.00
Rural area	.270	.444	.263	.440	.273	.445	-1.07	.29
Neighbourhood	3.577	.636	3.623	.609	3.561	.645	-4.50	.00
Education								
No qualifications	.153	.360	.147	.355	.155	.362	.99	.32
Lower secondary	.378	.485	.375	.484	.380	.485	.42	.67
Higher secondary	.166	.372	.165	.371	.166	.372	.12	.90
Lower tertiary	.075	.263	.079	.270	.073	.260	-1.09	.27
Higher tertiary	.228	.420	.233	.423	.227	.419	-.77	.44
Economic activity								
Employed	.599	.490	.593	.491	.601	.490	.85	.40
Unemployed	.059	.236	.043	.203	.065	.247	4.34	.00
Retired	.240	.427	.290	.454	.223	.416	-7.35	.00
Student	.051	.219	.030	.171	.058	.233	5.88	.00
Inactive	.050	.219	.044	.205	.053	.223	1.86	.06
Lifestyle								
Smoking	.693	.461	.691	.462	.695	.461	.35	.73
Drinking	4.062	1.780	3.925	1.745	4.109	1.790	4.85	.00
Health								
General health	2.612	1.032	2.590	1.009	2.620	1.039	1.38	.17
Depressive sympt	10.520	5.089	10.293	4.893	10.597	5.152	2.79	.01
No. of participants	*11,659*		*2,943*		*8,626*		*df* = *11,657*	

A.2 Factors associated with the probability to be genotyped

Table A.2.2: Weighted descriptive statistics for **males**, by genotyping status with t-statistics of differences

	All males		Genotyped		Non-genotyped			
	Mean	SD	Mean	SD	Mean	SD	*t*	*p*
Age	47.96	18.45	47.98	18.15	47.89	19.39	-.10	.92
Ever married	.744	.436	.747	.435	.737	.441	-.46	.64
Ever had child	.699	.459	.708	.454	.668	.471	-1.83	.07
Rural area	.231	.422	.234	.424	.221	.415	-.68	.49
Neighbourhood	3.529	.645	3.534	.647	3.516	.640	-.57	.57
Education								
No qualifications	.152	.359	.144	.351	.178	.382	2.09	.04
Lower secondary	.384	.486	.392	.488	.358	.480	-1.59	.11
Higher secondary	.166	.372	.167	.373	.164	.371	-.18	.86
Lower tertiary	.070	.255	.069	.254	.071	.258	.19	.85
Higher tertiary	.228	.420	.228	.420	.229	.420	.04	.97
Economic activity								
Employed	.617	.486	.624	.485	.597	.491	-1.25	.21
Unemployed	.064	.244	.062	.240	.070	.256	.63	.53
Retired	.231	.421	.227	.419	.243	.429	.95	.34
Student	.043	.202	.040	.195	.051	.221	1.07	.28
Inactive	.046	.209	.048	.213	.038	.192	-1.02	.31
Lifestyle								
Smoking	.690	.463	.693	.461	.680	.467	-.58	.57
Drinking	4.056	1.747	4.034	1.726	4.128	1.811	1.21	.23
Health								
General health	2.568	1.024	2.576	1.022	2.540	1.028	-.79	.43
Depressive sympt	10.526	5.129	10.494	5.079	10.627	5.289	.55	.59
No. of participants	*11,569*		*2,943*		*8,626*		*df* = 1,077*	

**Design df which takes into account survey stratification*

Table A.2.3: Unweighted descriptive statistics for **females**, by genotyping status with t-statistics of differences

	All females		Genotyped		Non-genotyped			
	Mean	SD	Mean	SD	Mean	SD	*t*	*p*
Age	48.98	17.61	51.47	16.32	46.79	17.88	-14.20	.00
Ever married	.783	.412	.842	.365	.763	.425	-10.20	.00
Ever had child	.788	.409	.832	.374	.773	.419	-7.61	.00
Rural area	.268	.443	.254	.435	.273	.446	2.35	.02
Neighbourhood	3.652	.649	3.713	.623	3.631	.656	-6.69	.00
Education								
No qualifications	.173	.378	.169	.375	.174	.379	.63	.53
Lower secondary	.361	.480	.373	.484	.357	.479	-1.71	.09
Higher secondary	.131	.337	.115	.319	.136	.343	3.27	.00
Lower tertiary	.126	.331	.143	.350	.120	.325	-3.73	.00
Higher tertiary	.210	.407	.200	.400	.213	.410	1.76	.08
Economic activity								
Employed	.530	.500	.528	.499	.531	.499	.42	.68
Unemployed	.039	.193	.032	.174	.041	.199	2.67	.01
Retired	.241	.428	.287	.452	.225	.418	-7.68	.00
Student	.053	.223	.028	.164	.061	.239	7.98	.00
Inactive	.137	.344	.126	.332	.141	.348	2.26	.02
Lifestyle								
Smoking	.603	.489	.602	.490	.603	.489	0.06	.95
Drinking	4.745	1.819	4.595	1.845	4.796	1.807	5.87	.00
Health								
General health	2.606	1.040	2.581	1.022	2.615	1.0045	1.74	.08
Depressive sympt	11.571	5.533	11.552	5.601	11.578	5.511	.24	.81
No. of participants	*14,813*		*3,776*		*11,037*		*df* = *14,811*	

Table A.2.4: Weighted descriptive statistics for **females**, by genotyping status with t-statistics of differences

	All females		Genotyped		Non-genotyped			
	Mean	SD	Mean	SD	Mean	SD	*t*	*p*
Age	48.28	17.97	48.36	17.80	48.00	18.55	-.45	.66
Ever married	.768	.422	.772	.420	.757	.429	-.73	.47
Ever had child	.789	.408	.802	.398	.741	.438	-3.08	.00
Rural area	.225	.418	.224	.417	.228	.420	.24	.81
Neighbourhood	3.634	.643	3.647	.640	3.589	.653	-2.05	.04
Education								
No qualifications	.175	.380	.174	.379	.178	.383	.25	.81
Lower secondary	.370	.483	.377	.485	.346	.476	-1.54	.12
Higher secondary	.129	.335	.130	.337	.124	.330	-.41	.68
Lower tertiary	.127	.332	.128	.335	.122	.327	-.54	.59
Higher tertiary	.199	.399	.190	.393	.230	.421	2.35	.02
Economic activity								
Employed	.522	.500	.525	.499	.509	.500	-.77	.44
Unemployed	.041	.197	.042	.200	.037	.188	-.55	.59
Retired	.248	.432	.247	.431	.252	.434	.32	.75
Student	.045	.208	.042	.200	.058	.235	1.34	.18
Inactive	.145	.352	.145	.352	.144	.352	-.04	.97
Lifestyle								
Smoking	.621	.485	.619	.486	.629	.483	.53	.59
Drinking	4.739	1.821	4.719	1.819	4.807	1.826	1.17	.24
Health								
General health	2.618	1.048	2.620	1.046	2.610	1.056	-.23	.82
Depressive sympt	11.662	5.738	11.675	5.767	11.617	5.634	-.23	.82
No. of participants	*11,569*		*2,943*		*8,626*		*df* = 1,077*	

**Design df which takes into account survey stratification*

Table A.2.5: Estimated regression coefficients (robust SE) for all respondents and by sex following weighted and unweighted schemes

	All		Men		Women	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Sex: female	.029	.084	-	-	-	-
	(.030)	(.063)	-	-	-	-
Age	.049***	.021	.035***	.031	.060***	.010
	(.006)	(.012)	(.009)	(.017)	(.008)	(.016)
Age-sq	-.000***	-.000	-.000*	-.000*	-.000***	.000
	.000	.000	.000	.000	.000	.000
Ever married	.087	-.171	.183*	-.168	.002	-.157
	(.048)	(.108)	(.074)	(.158)	(.064)	(.145)
Ever had child	.86*	.275**	.049	.187	.111*	.358**
	(.041)	(.087)	(.061)	(.121)	(.057)	(.121)
Type of area: rural	-.159***	.005	-.119**	.082	-.191***	-.051
	(.033)	(.090)	(.050)	(.118)	(.044)	(.107)
Neighbour. cohesion	.032	.071	-.009	.012	.061	.115
	(.024)	(.059)	(.038)	(.086)	(.032)	(.073)
No qualifications (ref.)						
Lower secondary	.198***	.170	.173*	.293***	.214***	.077
	(.045)	(.089)	(.069)	(.138)	(.060)	(.120)
Higher secondary	.227***	.157	.257***	.263	.196*	.077
	(.057)	(.124)	(.082)	(.173)	(.079)	(.169)
Lower tertiary	.243***	.094	.207*	.169	.259***	.019
	(.059)	(.124)	(.099)	(.202)	(.073)	(.147)
Higher tertiary	.103*	-.021	.120	.201	.087	-.202
	(.052)	(.109)	(.078)	(.158)	(.072)	(.142)
Employed (ref.)						
Unemployed	-.130	-.044	-.178	-.119	-.064	.039
	(.076)	(.184)	(.106)	(.233)	(.110)	(.278)
Retired	.053	.027	.031	.053	.063	.003
	(.053)	(.101)	(.079)	(.146)	(.071)	(.132)
Student	-.018	-.160	-.018	-.077	-.025	-.272
	(.094)	(.206)	(.141)	(.297)	(.126)	(.279)
Inactive	-.009	-.023	-.025	.200	-.004	-.100
	(.053)	(.118)	(.112)	(.258)	(.062)	(.130)

Table A.2.6: Estimated regression coefficients for all respondents and by sex following weighted and unweighted schemes (continued)

	All		Men		Women	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Smoking	-.014	-.041	-.047	.036	.007	-.094
	(.031)	(.070)	(.048)	(.106)	(.040)	(.086)
Drinking	-.040***	-.026	-.034**	-.018	-.044***	-.035
	(.008)	(.019)	(.013)	(.026)	(.011)	(.025)
General health	-.081***	-.037	-.065**	.055	-.090***	.017
	(.016)	(.035)	(.025)	(.053)	(.022)	(.046)
Depressive sympt	.002	-.003	-.006	-.010	.008	.002
	(.003)	(.007)	(.005)	(.010)	(.004)	(.008)
Intercept	-2.584***	.416	-2.095***	.243	-2.929***	.734
	(.169)	(.375)	(.255)	(.521)	(.228)	(.497)
R-squared	*.020*	*.019*	*.018*	*.018*	*.022*	*.021*
N	*26,382*	*26,382*	*11,569*	*11,569*	*14,813*	*14,813*

Table A.2.7: Healthy volunteer hypothesis: descriptive statistics for **males**, by genotyping status with t-statistics of differences

	All males		Genotyped		Non-genotyped			
	Mean	SD	Mean	SD	Mean	SD	*t*	*p*
Unweighted								
BMI	26.71	4.666	26.85	4.424	26.67	4.738	-1.72	.09
High blood pressure	.211	.408	.239	.427	.202	.401	-4.20	.00
Diabetes	.069	.254	.069	.254	.069	.254	.00	.99
Heart condition	.076	.266	.070	.255	.078	.269	1.46	.14
Respiratory diseases	.175	.380	.170	.375	.177	.382	.87	.38
Cancer	.036	.185	.042	.201	.034	.180	-2.08	.04
Clinical depression	.052	.223	.053	.224	.052	.222	-.23	.82
Weighted								
BMI	26.56	4.611	26.63	4.588	26.36	4.676	-1.24	.21
High blood pressure	.202	.401	.206	.405	.188	.391	-1.23	.22
Diabetes	.062	.242	.059	.236	.072	.259	1.35	.18
Heart condition	.064	.244	.062	.241	.069	.253	.70	.48
Respiratory diseases	.175	.380	.175	.380	.175	.380	.01	.99
Cancer	.031	.174	.032	.177	.027	.163	-.87	.38
Clinical depression	.052	.223	.053	.224	.050	.219	-.28	.78
No. of participants	*9,373*		*2,712*		*3,667*			

Note. Degrees of freedom for unweighted t-tests = 9,371; design df for weighted tests = 1,065

Table A.2.8: Healthy volunteer hypothesis: descriptive statistics for **females**, by genotyping status with t-statistics of differences

	All females		Genotyped		Non-genotyped			
	Mean	SD	Mean	SD	Mean	SD	*t*	*p*
Unweighted								
BMI	26.08	5.287	26.23	5.123	26.03	5.339	-1.97	.05
High blood pressure	.191	.393	.207	.405	.187	.390	-2.64	.01
Diabetes	.046	.209	.050	.218	.044	.206	-1.49	.14
Heart condition	.039	.193	.037	.188	.039	.195	.79	.43
Respiratory diseases	.173	.378	.163	.369	.177	.381	1.87	.06
Cancer	.046	.209	.052	.221	.044	.205	-1.88	.06
Clinical depression	.092	.289	.096	.295	.091	.287	-.98	.33
Weighted								
BMI	26.03	5.231	26.10	5.158	25.78	5.469	-1.46	.15
High blood pressure	.190	.392	.194	.395	.177	.382	-1.21	.23
Diabetes	.050	.219	.050	.217	.053	.223	.39	.70
Heart condition	.038	.190	.037	.189	.040	.196	.36	.72
Respiratory diseases	.171	.376	.169	.375	.176	.381	.42	.68
Cancer	.045	.208	.047	.212	.038	.190	-1.48	.14
Clinical depression	.098	.297	.097	.296	.100	.300	.19	.85
No. of participants	*14,524*		*3,583*		*10,941*			

Note. Degrees of freedom for unweighted t-tests = 14,522; design df for weighted tests = 1,130

Table A.2.9: Estimated regression coefficients (robust SE) for healthy volunteer models, for all respondents and by sex following weighted and unweighted schemes

	All		Men		Women	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
BMI	-.003	.006	-.004	.005	-.003	.006
	(.003)	(.007)	(.005)	(.011)	(.004)	(.009)
High blood pressure	.004	.127	.046	.128	-.026	0.119
	(.039)	(.079)	(.057)	(.111)	(.053)	(.106)
Diabetes	-.048	-.196	-.154	-.268	.067	-.116
	(.065)	(.123)	(.090)	(.165)	(.094)	(.169)
Heart condition	-.283***	-.082	-.361***	-.097	-.196	-.063
	(.068)	(.145)	(.090)	(.166)	(.106)	(.224)
Respiratory diseases	-.026	-.000	.011	.049	-.054	-.044
	(.040)	(.085)	(.059)	(.127)	(.053)	(.112)
Cancer	.009	.208	.030	.167	-.008	.237
	(.071)	(.137)	(.114)	(.206)	(.090)	(.180)
Clinical depression	.045	-.067	.010	-.031	.056	-.086
	(.055)	(.113)	(.099)	(.194)	(.067)	(.141)
Covariates						
Sex: female	.058	.143*	-	-	-	-
	(.030)	(.060)	-	-	-	-
Age	.069***	.037***	.057***	.039**	.078***	.035*
	(.005)	(.010)	(.008)	(.014)	(.007)	(.014)
Age-sq	-.000***	-.000***	-.000***	-.000**	-.000***	-.000*
	.000	.000	.000	.000	.000	.000
No qualifications (ref.)						
Lower secondary	.264***	.200*	.276***	.288*	.247***	.125
	(.044)	(.088)	(.068)	(.137)	(.059)	(.117)
Higher secondary	.326***	.103	.392***	.172	.251**	.040
	(.056)	(.120)	(.082)	(.168)	(.079)	(0.165)
Lower tertiary	.337***	.132	.323**	.223	.330***	.057
	(.058)	(.122)	(.098)	(.201)	(.072)	(0.143)
Higher tertiary	.242***	-.008	.292***	.134	.185**	-.146
	(.050)	(.104)	(.074)	(.146)	(.069)	(.137)
Intercept	-3.377***	-.024	-3.178***	-.168	-3.467***	.275
	(.144)	(.303)	(.215)	(.428)	(.191)	(.393)
R-squared	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*
N	*26,292*	*26,292*	*11,768*	*11,768*	*14,524*	*14,524*

A.3 Mortality selection

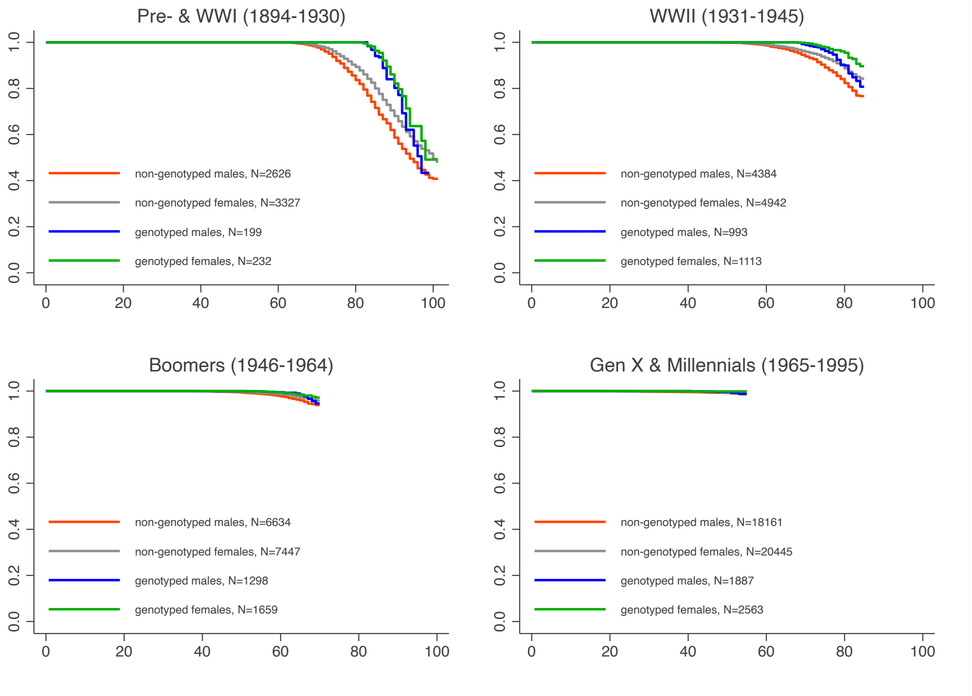


Figure A.3.1: Kaplan-Meier survival curves for genotyped and non-genotyped UKHLS respondents, by birth cohorts and by sex.

Table A.3.1: Estimates from separate Cox models (equation above) for each birth cohort. Age at first interview is mean-centered. Generation X and Millennials cohorts are merged due to small number of deaths in these cohorts

	Coef.	Std. Err.	z	P	[95% Conf. Interval]	
WWI cohort (1894-1930), N=6081						
Genotyped	-0.402	0.1	-4.04	0	-0.598	-0.207
Female	-0.35	0.039	-8.92	0	-0.427	-0.273
Age	0.091	0.003	32.55	0	0.086	0.097
WWII cohort (1931-1945), N=11432						
Genotyped	-0.584	0.087	-6.71	0	-0.755	-0.414
Female	-0.48	0.056	-8.64	0	-0.589	-0.371
Age	0.096	0.006	15.04	0	0.084	0.109
Boomers (1946-1964), N=17038						
Genotyped	-0.466	0.136	-3.42	0.001	-0.733	-0.199
Female	-0.454	0.089	-5.11	0	-0.628	-0.28
Age	0.081	0.01	8.01	0	0.062	0.101
Generation X & Millennials cohorts (1965-1995), N=43155						
Genotyped	-0.71	0.298	-2.38	0.017	-1.294	-0.126
Female	-0.562	0.131	-4.3	0	-0.818	-0.306
Age	0.058	0.007	8.84	0	0.045	0.071

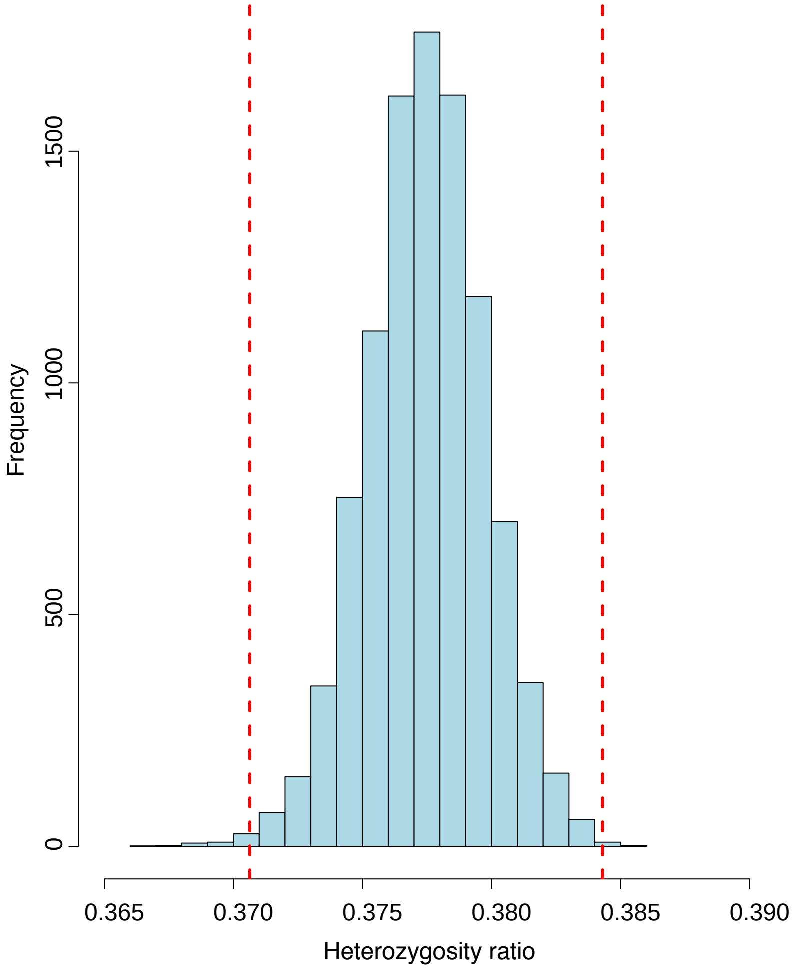
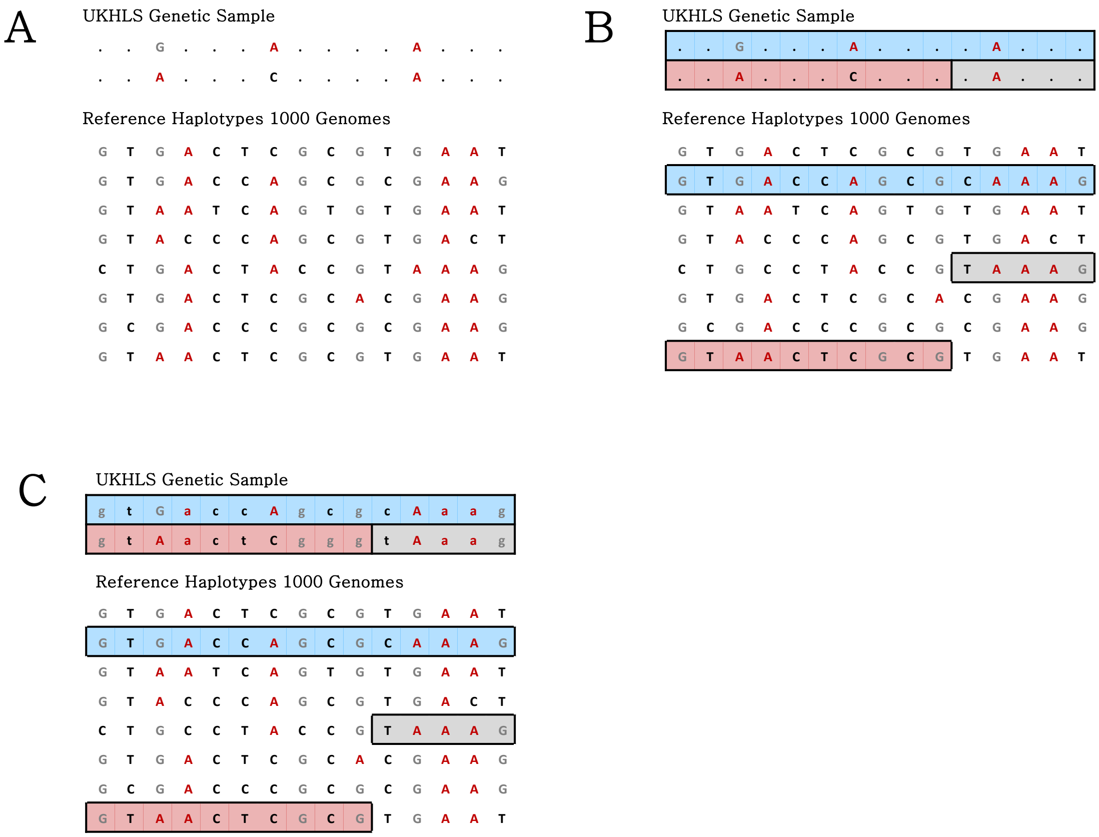


Figure A.4.1: Histogram of heterozygosity statistics

A.4 Polygenic prediction of depression



UKHLS Genetic Sample

. . G . . . A . . . . A . . .
. . A . . . C . . . . A . . .

Reference Haplotypes 1000 Genomes

G T G A C T C G C G T G A A T
G T G A C C A G C G C G A A G
G T A A T C A G T G T G A A T
G T A C C C A G C G T G A C T
C T G A C T A C C G T A A A G
G T G A C T C G C A C G A A G
G C G A C C C G C G C G A A G
G T A A C T C G C G T G A A T

UKHLS Genetic Sample

. . G . . . A . . . . A . . .
. . A . . . C . . . . A . . .

Reference Haplotypes 1000 Genomes

G T G A C T C G C G T G A A T
G T G A C C A G C G C A A A G
G T A A T C A G T G T G A A T
G T A C C C A G C G T G A C T
C T G C C T A C C G T A A A G
G T G A C T C G C A C G A A G
G C G A C C C G C G C G A A G
G T A A C T C G C G T G A A T

UKHLS Genetic Sample

g t G a c c A g c g c A a a g
g t A a c t C g g g t A a a g

Reference Haplotypes 1000 Genomes

G T G A C T C G C G T G A A T
G T G A C C A G C G C A A A G
G T A A T C A G T G T G A A T
G T A C C C A G C G T G A C T
C T G C C T A C C G T A A A G
G T G A C T C G C A C G A A G
G C G A C C C G C G C G A A G
G T A A C T C G C G T G A A T

Figure A.4.2: Initial Steps of UKHLS Genetic Sample Genotype Imputation.
Note: The diagram based on conceptual scheme demonstrated in Li et al., 2009, p. 17. Panel A demonstrates initial state of data – missing markers in the study sample and a respectful region from reference panel containing detailed information on SNPs. Panel B demonstrates identification of shared region of a chromosome between study sample and reference panel. Panel C demonstrates filling in of the series unobserved in a study sample from a reference panel as a baseline information.

Table A.4.1: Imputation summary.

Step	Tool
Reference panel	1000 Genomes Phase 3
Data preparation	PLINK 2.0; VCFtools
Pre-phasing (estimation of haplotypes)	MaCH
Phasing	Eagle2
Phasing server	University of Michigan
Imputation	minimac3
Imputation server	University of Michigan
Initial #SNPs	248,587
Imputed #SNPs (post QCs)	23,439,105

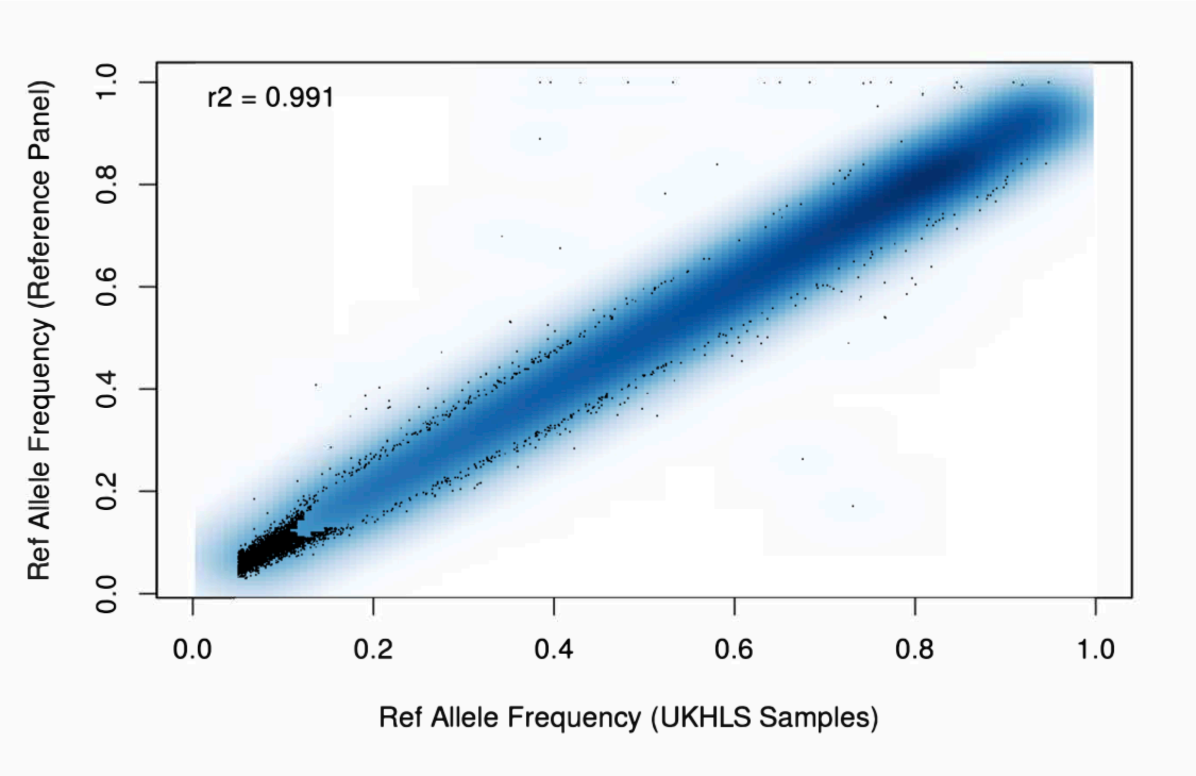


Figure A.4.3: Imputation QC-report. Allele-frequency correlation.
UKHLS Samples vs. Reference Panel. The plot shows the densities of frequencies falling into each part (excluding chromosome X). The first 5000 points from areas of lowest regional densities plotted.

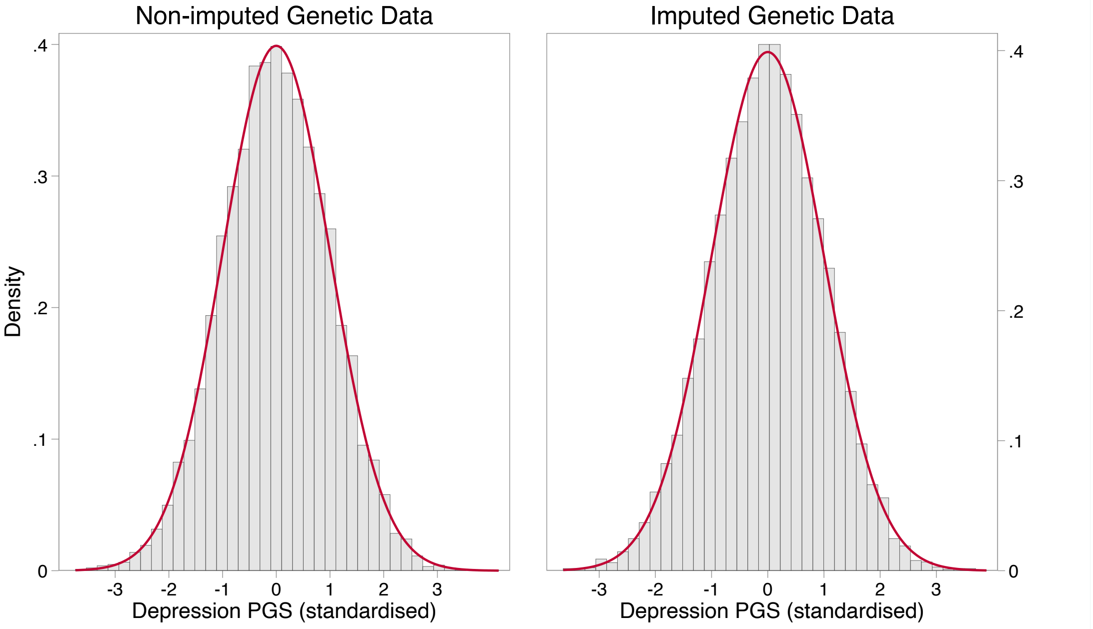


Figure A.4.4: Distribution of PGS for Depression based on Howard et al. [2019] GWAS.

Table A.4.2: Depressive symptoms score questionnaires

Self-Completion GHQ Module	a. Have you recently been able to concentrate on whatever you're doing? *(1=Better than usual; 2=Same as usual; 3=Less than usual; 4=Much less than usual)*
	b. Have you recently lost much sleep over worry? *(1=Not at all; 2=No more than usual; 3=Rather more than usual; 4=Much more than usual)*
	c. Have you recently felt that you were playing a useful part in things? *(1=More than usual; 2=Same as usual; 3=Less so than usual; 4=Much less than usual)*
	d. Have you recently felt capable of making decisions about things? *(1=More so than usual; 2=Same as usual; 3=Less so than usual; 4=Much less capable)*
	e. Have you recently felt constantly under strain? *(1=Not at all; 2=No more than usual; 3=Rather more than usual; 4=Much more than usual)*
	f. Have you recently felt you couldn't overcome your difficulties? *(1=Not at all; 2=No more than usual; 3=Rather more than usual; 4=Much more than usual)*
	g. Have you recently been able to enjoy your normal day-to-day activities? *(1=More than usual; 2=Same as usual; 3=Less so than usual; 4=Much less than usual)*
	h. Have you recently been able to face up to problems? *(1=More so than usual; 2=Same as usual; 3=Less able than usual; 4=Much less able)*
	i. Have you recently been feeling unhappy or depressed? *(1=Not at all; 2=No more than usual; 3=Rather more than usual; 4=Much more than usual)*
	j. Have you recently been losing confidence in yourself? *(1=Not at all; 2=No more than usual; 3=Rather more than usual; 4=Much more than usual)*

Table A.4.3: Depressive symptoms score questionnaires (continued)

	k. Have you recently been thinking of yourself as a worthless person? *(1=Not at all; 2=No more than usual; 3=Rather more than usual; 4=Much more than usual)*
	l. Have you recently been feeling reasonably happy, all things considered? *(1=More so than usual; 2=Same as usual; 3=Less able than usual; 4=Much less able)*

Table A.4.4: Coefficients (robust SE) of Poisson models assessing non-linearity of polygenic score prediction of GHQ depressive symptoms score.

	Weighted	Unweighted
PGS	**.043** (.00)	**.045** (.00)
PGS-sq.	.003 (.00)	.003 (.00)
Sex (female)	**.118** (.01)	**.117** (.01)
Age	**.007** (.00)	**.007** (.00)
Age-sq.	**-.000** (.00)	**-.000** (.00)
Intercept	**2.10** (.04)	**2.10** (.04)
Pseudo R-sq.	*.015*	*.015*
No. of participants	*9,113*	*9,113*

Bold values are significant at 99.9% level; Genetic score is standardised;All models include the largest 20 PCs

Table A.4.5: Coefficients (Robust SE) of Logistic Models Assessing Probability of Depressive Symptoms Threshold Crossing and its Polygenic Score Prediction.

	Weighted	Unweighted
PGS	**.160** (.02)	**.156** (.03)
Sex (female)	**.403** (.04)	**.400** (.04)
Age	.006 (.00)	.007 (.00)
Age-sq.	-.000 (.00)	-.000 (.00)
Intercept	**-.95** (.21)	**-1.36** (.26)
Pseudo R-sq.	*.015*	*.015*
No. of participants	*9,113*	*9,113*

Bold values are significant at 99.9% level; Genetic score is standardised; All models include the largest 20 PCs

Table A.4.6: Coefficients (Robust SE) of Logistic Models Assessing Probability of Clinical Depression Diagnosis and its Polygenic Score Prediction.

	Weighted	Unweighted
PGS	**.307** (.01)	**.267** (.01)
Sex (female)	**.553** (.20)	**.644** (.22)
Age	.058 (.04)	.071 (.04)
Age-sq.	**-.001** (.00)	**-.001** (.00)
Intercept	**-5.87** (.88)	**-6.28** (.99)
Pseudo R-sq.	*.034*	*.035*
Incremental R-sq.	*1.5%*	*1.5%*
No. of participants	*8,676*	*8,676*

Bold values are significant at 99.9% level; Genetic score is standardised; All models include the largest 20 PCs

Supplementary material - Chapter 3

Table B.0.1: Non-response statistics of analytic sample

	All	
	N part.	N obs.
Initial genetic sample	9944	116822
QC'ed & depression PGS	9237	104009
sex	9237	104009
Age	9237	104009
GHQ depressive symptoms	9113	81246
% reduction	8.4	30.5

Table B.0.2: Coefficients and standard errors of Poisson multilevel models assessing the moderation of birth cohorts and recessions on the genetic association with depressive symptoms correcting for differential probability to be genotyped

	Model 1		Model 2		Model 3	
	Beta	Std. Err.	Beta	Std. Err.	Beta	Std. Err.
Cohorts (World Wars - ref.)						
Boomers	.057***	.012	.056***	.012	.056***	.012
Generation X	.036*	.018	.034*	.016	.037*	.018
Millennials	.069**	.026	.064*	.023	.070*	.027
PGS depression			.028***	.007	.031***	.007
PGS × Cohort (World Wars - ref.)						
Boomers			.026**	.009	.023*	.009
Generation X			.010	.009	.010	.010
Millennials			.012	.015	.012	.016
Recession					.001	.006
Recession × Cohort (World Wars - ref.)						
Boomers					.010	.009
Generation X					.006	.011
Millennials					-.012	.020
Recession × PGS					-.009	.006
Recession × PGS × Cohort (World Wars - ref.)						
Boomers					.012	.008
Generation X					.002	.010
Millennials					.001	.022
Female	.117***	.007	.117***	.007	.117***	.007
Age	.037***	.005	.037***	.005	.036***	.005
Age^2	-.001***	.000	-.001***	.000	-.001***	.000
Age^3	.000***	.000	.000***	.000	.000***	.000
Random-Effect Variance						
$\sigma^2 u$	.091	.002	.089	.002	.089	.002
AIC	343391.8		343261.0		343263.0	
BIC	343661.5		343568.0		343644.4	
Sample Size						
No. of participants	9,113		9,113		9,113	
No. of observations	81,246		81,246		81,246	

+p<0.1, *p<0.05, **p<0.01, ***p<0.001

Genetic score is standardised; Models include first 20 PCs as covariates

Supplementary material - Chapter 4

Table C.0.1: Non-response statistics of analytic sample [i.e. in working age]

	All		Men		Women	
	N part.	N obs.	N part.	N obs.	N part.	N obs.
Initial genetic sample	8166	91313	3541	39784	4625	51529
QC'ed & depression PGS	7490	81197	3239	35257	4251	45940
sex	7490	81197	3239	35257	4251	45940
Age	7490	81197	3239	35257	4251	45940
Labour force status	7211	58891	3141	26031	4070	32860
% reduction	11.7	35.5	11.3	34.6	12.0	36.2

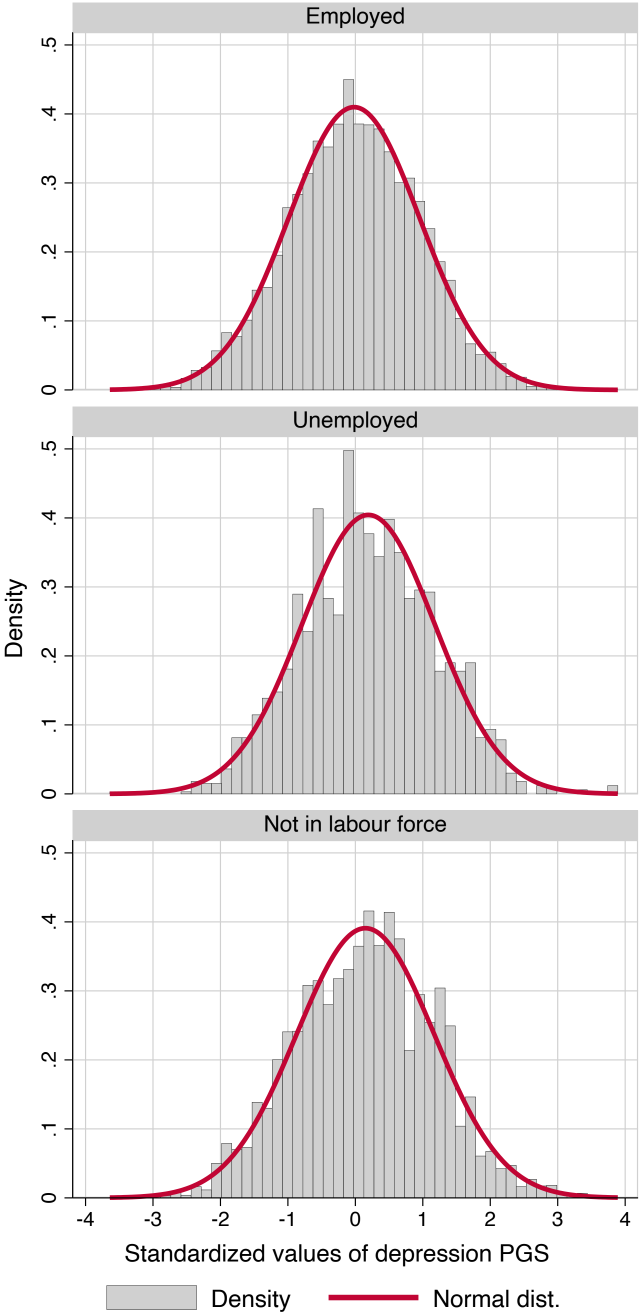


Figure C.0.1: Distribution of depression polygenic score by worklessness status in the UKHLS genetic sample of working age participants

Table C.0.2: Coefficients and standard errors of **weighted** multinomial multilevel model assessing the association between depression polygenic score and worklessness status

	Parameters	β	Std. Err.
$\frac{P(Y_i=unemployed)}{P(Y_i=employed)}$	PGS depression	0.45***	0.05
	PGS depression^2	0.04	0.04
	Female	0.52***	0.11
	Age	-0.25***	0.03
	Age^2	0.00***	0.00
	Intercept	-0.41	0.61
$\frac{P(Y_i=not\ in\ lab.\ frc.)}{P(Y_i=employed)}$	PGS depression	0.56***	0.08
	PGS depression^2	0.12*	0.05
	Female	3.64***	0.16
	Age	-0.24***	0.04
	Age^2	0.00***	0.00
	Intercept	-4.65***	0.82
Random-Effect Variance			
	σ^2_{u1}	19.25***	1.60
	σ^2_{u2}	36.57***	2.60
	AIC	29,367.36	
	BIC	29,861.25	
Sample Size			
No. of participants		7,211	
No. of observations		58,891	

*+p<0.1, *p<0.05, **p<0.01, ***p<0.001*

Genetic score is standardised; Model includes first 20PCs as additional covariates.

Table C.0.3: Coefficients and standard errors of **weighted** multinomial multilevel model assessing the association between depression polygenic score and worklessness status, by sex

		Men		Women	
	Parameters	β	Std. Err.	β	Std. Err.
$\frac{P(Y_i\text{=}\mathit{unemployed})}{P(Y_i\text{=}\mathit{employed})}$	PGS depression	0.37***	0.06	0.52***	0.09
	PGS depression^2	0.03	0.05	0.03	0.07
	Age	-0.18***	0.04	-0.29***	0.04
	Age^2	0.00***	0.00	0.00***	0.00
	Intercept	-2.46**	0.93	1.14	0.82
$\frac{P(Y_i\text{=}\mathit{not\ in\ lab.\ frc.})}{P(Y_i\text{=}\mathit{employed})}$	PGS depression	0.90***	0.16	0.49***	0.09
	PGS depression^2	0.02	0.09	0.14*	0.07
	Age	-0.11	0.09	-0.26***	0.04
	Age^2	0.00*	0.00	0.00***	0.00
	Intercept	-11.82***	2.24	0.45	0.90
Random-Effect Variance					
	σ^2_{u1}	23.66***	3.18	17.21***	1.74
	σ^2_{u2}	68.87***	7.30	27.13***	2.16
	AIC	9,766.27		19,519.61	
	BIC	10,198.89		19,964.65	
Sample Size					
No. of participants		3,141		4,070	
No. of observations		26,031		32,860	

*+p<0.1, * p<0.05, ** p<0.01, *** p<0.001*

Genetic score is standardised; Model includes first 20PCs as additional covariates.

Table C.0.4: Coefficients and standard errors of **weighted** multinomial multilevel model assessing the association between depression polygenic score and worklessness with detailed profile of economic inactivity

	Parameters	β	Std. Err.
$\frac{P(Y_i=unemployed)}{P(Y_i=employed)}$	PGS depression	0.41***	0.06
	PGS depression^2	0.03	0.04
	Female	0.43***	0.11
	Age	-0.24***	0.03
	Age^2	0.00***	0.00
	Intercept	-0.81	0.62
$\frac{P(Y_i=family\ care)}{P(Y_i=employed)}$	PGS depression	0.30***	0.08
	PGS depression^2	0.08	0.06
	Female	5.45***	0.22
	Age	-0.27***	0.04
	Age^2	0.00***	0.00
	Intercept	-4.86***	0.82
$\frac{P(Y_i=sick\ /\ disable)}{P(Y_i=employed)}$	PGS depression	1.04***	0.12
	PGS depression^2	0.08	0.08
	Female	1.28***	0.25
	Age	0.13	0.07
	Age^2	0.00	0.00
	Intercept	-19.63***	2.15
Random-Effect Variance			
	σ^2_{u1}	18.46***	1.42
	σ^2_{u2}	28.14***	2.10
	σ^2_{u3}	64.14***	5.92
	AIC	32,096.78	
	BIC	32,851.09	
Sample Size			
No. of participants		7,211	
No. of observations		58,891	

*+p<0.1, *p<0.05, **p<0.01, ***p<0.001*

Genetic score is standardised; Model includes first 20PCs as additional covariates.

Table C.0.5: Coefficients and standard errors of **weighted** multinomial multilevel model assessing the association between depression polygenic score and worklessness with detailed profile of economic inactivity, by sex

		Men		Women	
	Parameters	β	Std. Err.	β	Std. Err.
$\frac{P(Y_i{=}unemployed)}{P(Y_i{=}employed)}$	PGS depression	0.38***	0.09	0.48***	0.09
	PGS depression^2	0.03	0.06	0.03	0.06
	Age	-0.29***	0.02	-0.28***	0.04
	Age^2	0.00***	0.00	0.00***	0.00
	Intercept	0.32	0.72	0.54	0.84
$\frac{P(Y_i{=}family\ care)}{P(Y_i{=}employed)}$	PGS depression	0.25	0.11	0.32**	0.09
	PGS depression^2	0.04	0.09	0.09	0.07
	Age	-0.31***	0.06	-0.30***	0.04
	Age^2	0.00***	0.00	0.00***	0.00
	Intercept	-32.56***	7.23	1.32	0.90
$\frac{P(Y_i{=}sick\ /\ disable)}{P(Y_i{=}employed)}$	PGS depression	1.92***	0.34	1.04***	0.16
	PGS depression^2	-0.15	0.23	0.12	0.11
	Age	-0.26*	0.15	0.16	0.09
	Age^2	0.00***	0.00	0.00	0.00
	Intercept	-15.44***	3.10	-18.12***	2.66
Random-Effect Variance					
	σ^2_{u1}	24.08***	2.75	16.91***	1.70
	σ^2_{u2}	30.00***	3.73	25.11***	2.05
	σ^2_{u3}	65.78***	8.02	51.00***	5.79
	AIC	12,632.11		22,276.89	
	BIC	13,320.60		22,600.76	
Sample Size					
No. of participants		3,141		4,070	
No. of observations		26,031		32,860	

*+p<0.1, * p<0.05, ** p<0.01, *** p<0.001*

Genetic score is standardised; Model includes first 20PCs as additional covariates.

Table C.0.6: Coefficients and standard errors of multilevel logistic regressions assessing the differences between unemployed and not in labour force groups. (Weighted coefficients and robust SE)

	Parameters	β	Std. Err.
$\frac{P(Y_i = \textit{not in lab. frc.})}{P(Y_i = \textit{unemployed})}$	PGS depression	-0.03	0.08
		-0.03	**0.07**
	PGS depression^2	0.07	0.05
		0.08	**0.05**
	Female	3.65^{***}	0.20
		$\mathbf{3.57^{***}}$	**0.21**
	Age	0.07^{*}	0.05
		$\mathbf{0.09^{**}}$	**0.04**
	Age^2	0.00	0.00
		0.00	**0.00**
	Intercept	-4.01^{***}	0.62
		$\mathbf{-4.61^{***}}$	**0.77**
Random-Effect Variance			
	σ^2_{u1}	9.09^{***}	0.86
		$\mathbf{7.82^{***}}$	**0.84**
	AIC	7296.77	
		6495.67	
	BIC	7488.90	
		6687.70	
Sample Size			
No. of participants		2,912	
No. of observations		9,098	

$+p<0.1$, $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

Genetic score is standardised; Model includes first 20PCs as additional covariates.

Table C.0.7: Coefficients and standard errors of multilevel logistic regressions assessing the differences between unemployed and not in labour force groups, by sex. (Weighted coefficients and robust SE)

		Men		Women	
	Parameters	β	Std. Err.	β	Std. Err.
$\frac{P(Y_i=not\ in\ lab.\ frc.)}{P(Y_i=unemployed)}$	PGS depression	0.44**	0.17	-0.15	0.09
		0.41**	**0.16**	**-0.15**	**0.09**
	PGS depression^2	-0.04	0.11	0.10	0.06
		-0.02	**0.10**	**0.11**	**0.06**
	Age	0.05	0.06	0.07*	0.03
		0.11	**0.08**	**0.08***	**0.04**
	Age^2	0.00	0.00	0.00	0.00
		0.00	**0.00**	**0.00**	**0.00**
	Intercept	-5.32***	1.25	-0.16	0.65
		-6.54***	**1.53**	**-0.55**	**0.81**
Random-Effect Variance					
	σ^2_{u1}	14.21***	2.51	6.58***	0.77
		11.64***	**2.44**	**5.52*****	**0.71**
	AIC	2346.96		4928.27	
		2118.46		**4356.74**	
	BIC	2498.07		5105.05	
		2269.51		**4533.40**	
Sample Size					
No. of participants		917		1,995	
No. of observations		2,470		6,628	

*+p<0.1, * p<0.05, ** p<0.01, *** p<0.001*

Genetic score is standardised; Model includes first 20PCs as additional covariates.

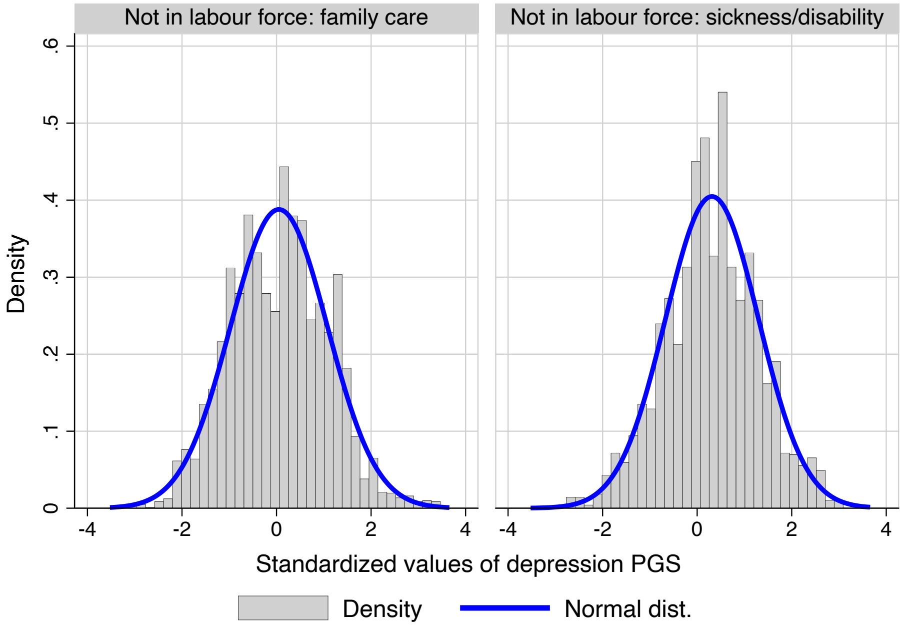


Figure C.0.2: Distribution of depression polygenic score by detailed profile of economic inactivity in the UKHLS genetic sample of working age participants

Supplementary material - Chapter 5

D.1 Code for simulation analyses and figures

D.1.1 Packages needed

```
libraries("dplyr", "broom", "purrr", "mvtnorm", "ggplot2", "tidyr")
```

D.1.2 Additive model

```
n <- 1000000

sim_data <- function(r, n = 1000000) {

epsilon <- rmvnorm(n, c(0, 0),

  matrix(c(1, 0.2, 0.2, 1), nrow = 2))

  G <- rnorm(n, 0, 1) #G = Polygenic score

  E <- r*G + epsilon[, 1] #E = covariate

  Y <- 0.6 * G + 0.6 * E + epsilon[, 2]

  tibble(PGScore = G, E = E, Outcome = Y)}


run_model <- function(df) {lm(Outcome ~ PGScore + E, df)}

get_rsquared <- function(ml) {glance(ml)$r.squared[[1]]}

get_pgscore_beta <- function(ml) {tidy(ml)[[2, 2]]}

get_e_beta <- function(ml) {tidy(ml)[[3, 2]]}
```

```
r_range <- seq(0, 0.5, 0.05)
dat <- map(r_range, sim_data) %>% map(run_model)
rsqs <- map(dat, get_rsquared) %>% unlist()
pgs_betas <- map(dat, get_pgscore_beta) %>% unlist()
e_betas <- map(dat, get_e_beta) %>% unlist()

# R-square | Fifure 1C
plot_rsq_inflation <- tibble(r = r_range,
         'R-Squared' = rsqs) %>%
         gather('key', 'value', 'R-Squared')
# The true share of the variance in Y explained by G and E
# in the context of this simulation is 0.461178,
plot_rsq_inflation <- transform(
                      plot_rsq_inflation, base = 0.461178)
plot_rsq_inflation <- transform(
                      plot_rsq_inflation,
                      inflation = value*100/base-100)

plt <- ggplot(plot_rsq_inflation, aes(x=r, y=inflation,
  group=key))+ geom_line(size = 1, color="violetred4")+
  geom_point(size = 4, color="violetred4", shape=15) +
  geom_hline(yintercept = 0, linetype = "longdash", size = 1,
  color="violetred4") + scale_y_continuous(limits=c(0, 40)) +
  cowplot::theme_cowplot()
plt  + theme(legend.position = "bottom") +
  labs(title="R-square change depending on rGE values",
  x ="Gene-environment correlation (rGE)", y = "% change") +
  theme(panel.background = element_rect(fill = "white",
        colour = "white", size = 0.45, linetype = "solid"),
        panel.grid.major = element_line(size = 0.45,
        linetype = 'solid', colour = "gray92"),
        panel.grid.minor = element_line(size = 0.25,
        linetype = 'solid', colour = "gray92"))

# Coefficients | Figure 1B
plot_ebeta_inflation <- tibble(r = r_range,
  'E Beta' = e_betas) %>% gather('key', 'value', 'E Beta')
plot_ebeta_inflation <- transform(
  plot_ebeta_inflation, base = 0.6)
plot_ebeta_inflation <- transform(
  plot_ebeta_inflation, inflation = value*100/base-100)
plot_ebetg_inflation <- tibble(r = r_range,
  'PGS Beta' = pgs_betas) %>%
  gather('key', 'value', 'PGS Beta')
plot_ebetg_inflation <- transform(
  plot_ebetg_inflation, base = 0.6)
plot_ebetg_inflation <- transform(
  plot_ebetg_inflation, inflation = value*100/base-100)

combinebetas = rbind(plot_ebeta_inflation, plot_ebetg_inflation)
plt <- ggplot(combinebetas, aes(x=r, y=inflation, group=key))+
  geom_line(size = 1, aes(color=key))+
  geom_point(size = 4, aes(color=key, shape=key)) +
  geom_hline(yintercept = 0, linetype = "longdash", size = 1,
  color="gray54") + scale_y_continuous(limits=c(-20, 40)) +
  cowplot::theme_cowplot() + theme(legend.position = "bottom") +
  labs(title="Coefficients inflation depending on rGE values",
  x ="Gene-environment correlation (rGE)", y = "% inflation") +
  scale_colour_discrete(name  ="", breaks=c("E Beta",
  "PGS Beta"), labels=c("Environment", "PGScore")) +
  scale_shape_discrete(name  ="", breaks=c("E Beta",
  "PGS Beta"), labels=c("Environment", "PGScore")) +
  theme(panel.background = element_rect(fill = "white",
  colour = "white", size = 0.45, linetype = "solid"),
  panel.grid.major = element_line(size = 0.45,
  linetype = 'solid', colour = "gray92"),
  panel.grid.minor = element_line(size = 0.25,
  linetype = 'solid', colour = "gray92"))
plt
```

D.1.3 Gene-environment model

```
n <- 1000000
sim_data <- function(r, n = 1000000) {
  epsilon <- rmvnorm(n, c(0, 0, 0),
  matrix(c(1, 0.2, 0.2, 0.2, 1, 0.2, 0.2, 0.2, 1), nrow = 3))
  G <- rnorm(n, 0, 1) #G = polygenic score
  E <- r*G + epsilon[, 1] #E = exposure
  GxE <- E*G # Interaction term
  Y <- 0.6 * G + 0.6 * E + 0.1 * GxE + epsilon[, 2]
  tibble(PGScore = G, E = E, GxE = GxE, Outcome = Y)}

run_model <- function(df) {lm(Outcome ~ PGScore + E + GxE, df)}
```

```r
get_rsquared <- function(ml) {glance(ml)$r.squared[[1]]}
get_pgscore_beta <- function(ml) {tidy(ml)[[2, 2]]}
get_e_beta <- function(ml) {tidy(ml)[[3, 2]]}
get_gxe_beta <- function(ml) {tidy(ml)[[4, 2]]}
r_range <- seq(0, 0.5, 0.05)
dat <- map(r_range, sim_data) %>% map(run_model)
rsqs <- map(dat, get_rsquared) %>% unlist()
pgs_betas <- map(dat, get_pgscore_beta) %>% unlist()
e_betas <- map(dat, get_e_beta) %>% unlist()
gxe_betas <- map(dat, get_gxe_beta) %>% unlist()

# R-square | Fifure 2C
plot_rsq_inflation <- tibble(r = r_range,
  'R-Squared' = rsqs) %>% gather('key', 'value', 'R-Squared')
# The true share of the variance in Y explained by G and E
# in the context of this simulation is 0.4627.
plot_rsq_inflation <- transform(
  plot_rsq_inflation, base = 0.4627)
plot_rsq_inflation <- transform(
  plot_rsq_inflation, inflation = value*100/base-100)
plt <- ggplot(plot_rsq_inflation, aes(x=r, y=inflation,
  group=key))+ geom_line(size = 1, color="violetred4")+
  geom_point(size = 6, color="violetred4", shape=18) +
  geom_hline(yintercept = 0, linetype = "longdash", size = 1,
  color="violetred4") + scale_y_continuous(limits=c(0, 35)) +
  cowplot::theme_cowplot()
plt  + theme(legend.position = "bottom") +
  labs(title="R-square inflation depending on rGE values",
  x ="Gene-environment correlation (rGE)", y = "% inflation") +
  theme( panel.background = element_rect(fill = "white",
  colour = "white", size = 0.45, linetype = "solid"),
  panel.grid.major = element_line(size = 0.45,
  linetype = 'solid', colour = "gray92"),
  panel.grid.minor = element_line(size = 0.25,
  linetype = 'solid', colour = "gray92"))

# Coefficients | Figure 2B
plot_ebeta_inflation <- tibble(r = r_range,
  'E Beta' = e_betas) %>% gather('key', 'value', 'E Beta')
plot_ebeta_inflation <- transform(
  plot_ebeta_inflation, base = 0.6)
plot_ebeta_inflation <- transform(
  plot_ebeta_inflation, inflation = value*100/base-100)
plot_ebeta_inflation <- transform(
  plot_ebeta_inflation, case = 'beta E = beta G')
plot_ebetg_inflation <- tibble(r = r_range,
  'PGS Beta' = pgs_betas) %>% gather('key', 'value', 'PGS Beta')
plot_ebetg_inflation <- transform(
  plot_ebetg_inflation, base = 0.6)
plot_ebetg_inflation <- transform(
  plot_ebetg_inflation, inflation = value*100/base-100)
plot_ebetg_inflation <- transform(
  plot_ebetg_inflation, case = 'beta E = beta G')
plot_gxe_inflation <- tibble(r = r_range,
  'GxE Beta' = gxe_betas) %>% gather('key', 'value', 'GxE Beta')
plot_gxe_inflation <- transform(
  plot_gxe_inflation, base = 0.1)
plot_gxe_inflation <- transform(
  plot_gxe_inflation, inflation = value*100/base-100)
plot_gxe_inflation <- transform(
  plot_gxe_inflation, case = 'beta E = beta G')

combinebetas = rbind(plot_ebeta_inflation, plot_ebetg_inflation,
  plot_gxe_inflation)
plt <- ggplot(combinebetas, aes(x=r, y=inflation, group=key))+
  geom_line(size = 1, aes(color=key))+
  geom_point(size = 4, aes(color=key, shape=key)) +
  geom_hline(yintercept = 0, linetype = "longdash", size = 1,
  color="gray54") + scale_y_continuous(limits=c(-20, 40)) +
  cowplot::theme_cowplot() + theme(legend.position = "bottom") +
  labs(title="Coefficients inflation depending on rGE values",
  x ="Gene-environment correlation (rGE)", y = "% inflation") +
  scale_colour_discrete(name  ="",
       breaks=c("E Beta", "PGS Beta", "GxE Beta"),
       labels=c("Environment", "PGScore", "GxE interaction")) +
  scale_shape_discrete(name  ="",
       breaks=c("E Beta", "PGS Beta", "GxE Beta"),
       labels=c("Environment", "PGScore", "GxE interaction")) +
  theme(panel.background = element_rect(fill = "white",
       colour = "white", size = 0.45, linetype = "solid"),
  panel.grid.major = element_line(size = 0.45,
          linetype = 'solid', colour = "gray92"),
   panel.grid.minor = element_line(size = 0.25,
          linetype = 'solid', colour = "gray92"))
plt
```