PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

Genetic Prediction and Adverse Selection*

Eduardo Azevedo† Jonathan Beauchamp‡ Richard Karlsson Linnér§

May 8, 2024

Abstract

The predictive power of genetic data has been increasing rapidly and is reaching levels of clinical utility for many diseases. Meanwhile, many jurisdictions have banned insurers from utilizing genetic information. This has led to concerns that further improvements in the accuracy and availability of genetic prediction technology will lead to adverse selection and failure in insurance markets. We make three contributions to this debate. First, we develop a method to measure the amount of selection that would take place if consumers had widespread access to current-day genetic information. Second, we extend the method to estimate the amount of selection that will occur with future genetic prediction technology. Third, using the rich UK Biobank dataset with nearly 500,000 genotyped individuals, we predict the risk of eight critical illnesses and apply the method to measure selection in the critical illness insurance market with both current and future genetic prediction technology. Our estimates suggest that expected improvements in genetic prediction will lead to unsustainably high levels of selection and thus threaten the viability of the market. We discuss policy implications.

Keywords: adverse selection, asymmetric information, complex disease, insurance, polygenic indexes, genomic risk prediction, genetic testing

JEL classifications: D82, G22, G28, I13, I18

*For helpful comments and suggestions, we thank conference and seminar participants at Stanford University, Uppsala University, Washington University in St-Louis, the Yale School of Management, the American Enterprise Institute, The Advances in Social Genomics Conference, and the Integrating Genetics in the Social Sciences conference. For outstanding research assistance, we are grateful to Gareth Markel. This research has been conducted using the UK Biobank Resource under application number(s) 99086, 217716 (approved, pending), and (for preliminary analyses) 11425. This research was supported by the Dutch Research Council (NWO; project number Vl.Veni.221E.080) and by a Faculty Research and Development Award (FRDA) at George Mason University. For extended acknowledgments, see the Online Appendix.

†The Wharton School, University of Pennsylvania. Email: eazevedo@wharton.upenn.edu

‡Interdisciplinary Center for Economic Science and Department of Economics, George Mason University. Email: jonathan.pierre.beauchamp@gmail.com

§Department of Economics, Leiden University: Email: r.karlsson.linner@law.leidenuniv.nl
School of Business and Economics, Vrije Universiteit Amsterdam: Email: r.karlssonlinner@vu.nl

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

1 Introduction

Over the past two decades, technological and scientific advances have led to substantial improvements in the predictive power of genetic data (Abdellaoui et al., 2023). Scientists can now construct **"polygenic indexes"** (**PGIs**) that partially predict an individual's risk of various diseases from that individual's genetic data. The predictive accuracy of these PGIs has reached levels of clinical and personal utility for many diseases (Khera et al., 2018) and is expected to continue increasing and approach its theoretical upper limits (Daetwyler et al., 2008) in the coming decades. Clinical trials are underway to evaluate the benefits of integrating PGIs in routine screening procedures (Klarin and Natarajan, 2022; Saya et al., 2022; Widén et al., 2022). By introducing risk assessments for common diseases, PGIs are complementary to the existing class of genetic tests for rare single-gene disorders (e.g., Tay-Sachs disease) that are already used routinely in healthcare.

Private companies, such as 23andMe and MyHeritage, already offer consumer genetic testing with PGI-based risk assessments (23andMe, 2023; MyHeritage, 2021). The price of such consumer genetic health reports is now comparable to the price of blood or cholesterol tests (and are still plummeting). MyHeritage charges $199 for genotyping plus PGI-based risk assessment, while Labcorp charges $169 for their comprehensive blood panel test. Many experts predict that patients will be routinely informed of their PGI-based risk of various diseases in the foreseeable future (Zeggini et al., 2019), much as with blood test results today.

Meanwhile, concerns that individuals may be treated inequitably on the basis of their genetic data have led many jurisdictions to ban genetic discrimination in various sectors (Golinghorst et al., 2022).[1] These bans often prohibit insurers from requesting genetic information known to applicants, e.g., from a consumer genetic test, or from utilizing any genetic information at all, when they assess consumers' risks in underwriting—creating a distinct information asymmetry. Scholars, policymakers, and insurance industry stakeholders have voiced concerns that these developments will lead to high levels of adverse selection in some segments of the insurance industry (Joly et al., 2014; Rechfeld et al., 2019; Maxwell et al., 2021). This, in turn, could lead to spiraling insurance premiums, or even the unraveling of these market segments, which could leave all their consumers, including the most vulnerable ones, uninsured (Akerlof, 1970; Einav et al., 2023).

This study combines large-scale biobank data, economic theory, and econometrics to

[1] For instance, in the US, the Genetic Information Nondiscrimination Act (GINA) of 2008 prohibits the use of genetic information by health insurers, but not by other types of insurers. In Canada, the Genetic Non-Discrimination Act (GNDA) of 2017 broadly prohibits genetic discrimination, including in all sectors of the insurance industry. Similarly, in France, the Penal Code prohibits the use of genetic information other than for medico-scientific purposes, and thus by all insurers (Bélisle-Pipon et al., 2019).

inform the debate on genetics and adverse selection in insurance markets. We focus on the market for **critical illness insurance** (**CII**), which is worth more than $100 billion globally and is growing (Allied Market Research, 2022). CII pays out a lump sum in the event that the insured person gets diagnosed with any of a number of medical conditions listed on the policy (Brackenridge et al., 2006).[2] We focus on the CII market because it is an ideal testbed for the methodology we develop.

This study makes three key contributions. First, we develop a method to measure the amount of adverse selection that would take place if consumers had widespread access to current-day genetic information. The method uses large-scale datasets with information on genetics and on insured events. Second, we extend the method to estimate the amount of selection that will occur with future genetic prediction technology. The idea is to combine information from heritability studies—which are informative of the likely accuracy of future PGIs—with estimates of current PGIs' correlations with covariates that are used by insurers in underwriting.

Third, we predict the risk of eight critical illnesses and apply the method to the critical illness insurance market to measure selection with both current- and future-day PGIs. To estimate the accuracy of future genetic prediction technology, we use data from the UK Biobank (UKB) (Bycroft et al., 2018), a large dataset with genotypic and rich health-related data, including linked data from the UK's National Health Service (NHS), for 500,000 individuals.

We have three main substantive findings. First, if genetic testing were widespread, there would be a noticeable degree of selection even with the current prediction technology. We find meaningful selection for most of the single-disease CII contracts we study. For some contracts, like those that insure against the onset of prostate cancer and against the onset of schizophrenia, we find a degree of selection similar to those in previously studied market segments that have unraveled. Second, if we consider a scenario with the expected future prediction technology, the degree of selection is extremely high. For most contracts we find degrees of selection that are within or above the range observed in market segments that have unraveled. Third, we find that there is subtle variation in the amount of selection for different contracts. The amount of selection is mostly determined by the expected accuracy of future prediction technology and by how well observable non-genetic covariates can predict the diseases.

A key driver of these findings is that we consider genetic predictions based on PGIs, which aggregate information from over one million genetic variants to predict a substantial

[2]Commonly covered conditions include different types of cancers, coronary artery disease, multiple sclerosis, and stroke, paralysis, blindness, and unspecified terminal illness (Gatzert and Maegebier, 2015).

share of the variation in disease risk. For instance, the current-day PGI of prostate cancer accounts for 11.2% of the variation and—based on heritability studies—we expect future-day PGIs to account for 18 to 57% of the variation.[3] In contrast, popular discussion of genetic prediction often focuses on rare mutations that can individually have large effects on some rare diseases. Well-known examples include mutations in the *BRCA* genes for breast cancer and in the *PKD* genes for polycystic kidney disease. Interestingly, these rare mutations would generate little adverse selection. The reason is that, even though they are highly predictive, they are extremely rare and so hardly affect insurers' average costs.

Our findings complement a literature that finds that consumers' insurance choices are responsive to genetic information. Oster et al. (2010) find that carriers of the genetic variant that causes Huntington's disease (a rare single-gene disorder affecting less than 0.01% of the population) were 5 times more likely to own long-term care insurance compared to the general population. Taylor Jr et al. (2010) find that carriers of the *APOE4* variant, which increases the risk of Alzheimer's by a factor of 2 to 3, were 2.3 times more likely to buy (or plan to buy) long-term care insurance relative to carriers of the more common *APOE3* variant.

Despite the evidence from rare diseases and the potential importance of the problem, there is little empirical evidence about whether genetic adverse selection driven by increasingly accurate PGIs for common diseases will be a serious threat to insurance markets. Scholars and policymakers have called for more research (Golinghorst et al., 2022; European Commission, 2019). Our study is the first (to our knowledge) to use large-scale genetic and health data, state-of-the-art genetic predictions and econometric modeling to examine this question.

Given our findings of considerable potential for selection, we discuss policy remedies in the end of the paper. Our view is that genetic testing is a powerful new technology that can bring about large benefits. To harness these benefits, it is important to understand how to mitigate unintended harms from the technology. Our empirical analysis is a step in this direction.

1.1 Illustrative example: CII for prostate cancer

It is possible to understand the gist of our contribution with a simple example. Consider a simplified version of a CII contract: a one-time contract that pays $100,000 if the buyer ever incurs prostate cancer by age 65. This contract is a simplification of real CII contracts but, as we will discuss later, it captures their key features. The contract is purchased in

[3]These figures are PGI R^2's on the liability scale, which we define in Section 2.1.

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

young adulthood, many years before the usual age of onset of prostate cancer. The cost of selling this contract is roughly proportional to the buyer's risk of contracting prostate cancer by age 65. Our methodology allows us to calculate the distribution of risk in a population conditional on various subsets of information. We will focus on information subsets that comprise observable covariates used by insurers in underwriting, as well as combinations of these covariates with a current-day or future-day PGI.

Figure 1, left panel, shows results based on UKB data. Scenario 1 shows the distribution of the risk of contracting prostate cancer by age 65 in an all-male population of typical consumers, conditional on non-genetic covariates only. This information is observed by both consumers and insurers and is used by insurers to set prices. There would be a non-trivial amount of selection if insurers were not allowed to use this information. But in practice, insurers do underwrite and define various consumer risk classes which they price differently. Following standard practice in the insurance industry, we define the **standard risk class** as the set of consumers whose risk of contracting prostate cancer is between 0.75 and 1.25 times the average population risk ($\bar{r} = 0.053$) (Brackenridge et al., 2006). These consumers are shown in blue in the figure and face the same price of insurance. By definition, these consumers have almost the same risk conditional on the non-genetic covariates, and so there is almost no selection within the standard risk class in Scenario 1.

Scenario 2 shows the distribution of risk conditional on both the non-genetic covariates and the current-day PGI of prostate cancer. The standard risk class now has a wider spread. Because insurers cannot observe genetic information, these consumers look like nearly identical risks to insurers and are therefore charged the same price. However, because consumers can observe their PGIs, their privately known risk varies widely.

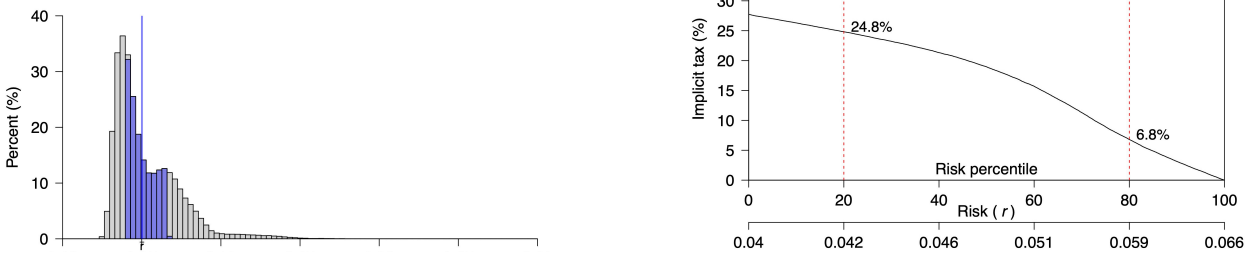
Scenarios 3L and 3U show the distribution of risk from two future-day PGIs with different assumed accuracies, both informed by existing estimates of the heritability of prostate cancer. We consider two future-day PGIs with accuracies corresponding to what are considered reasonable lower bounds (Scenarios 3L) and upper bound (Scenarios 3U). In Scenario 3L, the standard risk class now has an even wider spread, with consumers at the 5th and 95th percentiles facing risks of 0.4% and 15.9% of contracting prostate cancer, respectively. And, in Scenario 3U, consumers at the 5th percentile face a virtually null risk (as can be seen from the spike at 0), vs. a 30.2% risk for consumers at the 95th percentile. Thus, consumers who face the same price of insurance vary widely with respect to their privately known risk. This suggests that adverse selection could be a serious issue.

To quantify the degree of asymmetric information and adverse selection that would result from this situation, we employ metrics from economic theory developed by Hendren (2013, 2017). Figure 1, right panel, shows results for prostate cancer that imply very high levels

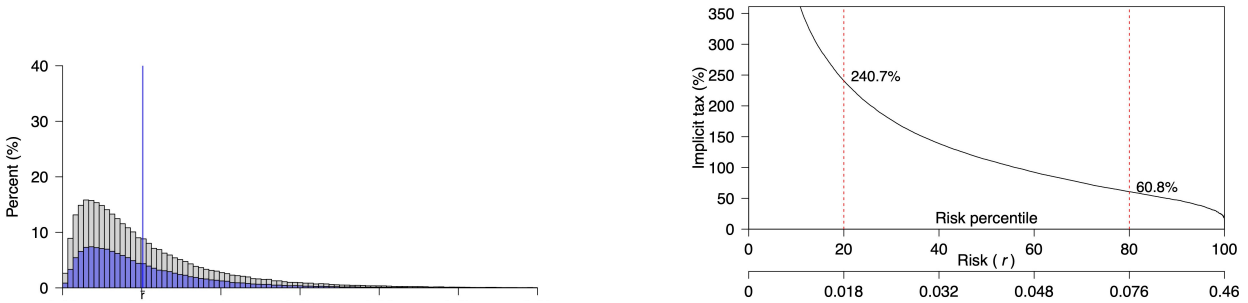
PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

Risk **Implicit Tax**

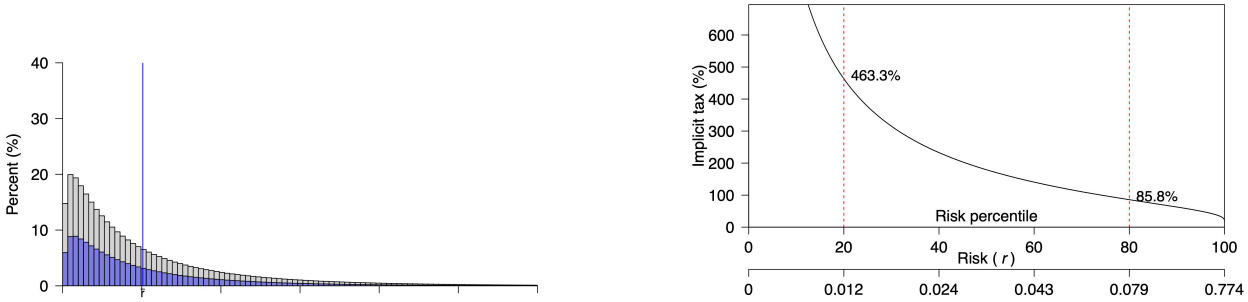
Scenario 1: non-genetic covariates



Scenario 2: non-genetic covariates and current PGI



Scenario 3L: non-genetic covariates and future PGI with lower-bound SNP heritability



Scenario 3U: non-genetic covariates and future PGI with upper-bound twin heritability

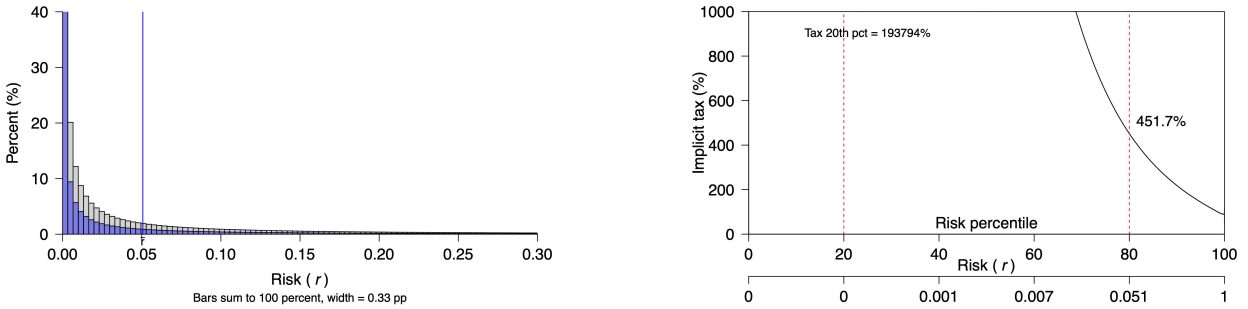


Figure 1: Single-disease Prostate Cancer CII Contract: Risk Distribution and Implicit Tax
Notes: Left panel: distribution of the risk of contracting prostate cancer by age 65, conditional on various information scenarios and computed in the UKB data ($N = 181,642$). The standard risk class is shaded blue. The vertical blue line marks the average population risk ($\bar{r} = 0.053$). Right panel: implicit tax for consumers in the standard risk class. The sub-figures show the implicit tax, as a function of the consumer's risk.

of selection—higher even than what has been estimated for insurance market segments that insurers currently refuse to serve. This, along with the results we present below for other contracts, suggests selection due to future genetic prediction technology will be a first-order concern and may make some insurance market segments non-viable.

1.2 Paper structure

The remainder of the paper is structured as follows. Section 2 provides background information on genetics and on the CII market. Section 3 describes our theory, introducing our economic and econometric models. Section 4 describes our data and our empirical specification. Section 5 presents our results, in a detailed study of a prostate cancer CII contract. Section 6 presents the results for CII contracts of many other diseases and for a bundle of diseases. Section 7 discusses extensions and policy implications. Section 8 concludes.

2 Background

2.1 Genetics background

A few disorders, such as Huntington's disease and sickle cell anemia, are caused by single genetic variants. These disorders are called monogenic. Some traits, like eye color, are principally influenced by a relatively small number of variants (Simcoe et al., 2021). However, the vast majority of traits are polygenic: they are influenced by a very large number of causal genetic variants (as well as by environmental factors), with each causal genetic variant having a tiny effect (Chabris et al., 2015; Holland et al., 2020; O'Connor et al., 2019). For example, coronary artery disease, colorectal cancer, prostate cancer, schizophrenia, and type 2 diabetes are polygenic diseases, and height, BMI, and cognitive ability are polygenic traits.[4] We focus on polygenic diseases.

To understand how genetic data are used to predict a disease such as prostate cancer, suppose we have data on a set of genetic variants from across the genome. The entire genome contains more than 3 billion pairs of elements (associated with the letters "A", "C", "G", and "T"), but for most diseases, genotyping data on ∼1 million variants capture an important share of the relevant genetic variation.[5,6] Following standard practice in epidemiological

[4] Alzheimer's disease and breast cancer are also polygenic, but in addition to being affected by a large number of variants with tiny effects, they are affected by one common variant with moderate effect (in the APOE gene, for Alzheimer's) or by rare variants with large effects (in the BRCA genes, for breast cancer).

[5] Genotyping is the process of identifying a subset of an individual's DNA that tags much of the genome's variation; sequencing is the process of identifying every single element of an individual's DNA (and is costlier).

[6] For a more detailed overview, see Beauchamp et al. (2011) and Benjamin et al. (2012).

genetics, we will model the **disease** D with a liability threshold model, according to which one contracts the disease if one's **liability** $\mathcal{L}$ is positive. Based on the additive genetic model, which assumes no interaction across variants or between variants and the environment (Falconer and Mackay, 1983), $\mathcal{L}$ is the weighted sum of an individual's genetic variants, where the weights are the variants' effect sizes on $\mathcal{L}$:

$$\begin{aligned} D &= \{\mathcal{L} > 0\}, \\ \mathcal{L} &= k + G^* + \psi = k + \mathbf{X}\,\mathbf{b}^* + \psi, \end{aligned}$$

where k is a constant; $\mathbf{X}$ is a row vector that contains the measured variants and $\mathbf{b}^*$ is a column vector that contains their true effect sizes on D; $G^* = \mathbf{X}\,\mathbf{b}^*$ is the individual's **true additive genetic factor** for D; ψ captures all other influences on D; and the operator $\{..\}$ is equal to 1 if its argument is true and to 0 otherwise. Despite its simplicity, the additive genetic model is commonly used in genetics, and both theory and empirical evidence suggest that additive effects account for the bulk of the genetic variation for most polygenic diseases (Hill et al., 2008; Polderman et al., 2015).

In practice, neither $\mathbf{b}^*$ nor G^* are observed. Instead, one can obtain an estimate $\widehat{\mathbf{b}}$ from a previously published genome-wide association study (GWAS) of the disease and use it to construct a genetic predictor $G = \mathbf{X}\,\widehat{\mathbf{b}}$. G is the **polygenic index** (**PGI**) of D.[7] Observe that

$$G = \mathbf{X}\,\widehat{\mathbf{b}} = \mathbf{X}\,(\mathbf{b}^* + \boldsymbol{\epsilon}) = \mathbf{X}\,\mathbf{b}^* + \mathbf{X}\,\boldsymbol{\epsilon} = G^* + \epsilon, \tag{1}$$

where $\boldsymbol{\epsilon}$ is the measurement error in $\widehat{\mathbf{b}}$ and $\epsilon = \mathbf{x}\,\boldsymbol{\epsilon}$ is the error in the PGI G. Measurement error $\boldsymbol{\epsilon}$ in $\widehat{\mathbf{b}}$ arises because of sampling variation in the GWAS, which is conducted in a sample that is independent of the dataset we are analyzing. Thus, $\mathbf{X} \perp \boldsymbol{\epsilon}$ and $G^* \perp \epsilon$.

As mentioned, it is common for human traits—including the diseases we study in this research—to be influenced by a large number of genetic variants, each with a tiny effect. This implies, by the central limit theorem, that the true additive genetic factor for a disease is approximately normally distributed. By the same reasoning, so is ϵ. As a result, so is G—i.e., PGIs also are approximately normally distributed (shown in the Online Appendix).

As GWASs have rapidly grown in sample size over the past decade (Loos, 2020), estimates of $\mathbf{b}^*$ have become much more precise; as a result, the predictive power of PGIs has also

[7] A GWAS is a large-scale genetic study in which a trait of interest is regressed on each of millions of genetic variants, separately. Because nearby, correlated variants are not included in a given variant's GWAS regression, GWAS coefficient estimates capture both that variant's effect and those of correlated variants. When constructing PGIs for this study, we employ a method, PRS-CS (Ge et al., 2019), that models patterns of correlation across variants and use these to transform the GWAS coefficients into estimates of $\mathbf{b}^*$ (i.e., into $\widehat{\mathbf{b}}$).

increased dramatically (see, e.g., Okbay et al., 2022). As GWAS sample sizes keep increasing, and as the set of genetic variants they analyze is augmented to include rarer variants, PGI predictive power should increase further (Abdellaoui et al., 2023; Daetwyler et al., 2008). In theory, the upper bound for a PGI's predictive power for a given trait is the trait's narrow sense heritability (So and Sham, 2010; Dudbridge, 2013), defined as the share of the trait variation that is attributable to additive genetic factors.

Table 1 shows, for the eight diseases we study, our estimates of their current-day PGIs' R^2's together with previous studies' estimates of their "SNP heritabilities" and of their "twin heritabilities". Since the diseases are binary traits, we report the R^2 **on the liability scale**. This R^2 is also known as the McKelvey & Zavoina pseudo-R^2 and is defined as $\text{Var}[G\beta]/\text{Var}[\mathcal{L}_r]$, where β is the coefficient from a probit regression of D on G and $\text{Var}[\mathcal{L}_r] = \text{Var}[G\beta] + 1$ is the variance of the liability in that regression (Lee et al., 2012; McKelvey and Zavoina, 1975).[8] A disease's SNP heritability is the share of the variance in its liability $\mathcal{L}$ that is attributable to the set of common genetic variants used in the SNP heritability estimation.[9] Because not all variants are used (as most were not measured, and for computational reasons), SNP heritability is necessarily lower than the true heritability. A trait's twin heritability is computed by comparing the resemblance of monozygotic twins—who share 100% of their genomes—to that of dizygotic twins—who share 50% on average. Under some assumptions, it is a consistent estimate of the trait's true heritability.

Table 1: Descriptive statistics for selected diseases

Disease	**Prevalence**	**Lifetime risk**	**Current PGI R^2**	**SNP h^2**	**Twin h^2**
Alzheimer's disease	7.1%	6.3% (M), 12% (F)	6.5%	31%	58%
Breast cancer (F only)	8.1%	12.3%	6.8%	14%	31%
Colorectal cancer	2.9%	4.6%	2.5%	9%	40.2%
Coronary artery disease	3%	51.7% (M), 39.2% (F)	3.0%	22%	40%
Lung cancer	3.2%	5.7% (smokers)	1.1%	7.5%	38%
Prostate cancer (M only)	0.9%	12.7%	11.2%	18%	57%
Schizophrenia	0.46%	0.7%	5.8%	24%	79%
Type 2 diabetes	3.9%	27% (M), 18.6% (F)	8.6%	19.6%	72%

Notes: Reported R^2's are on the liability scale and were estimated in the UKB (further described below). All other statistics come from earlier, separate studies of the diseases (references available upon request) and were not produced with a uniform methodology across diseases. Prevalence and lifetime risk include various age-corrected measures (e.g., the prevalence of Alzheimer's disease is reported for ages above 65).

While it is impossible to know with certainty how predictive a disease's future PGI will be, it is likely that it will explain at least as much as the disease's SNP heritability estimated

[8]The R^2 on the liability scale is intuitively analogous to the usual R^2, but indicates the share of the variance of the liability—instead of the variance of a continuous dependent variable—that is accounted for by the covariates.

[9]The expression "SNP heritability" comes from the fact that the variants used are typically single nucleotide polymorphisms (SNPs).

with common variants. This is because future PGIs, in addition to being constructed with more precise estimates of common variants' effect sizes, will also account for the effects sizes of less common variants. And, as mentioned, the theoretical upper limit on a PGI's predictive power for a disease is the disease's heritability. In our analyses, we report results both under the assumption that future PGIs' accuracies will be equal to their SNP heritabilities (as lower bounds) and that they will be equal to their twin heritabilities (as upper bounds).

As is currently standard practice in applied genetic research, our analysis only uses data from individuals of European ancestry, for two main reasons. First, available genetic data for other ancestries are still scarce (Mills and Rahal, 2020); as a result, most large-scale GWASs whose estimates we can use to construct our PGIs were conducted among European ancestry individuals. Second, PGIs based on European ancestry GWAS estimates tend to perform poorly among individuals of other ancestries (Martin et al., 2017; Privé et al., 2022).[10] There are numerous ongoing efforts to diversify the world's biobanks (Zhou et al., 2022) and to develop statistical methods to improve polygenic prediction across ancestries. As a result, future PGIs should eventually have predictive powers of similar magnitudes across different ancestries. Our results regarding the implications of improved genetic prediction for insurance markets are thus equally relevant for non-European ancestry individuals.

2.2 Critical illness insurance

Critical illness insurance (**CII**) pays out a lump sum in the event that the insured person gets diagnosed with any of the medical conditions listed on the policy (Brackenridge et al., 2006). The lump sum can be used as the policyholder wishes. The policy pays out once and is thereafter terminated. The typical policy is renewable at a guaranteed rate, which is determined before contract initiation by medical underwriting. Typical contracts cover about thirty core diseases or disorders, including different types of cancers (including breast, prostate, and colorectal cancer), coronary artery disease, multiple sclerosis, stroke, as well as serious conditions like paralysis, blindness, and unspecified terminal illness (Gatzert and Maegebier, 2015). Enhanced plans exist that cover additional conditions, such as dementia, diabetes, or severe psychiatric illness. The Canadian Institute of Actuaries (2012) reports that, in Canada, 68% of claims were paid to cancers and 13% to heart attacks.

CII originated in South Africa in 1983. It proved popular because expensive medical treatments were not covered by public insurance programs. Presently, major markets include Canada, the United Kingdom, Japan, Australia, India, China, and Germany. It is estimated that 20% of British workers were covered by a CII policy in 2009 (Gatzert and Maegebier,

[10] This is likely due to population genetic reasons and differential genetic effects in different environments.

2015). The global CII market has been valued at over $100 billion in 2021 and was projected to grow to over $350 billion by 2031 (Allied Market Research, 2022).

Contract details vary across markets. One dimension of variation is the set of covered illnesses, which has tended to increase over time and is a point of differentiation between competing firms' offerings. Some contracts only cover a specific disease or a specific narrow group of diseases. One example is cancer insurance in the US and Japan (Gatzert and Maegebier, 2015; Canadian Institute of Actuaries, 2012). Another dimension of variation is whether policies include other benefits, such as paying back a fraction of the premium in case of death of the policyholder. Some contracts are standalone, whereas others are riders to another insurance policy such as life insurance.

Underwriting for individual policies is similar to life insurance underwriting. For individual policies, insurers request detailed information, and may ask for medical records, blood or other medical tests, and medical examinations (Canadian Institute of Actuaries, 2012). Industry reports often discuss concerns with adverse selection, and some contract features like waiting periods are meant to curb adverse selection (Gatzert and Maegebier, 2015).

The existing economic literature on the impact of genetic testing on CII is scarce. Our study relates most to a series of studies in the actuarial science literature by MacDonald and colleagues (e.g., Macdonald et al., 2005; Macdonald and McIvor, 2009; Macdonald and Tapadar, 2010; Macdonald and Yu, 2011; Adams et al., 2015). They investigate the impact genetic information will have on CII and generally find that adverse selection due to private genetic information is unlikely to be a problem. However, most of these studies are from more than a decade ago and, consistent with the genetic prediction technology at the time, mainly consider single-gene mutations that account for only small shares of the population variation in the studied disorders. Thus, by largely ignoring expected improvements in the predictive accuracy, this literature has found that selection will not be a problem.

More recently, Maxwell et al. (2021) found that current-day PGIs predict breast cancer and coronary artery disease above and beyond normal underwriting factors, and discuss the possibility that this may lead to adverse selection. In a study for the Canadian Institute of Actuaries, Howard (2016) considers only some rare genetic variants associated with six conditions and finds that barring insurers from using genetic data would have a sizeable impact on CII claims rates (and thus on premiums) due to adverse selection.

3 Theory

3.1 Economic model

The model. Consider a population of consumers who may incur a binary loss and can purchase an insurance contract for it. For example, in a single-disease CII contract, the loss corresponds to developing the disease. In a multiple-disease CII contract, the loss corresponds to developing one of the diseases listed in the contract. We will use the terms "loss" and "disease" interchangeably.

Throughout this section, we consider a population that is treated equally by insurers, such as consumers in the same risk class. We assume that consumers have additional private information, so that there may be selection. This additional private information can include both genetic and non-genetic information.

Each consumer is characterized by the realization of four random variables $(D, G_c, G_f, \mathbf{W})$. The **disease** (or **loss**) D is a random variable equal to zero or one depending on whether a consumer contracts a disease. G_c, the **current polygenic index** (**PGI**) of the disease D, is a sufficient statistic of the consumer's DNA for the purpose of estimating the probability of contracting D, conditional on the current genetic prediction technology. G_f, the future PGI of D, is a sufficient statistic of the consumer's DNA conditional on the future genetic prediction technology. The **non-genetic covariates** $\mathbf{W}$ are a row vector of covariates. The variables $(D, G_c, G_f, \mathbf{W})$ are independently identically distributed across consumers, with joint distribution $\mathbb{P}$. G_c and G_f take real values. $\mathbf{W}$ takes values in Euclidean space.

Private information about risk. Each consumer has private information about her non-genetic factors $\mathbf{W}$ and her PGI G. We will consider three counterfactual scenarios with different sets of consumer private information that vary as a function of the assumed genetic prediction technology: with no genetic prediction technology; with the current-day PGI $(G = G_c)$; and with the future-day PGI $(G = G_f)$.

Denote consumers' private information about risk as follows. The **non-genetic private risk function** is defined as the probability $\pi(\mathbf{w})$ of a disease conditional on non-genetic factors $\mathbf{W} = \mathbf{w}$:

$$\pi(\mathbf{w}) := \mathbb{E}\left[D|\mathbf{W} = \mathbf{w}\right].$$

The **private risk function** is defined as the probability $\rho(g, \mathbf{w})$ of a loss conditional on the PGI observed by the consumer, $G = g$, and non-genetic factors $\mathbf{W} = \mathbf{w}$:

$$\rho(g, \mathbf{w}) := \mathbb{E}\left[D|G = g, \mathbf{W} = \mathbf{w}\right].$$

Define the random variable for **non-genetic private risk** as $P := \pi(\mathbf{W})$, and the random variable for **private risk** as $R := \rho(G, \mathbf{W})$. We henceforth refer to R as **risk**. The realization of losses as well as both non-genetic private risk and private risk are independent across consumers.

Measures of adverse selection. The distribution of risk R across the population is informative about the degree of adverse selection when consumers have genetic information. Figure 1, left panel, displays the possible distribution of the risk of contracting prostate cancer by age 65 under the different information scenarios. The blue region represents consumers in the standard risk class. In Scenario 1, where consumers only have non-genetic private risk, the distribution is concentrated around the mean. Therefore, consumers have little private information about their own risk, and consumers who purchase insurance are not likely to be particularly high-risk, making it easier for an insurance market to perform well. By contrast, the distributions in Scenarios 3L and 3U, where consumers observe also observe future-day PGIs, are highly dispersed. Therefore, some consumers know that they have low risk and other consumers know that they have high risk. So consumers who purchase more insurance are likely to be those with particularly high risk. This makes it more likely that an insurance market will perform poorly. In standard adverse selection models, few or no consumers may end up buying insurance, because firms expect that only high-risk consumers would want to purchase and so set a high price, which further deters low-risk consumers, which leads firms to set an even higher price (Akerlof, 1970).

Adverse selection depends on the entire distribution of risk R. Therefore, we will report all the key distributions with histograms like those in Figure 1. Moreover, we will also report an economic measure of selection, the **implicit tax** (Hendren, 2013). The implicit tax $t(r)$ is defined as the extra amount a consumer with risk r would have to pay if she had to buy insurance at the actuarially fair price for all consumers with risk r or greater, versus what her own actuarially fair price is. The formal definition is in Appendix A.1, Equation (A.1). Intuitively, the implicit tax tells us how much more the marginal consumer has to pay to buy insurance due to adverse selection, and gives this number as a percentage tax that is easy to interpret.

The implicit tax measure has three advantages:

1. Hendren (2013) proposed a no-trade theorem that shows that sufficiently high values of the minimum of the implicit tax cause insurance markets to break down under certain assumptions. Therefore, at least in some settings, the implicit tax is a measure of selection problems.[11]

[11] A key assumption of this no-trade theorem is that consumers are assumed to have identical risk pref-

2. Regardless of whether the specific assumptions hold in any given example, the implicit tax is intuitive, and likely to be a relevant statistic even in more general settings.

3. We can benchmark the implicit taxes with the values calculated for other markets. Hendren (2013) estimated the implicit tax in market segments that had unraveled (in the sense of not being served by insurers) and in market segments that had not unraveled. He reports the minimum implicit tax up to the 80th percentile of risk, which we denote t_{80}. If t_{80} were to occur at the 70th percentile, say, and were sufficiently low, then insurance in the market segment comprising the top 30% riskiest consumers should be viable. Hendren (2013) considers the *minimum* up to the 80th percentile because if the implicit tax is high at that and at all lower percentiles, then any market segment that comprises more than the top 20% of the riskiest consumers would not be viable, and any segment that comprises less may be too small and thus also unviable. So a high t_{80} is indicative that a market might unravel. For market segments that had not unraveled, he estimated t_{80} to be between 7% and 35%. For market segments that had unraveled, he estimated t_{80} to be between 43% and 83%.

3.2 Econometric model and identification

Our model is fully specified by $\mathbb{P}$, the joint distribution of $(D, G_c, G_f, \mathbf{W})$. While the current-day PGI G_c can be observed in the UKB data by the econometrician, the future-day PGI G_f cannot. Define the distribution of the observed data as a joint distribution $\mathbb{P}_{\text{data}}$ over $(D, G_c, \mathbf{W})$. We say that the model is identified if there is a unique value of $\mathbb{P}$ consistent with $\mathbb{P}_{\text{data}}$.

It may at first appear that the model cannot be identified. Identification implies that we can use the UKB data to estimate the degree of selection in a world in which consumers can observe their future-day PGIs. It turns out that the model is identified under reasonable assumptions. The key idea is that empirical regularities from quantitative genetics considerably restrict the data, making identification possible. Our assumptions, detailed below, are based on two empirical regularities, which we discussed in Section 2.1. First, PGIs are approximately normally distributed. And second, existing heritability estimates for a disease are informative about the likely predictive power of its future PGI.

erences. However, in equilibrium selection models, the degree of market failure depends both on how risk is distributed and on its joint distribution with risk preferences (Einav et al., 2010a). Therefore, studies of the costs of adverse selection often take into account how risk is correlated with demand for insurance. Due to data limitations, we cannot reliably measure insurance demand from the UKB data. Therefore, we follow Hendren (2013) in measuring only risk distributions. Developing a full structural model for insurance demand and risk would be an interesting topic of future research.

We now state the necessary assumptions. The first assumption is a regularity condition on the distribution of covariates.

Assumption 1. *[Covariates' distribution]* $\mathbf{W}$ *is distributed according to a distribution* $\mathbb{P}_{\mathbf{W}}$ *in Euclidean space. The support of* $\mathbf{W}$ *spans the entire space.*

The next two assumptions are conditional normality assumptions on the PGIs G_c and G_f. They are motivated by the two regularities mentioned above and by the fact that what distinguishes a current-day PGI from a future-day PGI is simply the size of the independent and approximately normally distributed error. Further, because G_c does not have a natural scale, it is standardized to have unit variance (and zero mean) in the data.

Assumption 2. *[Gaussian future-day PGI] Conditional on observables, the future-day PGI has a Gaussian distribution. The variance is constant and the mean is a linear function of observables:*

$$G_f := \mathbf{W}\theta + V, \tag{2}$$

with $V \sim N(0, \sigma_V^2)$ *and* $V \perp \mathbf{W}$.

Assumption 3. *[Noisy current-day PGI] The observed current-day PGI* G_c *has unit variance,* $\mathrm{Var}[G_c] = 1$*, and is a noisy estimate of* G_f*:*

$$G_c := G_f + \epsilon, \tag{3}$$

with $\epsilon \sim N(0, \sigma_\epsilon^2)$ *and* $\epsilon \perp (\mathbf{W}, V)$.

We note that when we assume that a future PGI's accuracy is equal to the disease's twin heritability (as an upper bound), Assumption 3 is likely to be violated. Specifically, in that scenario, it is unlikely to be the case that $\epsilon \perp \mathbf{W}$.[12] Here, we ignore these subtleties, as they are unlikely to impact our main conclusions. Online Appendix E further discusses this issue and introduces a more complicated model that deals with it, albeit at the cost of additional assumptions.

The fourth assumption is that D follows a probit model, which is the standard functional form for binary diseases in genetic epidemiology. The assumption also implies that the current-day PGI has no predictive power beyond the future-day PGI.

Assumption 4. *[Probit disease model] Disease probabilities follow a linear probit model:*

$$\Pr[D = 1 | G_f = g_f, G_c = g_c, \mathbf{W} = \mathbf{w}] = \Phi(g_f\beta_g + \mathbf{w}\beta_w). \tag{4}$$

[12] The reason is that G_c, being constructed with currently available SNP data, is then not a noisy signal of G_f, but only of the component of G_f that can be predicted by SNPs.

Under assumptions (1)-(4), the distribution $\mathbb{P}$ has parameters

$$\Gamma := (\mathbb{P}_{\mathbf{W}}, \beta, \theta, \sigma_V^2, \sigma_\epsilon^2).$$

The data can be generated by taking independent draws of $\mathbf{W} \sim \mathbb{P}_{\mathbf{W}}$, $V \sim N(0, \sigma_V^2)$, and $\epsilon \sim N(0, \sigma_\epsilon^2)$. Equation (2) then determines the future PGI G_f; Equation (3) determines the current PGI G_c; and the disease D is realized with probability given by Equation (4).

For the model to be identified, we need one more assumption. We need to assume that we know the predictive power of the future PGI G_f, which we denote R_f^2, from heritability studies. (We can estimate the predictive power of the current PGI, R_c^2, in the data.)

Assumption 5. *[Future PGI predictive power] The pseudo R^2 on the liability scale of a probit regression of D on G_f equals a known constant R_f^2.*

We are now ready to state the identification result.

Theorem 1. *Under assumptions (1)-(5), the model is identified.*

Appendix A.2 provides the proof. The intuition for this is as follows. $\mathbb{P}_{\mathbf{W}}$ is identified because $\mathbf{W}$ is observed. The coefficients θ are the coefficients of a regression of G_c on $\mathbf{W}$. The variances σ_ϵ^2 and σ_V^2 are identified from G_f's assumed future predictive power, the estimated predicted power of G_c in the data, and the variance of $\mathbf{W}\theta$. Finally, the coefficients β are related to the coefficients of a probit regression of D on G_c and $\mathbf{W}$. Once the other parameters are known, β can be computed from that regression's coefficients.

4 Data and empirical specification

4.1 Data

We will analyze data from the UK Biobank (UKB) (Bycroft et al., 2018; Sudlow et al., 2015). The UKB contains genotypic and rich health-related data for over 500,000 individuals from across the United Kingdom who were between 40 and 69 years old at recruitment (between 2006 and 2010). UKB data is linked to the UK's National Health Service (NHS), which maintains detailed records on health events from birth to death and with whom 98% of the UK population is registered. In addition, all UKB participants took part in a baseline assessment, in which they provided rich environmental, family history, health, lifestyle, physical, and sociodemographic data, as well as blood, saliva, and urine samples.

The UKB contains genome-wide array data for ∼800,000 genetic variants, for ∼488,000 participants. Following standard practices, all genotyping data were imputed (by UKB

researchers) using reference genomes. In our analyses, we use these imputed data for the circa 1 million well-characterized variants from the HapMap 3 Project that are commonly used to compute PGI (Altshuler et al., 2010; Yengo et al., 2022). Our analysis sample comprises the 446,252 UKB participants of European ancestry who have been genotyped and whose genetic data passed a number of quality control filters. Table 2 reports sample descriptive statistics after having filtered on variable missingness over the baseline risk factors and the disease-specific risk factors (listed in Table 3). For each of a suite of diseases, we use these genetic data and estimates of the variants' effect sizes, together with the software PRS-CS (Ge et al., 2019), to compute the disease's current-day PGI. The estimates of the variants' effect sizes come from the largest available GWAS of the disease whose analysis sample did not contain the UKB (see the Online Appendix).[13,14,15]

The UKB's linked electronic health record (EHR) data cannot be directly analyzed because it is large and unstructured. To code each disease, we rely on the established and widely used "Phecode" mapping system (Denny et al., 2013), which was designed specifically to handle the complexities of defining clinically meaningful case-control status from various complex EHR sources (Bastarache, 2021). The mapping system uses the categories of the International Classification of Diseases (ICD9/10) as reference point.[16]

For each disease, we reviewed the epidemiological literature to identify the main disease-specific non-genetic risk factors. These risk factors are typically used both by medical practitioners and by insurers in underwriting to assess individual disease risk. We have coded variables that capture most of these risk factors in the UKB. Table 3 lists these disease-specific risk factors, together with a set of baseline risk factors for which we account in our analysis of all diseases; the few risk factors we were not able to code with the UKB data are in square brackets. Accounting for conventional epidemiological risk factors in our analysis is critical, because they are typically observed and used by insurers in underwriting. We use

[13]We focus on the European ancestry participants for the reasons outlined in Section 2.1. Online Appendix Section B provides more details on the specific GWASs we used as well as on the quality control filters and how we constructed the PGIs.

[14]The GWAS sample must not overlap with the UKB to avoid overfitting (Wray et al., 2013)—i.e., to ensure that $G^* \perp \epsilon$.

[15]As mentioned, Alzheimer's disease and breast cancer are polygenic diseases, but they are additionally affected by variants with sizeable effects in the *APOE* and *BRCA* genes, respectively. In Online Appendix Section B, we show that our PGI for Alzheimer's disease could capture completely the effect of *APOE* risk type. Our PGI also includes common variants in the *BRCA* genes, but we did not investigate the overlap with rare *BRCA* variants because doing so would require sequencing data.

[16]A key issue addressed by this Phecode mapping system is the high level of heterogeneity among codes in the ICD hierarchy. For example, the general breast cancer ICD10 code C50 (malignant neoplasm of breast) has some ten subcodes to identify what part of the tissue is affected. The Phecode system groups these subtypes together under a single code, 174.1 (Breast cancer [female]). In the latest version, the Phecode mapping system condenses some 90,000 ICD-10 codes into about 1,900 Phecodes (Bastarache, 2021).

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

Table 2: Sample descriptive statistics

Panel A. Case proportion and age of onset

			Age of onset			
Disease	**N**	**% Cases**	**Mean**	**SD**	**5th pct**	**95th pct**
Alzheimer's disease	405,730	0.6%	74.3	5.2	64.4	81.1
Breast cancer	184,142	6.4%	59.9	10.5	42.2	75.9
Coronary artery disease	410,858	12.0%	62.9	9.7	45.9	77.3
Colorectal cancer	410,688	2.3%	64.9	9.2	48.4	77.8
Lung cancer	399,843	1.3%	68.4	7.7	54.2	78.8
Prostate cancer	177,803	7.7%	66.9	6.6	55.4	77.2
Schizophrenia	410,858	0.2%	47.2	17.1	20.5	74.6
Type 2 diabetes	410,858	7.7%	63.7	9.0	47.9	77
Any disease (males)	170,170	30.8%	62.9	9.4	46.4	76.5
Any disease (females)	177,803	19.3%	62.3	10.1	44.7	77.1

Panel B. Descriptive statistics baseline risk factors (N = 416,261)

Variable	**Mean**	**SD**	**Min**	**Max**
Age (at last observation)	65	9.67	39	85.7
Sex (Male = 1)	0.46	0.5	0	1
Deprivation index (Townsend)	-1.52	2.94	-6.26	10.88
Education (years of schooling)	13.98	4.95	7	30
BMI	27.36	4.74	12.12	74.68
Drinks per week	8.35	9.27	0	52
Smoking status: - Current	0.1	0.3	0	1
- Ex smoker	0.36	0.48	0	1
- Never	0.54	0.5	0	1
Physical inactivity	0.01	0.11	0	1
Systolic blood pressure (SBP)	138.3	18.43	72	253.5

Notes: The sample sizes in panel A reflect the number of participants with no missing values for (i) the baseline risk factors (Panel B) and (ii) the disease-specific risk factors (Table 3).

these factors, together with ten genetic principal components (PCs) and a genotyping array dummy, as the covariates $\mathbf{W}$ in our prediction models.

4.2 Contracts and specification

We will apply our model to CII contracts. We consider a loss to be the occurrence of a covered disease and consider contracts that pay a lump sum at the onset of the disease, thus ignoring the uncommon additional benefits some contracts provide. Actual contracts vary across several dimensions (see Section 2.2). So we narrow down our analysis to particular types of contracts, especially in two key contract dimensions: the set of covered diseases and the coverage period. Furthermore, we specify a population of interest and the covariates. With this, we have a full specification to take our theoretical model variables (current PGI G_c, covariates $\mathbf{W}$, and loss D) to the data.

Set of covered diseases. CII contracts vary in the number of diseases covered. Most contracts cover about thirty diseases, and the set of diseases has been a point of differentiation between different companies. Nevertheless, most claims are due to cancers (Canadian

Table 3: Non-genetic risk factors for selected diseases

Disease	**Key risk factors**
Baseline risk factors (for all diseases)	Age, sex, Townsend Deprivation index, education (years of schooling), body mass index (BMI), alcohol consumption, smoking (current, ex, or never), physical inactivity, systolic blood pressure, family history* (did each of father, mother, or siblings ever suffer from the disease?)
Alzheimer's disease	Air pollution, APOE risk type**, [brain injury], coronary artery disease, depression, diabetes, hearing loss, hypertension, physical inactivity, social isolation
Breast cancer	Diabetes, ever had breast cancer screening / mammogram, ever had hormonal replacement therapy, hormonal contraceptives, age at menarche, age at menopause, age at first birth, number of live births
Colorectal cancer	Crohn's disease, diabetes, [diet], ever had bowel cancer screening, irritable bowel syndrome (IBS), polyp of colon, ulcerative colitis (and other non-infective gastro-enteritis and colitis)
Coronary artery disease	Depression, diabetes, [diet], hypercholesterolemia, hyperglycemia, hypertension, social isolation
Lung cancer	Air pollution, asthma, [diet], Human papillomavirus (HPV) infection, lung or respiratory disease due to environmental or occupational exposure (e.g., asbestos, diesel, radon, or other external agents), other history of lung or respiratory disease, smoking, smoking intensity (peak adulthood number of cigarettes per day), social isolation
Prostate cancer	Ever had prostate specific antigen (PSA) test
Schizophrenia	Anxiety, bipolar disorder, [cannabis use], depression, neuroticism, [psychopathology]***, [substance use]
Type 2 diabetes	[Diet], hypercholesterolemia, hyperglycemia, hypertension, social isolation

Notes: The risk factors for each disease were identified by a review of the epidemiological literature and the reviewed references are reported in Supplementary Table [Risk factors]. Variables capturing all identified factors were coded in the UKB, except for a few that were not clearly defined in the literature (e.g., diet) or for which data was not available (indicated in square brackets). * For schizophrenia, family history of "severe depression" was used as proxy. ** A covariate for APOE risk type was omitted because of perfect overlap with the PGI. *** Psychopathology was proxied by neuroticism score, bipolar disorder, and depression.

Institute of Actuaries, 2012). Some insurers offer enhanced policies with multiple tiers of covered diseases and medical conditions. There are also less comprehensive contracts such as cancer insurance and even single-disease CII contracts.

We first consider single-disease contracts. While these are uncommon, they have three advantages. First, implementation is simpler because PGIs and heritability estimates are available from the literature. Second, these simple contracts make the interpretation of the results more transparent, which is useful for evaluating our methodology. Third, CII claims are dominated by a few diseases, so single-disease contracts may be representative of typical contracts. The single-disease CII contracts can be seen as the smallest unit of coverage, and thus, a useful Petri dish with which to illustrate and test our methodology.

We consider the eight diseases in Table 1. These include some of the most common cancers (prostate, breast, lung, colorectal), which by themselves account for a large share of CII claims. We include coronary artery disease, which is the next most common non-cancer source of claims. We include Alzheimer's disease because it is important for long-term care insurance, so the results might be useful for future research on that market. Schizoprenia is included as an example of a highly heritable but rare condition. We included type 2 diabetes because it is common and of public health concern, although it is not covered by most CII policies.

We also consider a bundled CII contract, that pays the lump sum upon occurrence of any of the eight diseases we study, and then terminates. Our eight diseases include the most common cancers and some of the most common non-cancer critical illnesses. So the bundled contract is similar to a realistic CII contract. In fact, many of the diseases and medical conditions covered by real CII contracts but not by our analysis are rare. Some industry observers believe that insurers include rare diseases because consumers value the more comprehensive policies, perhaps due to poor statistical reasoning, even though the costs are negligible from an actuarial perspective.

Coverage period. Typical CII contracts are guaranteed renewable, at a predetermined rate, up to a maximum age. It is common for the maximum age to be 65. We will restrict attention to this type of contract. The simplest way to model this contract is to assume that consumers will renew until age 65. Ignoring interest rates, this means that firm's costs are proportional to the probability that the consumer will incur the disease up to age 65. This is the approximation we will use. In the terminology of our model, we set the loss $D = 1$ if the consumer ever incurs the disease by age 65.

We stress that this is an approximation; it departs from reality in two important ways. First, in practice, consumers do let policies lapse. Lapsation is a secondary but important driver of insurance premiums. For example, in life insurance it has been documented that older consumers are more likely to let their policies lapse, even though they have the greatest probability of dying. This is crucial for firms to make positive profits (Gottlieb and Smetters, 2021). Second, the timing of losses also influences insurers' costs, due to time discounting.

Population of interest. We are interested in a population of young adults at a typical age of first CII purchase (and younger than the usual age of disease onset) and who may purchase CII contracts. Our model considers a population of consumers who are treated similarly by insurers. We follow standard underwriting practices to define this group. Insurers often define the **standard risk class** as all consumers who are within 0.75 to 1.25 times the average population-wide risk, based on underwriting using the covariates insurers can observe. It is typical for insurers to charge the same rates to all consumers in the standard risk class. We take the standard risk class to be the population of interest in the analyses we report in the main text. This fits our model, where we measure selection within a group of consumers who are treated identically by insurers.

PGIs and covariates. For each disease, we construct the current PGI G_c based on the genetic data from the UKB and prior genome-wide association studies, as described in Section 4.1. Table 3 lists the covariates $\mathbf{W}$ for each disease.

Scenarios and information sets. For each contract, we produce results under four Scenarios. In all Scenarios, insurers only observe a number of discrete consumer risk classes (including the standard and high risk classes).[17] The scenarios differ with respect to the consumers' information sets. In Scenario 1, consumers observe only their non-genetic covariates and thus observe a moderate amount of non-genetic private risk P. In the other scenarios consumers observe both their non-genetic covariates and their current or a future PGI, and thus observe their private risk R. In Scenario 2, consumers observe their current-day PGI (G_c). In Scenarios 3L and 3U, they observe a future-day PGI (G_f). In Scenario 3L, the PGI's predictive power is equal to the disease's SNP heritability ($R^2_f = h^2_{SNP}$), and in Scenario 3U it is equal to the disease's twin heritability ($R^2_f = h^2_{twin}$).

Scenarios 2 and 3 assume that consumers have widespread access to their genetic risk predictions. As discussed in the Introduction, it is widely believed that this may be the case in the future. Currently, however, there is no widespread consumer access to genetic risk predictions. Therefore, though Scenario 2 is based on the current PGI, it does not reflect the current state of affairs.

4.3 Estimation

For each contract, estimation is based on the identification Theorem 1. The theorem's proof shows that all parameters of the econometric model are simple functions of moments in the data. While we could simply read off these moments in the data, our actual estimator estimates these moments jointly using the generalized method of moments (GMM) to properly estimate standard errors.

One minor complication is that, in the UKB, we do not have long-term follow-up data. If we had such follow-up data, we would estimate the model off a population of consumers that we would observe both at a typical age of first CII purchase, say 30 years old, and then again at 65 years old. In practice, because of this data limitation, for each disease we use data from individuals of different ages to estimate the risk each individual develops the disease by age 65.

For each contract, our method proceeds as follows:

Step 1. Estimating the econometric model. We begin by estimating our econometric model for the contract, using our sample of European ancestry UKB participants who have been genotyped. We pool data from females and males, except for breast cancer,

[17] In reality, insurers observe all the non-genetic covariates. However, they only use these to classify consumers across the risk classes, and so they effectively only observe the risk classes. Because the risk classes are somewhat coarse, there is still non-genetic private risk within each class, but not enough to lead to high levels of selection (that's why insurers are content to use these coarse risk classes to set prices).

prostate cancer, and the bundled contract. Estimation for single-disease CII contracts works as follows. For Scenarios 1 and 2, this involves only a probit regression of the disease on the non-genetic covariates (Scenario 1) and of the disease on both the non-genetic covariates and the current-day PGI (Scenario 2) (see Appendix A.2).

For Scenarios 3L and 3U, we estimate the current PGI's predictive power (R_c^2) in the UKB data, and assume that the future PGI's predictive power (R_f^2) is equal to the disease's SNP heritability and twin heritability for Scenarios 3L and 3U, respectively. Next, we estimate our econometric model, which involves first regressing G_c on $\mathbf{W}$ to estimate θ (Equation A.2); estimating σ_ϵ^2 and σ_V^2 from the assumed R_f^2, the estimated R_c^2, and the variance of $\mathbf{W}\theta$ (Equations A.3 and A.4); and estimating a probit regression of the disease on the non-genetic covariates and the current-day PGI and then transforming the coefficients into estimates of β_g and β_w (Equations A.6-A.7). We estimate all the corresponding moments together with the generalized method of moments (GMM).

Estimation for the bundled CII contract is nearly identical. We construct a PGI for the loss by running a regression of the loss on the individual PGIs for each disease. We obtain the current predictive power in the same way and estimate the SNP heritability of the loss in the UK Biobank with a standard approach used in genetic epidemiology (genome-based restricted maximum likelihood) (Yang et al., 2011). Twin heritability is not available in the literature and we do not have sufficient twin pairs in our data. Therefore, we cannot perform the analysis for twin h^2 as in the single-disease contracts. To have a sense of what selection would look like with higher heritability, we report results under a back-of-the-envelope assumption: that the ratio of twin h^2 to the SNP h^2 equals the average ratio across the eight diseases. We stress that the particular value of h^2 should be taken with a grain of salt.

Step 2. Generating the private risk distributions. For all four scenarios, we generate the distribution of private risk for the disease for all the individuals in our sample. We use Equation (4) along with the parameter estimates obtained from Step 1, the PGI (for Scenarios 2-3), and the age-adjusted covariates (described next) to estimate each individual's risk of contracting the disease by age 65. For Scenarios 3L and 3U, we do not observe G_f, so for each individual we draw 100 values from its probability distribution conditional on G_c and $\mathbf{W}$ (given by Lemma 1 in the Appendix) and use these to generate the risk distribution.

Individuals in the data range in age from 39 to 86. Our set of covariates $\mathbf{W}$ (see Table 3) includes age as well as a number of age-dependent covariates, such as BMI, blood pressure, hypertension, and whether one ever had menopause (for breast cancer for females) or a prostate specific antigen (PSA) test (for prostate cancer for males). Because age and a number of these covariates are important predictors of risk for all the diseases, we set age

to 65 and adjust the most age-dependent covariates to their expected values at age 65 when using Equation (4) to estimate the risk of contracting the disease by age 65. The Online Appendix describes how we adjust these age-dependent covariates for age.

Step 3. Identifying the individuals in the risk classes. To identify the individuals in the standard and high risk classes, we use the risk distribution we generated for Scenario 1. That distribution is based on the non-genetic covariates only, which is what insurers use to construct the risk classes. Individuals whose risk is between 0.75 times to 1.25 times the average population-wide risk are assigned to the standard risk class, and those whose risk is between 1.75 to 2.25 times that average risk are assigned to the high risk class.

Step 4. Estimating the implicit tax. For each scenario, we use the scenario's risk distribution from Step 2, limited to the individuals in the standard (or high) risk class. We compute Hendren's implicit tax measure for all percentiles of the distribution. Our main measure of the amount of selection that is likely to ensue is the minimum implicit tax up to the 80th percentile, which we denote t_{80}. We compare that measure with previous estimates for other market segments of the insurance industry.

5 Case Study: prostate cancer

We now apply the methodology to a detailed case study: CII for prostate cancer. We consider typical male CII buyers in young adulthood. Each buyer can buy a one-time insurance contract that pays a lump sum if he contracts prostate cancer by age 65.

The risk distributions for prostate cancer under Scenarios 1-3 is illustrated in the left panel of Figure 1, Section 1.1. Consider first Scenario 1, with only non-genetic covariates. The distribution is bimodal because family history is an important explanatory variable. However, there is still a significant amount of variation due to other covariates.

Overall, there is considerable variation in non-genetic risk. Therefore, there would be a significant amount of selection if insurers had to treat all consumers equally. In practice, insurers underwrite policies based on non-genetic covariates, classify consumers into the aforementioned risk classes, and price these differentially. As mentioned, we will focus on consumers in the standard risk class, shown in blue in Figure 1. Since insurers treat these equally, any variation in risk within the risk class effectively is private information. Thus, the blue area in Figure 1, Scenario 1, corresponds to the distribution of non-genetic private risk P for the standard risk class.

We now consider selection under the current prediction technology. Figure 1, Scenario 2, shows in blue the distribution of private risk R (the expectation of a loss conditional

on covariates and the current PGI) for the standard risk class consumers. The current PGI adds considerable explanatory power (it explains an additional 9.3% of the variance in liability in addition to the 23.6% already explained by $\mathbf{W}$). The distribution of private risk is more spread out and is now one-humped because the explanatory power of the coarse family history variable is mostly superseded by that of the fine-grained PGI. Thus, *even for the consumers in the standard risk class*, who look like average risks based on observable covariates, the current PGI introduces significant spread in private risk R. This suggests that, were these consumers to have access to their genetic predictions, there would be a non-negligible amount of selection.

The amount of selection grows considerably with future prediction technology. Figures 1 Scenarios 3L and 3U plot, in blue, the distribution of private risk for the standard risk class based on future-day PGIs. Scenario 3L uses the lower bound for the future PGI's R^2 based on prostate cancer's SNP heritability, and Scenario 3U uses the upper bound based on the twin heritability.

We can use the implicit tax to quantify the amount of selection and compare it to that in previously studied insurance markets. Figure 1, right panel, plots the implicit taxes in the four scenarios for consumers in the standard risk class. Following Hendren (2013), we focus on the minimum implicit tax up to the 80th percentile, t_{80}.

In Scenario 1, with the non-genetic covariates only, t_{80} is equal to 6.8%. This minimum occurs at the 80th percentile.[18] This means that a consumer at that risk percentile would have to pay 6.8% more than his actuarially fair price, if he were charged the actuarial fair price for the segment of consumers with equal or higher risk than his. All consumers at lower risk percentiles face higher implicit taxes. This level of t_{80} has been observed in many well functioning insurance markets (Hendren, 2013), consistent with insurers not using the non-genetic covariates to define narrower risk classes in underwriting.

In Scenario 2, with the current-day PGI (with an R^2_c of 11.2%—see Table 1), t_{80} is now 60.8%. This implies that adverse selection would already be significant with today's genetic prediction technology, were all consumers to have routine access to their genetic prediction. This situation is completely different from the current reality. Today, the majority of consumers have never been genotyped do not have detailed predictions of their risk of contracting prostate cancer and other diseases. Scenario 2 should be interpreted as the amount of selection in a counterfactual with the current prediction technology, but also with widely used genetic prediction.

In Scenario 3L, with a future PGI that explains as much as prostate cancer's SNP heri-

[18] In all four scenarios, the implicit tax is a strictly decreasing function of percentile, so the minimum up to the 80th percentile occurs exactly at the 80th percentile.

tability ($R^2_f = 18\%$), adverse selection increases considerably, with t_{80} now reaching 85.8%. And in Scenario 3U, with a future-day PGI with predictive power equal to the twin heritability ($R^2_f = 57\%$), adverse selection increases dramatically, with t_{80} reaching 451.7%. The implicit taxes in Scenarios 3L and 3U—which we see as lower and upper bounds for future genetic prediction technology—are considerably higher than what Hendren (2013) found even in market segments that had completely unraveled and that insurers refused to cover. This suggests the market for CII for prostate cancer may unravel if future genetic prediction technology becomes widely available.

The figures also plot the implicit tax at low risk percentiles, where the implicit tax explodes. The implicit tax at the 20th percentile—that for marginal consumers in the market segment that includes the most risky 80% of the consumers—is very high in all three scenarios with private genetic information. The reason is that, with future genetic prediction, there is a considerable mass of consumers who know they have virtually null risk of ever contracting prostate cancer. To cover most consumers, we need consumers with negligible risks to pay the price of an average risk. This division by a small number creates the exploding implicit tax.

We can summarize the results for this case study of prostate cancer as follows. With the current prediction technology, if all consumers were informed of their genetic predictions, there would already be a significant amount of selection. If tests were widely available, selection would at least be clearly noticeable to market participants. With future prediction technology, there would be a crippling amount of selection. It is unlikely that this type of insurance product could survive unravelling under the current genetic information ban for insurers.

6 Main results

We now discuss our results for all eight single-disease CII contracts, and for the bundled CII contract. Figure 2 reports our estimates of the minimum implicit tax up to the 80th percentile of risk (t_{80}) for each disease and under each scenario, and Table 4 reports detailed results. Though there are some important nuances, the big picture is broadly consistent with the results for prostate cancer. There are three main findings.

First, the level of selection under current genetic prediction technology and with widespread testing (Scenario 2) would be noticeable. The implicit tax t_{80} ranges from 14% for lung cancer to 85% for Alzheimer. The next lowest minimum implicit taxes after lung cancer are the ones for coronary artery disease and the bundled contracts, at around 20%, in the mid range of market segments where Hendren (2013) found no unraveling. The t_{80} for breast cancer

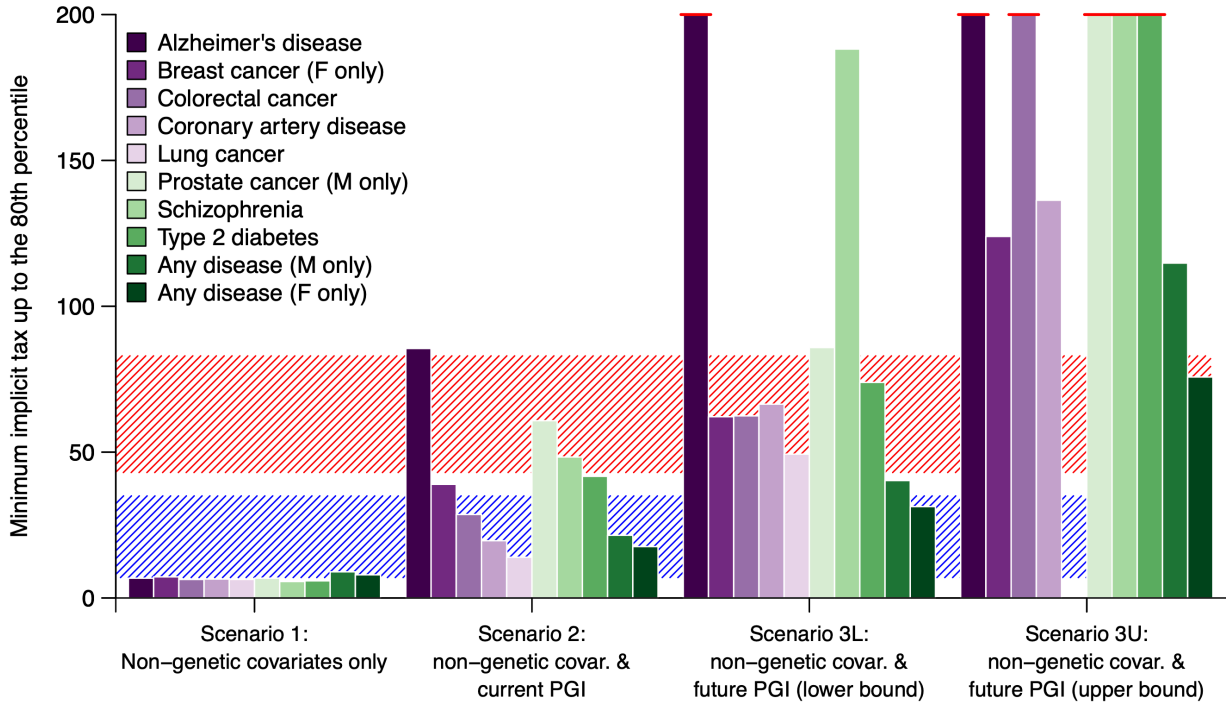


Figure 2: Minimum implicit tax up to the 80th percentile (t_{80}). The figure displays t_{80} within the standard risk class for each of the single-disease CII contracts, as well as the bundled contracts, in the four scenarios. The striped blue area corresponds to the range of t_{80} observed by Hendren (2013) in market segments that had not unraveled, and the striped red area corresponds to the range for market segments that had unraveled. The red horizontal lines for the tallest bars indicate that these bars' heights exceed the figure's.

and type 2 diabetes are between the ranges of no unraveling and unravelling. The contracts for the other three diseases—schizophrenia, prostate cancer, and Alzheimer's—have minimum implicit taxes that are all within or above the range observed in market segments that had unraveled. Again, though Scenario 2 assumes today's genetic prediction technology, it also assumes that all consumers have access to their genetic risk predictions, which does not reflect the situation today. These Scenario 2 results suggest that current prediction technology is already sufficiently accurate to make adverse selection noticeable, and that the main question is when genetic predictions will become widely available.

The second main finding is that, with the expected predictive power of future PGIs, adverse selection will become crippling for many contracts. In Scenario 3L—where PGIs' future R^2's are conservatively assumed to be equal to the diseases' SNP heritabilities—the minimum implicit tax t_{80} ranges from 31% for the bundled contract for females to >1,000% for Alzheimer's. Only the bundled contracts are below the level where Hendren found unravelling, and for males the implicit tax of 40% is close to the unraveling region. Some contracts have implicit taxes considerably higher than what Hendren (2013) found even in market seg-

ments that had unravelled. The detailed plots in the Online Appendix show that implicit taxes are high throughout the distribution. The raw histograms of private risk display considerable skewness even for consumers who are in the standard risk class. In Scenario 3U—in which future PGIs explain as much as the diseases' twin heritabilities, which we consider upper bounds—minimum implicit taxes are even higher, exceeding 100% for seven out of eight single-disease contracts, and exceeding 1,000% for three of the eight diseases. Even the bundled contracts display high implicit taxes of 76% (females) and 115% (males). These findings suggest that, with widespread testing and future prediction technology, many CII contracts will not be viable.

The third main finding is that there is substantial variation between the contracts. The minimum implicit taxes seem to be mainly driven by two factors: the predictive power of the PGI, and how much predictive power is additional to covariates (Table 4). To see this, consider the following examples in Scenario 3L. Alzheimer's and schizophrenia have the highest SNP h^2. Accordingly, they have the highest implicit taxes. In the opposite end of the spectrum, the bundled contracts and lung cancer have the lowest SNP h^2, the the lowest increase in R^2 when G_f is added to the model (Table 4, Scenario 3L, ΔR^2). Accordingly, these three contracts display the lowest implicit taxes.

7 Discussion and Extensions

7.1 Bundled Contracts

The bundled contracts are the most representative of typical CII contracts. For that reason, we now describe the bundled contract results in more detail. Figure 3 displays the risk distributions and implicit taxes for males. There are three main takeaways.

First, the broad pattern from the single-disease CII contracts holds. There is noticeable selection with the current PGI (Scenario 2), and serious, potentially crippling selection with future PGI depending on the assumed heritability (Scenarios 3L and 3U).

Second, bundled contracts display considerably less adverse selection than the single-disease CII. For example, under scenario 3L, implicit taxes are below Hendren's unravelling range for the bundled contracts, but in the unravelling range for all single-disease contracts. This suggests that bundling may be sufficient to keep selection to an acceptable level, and that this type of CII contract could be offered in a relatively well-functioning market even if genetic predictions are widely available. This seems to be especially true if the predictability of future PGIs is close to the SNP lower bound.

Third, the bundled contract for females is considerably less selected than that for males.

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

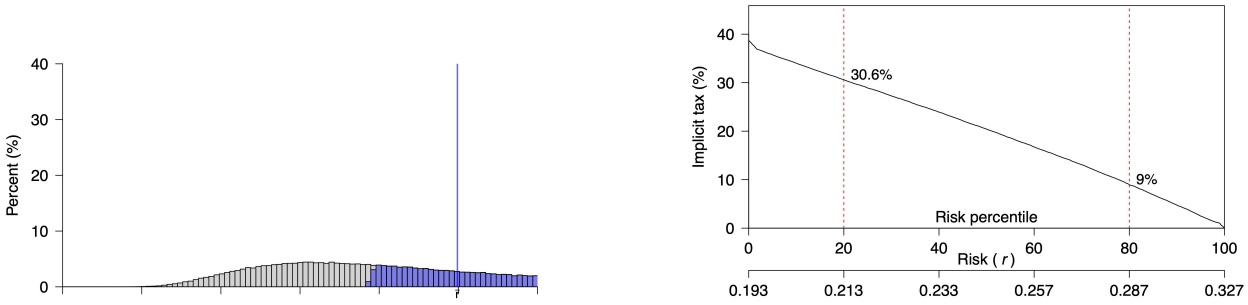
Table 4: Summary of main results

	ALZ	BRC	CAD	CRC	LUC	PRC	SCZ	T2D	anyDx	anyDx
Sex	M & F	F only	M & F	M & F	M & F	M only	M & F	M & F	F only	M only
Sample size	405,730	184,142	410,858	410,688	399,843	181,642	410,858	410,858	177,803	170,170
Cases	2,628	11,697	49,174	9,340	5,098	13,927	925	31,759	34,392	52,467
Panel A. Explanatory power of non-genetic covariates (**W**) and PGIs										
Scenarios 1-3										
R^2 of **W**	36.4%	7.7%	37.2%	12.0%	22.2%	23.6%	19.8%	40.0%	30.4%	42.5%
Scenarios 2										
R^2 of G_c	6.5%	6.8%	3.0%	2.5%	1.0%	11.2%	5.8%	8.6%	2.5%	3.7%
R^2 of G_c & **W**	41.3%	13.7%	38.4%	14.3%	22.6%	32.9%	23.1%	44.3%	31.6%	44.2%
ΔR^2 of G_c	4.9 pp	6.0 pp	1.2 pp	2.3 pp	0.3 pp	9.3 pp	3.4 pp	4.4 pp	1.2 pp	1.7 pp
Scenarios 3L										
R^2 of G_f	33.1%	14.0%	22.0%	9.0%	7.5%	18.0%	24.0%	19.6%	8.1%	12.5%
R^2 of G_f & **W**	63.2%	20.4%	48.0%	20.5%	26.3%	38.8%	35.9%	50.5%	34.4%	50.3%
ΔR^2 of G_f	26.8 pp	12.7 pp	10.8 pp	8.6 pp	4.0 pp	15.2 pp	16.1 pp	10.5 pp	4.0 pp	7.7 pp
Scenarios 3U										
R^2 of G_f	58.0%	31.0%	40.0%	40.2%	38.0%	57.0%	79.0%	72.0%	25.6%*	39.5%*
R^2 of G_f & **W**	86.3%	36.9%	61.6%	55.1%	–	73.6%	86.2%	91.6%	44.5%	77.8%
ΔR^2 of G_f	49.9 pp	29.2 pp	24.4 pp	43.2 pp	–	50.0 pp	66.5 pp	51.7 pp	14.1 pp	35.3 pp
Panel B. Risk of contracting disease (r) by age 65, conditional on non-genetic covariates (**W**) only										
Full analysis sample										
Mean	0.2%	5.7%	8.2%	1.8%	1%	5.3%	0.2%	5.7%	15.1%	26%
SD	0.1 pp	1.7 pp	8.9 pp	1.1 pp	1.9 pp	2.1 pp	1.1 pp	8.1 pp	10 pp	14.7 pp
Standard risk class (r = 0.75–1.25 of sample mean)										
Mean	0.1%	5.4%	8%	1.7%	1%	5%	0.2%	5.5%	14.4%	25%
SD	0.0 pp	0.7 pp	1.2 pp	0.2 pp	0.1 pp	0.8 pp	0.0 pp	0.8 pp	2.1 pp	3.7 pp
Panel C. Minimum implicit tax up to the 80th percentile of the disease risk distribution (t_{80})										
No individual underwriting by insurers (same premium for all)										
Scenario 1	75.1%	28.6%	114.8%	42.9%	206.5%	27.9%	389.4%	134.6%	79.3%	81.9%
Standard risk class (r = 0.75–1.25 of full sample mean)										
Scenario 1	6.8%	7.2%	6.5%	6.4%	6.4%	6.8%	5.6%	5.9%	7.9%	9.0%
Scenario 2	85.5%	39.0%	19.6%	28.6%	14.0%	60.8%	48.3%	41.7%	17.9%	21.6%
Scenario 3L	1060.6%	62.1%	66.4%	62.4%	49.3%	85.8%	188.2%	73.9%	31.3%	40.3%
Scenario 3U	>1,000%	123.9%	136.3%	404.7%	–	451.7%	>1,000%	>1,000%	75.8%	114.8%

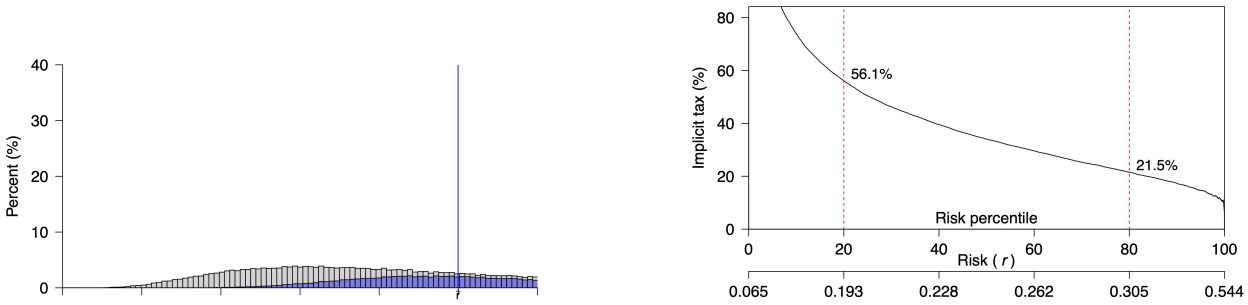
Notes: In Panel A, the reported R^2 are on the pseudo-R^2 on the liability scale (defined in the text). The R^2's of the non-genetic covariates and of the current PGI (G_c) are estimated in the data; in Scenarios 3L and 3U, the R^2 of the future PGI (G_f) is assumed to be equal to the disease's SNP heritability and twin heritability, respectively, and the joint R^2 of the future PGI and the covariates is obtained from our econometric model. The ΔR^2 of a PGI is its incremental predictive power, defined as the difference between the R^2 of the PGI and the covariates jointly and the R^2 of the covariates only. References for the heritability estimates are listed in Supplementary Table 1. The four scenarios are defined in the text.

Risk **Implicit Tax**

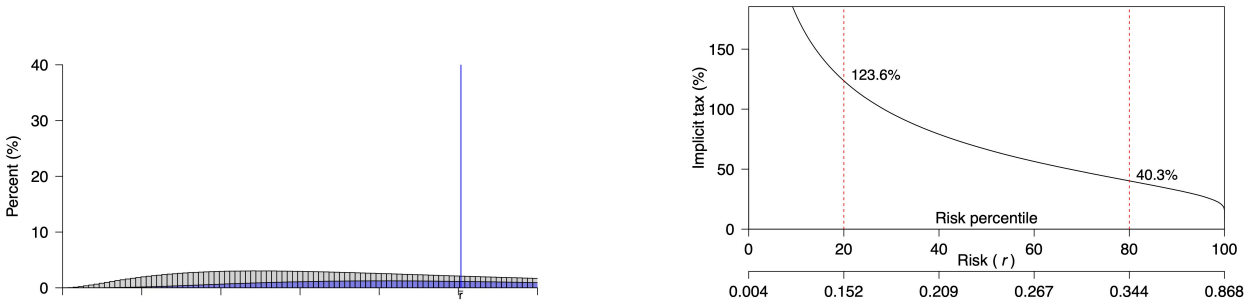
Scenario 1: non-genetic covariates



Scenario 2: non-genetic covariates and current PGI



Scenario 3L: non-genetic covariates and future PGI with lower-bound SNP heritability



Scenario 3U: non-genetic covariates and future PGI with upper-bound twin heritability

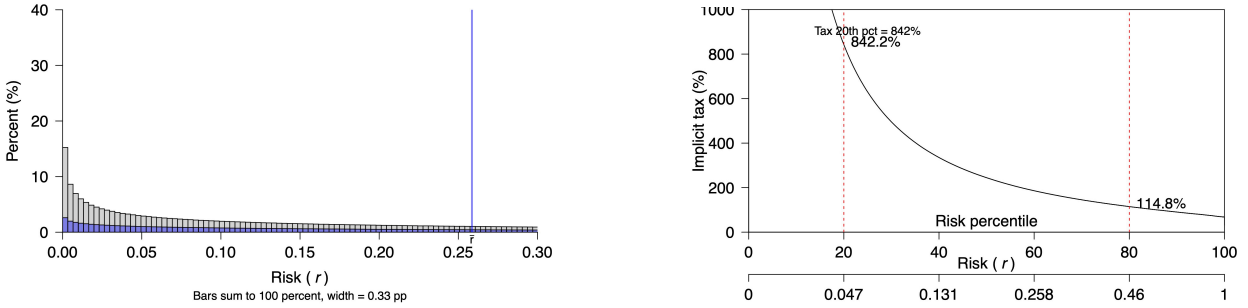


Figure 3: Bundled CII Contract: Risk Distribution and Implicit Tax

Notes: Left panel: distribution of the risk of contracting any of the eight diseases (anyDx) by age 65, conditional on various information scenarios and computed in the UKB data ($N = 170,170$). The standard risk class is shaded blue. The vertical blue line marks the average population risk ($\bar{r} = 0.25$). Right panel: implicit tax for consumers in the standard risk class. The sub-figures show the implicit tax, as a function of the consumer's risk.

The bundled contract for females is the only contract that is within the range of markets where Hendren found no unravelling in Scenario 3L. The difference between males and females is consistent with the higher heritability of common diseases that affect males disproportionately (prostate cancer and cardiovascular disease), versus the common disease that affects females (breast cancer). The difference also gives an example that the amount of selection in a given contract is subtle, and that even contracts that are superficially similar can have significant differences in risk distributions and hence in implicit taxes.

7.2 Incomplete Takeup of Testing

We reported results assuming that 100% of consumers will perform genetic testing and be aware of their probabilistic predictions. However, there is great uncertainty and broad debate about how common genotyping will be in the future. The modal expert opinion seems to be that genotyping will be extremely common, so that near 100% takeup is a reasonable scenario. The motivation for this standard view is the following. The costs of genotyping are already in the same order of magnitude as a blood test, and likely to go down further. Moreover, genotyping only has to be done once in a patient's lifetime. And, with improved prediction technology, the medical utility of genotyping will be so high that it is done for virtually the entire population. Consistent with the standard view, there are currently a large number of venture capital funded startups trying to offer genotyping and probabilistic predictions to consumers.

To understand how the results change under lower takeup, we calculate our results under different takeup rates. The analysis has three main takeaways. First, the amount of selection usually goes down with lower takeup. In particular, with very low takeup, as is the case today, there would not be much selection regardless of the available prediction technology. Second, under the standard view that takeup will be close to 100%, the results are similar to those reported in the main text.

The third main takeaway is that incomplete takeup generates kinks and nonmonotonicities in the implicit tax function. The main reason is that the risk distributions are not continuous. Consider the counterfactual where 50% of consumers in the standard risk class perform the genetic test. Then the histogram of risks is a mix of a distribution with support in the standard risk class (for untested consumers) and another histogram with full support (for tested consumers). The discontinuity induces kinks and non-monotone regions in the implicit tax.

7.3 Policy implications

Market failures under genetic information bans. Most developed countries ban insurers from using genetic information. Our findings suggest that, if genetic information becomes widely available, the bans will lead to a high degree of adverse selection in CII. In particular, the measured implicit taxes suggest complete unravelling of the markets for most of the single disease contracts we studied.

On the demand side, genetic information is also likely to make this market failure larger and more noticeable. For example, our results for prostate cancer imply that between 5.5% and 10% of the standard risk class consumers will learn from future genetic tests that their risk of contracting prostate cancer by age 65 is over 15%, and between 1% and 6% will learn that their risk is over 25%. Many of these consumers would rationally demand exactly the kind of CII that genetic testing will make difficult to provide.

Optimal Policy Responses. While genetic predictions are new, selection is a standard issue in insurance markets. There is a large amount of research and practical experience on how to regulate selection markets to address the market failures due to asymmetric information (Einav et al., 2010b, 2021, 2023). The problem of regulating genetic information in insurance is similar. Therefore, we can draw on extensive research and practical experience from other selection markets.

Appendix **??** gives a brief overview of the literature and of standard policy instruments. In the case of CII, a natural policy is a combination of community rating, subsidies, and risk adjustment. Community rating are rules that limit price discrimination. The current ban on genetic underwriting is a community rating regulation. Community rating improves redistribution across consumers, but generates adverse selection. Some level of community rating is desirable, to help high-risk consumers, but perhaps less than a full ban on genetic price discrimination, due to the large amount of adverse selection. Likewise, subsidies might be desirable to increase efficiency and help vulnerable populations, at the cost of spending tax revenue. Finally, risk adjustment are cross subsidies to insurers based on the populations they enroll. Risk adjustment is an important policy to guarantee high-quality insurance offerings in settings with adverse selection. See Appendix **??** for further discussion.

Political Economy Issues. Improving on the status quo policy of banning insurers from using genetic information is politically complex. For example, the existing bans on insurers' use of genetic testing is aimed at helping consumers. However, if a market such as CII unravels, all consumers will be worse off, because they will not get any insurance at all. A law that allows insurers to underwrite based on genetic information would benefit all

consumers—including the high risk ones—because they would then at least be able to get some insurance. However, high-risk consumers may be upset about the high premiums they would have to pay. Thus, even consumers who would benefit from the law would probably oppose it after the fact.

Policies specific to genetic testing. Some existing policy proposals are specifically aimed at regulating genetic information in insurance. The most well-known is Tabarrok's (1994) idea for genetic insurance. Tabarrok (1994) proposes a mandatory insurance contract to anyone taking a genetic test for the first time. The proposed insurance would calculate, for each consumer, some expected loss due to the makeup of their genes. Tabarrok (1994) shows that, under certain assumptions, this policy is second-best optimal.

Interestingly, our risk histograms suggest that the premiums for Tabarrok's proposal might be large. Consider a Tabarrok contract that values incurring prostate cancer at \$100,000. Before taking a genetic test, each consumer would have to deposit ∼\$5,000, based on the average probability of contracting prostate cancer by age 65 (and neglecting interest rates). Moreover, with the future prediction technology, a fair number of these consumers would have a negligible chance of contracting prostate cancer, and therefore would get almost none of their deposit back after taking the genetic test. The premiums will be much higher if the Tabarrok contract includes many diseases. The main reason for these large numbers is that many diseases are polygenic, and that PGIs have high and increasing predictive power. In contrast, at the time Tabarrok (1994) developed the idea, it was thought that genetic testing would involve only a handful of rare monogenic diseases for which genetic tests give either a positive or negative result. For example, Huntington's disease affects less than 1 in 10,000 people. So the deposit would be less than ∼\$10. Most consumers would get the deposit back, and the few unlucky consumers would receive the ∼\$100,000 lump sum.

While genetic insurance is an interesting idea in such a setting, it would be challenging to implement it in a setting with common polygenic diseases and continuous PGIs. It may be politically fraught to ask consumers with healthy genes to foot the bill for all the premiums being paid. This will be especially unfair to the consumers who have a low genetic predisposition to serious diseases, but eventually have the bad luck of developing a serious disease like cancer. These consumers will both contract a serious disease and have get no payout from genetic insurance. Moreover, it would be difficult to agree on a set of diseases to be covered and on the covered amounts.

Another possibility is to consider industry-led solutions. For example, one natural idea is to allow consumers to disclose that they have not taken a genetic test at the time of purchasing a CII policy. This would allow at least some consumers to purchase CII with little genetic adverse selection. Whether this idea would be feasible depends on overcoming

two main practical hurdles. The first is that it must be detectable and enforceable to void policies of consumers who get tested in secret. The second is that strategic lapsation does not lead to too much adverse selection.

7.4 Advantageous Selection

So far, we measured selection by looking exclusively at the distribution of risk in the population. However, in standard equilibrium models, selection depends both on heterogeneity in risk and in how risk correlates with insurance demand (Akerlof, 1970; Azevedo and Gottlieb, 2017). de Meza and Webb (2001) shows that it is possible to have advantageous selection if, for example, low-risk consumers also have higher demand for insurance, because of risk aversion for example. Fang et al. (2008) find clear evidence of advantageous selection in Medigap insurance. And Finkelstein and McGarry (2006) find that demand for long-term care insurance varies on dimensions unrelated to risk. Therefore, non-risk dimensions of private information can be important, and in principle could affect our conclusions.

Fortunately, the UKB includes a survey measure of risk aversion. Appendix XXX reports the correlation of risk aversion with the diseases we study, the bundle of diseases, and their PGIs. In all cases, we find extremely low correlations. This suggests that it is unlikely that there exists a negative strong correlation between risk and risk aversion. So it is unlikely that such correlations could have a large impact on our main findings of significant levels of selection in Scenario 3.

More generally, the high levels of selection that we find make it somewhat unlikely that non-risk determinants of insurance demand are sufficiently negatively correlated with risk to overturn the main qualitative findings. That said, the UKB risk aversion measure is a very limited proxy for insurance demand, so it is always possible that other dimensions of private information would significantly dampen the amount of selection. We also stress that our results are specific for the CII contracts we study. It is plausible that other dimensions of private information reduce the amount of genetic selection in different insurance markets.

7.5 Changes in Consumer Behavior after Testing

One of the main potential benefits of genetic testing is preventive medicine. The idea is that genetic tests should be used to treat patients who are prone to a disease even before the disease manifests itself. A particularly active area today is preventive cardiology.

This kind of change in behavior would, in principle, reduce the amount of selection due to genetics. Consider a single-disease CII contract. If we assume that patients with a higher PGI for the disease take more preventive actions that reduce the disease probability, we will

find less selection.

In practice, this effect is unlikely to substantively affect our main findings. We find very high levels of selection for most contracts. So, given plausible assumptions about the takeup of preventive actions and the effectiveness of these actions, this is unlikely to reduce selection so much as to overturn the main findings. This is especially true of cancers because known preventive actions have limited efficacy.

Another point in this discussion is that genetic tests are potentially an extremely valuable technology. Understanding the unintended consequences of this technology is important because society can mitigate any harms, while still taking advantage of the benefits of the technology.

7.6 Firms with an Informational Advantage

There are a few theoretical papers about insurance markets where insurers have private information.[19] Genetic testing could make such theoretical assumptions plausible in some institutional settings. For example, consider a future where insurers are allowed to underwrite based on genetics, tests are widely available, but consumers do not have the sophistication to interpret the tests. This would correspond to a model where insurers receive private information but consumers do not.

The simplest and most robust prediction of this type of model is in the competitive case. If insurers just compete their profits to zero, they will offer actuarially fair insurance to each consumer. Thus, even if consumers couldn't interpret the genetic tests, they will be able to assess their risk by getting insurance quotes. Thus, it is irrelevant whether consumers can or cannot interpret the genetic tests. Equilibrium is the same regardless. Equilibrium is efficient, in the sense that each consumer is charged her true cost for insurance. But equilibrium can be seen as very unfair, as a consumer who is unlucky in the genetic lottery will face high prices.

Interestingly, assuming that the level of selection is so high that a ban on testing would lead to complete unravelling, then allowing insurers to do genetic underwriting benefits all consumers. Nevertheless, a planner who cares about redistribution can likely do better with some community rating and subsidies (see Section 7.3).

[19] See Brunnermeier et al. (2021) and references therein.

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

8 Conclusion

We developed a methodology to measure adverse selection in insurance markets due to genetic information. There are three key innovations to our approach. First, we use large-scale genetic and loss micro-data as inputs. Second, we consider genetic prediction based on PGIs, which depend on many common genetic variants, as opposed to some rare mutations. Whereas rare mutations can have large effects, they are too rare to generate substantial adverse selection. Third, our methodology uses heritability estimates as reasonable bounds for the future predictive power of genetic prediction, which allows us to estimate how selection will change as prediction technology improves.

We apply our methodology to CII contracts using data from the UKB. We find that, if genetic testing becomes widespread, there would be noticeable selection with the current prediction technology. The amount of selection would be crippling with the expected future prediction technology.

We also highlight some limitations to our findings. First, our methodology ignores lapsation and interest rates: it assumes consumers can buy one-time contracts at a young age, whereas in reality consumer can renew their contracts and must pay premiums annually until they contract a covered disease. Second, our methodology does not take into account heterogeneity in risk preferences and insurance demand. It would be interesting to measure selection in equilibrium models that take these into account. Third, our counterfactual scenarios assume that genetic testing will be commonplace and that test results will be widely available and understood. Section 7.2 shows that there should be less selection with lower takeup of genetic testing.

Future work should seek to address these limitations and to apply our methodology to larger markets like life, health, and long-term care insurance.

References

23andMe. Information about Genetic Health Risk Reports - 23andMe. Available online, 2023.

Abdellaoui, A., Yengo, L., Verweij, K. J., and Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics*, 2023.

Adams, C., Donnelly, C., and Macdonald, A. The impact of known breast cancer polygenes on critical illness insurance. *Scandinavian Actuarial Journal*, 2015(2):141–171, 2015.

Akerlof, G. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970.

Allied Market Research. Critical illness insurance market, 2022.

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–58, 2010.

Azevedo, E. M. and Gottlieb, D. Perfect competition in markets with adverse selection. *Econometrica*, 85(1):67–105, 2017.

Bastarache, L. Using phecodes for research with the electronic health record: From PheWAS to PheRS. *Annual Review of Biomedical Data Science*, pages 1–19, 2021.

Beauchamp, J. P., Cesarini, D., Johannesson, M., van der Loos, M. J., Koellinger, P. D., et al. Molecular genetics and economics. *Journal of Economic Perspectives*, 25(4):57–82, 2011.

Bélisle-Pipon, J.-C., Vayena, E., Green, R. C., and Cohen, I. G. Genetic testing, insurance discrimination and medical research: what the united states can learn from peer countries. *Nature Medicine*, 25(8):1198–1204, 2019.

Benjamin, D. J., Cesarini, D., Chabris, C. F., Glaeser, E. L., Laibson, D. I., et al. The promises and pitfalls of genoeconomics. *Annu. Rev. Econ.*, 4(1):627–662, 2012.

Brackenridge, R. D. C., Croxson, R. S., and Mackenzie, R., editors. *Medical selection of life risks*. Palgrave Macmillan, New York, fifth edition, 2006.

Brunnermeier, M. K., Lamba, R., and Segura-Rodriguez, C. Inverse selection. *Available at SSRN 3584331*, 2021.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Canadian Institute of Actuaries. Critically canadian: Canadian critical illness standalone base incidence tables. *Mimeo*, 2012.

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

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS One*, 3(10):e3395, 2008.

de Meza, D. and Webb, D. C. Advantageous selection in insurance markets. *RAND Journal of Economics*, 32(2):249–262, 2001.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, pages 1102–1111, 2013.

Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9: e1003348, 2013.

Einav, L., Finkelstein, A., and Cullen, M. R. Estimating welfare in insurance markets using variation in prices. *The Quarterly Journal of Economics*, 125(3):877–921, 2010a.

Einav, L., Finkelstein, A., and Levin, J. Beyond testing: Empirical models of insurance markets. *Annu. Rev. Econ.*, 2(1):311–336, 2010b.

Einav, L., Finkelstein, A., and Mahoney, N. The io of selection markets. In *Handbook of Industrial Organization*, volume 5, pages 389–426. Elsevier, 2021.

Einav, L., Finkelstein, A., and Fisman, R. *Risky Business: Why Insurance Markets Fail and What to Do About It*. Yale University Press, 2023.

European Commission. Genome-wide association studies, polygenic scores and social science genetics: Overview and policy implications. Technical Report JRC117414, 2019.

Falconer, D. S. and Mackay, T. F. *Quantitative Genetics*. Longman London, 1983.

Fang, H., Keane, M. P., and Silverman, D. Sources of advantageous selection: Evidence from the medigap insurance market. *Journal of Political Economy*, 116(2):303–350, 2008.

Finkelstein, A. and McGarry, K. Multiple dimensions of private information: Evidence from the long-term care insurance market. *American Economic Review*, 96(4):938–958, 2006.

Gatzert, N. and Maegebier, A. Critical illness insurances: Challenges and opportunities for insurers. *Risk Management and Insurance Review*, 18(2):255–272, 2015.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10:1776, 2019.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian data analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

Golinghorst, D., De Paor, A., Joly, Y., Macdonald, A. S., Otlowski, M., et al. Anti-selection & genetic testing in insurance: An interdisciplinary perspective. *Journal of Law, Medicine & Ethics*, 50(1):139–154, 2022.

Gottlieb, D. and Smetters, K. Lapse-based insurance. *American Economic Review*, 111(8): 2377–2416, 2021.

Hendren, N. Private information and insurance rejections. *Econometrica*, 81(5):1713–1762, 2013.

Hendren, N. Knowledge of future job loss and implications for unemployment insurance. *American Economic Review*, 107(7):1778–1823, 2017.

Hill, W. G., Goddard, M. E., and Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genetics*, 4(2):e1000008, 2008.

Holland, D., Frei, O., Desikan, R., Fan, C.-C., Shadrin, A. A., et al. Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate gaussian mixture model. *PLOS Genetics*, 16(5):e1008612, 2020.

Howard, R. C. Genetic testing model for ci: if underwriters of individual critical illness insurance had no access to known results of genetic tests. *Canadian Institute of Actuaries Report to CIA Research Committee*, 2016.

Joly, Y., Burton, H., Knoppers, B. M., Feze, I. N., Dent, T., et al. Life insurance: Genomic stratification and risk classification. *European Journal of Human Genetics*, 22:575–579, 2014.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, 2018.

Klarin, D. and Natarajan, P. Clinical utility of polygenic risk scores for coronary artery disease. *Nature Reviews Cardiology*, 19(5):291–301, 2022.

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. A better coefficient of determination for genetic profile analysis. *Genetic Epidemiology*, 36(3):214–224, 2012.

Loos, R. J. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, 11(1):1–3, 2020.

Macdonald, A. and Tapadar, P. Multifactorial genetic disorders and adverse selection: Epidemiology meets economics. *Journal of Risk and Insurance*, 77(1):155–182, 2010.

Macdonald, A. and Yu, F. The impact of genetic information on the insurance industry: Conclusions from the 'bottom-up' modelling programme. *ASTIN Bulletin*, 41(2):343–376, 2011.

Macdonald, A. S. and McIvor, K. R. Modelling adverse selection in the presence of a common genetic disorder: the breast cancer polygene. *ASTIN Bulletin*, 39(2):373–402, 2009.

Macdonald, A. S., Waters, H. R., and Wekwete, C. T. A model for coronary heart disease and stroke with applications to critical illness insurance underwriting I: The model. *North American Actuarial Journal*, 9(1):13–40, 2005.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., et al. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

Maxwell, J. M., Russell, R. A., Wu, H. M., Sharapova, N., Banthorpe, P., et al. Multifactorial disorders and polygenic risk scores: predicting common diseases and the possibility of adverse selection in life and protection insurance. *Annals of Actuarial Science*, 15(3): 488–503, 2021.

McKelvey, R. D. and Zavoina, W. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4(1):103–120, 1975.

Mills, M. C. and Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nature Genetics*, 52(3):242–243, 2020.

MyHeritage. Understanding Your Polygenic Risk Reports. Available online, 2021.

O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., et al. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105:456–476, 2019.

Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., et al. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*, 54(4):437–449, 2022.

Oster, E., Shoulson, I., Quaid, K., and Dorsey, E. R. Genetic adverse selection: Evidence from long-term care insurance and huntington disease. *Journal of Public Economics*, 94 (11-12):1041–1050, 2010.

Polderman, T. J. C., Benyamin, B., de Leeuw, C., Sullivan, P. F., van Bochoven, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7):702–709, 2015.

Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, 109(1):12–23, 2022.

Rechfeld, F., Rischatsch, M., Nemoto, T., and Nabholz, C. Can life insurance pass the genetic test? Technical report, Swiss Re Institute, Zurich, 2019.

Saya, S., Boyd, L., Chondros, P., McNamara, M., King, M., et al. The SCRIPT trial: study protocol for a randomised controlled trial of a polygenic risk score to tailor colorectal cancer screening in primary care. *Trials*, 23(1):810, 2022.

Simcoe, M., Valdes, A., Liu, F., Furlotte, N. A., Evans, D. M., et al. Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. *Science Advances*, 7(11):eabd1239, 2021.

So, H.-C. and Sham, P. C. A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained. *PLoS Genetics*, 6(12):e1001230, 2010.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015.

Tabarrok, A. Genetic testing: an economic and contractarian analysis. *Journal of Health Economics*, 13(1):75–91, 1994.

Taylor Jr, D. H., Cook-Deegan, R. M., Hiraki, S., Roberts, J. S., Blazer, D. G., et al. Genetic testing for Alzheimer's and long-term care insurance. *Health Affairs*, 29(1):102–108, 2010.

Widén, E., Junna, N., Ruotsalainen, S., Surakka, I., Mars, N., et al. How communicating polygenic and clinical risk for atherosclerotic cardiovascular disease impacts health behavior: an observational follow-up study. *Circulation: Genomic and Precision Medicine*, 15 (2):e003459, 2022.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., et al. Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(7):507–515, 2013.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, 88:76–82, 2011.

Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., et al. A saturated map of common genetic variants associated with human height. *Nature*, 610:704–712, 2022.

Zeggini, E., Gloyn, A. L., Barton, A. C., and Wain, L. V. Translational genomics and precision medicine: Moving from the lab to the clinic. *Science*, 365(6460):1409–1413, 2019.

Zhou, W., Kanai, M., Wu, K.-H. H., Rasheed, H., Tsuo, K., et al. Global biobank meta-analysis initiative: Powering genetic discovery across human disease. *Cell Genomics*, 2 (10), 2022.

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

APPENDIX

A Theory

A.1 Definition of the Implicit Tax

Consider an insurance contract that pays \$1 if a loss happens, and otherwise charges a premium. For the contract to break even with a consumer with risk r, the premium has to be

$$\frac{r}{1-r}.$$

Consider now the case where this policy is purchased by all consumers with risk r or greater. In that case, for the policy to break even the premium would have to be

$$\frac{\mathbb{E}[R|R \geq r]}{1-\mathbb{E}[R|R \geq r]}.$$

The consumer with risk r would have to pay $T(r)$ times the actuarially fair price, where

$$T(r) := \frac{\mathbb{E}[R|R \geq r]}{1-\mathbb{E}[R|R \geq r]} \Big/ \frac{r}{1-r}.$$

Hendren (2013) calls $T(r)$ the "pooled price ratio". To make exposition more intuitive, we define the **implicit tax** $t(r)$ as

$$1 + t(r) := T(r). \tag{A.1}$$

So $t(r)$ is the implicit tax that a risk r consumer would have to pay due to selection if she were charged an actuarially fair price for all higher-risk consumers.

A.2 Proof of Theorem 1

Here we prove that the model is identified under Assumptions (1)-(5) (Theorem 1). The proof uses the following Lemma.

Lemma 1. *Conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$, G_f is normally distributed with mean*

$$ag_c + b\mathbf{w}\theta$$

and variance c^2*, where*

$$a = \frac{\frac{1}{\sigma_\epsilon^2}}{\frac{1}{\sigma_\epsilon^2} + \frac{1}{\sigma_V^2}},$$
$$b = \frac{\frac{1}{\sigma_V^2}}{\frac{1}{\sigma_\epsilon^2} + \frac{1}{\sigma_V^2}},$$
$$\frac{1}{c^2} = \frac{1}{\sigma_\epsilon^2} + \frac{1}{\sigma_V^2}.$$

Proof. By Assumption 2, conditional on $\mathbf{W} = \mathbf{w}$, $G_f = \mathbf{w}\theta + V$, with $V \sim N(0, \sigma_V^2)$. And by Assumption 3, $G_c = G_f + \epsilon$, with $\epsilon \sim N(0, \sigma_\epsilon^2)$. Lemma 1 then follows from the formula in Gelman et al. (2013, p. 40, Equation 2.10). This is a standard formula from Bayesian statistics. □

Proof of Theorem 1. We will show that the parameters $\Gamma = (\mathbb{P}_{\mathbf{W}}, \beta, \theta, \sigma_V^2, \sigma_\epsilon^2)$ are functions of $\mathbb{P}_{\text{data}}$. $\mathbb{P}_{\mathbf{W}}$ is trivial because $\mathbf{W}$ is observed. We now turn to the other parameters.

Part 1: θ

Assumption 4 implies that

$$\mathbb{E}[G_c|\mathbf{W} = \mathbf{w}] = \mathbf{w}\theta. \tag{A.2}$$

Assumption 1 implies that there is a set of $\mathbf{w}$'s in the support of $\mathbb{P}_{\mathbf{W}}$ that span the entire space. So θ is the vector of coefficients of a regression of G_c on $\mathbf{W}$.

Part 2: σ_ϵ^2

By Assumption 4, there exists a latent variable L for disease such that

$$L := G\beta_g + w\beta_w + \eta,$$

where $\eta \sim N(0, 1)$, $\eta \perp (\boldsymbol{W}, V, \epsilon)$, and $D = \{L > 0\}$.

Let the pseudo-R^2 of a probit regression of D on the current PGI be R_c^2. R_c^2 is a function of the data. R_c^2 equals

$$R_c^2 = \frac{\text{Cov}[L, G_c]^2}{\text{Var}[G_c]\text{Var}[L]}.$$

The pseudo-R^2 of a probit regression of D on the future PGI is

$$R_f^2 = \frac{\text{Cov}[L, G_f]^2}{\text{Var}[G_f]\text{Var}[L]}.$$

Therefore, using the fact that $\text{Cov}[L, G_f] = \text{Cov}[L, G_c]$ and $\text{Var}[G_c] = 1$,

$$\frac{R_c^2}{R_f^2} = \frac{\text{Var}[G_f]}{\text{Var}[G_c]} = \text{Var}[G_f].$$

Assumption 3 implies that

$$\text{Var}[G_c] = \text{Var}[G_f] + \sigma_\epsilon^2.$$

Plugging in the formula above we have

$$\sigma_\epsilon^2 = 1 - \frac{R_c^2}{R_f^2}. \tag{A.3}$$

Part 3: σ_V^2

Assumptions 3 and 2 yield

$$G_c = \mathbf{W}\theta + V + \epsilon.$$

Taking the variance we find

$$1 = \text{Var}[\mathbf{W}\theta] + \sigma_V^2 + \sigma_\epsilon^2,$$

so that

$$\sigma_V^2 = 1 - \text{Var}[\mathbf{W}\theta] - \sigma_\epsilon^2. \tag{A.4}$$

Part 4: β_g

By Assumption 4, the probability of a loss is

$$\Phi(G_f\beta_g + \mathbf{W}\beta_w).$$

Equivalently,

$$D = \{G_f\beta_g + \mathbf{W}\beta_w + \eta > 0\},$$

where η is a standard normal random variable orthogonal to $\mathbf{W}, V, \epsilon$. Consider the distribution of D conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$. Lemma 1 implies that

$$G_f = ag_c + b\mathbf{w}\theta + c\nu,$$

where ν is a standard normal random variable independent of η.

Therefore, conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$,

$$\begin{aligned} D &= \{(ag_c + b\mathbf{w}\theta + c\nu)\beta_g + \mathbf{w}\beta_w + \eta > 0\} \\ &= \{g_c a\beta_g + \mathbf{w}(b\theta\beta_g + \beta_w) + \eta + c\nu\beta_g > 0\} \\ &= \left\{ g_c \frac{a\beta_g}{\sqrt{1 + c^2\beta_g^2}} + \mathbf{w}\frac{b\theta\beta_g + \beta_w}{\sqrt{1 + c^2\beta_g^2}} + \frac{\eta + c\nu\beta_g}{\sqrt{1 + c^2\beta_g^2}} > 0 \right\}. \end{aligned}$$

The last term is distributed as $N(0, 1)$. Therefore

$$\Pr(D = 1 | G_c = g_c, \mathbf{W} = \mathbf{w}) = \Phi\left[g_c \frac{a\beta_g}{\sqrt{1 + c^2\beta_g^2}} + \mathbf{w}\frac{b\theta\beta_g + \beta_w}{\sqrt{1 + c^2\beta_g^2}} \right]. \tag{A.5}$$

Let γ be the coefficient vector of a probit regression of D on G_c and $\mathbf{W}$. Then

$$\frac{a\beta_g}{\sqrt{1+c^2\beta_g^2}} = \gamma_g.$$

This equation implies that β_g has the same sign as γ_g. Taking the square of both sides and solving, we find

$$\beta_g = \frac{\gamma_g}{\sqrt{a^2 - \gamma_g^2 c^2}}. \tag{A.6}$$

This result implies that $\gamma_g < a/c$.

Part 5: β_w

By equation (A.5) and using the fact that the support of $\mathbf{W}$ includes a set that spans the entire space (Assumption 1),

$$\frac{b\theta\beta_g + \beta_w}{\sqrt{1+c^2\beta_g^2}} = \gamma_w.$$

Then

$$\beta_w = \gamma_w\sqrt{1+c^2\beta_g^2} - b\theta\beta_g. \tag{A.7}$$

□

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

Genetic Prediction and Adverse Selection

ONLINE APPENDIX

Eduardo Azevedo* Jonathan Beauchamp† Richard Karlsson Linnér‡

May 8, 2024

*The Wharton School, University of Pennsylvania. Email: eazevedo@wharton.upenn.edu

†Interdisciplinary Center for Economic Science and Department of Economics, George Mason University. Email: jonathan.pierre.beauchamp@gmail.com

‡Department of Economics, Leiden University: Email: r.karlsson.linner@law.leidenuniv.nl
School of Business and Economics, Vrije Universiteit Amsterdam: Email: r.karlssonlinner@vu.nl

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

A Background on the genetics of complex disease

This study investigates common medical conditions that develop from the interplay of many genes and the environment, which genetic epidemiologists refer to as complex (or non-Mendelian) disease (Schork, 1997). Scientific advances have established that complex diseases are typically polygenic, i.e., there is a very large number of genetic variants involved and most have only a small or tiny individual effect (Visscher et al., 2021). However, the total genetic effect—the heritability—is often substantial and it explains, on average, about 50% of the variation in disease liability (Polderman et al., 2015). This knowledge implies that the risk of a complex disease cannot be predicted well from only a smaller and more easily characterized set of genes. Instead, a polygenic liability threshold model that measures a large number of genetic variants together with environmental variables is now considered the more appropriate model (Chatterjee et al., 2016).

To understand how genetic data can be used to predict a person's risk of a complex disease, recall that the human genome is a roughly 3-billion bases long sequence encoded in DNA (NHGRI, 2023). This sequence consists of four possible nucleotide bases: adenine ("A"), cytosine ("C"), guanine ("G"), and thymine ("T"). The two complementary strands that make up the DNA molecule are connected by bonding base pairs, like the rungs on a ladder, where A always bonds with T, and C with G. This rule means that the information content of both strands can be recorded perfectly from a single strand (e.g., the molecularly distinguishable "forward" strand).

Each parent passes on stretches of DNA organized as 22 chromosomes and a sex chromosome, which come together as 23 chromosome pairs in the offspring. Below, we exclude the sex chromosomes because most large genetic association studies omit these from analysis (Sun et al., 2023). This decision will not impact the results much because the sex chromosomes generally contribute the least to the heritability of complex traits and disease (<5%) (Visscher et al., 2006). Thus, each genomic position along the sequence on the (non-sex) chromosome pairs consists of two base pairs, one that gets passed on by the mother (e.g., A on the forward strand, and T on the reverse strand) and one by the father (e.g., G and C on the forward and reverse strand, respectively).

The nucleotides at most positions in the genome are identical for everyone (NHGRI, 2023). Nucleotides that do vary are called single-nucleotide polymorphisms (SNPs, "snips"), which can be viewed as the smallest unit of genetic variation.[1] Most SNPs are bi-allelic, meaning that only two possible nucleotide bases ("alleles") have ever been observed on the forward strand in human populations (in our example, A and G are the two possible alleles).[2] The two alleles are denoted as the major or minor based on their frequencies. SNPs with a minor allele frequency (MAF) greater than 1% are often described as "common SNPs", while SNPs with MAF <1% are described as "rare". About twenty million common SNPs have been observed in the global human population (Auton et al., 2015), which thus concerns <1% of all nucleotides in the genome.

[1] The SNPs tag variation within and outside of genes (protein-coding regions) and different types of structural variation (Sudmant et al., 2015), such as copy number variants (CNVs), but we ignore these. Focusing on the SNPs alone captures the bulk of the heritability of complex traits (Yengo et al., 2022).

[2] We exclude SNPs with more than two possible alleles from our analyses; multi-allelic SNPs are the exception.

To create a variable that captures the most relevant information at a given SNP, geneticists arbitrarily select one of the two possible nucleotides, label it as the reference allele, and then sum the number of non-reference (or "effect-coded") alleles; a SNP variable suitable for regression analysis thus takes the values 0, 1, or 2. Continuing our above example, if we select the A nucleotide to be the reference allele, then this person would be coded as having 1 effect-coded (G) allele, while the combinations A/A and G/G would be coded as 0 and 2, respectively. For more details, see Beauchamp et al. (2011), Benjamin et al. (2012), or Uffelmann et al. (2021).

B Details on the polygenic indexes (PGIs)

The genome-wide association study (GWAS) approach tests millions of SNPs, one-by-one, for statistical association with a trait or disease of interest. This relatively simple approach has proven much successful in discovering replicable SNP-associations. Large-scale GWAS are now uncovering novel SNP-associations at an unprecedented rate (Tam et al., 2019). To achieve sufficient sample size, the largest published GWAS are usually a meta-analysis of results derived in multiple biobanks and many smaller contributing cohort-level datasets. The GWAS regression coefficients (or effect sizes) for a set of SNPs can be used as weights to compute genetic predictor variables out-of-sample called polygenic indexes (PGIs).[3] Many experts predict that PGIs will play a major role in the future of genetic screening and precision medicine (Torkamani et al., 2018).

Here, we computed PGIs using a well-documented set of about 1 million SNPs, specifically the SNPs covered by the reference map of the International HapMap 3 Consortium (Altshuler et al., 2010). This is a standard set of SNPs used in the literature on PGI (Becker et al., 2021), and this set is known to capture the majority of the heritability attributable to common SNPs (MAF >1%) in populations of European (EUR) ancestry (Ge et al., 2019; Yengo et al., 2022). Our protocol for computing PGIs aligns with recent studies, e.g., Becker et al. (2021) or Ge et al. (2019), and with recent scientific guidelines (see, e.g., Figs 1–2. in Choi et al., 2020).

For each individual in the data that passed our sample quality-control procedure (described below), the PGI for disease D (e.g., breast cancer) is the weighted sum of the number of effect-coded alleles over a set of SNPs, with each allele weighted by its corresponding GWAS coefficient. In notation, the PGI (G) for a disease D is defined as:

$$G = \mathbf{X}\,\widehat{\mathbf{b}},$$

where $\mathbf{X}$ is a row vector with the number of effect-coded nucleotides of each SNP and $\widehat{\mathbf{b}}$ is a column vector with each SNP's GWAS coefficient on D. The next section describes how we searched and gathered GWAS "summary statistics" (i.e., regression coefficients, standard errors, and other per-SNP information), to use as PGI weights. Importantly, these weights must be estimated in data independent of the UK Biobank. Below, we apply PRS-CS to adjust the weights for the fact that GWAS do not fit the SNPs jointly in the same regression model.

[3] Other commonly seen names include "genetic/genomic/polygenic (risk) scores".

Because there is a lack of well-powered GWAS in non-European populations (Ruan et al., 2022), we restricted our analyses to European ancestry (EUR) samples. Nevertheless, all regressions below were adjusted for the possibility of population stratification by including the first ten genetic principal components (PCs) as control variables (Choi et al., 2020). We used the standard set of genetic PCs that are distributed with the genetic data release (Data Category 22009).

B.1 Details on the GWAS, quality control (QC), and meta-analysis

In summary, we computed PGIs using large-sample GWAS summary statistics for the following eight diseases of interest: Alzheimer's disease (ALZ, effective N (N_{eff}) = 126,275),[4] breast cancer (BRC, N_{eff} = 236,094), colorectal cancer (CRC, N_{eff} = 117,486), coronary artery disease (CAD, N_{eff} = 162,973), lung cancer (LUC, N_{eff} = 98,001), prostate cancer (PRC, effective N_{eff} = 137,933), schizophrenia (SCZ, effective N_{eff} = 146,348), and type 2 diabetes (T2D, N_{eff} = 193,440). The study references are listed in section G Acknowledgments. Additional details on the GWAS summary statistics are reported in Supplementary Table [ST. Prev h2].

To find the GWAS summary statistics, we searched four prominent public databases hosted by the scientific community. The four databases were the GWAS Catalog (Buniello et al., 2018), the GWAS Atlas (Watanabe et al., 2019), the Polygenic Score Catalog (Lambert et al., 2021), and the PGI Repository (Becker et al., 2021). The objective was to find the largest published GWAS meta-analysis on each of our diseases of interest. Because the UK Biobank is a standard dataset to include in GWAS meta-analysis, it has become common practice for authors to share, alongside their main results, a hold-out version of the meta-analysis that withholds the UK Biobank data, which we aimed to find. The latest search was done on January 19, 2024.

For breast cancer, coronary artery disease, prostate cancer, schizophrenia, and type 2 diabetes we were able to identify publicly available hold-out versions of the largest GWAS published on these diseases (minimum effective N = 137,933). However, for Alzheimer's disease, colorectal cancer, and lung cancer, we could not find any public hold-out summary statistics, and not all cohort-level files were publicly available. Therefore, for these three diseases, we meta-analyzed the subset of cohort-level files that we could find in the public domain, and compensated for the loss of sample size by including recent summary statistics derived by the FinnGen Biobank (Kurki et al., 2023). After the quality control described next, we combined the cohort-level files by fixed-effect meta-analysis using the METAL software (Willer et al., 2010).

Each downloaded summary statistics file first underwent GWAS quality control (QC), which is standard practice (Winkler et al., 2014). Our GWAS QC-protocol was based on the often-cited protocol developed by the Social Science Genetic Association Consortium (SSGAC) (Karlsson Linnér et al., 2019; Okbay et al., 2022), which is a continuation of the older "industry-standard" protocol of the GIANT consortium (Winkler et al., 2014). The QC-protocol serves two main purposes: (1) to remove SNPs that for technical reasons are likely to worsen the signal-to-noise ratio of the PGI

[4]The effective sample size metric penalises the total sample size of an imbalanced case-control regression so that it matches the expected statistical power of a balanced analysis with 50% cases/controls (Grotzinger et al., 2022).

(e.g., rare SNPs); and (2) to ensure that SNP coordinates, effect-coded alleles, and other per-SNP statistics get aligned across files (e.g., to align them all with the "forward strand" in the genome reference data).

The protocol was applied using the EasyQC software package (Winkler et al., 2014) and it removed (i) multi-allelic or non-SNP variants (e.g., indels), (ii) SNPs on the sex chromosomes, (iii) SNPs with MAF <0.5% (the PRS-CS procedure below additionally removes SNPs with MAF <1% in the reference data), (iv) SNPs that could not be matched with the genome reference data the Haplotype Reference Consortium (HRC) (McCarthy et al., 2016), and (v) SNPs with a MAF deviating more than 0.2 from the reference frequency in the HRC data. Because the imputation quality metric was missing from most of the public summary statistics files, we did not filter on this metric. However, we note that most of the downloaded files were already filtered on imputation quality by the original studies, and secondly, that the HapMap 3 are often imputed well. Lastly, the QC-procedure produces a series of diagnostic plots, see Winkler et al. (2014), which were inspected without finding any conspicuous issues.

B.2 Details on the PRC-CS adjustment

Before computing PGIs, it is standard practice to adjust for the correlation structure between SNPs ("linkage disequilibrium", LD) (Choi et al., 2020). The two main approaches used in the literature are either (i) to prune the set of SNPs in the PGI to remove correlated and/or weakly associated SNPs, but this approach discards predictive information, or (ii) to keep the correlated and/or weakly associated SNPs and instead adjust their GWAS coefficients with more sophisticated statistical procedures. Here, we decided for (ii) and adjusted the GWAS coefficients by applying the recommended PRS-CS technique and software (Ge et al., 2019; Choi et al., 2020).

This method aims to improve the signal-to-noise ratio by applying a continuous shrinkage prior in penalized regression to down-weight correlated and/or weakly associated SNPs. The PRS-CS adjustment has been shown to perform well for several of our diseases of interest and similar to alternative methods like LDpred2 (Privé et al., 2020). By default, PRS-CS restricts the set of SNPs to those covered by the reference map of the International HapMap 3 Consortium (Altshuler et al., 2010), followed by removing any SNPs with MAF <1% in the 1000 Genomes reference panel. We ran the software using its default parameters. Finally, we used the PLINK2 software (Chang et al., 2015) to compute the PGIs in the UK Biobank imputed genetic data, using the PRC-CS adjusted GWAS coefficients as the weights ($\widehat{\mathbf{b}}$).

B.3 Details on Alzheimer's disease and the *APOE* region

The *APOE* region on chromosome 19 (coordinates 45,384,477–45,432,606, human reference genome build 37/hg 19) has a well-established large effect on the risk of Alzheimer's disease, see e.g., Scheltens et al. (2021) or Lambert et al. (2013). There are six primary risk types of the gene product apolipoprotein E (*ApoE*), called $\epsilon 2/\epsilon 2$ through to $\epsilon 4/\epsilon 4$. These risk types can be assayed by measuring only two "tag-SNPs" (rs7412 and rs429358) (Faul et al., 2021). The greatest disease

risk is conferred by the C nucleotide of rs429358, which has a considerably larger effect than rs7412. Genetic testing for *APOE* status has been available for years and is standard information to receive in consumer genetic health reports (Ryan et al., 2021). To our knowledge, the insurance industry has not yet adopted *APOE* status as a routine underwriting factor. Therefore, we decided not to model *APOE* status as an observed covariate, but rather as a component of the PGI.

Because the tag-SNP rs429358 is missing from the HapMap3 reference set of SNPs, and because the *APOE* region has extensive correlation structure, we modeled *APOE* status as follows. During the GWAS QC, we excluded the genomic region chr19:43,000,000–48,000,000 from the summary statistics for Alzheimer's disease. This 5 Megabase (Mb) exclusion region, which covers ∼8.1% of chr 19, was decided by first identifying the furthest SNPs on each side of rs7412 and rs429358 with a pair-wise squared correlation of $r^2 >= 0.01$ (computed with the HRC reference data). Thereafter, to be extra cautious, we expanded the window on each side by one additional Mb, followed by rounding to the nearest preceding/succeeding Megabase (i.e., 43Mb and 48Mb). In genetic terminology, we excluded the entire "linkage-disequilibrium region", including a safety margin, from the PGI for Alzheimer's disease. Next, we scored the two tag SNPs separately using their (unadjusted) GWAS coefficients and then merged this two-SNP score with the genome-wide PGI (adjusted with PRS-CS) that excluded the *APOE* region.

To ensure that this "augmented" version of the PGI for Alzheimer's disease was successful in capturing *APOE* status, we benchmarked it against the standard approach of modeling *APOE* risk types with dummy variables. Online Appendix Figure 1 shows that the dummy variables could successfully capture *APOE* status in ad hoc logistic regression. The benchmark showed that the "augmented" PGI for Alzheimer's achieved marginally better fit than the approach using dummies. On their own, the two *APOE* tag SNPs explained about 4.2% of the variance (pseudo-R^2) of the liability of Alzheimer's disease. In total, the augmented PGI explained 6.5% of the variance. Thus, most of the signal of the final PGI for Alzheimer's disease comes from the two tag-SNPs.

C Genotype data, ancestry, and sample quality control

The source of the study genetic data is the harmonized genetic data resource collected, maintained, and distributed by the UK Biobank (Data Category 100319). This resource is described in detail by Bycroft et al. (2018). Imputed genetic data is available for about 487,000 participants. The total number of imputed genetic variants is about 97 million, but the vast majority of these have not yet been tested in GWAS because they are rare. There are about 10.3 million imputed variants with MAF >1% in the UK Biobank.

To restrict the UK Biobank sample to participants of EUR ancestry, we relied on the approach described in Karlsson Linnér et al. (2019). Specifically, participants were considered to be of EUR ancestry if they (i) self-reported to have "White", "White British", "White Irish", or "Any other white background" ethnic background (Data Field 21000), and (ii) have a value on the first genetic PC that is ≤0. This approach identifies 447,863 participants that both cluster tightly together on

the genetic PCs, and this cluster is distributed distinctly from the other major genetic population groups, suggesting shared EUR genetic ancestry.

Next, we applied the following sample-level quality control by analyzing the pre-computed quality metrics distributed with the imputed genetic data resource (Data Category 100313). Participants were removed in case of (1) a mismatch between self-reported and genetic sex (Data Fields 31, 22001), (2) evidence of sex-chromosome aneuploidy (Data Field 22019), (3) being classified as an outlier on either genotype heterozygosity or missingness rates (Data Field 22027).

C.1 Inspection of PGI distributions

We inspected the distributions of the PGIs (Online Appendix Figure 2) to examine (1) whether the distributions departed from normality and (2) whether the distributions showed any differences by sex. The only PGI that departed noticeably from normality was that of Alzheimer's disease, which had trimodal distribution because of the large risk conferred by *APOE*, and specifically the e3/e4 and e4/e4 risk types (see Online Appendix Figure 1). All other PGIs were approximately normally distributed. Next, we could not identify any meaningful differences in distribution by sex. Nevertheless, we standardized the PGIs to mean zero and unit variance separately by sex.

D Health data and control variables

The UK Biobank has collected self-reported information on a large number of health, disease, psychological, and socioeconomic variables by a touchscreen questionnaire administered at the baseline assessment. This resource is described in detail by Sudlow et al. (2015). The baseline assessment occurred in 2006–2010. A total of 9.2 million invitation letters were sent to people aged ∼40–69 years. It was also possible to enter the study by volunteering. The touchscreen procedure was followed by a structured follow-up interview by a trained nurse to verify and map the self-reported medical conditions to standardized diagnostic categories, as well as to run various physiological measurements and lab assays. The most recent version of the study data were refreshed and downloaded from the UK Biobank Access Management System on June 19, 2023.

Since the baseline assessment took place, the UK Biobank has been enriched by linkage with several electronic health record data sources. The main linked electronic health record sources are (i) primary care records (Category 3000), (ii) hospital inpatient records (Category 2000), (iii) national cancer registries (Category 100092), and (iv) national death registries (Category 100093). These linked data are still accruing. Unfortunately, because of the UK's decentralized system for primary care records, source (i) primary care records are not yet linked for all participants (about 45% of the participants have been linked thus far). Nevertheless, because the disease outcomes we study are fairly serious, and because the cancer and death registries are largely complete, we consider it unlikely that we are missing a substantial number of disease cases because of the missing primary care records. However, we acknowledge that the UK Biobank participants have been found to be healthier than the general population (Schoeler et al., 2023), and healthy volunteer bias is a

plausible source of missing disease cases in our data.

The data availability over time varies for the linked electronic health record sources. The primary care records go as far back as 1938 for some participants. The hospital inpatient records only go back to 1981 for Scotland, 1991 for Wales, and 1997 for England. The cancer registry goes back to 1957 for Scotland, and to 1971 for England and Wales. The death registry captures all deaths among UK Biobank participants after they joined the biobank. Lastly, the self-reported data has the best lifetime coverage because the participants were asked to recall their lifetime medical histories, but this source may suffer from imperfect recall and other such errors. Nevertheless, the combination of self-reported medical history with several distinct electronic health records sources provide the opportunity to study lifetime medical histories.

D.1 Coding of the disease outcomes and comorbidities

This section details how we coded the disease outcomes and control variables for statistical analysis. Our starting point for coding the disease outcomes for statistical analysis is the "First occurrence of health outcomes" resource (Category 1712), which is generated and distributed by the UK Biobank. This resource is a harmonization and merge of the self-reported medical data with the linked electronic health record sources. However, because the "First occurrence" resource excludes cancer codes, and because it is updated at periodic intervals that differ from the periodic updates of the underlying sources (i–iv), we merged in the most recent data from all sources (i–iv).

The disease outcomes are recorded in the data using the International Classification of Disease (ICD) version 10, censored to first three letters (e.g., C50 "Malignant neoplasm of breast"). To re-code the ICD10 codes into disease outcomes suitable for statistical analysis, we relied on the established and widely used "Phecode" mapping system (Denny et al., 2013; Bastarache, 2021). A feature of the Phecode system is to merge the many subcategories of the ICD-code system into a single general disease outcome. For example, the primary ICD10 code for Alzheimer's disease is G30, while there are sub-codes to distinguish early from late age of onset. The Phecode system merges these sub-types into a single code, 290.11 "Alzheimer's disease". Secondly, it combines redundant or alternative codes spread across ICD chapters, e.g., although the official chapter for breast cancer is C50, there are alternative codes that some doctor's may use, such as "D05 Carcinoma in situ of breast". In the latest version, the Phecode mapping system condenses some 90,000 ICD-10 codes into about 1,900 Phecodes (Bastarache, 2021).

The mapping of ICD10 codes to Phecodes for our eight diseases of interest is reported in Supplementary Table [ICD Phecode]. With the exception of coronary artery disease, our diseases of interest could all be captured by a single Phecode. To code coronary artery disease, which is defined as a collection of many underlying heart and circulatory conditions, we relied on the approach of the previous literature and considered the entire ICD10 Chapter I20–I25 "Ischaemic heart diseases" (Aragam et al., 2022), which maps to a total of six different Phecodes. The next section describes the selection of control variables and some of those are comorbidities, e.g., asthma or bipolar disorder. The comorbidities were coded analogously using the Phecode system. Online

Appendix Figure 4 shows the risk our eight diseases of interest as predicted by PGI deciles.

D.2 Epidemiological risk factors and the coding of control variables

An important contribution of this study is to control for existing information that overlaps with predictive signal of the PGI. Therefore, we carefully reviewed the epidemiological literature and clinical guidelines of public health organizations (e.g., the National Cancer Institute or the Centre for Disease Control) to identify the main observable risk factors that are used in consensus by clinicians for screening or diagnostic purposes. As such, this information is likely to be relevant and accessible to insurance companies during medical underwriting. Supplementary Table [Risk factors] report the disease-specific risk factors that were identified. The selection of epidemiological risk factors was verified independently by two medical doctors.

The UK Biobank data was rich enough for us to code almost all the risk factors and comorbidities identified by the review. A handful of risk factors were eventually omitted or proxied (see Panel B in Supplementary Table [Risk factors]), either because they were unspecific, e.g., diet, or because of limited data availability, e.g., substance use. Instead, the effects of diet were proxied by relevant anthropometric measures or biomarkers, such as BMI, hypertension, systolic blood pressure, and hypercholesterolemia. Similarly, general psychopathology was proxied by bipolar disorder, depression, and a neuroticism score. Only three risk factors were omitted completely: brain injury, cannabis use, and substance use. Notably, we were able to code family history (father, mother, or siblings) for the eight diseases except schizophrenia, which was proxied by family history of depression.

The data fields used to code each risk factor are reported in Supplementary Table [Covariates]. We defined a set of ten baseline covariates that were included in all the analyses: age (at the most recent observation), sex (omitted from any sex-specific analysis), Townsend's deprivation index, education (years of schooling), BMI, alcohol consumption (drinks per week), smoking (never vs. former vs. current smokers), physical inactivity, systolic blood pressure, and a genotyping array dummy to control for that part of the sample was genotyped on a different genotyping array. These ten baseline covariates were always accompanied by the first ten genetic PCs.

Next, Supplementary Table [Covariates] also reports the disease-specific covariates. The risk factors and comorbidities were coded either as continuous variables or as dummies, with the exception of BMI and alcohol consumption (drinks per week) that were coded as percentile ranks. The reason is that the percentile ranks remain more stable as people age (for more information, see the next section D.3). Also, because not all female participants have yet reached the menopause, we imputed the value zero in the covariate for age at menopause (or hysterectomy), which was modelled together with a dummy to control for menopause (or hysterectomy) status. Supplementary Table [Descriptive stats] reports sample descriptive statistics.

D.3 Interpolation of age-dependent covariate values

An objective of this study is to predict the lifetime risk of disease by a particular age (e.g., age a = 65). The data is informative of the age of onset for the medical conditions and comorbidities we study, but for most participants, we only observe their covariate values once (at the baseline assessment in 2006–2010). Therefore, in any exercise that predicts the risk of disease by a particular age, the values of age-dependent covariates were interpolated to expected values. We did not interpolate any time-fixed covariates (e.g., genotyping array dummy) or covariates that are mostly stable in older populations, such as education (years of schooling) or age at first menstruation, nor the handful of geographical variables based on the home address at the time of recruitment, such as the Townsend's deprivation index or air pollution. Supplementary Table [Covariates] reports which covariates were eventually interpolated.

To determinate whether each potentially age-dependent covariate depended on age, and thus, should be selected for interpolation, we regressed the covariates on a polynomial of age to the fourth degree. Whenever the polynomial explained more than 1% of the variation, the covariate was interpolated. Otherwise, we used the observed value both in the estimation and the prediction steps. Also, to be consistent across our eight diseases of interest, the family history variables were always interpolated, though we found that some did not depend on age.

The interpolation procedure was done separately by sex. For continuous covariates, we first ran a linear regression of each covariate on the age polynomial to model the covariate as a function of age: $f(a) = a_1 \times age + a_2 \times age^2 + a_3 \times age^3 + a_4 \times age^4 + e$. For binary covariates, we used logistic regression. The estimated regression coefficients were then used to project the values of the covariate for each year of age observed in the data (i.e., 39–86 years). For binary covariates, probabilities were projected. Below, we interpolate the observed values for a participant of age a_i by subtracting the difference between $f(a|a = a_i)$ and $f(a|a = 65)$.

We found that none of the age-dependent covariates in this study decrease as a function of age. Instead, in most cases, the age polynomial predicted a monotonic increase in the covariate. However, for a few covariates, the polynomial predicted slight non-monotonicity at the top and bottom of the age distribution (see, e.g., age >80 in the Online Appendix Figure 3, panel "Breast cancer screening" or "Prostate cancer screening"). Therefore, we decided to force monotonicity over the entire age range by fitting a cubic smoothing spline as implemented in the R package "mgcv". The smoothing had little to no effect on most age groups in the data (see the figure).

Lastly, for continuous covariates, we interpolated the age-dependent covariates by subtracting the difference between $f(a|a = a_i)$ and $f(a|a = 65)$ from the observed values. E.g., among women, the age polynomial predicted a blood pressure of 125 at age 50, and 135 at age 65. A woman of age 50 would thus be interpolated to x – (125 – 135) = x + 10, where x is the observed value. Similarly, a woman of age 80 would be interpolated to x – (145 – 135) = x – 10. In other words, because all covariates were observed to increases as a function of age, the procedure effectively increased the values of the age-dependent covariates for participants below 65, and decreased the values for those older than 65.

For binary covariates, we interpolated predicted probabilities among participants below age a who had not experienced the event of the covariate, as well as among individuals above age a who had experienced the event (but for which we could not observe the age of the event). Again, we subtracted the difference between $f(a|a = a_i)$ and $f(a|a = 65)$ from the observed values. For example, the probability that a woman of age 40 and who had not yet experienced hypertension was interpolated as 0 – (0.02 – 0.2) = 0.18 (i.e., the probability of experiencing the event between age a_i and 65). Similarly, the probability that a woman of age 80 and who had experienced hypertension was interpolated as 1 – (0.6 – 0.2) = 0.6. We did not modify the values of participants of $a_i < a$ and who had already experienced the event before age a, nor of participants of $a_i > a$ and who had not (yet) experienced the event.

E Extended econometric model: considering the SNP and non-SNP components of G_f separately

As discussed in the main text, when we assume that a future PGI's accuracy is equal to the disease's twin heritability (as an upper bound), Assumption 3 (*Noisy current-day PGI*) is likely to be violated, because it is unlikely to be the case that $\epsilon \perp \mathbf{W}$. In this Appendix, we explain in more detail why this happens. We also introduce an extended econometric model that deals with this issue by separately considering what we call the SNP and non-SNP components of the future PGI G_f, at the cost of additional assumptions.

To see why it is unlikely that $\epsilon \perp \mathbf{W}$ in a scenario where we assume a future PGI accuracy equal to the disease's twin heritability, recall that twin heritability captures the fraction of the variance in the disease liability $\mathcal{L}$ that is attributable to all genetic variants; by contrast, SNP heritability is the fraction of the variance attributable to SNPs in the data. To reach an accuracy that exceeds a disease's SNP heritability, a future PGI will necessarily have to rely on non-SNP variants.

Next, consider a future PGI G_f with accuracy equal to the twin heritability. G_f will is equal to the true additive genetic factor G^* and is the expected value of the disease liability $\mathcal{L}$ conditional on all genetic variants (X_{all}) under an additive genetic model:

$$G_f := E[\mathcal{L}|\mathbf{X}_{all}] = \mathbf{X}_{all}\,\mathbf{b}^*_{all}.$$

Let $G_{1,f}$ represent the part of G_f that is captured by the SNPs ($\mathbf{X}_{\text{SNPs}}$):

$$G_{1,f} := E[\mathcal{L}|\mathbf{X}_{\text{SNPs}}] = \mathbf{X}_{\text{SNPs}}\,\mathbf{b}^*_{\text{SNPs}}.$$

Finally, let $G_{2,f}$ represents part of G_f that is not captured by the SNPs:

$$G_{2,f} := G_f - G_{1,f}.$$

Under Assumption 3 from the econometric model in the main text, G_c is noisy estimate of G_f: $G_c = G_f + \epsilon$. Substituting $G_{1,f} + G_{2,f}$ for G_f yields

$$G_c = G_f + \epsilon = G_{1,f} + G_{2,f} + \epsilon = G_{1,f} + \tilde{\epsilon},$$

where $\tilde{\epsilon} = G_{2,f} + \epsilon$. This equation captures the reality that G_c is constructed with the available SNP data and is thus a noisy signal of $G_{1,f}$, rather than of G_f as per Assumption 3.

When the accuracy of G_f is equal to the SNP heritability, if one lets $G_{2,f} = 0$ and $G_f = G_{1,f}$, then $\tilde{\epsilon} = \epsilon \perp \mathbf{W}$ and Assumption 3 plausibly holds. But when the accuracy of G_f is equal to the twin heritability, then $G_{2,f} \neq 0$ and $\tilde{\epsilon} = G_{2,f} + \epsilon$ is unlikely to be orthogonal to $\mathbf{W}$, thus violating Assumption 3.

E.1 Assumptions and theorem

We now extend our econometric model to separately consider the SNP and non-SNP components $G_{1,f}$ and $G_{2,f}$ of G_f. The extended model does not assume that $\epsilon \perp \mathbf{W}$, but additional assumptions are needed to identify the model.

Our extended model is fully specified by $\mathbb{P}'$, the joint distribution of $(D, G_c, G_{1,f}, G_{2,f}, \mathbf{W})$. While the current-day PGI G_c can be observed in the UKB data by the econometrician, the future-day PGI $G_f = G_{1,f} + G_{2,f}$ cannot. Define the distribution of the observed data as a joint distribution $\mathbb{P}'_{\text{data}}$ over $(D, G_c, \mathbf{W})$. We say that the model is identified if there is a unique value of $\mathbb{P}'$ consistent with $\mathbb{P}'_{\text{data}}$.

We now state the assumptions of the extended model. We maintain Assumptions 1, 4, and 5, but modify Assumptions 2 and 3 and augment Assumption 5 to relate these to $G_{1,f}$ and/or $G_{2,f}$ (instead of only G_f). We maintain the assumption numbering from the main text but add a $'$ when an assumption is modified. In addition to the updated Assumptions 1-5, we need a novel assumption (Assumption 6) to describe the relationship between $G_{2,f}$ and the covariates $\mathbf{W}$; the relationship between $G_{1,f}$ and the $\mathbf{W}$ can be estimated from the PGI G_c, which is a noisy estimate of $G_{1,f}$, but there is no corresponding noisy estimate of $G_{2,f}$. For ease of reference, we state all the assumptions of the extended model, including those that are unchanged from the main text.

Assumption 1 is unchanged:

Assumption 1. *[Covariates' distribution]* $\mathbf{W}$ *is distributed according to a distribution* $\mathbb{P}_{\mathbf{W}}$ *in Euclidean space. The support of* $\mathbf{W}$ *spans the entire space.*

We replace Assumption 2 by the following:

Assumption 2'. *[Gaussian future-day PGI] Conditional on observables, the SNP and non-SNP components of the future PGI each have have a Gaussian distribution. For each component, the variance is constant and the mean is a linear function of observables:*

$$G_{1,f} := \mathbf{W}\theta_1 + V_1, \tag{1}$$

$$G_{2,f} := \mathbf{W}\theta_2 + V_2, \tag{2}$$

with $V_1 \sim N(0, \sigma^2_{V_1})$, $V_2 \sim N(0, \sigma^2_{V_2})$ *and* $V_1, V_2 \perp \mathbf{W}$.

We modify Assumption 3 to make it about $G_{1,f}$ rather than G_f:

Assumption 3'. *[Noisy current-day PGI] The observed current-day PGI* G_c *has unit variance,* $\text{Var}[G_c] = 1$*, and is a noisy estimate of* $G_{1,f}$*:*

$$G_c := G_{1,f} + \epsilon, \tag{3}$$

with $\epsilon \sim N(0, \sigma^2_\epsilon)$ *and* $\epsilon \perp (\mathbf{W}, V_1)$.

Assumption 4 is unchanged:

Assumption 4. *[Probit disease model] Disease probabilities follow a linear probit model:*

$$\Pr[D = 1|G_f = g_f, G_c = g_c, \mathbf{W} = \mathbf{w}] = \Phi(g_f\beta_g + \mathbf{w}\beta_w). \tag{4}$$

Under assumptions (1)-(4), the distribution $\mathbb{P}'$ has parameters

$$\Gamma' := (\mathbb{P}_{\mathbf{W}}, \beta, \theta_1, \theta_2, \sigma^2_{V_1}, \sigma^2_{V_2}, \sigma^2_\epsilon).$$

The data can be generated by taking independent draws of $\mathbf{W} \sim \mathbb{P}_{\mathbf{W}}$, $V_1 \sim N(0, \sigma^2_{V_1})$, $V_2 \sim N(0, \sigma^2_{V_2})$, and $\epsilon \sim N(0, \sigma^2_\epsilon)$. Equations (1) and (2) then determine the future PGI $G_f = G_{1,f} + G_{2,f}$; Equation (3) determines the current PGI G_c; and the disease D is realized with probability given by Equation (4).

As in the baseline econometric model from the main text, we need an additional assumption to identify the model. Here, we need to modify Assumption 5 to make it about both G_f and $G_{1,f}$ rather than about G_f only. This Assumption states that we know the predictive power of both the future PGI (R^2_f) as well as that of its SNP component ($R^2_{1,f}$); in practice, we can assume the former to be equal to the disease's twin heritability and the latter to its SNP heritability:

Assumption 5'. *[Future PGI predictive power] The pseudo R^2 on the liability scale of a probit regression of D on G_f equals a known constant R^2_f; and the pseudo R^2 on the liability scale of a probit regression of D on $G_{1,f}$ equals a known constant $R^2_{1,f}$.*

Finally, to identify the extended model, we also need one novel assumption. Because we do not observe a measure of $G_{2,f}$ (unlike for $G_{1,f}$, for which we observe G_c), we cannot estimate θ_2 by regressing such a measure on $\mathbf{W}$. We thus need one additional assumption to identify the model:

Assumption 6. *[Covariates and the non-SNP component of the future PGI] θ_2 is proportional to θ_1, and the R^2 of a regression of $G_{2,f}$ on $\mathbf{W}$ is equal to that of a regression of $G_{1,f}$ on $\mathbf{W}$:*

$$\theta_2 = \alpha\theta_1 \text{ for a scalar } \alpha, \text{ and}$$
$$\frac{\text{Var}[\mathbf{W}\theta_2]}{\text{Var}[G_{2,f}]} = \frac{\text{Var}[\mathbf{W}\theta_1]}{\text{Var}[G_{1,f}]}$$

We are now ready to state the identification result.

Theorem 1' *Under assumptions (1), (2'), (3'), (4), (5'), and (6), the model is identified.*

The proof is provided below. The intuition for it is similar to that for the proof of Theorem 1 in the main text. As for Theorem 1, $\mathbb{P}_{\mathbf{W}}$ is identified because $\mathbf{W}$ is observed; the coefficients θ_1 are the coefficients of a regression of G_c on $\mathbf{W}$; and the variance σ^2_ϵ is identified from $G_{1,f}$'s assumed future predictive power, the estimated predicted power of G_c in the data, and the variance of $\mathbf{W}\theta_1$.

The coefficients θ_2 as well as $\sigma^2_{V_1}$ can be obtained from the estimated coefficients θ_1, using the assumed predictive power of the future PGI and of its SNP component together with the assumed relationship between θ_1 and θ_2 (Assumption 6). Finally, as for Theorem 1, the coefficients β are related to the coefficients of a probit regression of D on G_c and $\mathbf{W}$. Once the other parameters are known, β can be computed from that regression's coefficients.

E.2 Proof of Theorem A1

Here we prove that the extended econometric model that considers $G_{1,f}$ and $G_{2,f}$ separately is identified under Assumptions (1), (2'), (3'), (4), (5'), and (6) (Theorem 1').

Proof. First, observe that if we consider only the SNP component $G_{1,f}$ of G_f and let—for the time being—$G_f = G_{1,f}$ (i.e., let $\theta_2 = \sigma^2_{V^2} = 0$ so that $G_{2,f} = 0$), then Assumptions (2'), (3'), and (5') become identical to Assumptions (2), (3), and (5) of the (non-extended) econometric model in the main text. From the proof of Theorem 1, under these Assumptions together with Assumption 1, we know that θ_1, σ^2_ϵ, and $\sigma^2_{V_1}$ are identified. Thus, under Assumptions (1), (2'), (3'), and (5'), these parameters are identified.

We still have to show that θ_2, $\sigma^2_{V_2}$, and β_g and β_w are identified in our extended econometric model.

Part 1: θ_2

The pseudo-R^2 on the liability scale of a probit regression of D on G_f is defined as the share of the variation in the probit regression liability that is accounted for by G_f:

$$\frac{Var[G_f\beta]}{\text{Var}[\mathcal{L}_r]},$$

where β is the coefficient from the probit regression of D on G_f and $\mathcal{L}_r = G_f\beta + \eta$ is the liability in that regression. Let $\mathcal{L}'_r := \mathcal{L}_r/\beta$ and $\eta' = \eta/\beta$. So

$$\mathcal{L}'_r = G_f + \eta' = G_{1,f} + G_{2,f} + \eta',$$

with $G_f, G_{1,f}, G_{2,f} \perp \eta'$ and $G_{1,f} \perp G_{2,f}$. It follows that the pseudo-R^2's on the liability scale in regressions of D on G_f, $G_{f,1}$, and $G_{f,2}$ are given by

$$R^2_f = \frac{\text{Var}[G_f]}{\text{Var}[\mathcal{L}'_r]},\ R^2_{1,f} = \frac{\text{Var}[G_{1,f}]}{\text{Var}[\mathcal{L}'_r]},\text{ and } R^2_{2,f} = \frac{\text{Var}[G_{2,f}]}{\text{Var}[\mathcal{L}'_r]}.$$

Thus, $R^2_{2,f} = R^2_f - R^2_{1,f}$.

By Assumption (6), $\theta_2 = \alpha\theta_1$ for a scalar α, and

$$\frac{\text{Var}[\mathbf{W}\theta_2]}{\text{Var}[G_{2,f}]} = \frac{\text{Var}[\mathbf{W}\theta_1]}{\text{Var}[G_{1,f}]} = \frac{\alpha^2\text{Var}[\mathbf{W}\theta_1]}{\text{Var}[G_{2,f}]}.$$

Observe that $\text{Var}[G_{1,f}] = R^2_{1,f}\text{Var}[\mathcal{L}'_r]$ and that $\text{Var}[G_{2,f}] = R^2_{2,f}\text{Var}[\mathcal{L}'_r]$.

It follows that

$$\alpha^2 = \frac{\text{Var}[G_{2,f}]}{\text{Var}[G_{1,f}]} = \frac{R^2_{2,f}}{R^2_{1,f}} = \frac{R^2_f - R^2_{1,f}}{R^2_{1,f}}.$$

Thus,

$$\theta_2 = \alpha\theta_1, \text{ with } \alpha = \sqrt{\frac{R^2_f - R^2_{1,f}}{R^2_{1,f}}}.$$

Under Assumption 5', R^2_f and $R^2_{1,f}$ are known quantities, so θ_2 is identified.

In practice, it will make sense to assume $R^2_f = h^2_{twin}$ and $R^2_{1,f} = h^2_{SNP}$. Then,

$$\alpha^2 = \frac{h^2_{twin} - h^2_{SNP}}{h^2_{SNP}}.$$

Part 2: $\sigma^2_{V_2}$

By Assumption (2'), $\text{Var}[G_{1,f}] = \text{Var}[\mathbf{W}\theta_1] + \sigma^2_{V_1}$. It follows from Assumption (6) that

$$\text{Var}[G_{2,f}] = \alpha^2\text{Var}[G_{1,f}] = \alpha^2\left\{\text{Var}[\mathbf{W}\theta_1] + \sigma^2_{V_1}\right\} = \text{Var}[\mathbf{W}\theta_2] + \alpha^2\sigma^2_{V_1}.$$

By Assumptions (2'), $\text{Var}[G_{2,f}] = \text{Var}[\mathbf{W}\theta_2] + \sigma^2_{V_2}$. So

$$\sigma^2_{V_2} = \alpha^2\sigma^2_{V_1}.$$

Part 3: β_g

By Assumption (4), the probability of a loss D is

$$\Phi(G_f\beta_g + \mathbf{W}\beta_w).$$

Equivalently,

$$D = \{G_f\beta_g + \mathbf{W}\beta_w + \eta > 0\},$$

where η is a standard normal random variable orthogonal to $\mathbf{W}, V_1, V_2, \epsilon$. Consider the distribution of D conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$. Lemma 1 from the proof of Theorem 1 implies that

$$G_{1,f} = ag_c + b\mathbf{w}\theta_1 + c\nu,$$

where ν is a standard normal random variable independent of η, $a = \frac{1}{\sigma^2_\epsilon}/(\frac{1}{\sigma^2_\epsilon} + \frac{1}{\sigma^2_{V_1}})$, $b = \frac{1}{\sigma^2_{V_1}}/(\frac{1}{\sigma^2_\epsilon} + \frac{1}{\sigma^2_{V_1}})$, and $\frac{1}{c^2} = \frac{1}{\sigma^2_\epsilon} + \frac{1}{\sigma^2_{V_1}}$.

And from Assumption (2),

$$G_{2,f} = \mathbf{w}\theta_2 + V_2.$$

So $G_f = G_{1,f} + G_{2,f} = (ag_c + b\mathbf{w}\theta_1 + c\nu) + (\mathbf{w}\theta_2 + V_2)$.

Therefore, conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$,

$$\begin{aligned} D &= \{[(ag_c + b\mathbf{w}\theta_1 + c\nu) + (\mathbf{w}\theta_2 + V_2)]\beta_g + \mathbf{w}\beta_w + \eta > 0\} \\ &= \{g_c a\beta_g + \mathbf{w}(b\theta_1\beta_g + \theta_2\beta_g + \beta_w) + \eta + c\nu\beta_g + V_2\beta_g > 0\} \\ &= \left\{ g_c \frac{a\beta_g}{\sqrt{1 + (c^2 + \sigma_{V_2}^2)\beta_g^2}} + \mathbf{w}\frac{b\theta_1\beta_g + \theta_2\beta_g + \beta_w}{\sqrt{1 + (c^2 + \sigma_{V_2}^2)\beta_g^2}} + \frac{\eta + (c\nu + V_2)\beta_g}{\sqrt{1 + (c^2 + \sigma_{V_2}^2)\beta_g^2}} > 0 \right\} \end{aligned}$$

The last term is distributed as $N(0, 1)$. Therefore

$$\Pr(D = 1 | G_c = g_c, \mathbf{W} = \mathbf{w}) = \Phi\left[g_c \frac{a\beta_g}{\sqrt{1 + (c^2 + \sigma_{V_2}^2)\beta_g^2}} + \mathbf{w}\frac{b\theta_1\beta_g + \theta_2\beta_g + \beta_w}{\sqrt{1 + (c^2 + \sigma_{V_2}^2)\beta_g^2}} \right]. \tag{5}$$

Let γ be the coefficient vector of a probit regression of D on G_c and $\mathbf{W}$. Then

$$\frac{a\beta_g}{\sqrt{1 + (c^2 + \sigma_{V_2}^2)\beta_g^2}} = \gamma_g.$$

This equation implies that β_g has the same sign as γ_g. Taking the square of both sides and solving, we find

$$\beta_g = \frac{\gamma_g}{\sqrt{a^2 - \gamma_g^2(c^2 + \sigma_{V_2}^2)}}. \tag{6}$$

This result implies that $\gamma_g^2 < a^2/(c^2 + \sigma_{V_2}^2)$.

Part 4: β_w

By equation (5) and using the fact that the support of $\mathbf{W}$ includes a set that spans the entire space (Assumption 1),

$$\frac{b\theta_1\beta_g + \theta_2\beta_g + \beta_w}{\sqrt{1 + (c^2 + \sigma_{V_2}^2)\beta_g^2}} = \gamma_w.$$

Then

$$\beta_w = \gamma_w \sqrt{1 + (c^2 + \sigma_{V_2}^2)\beta_g^2} - b\theta_1\beta_g - \theta_2\beta_g. \tag{7}$$

□

F Overview of Policy Responses

We will give a brief overview of the literature on the regulation of selection markets. We also discuss how the ideas relate to genetic private information. Regulation of selection markets balances two goals: efficiency and redistribution. Efficiency aims to maximize total economic surplus. Redistribution tries to help particular groups. Redistribution includes solidarity towards consumers with lower wealth, worse health status, or genes that are predictive of serious illnesses. The standard regulatory playbook can be divided in three approaches.

The first approach is **laissez-faire regulation**. That is, allowing firms to set prices as they wish, with minimal regulation focused on the financial solvency of insurers. The standard example is life insurance. Life insurance costs have a large predictable component because age and health status are highly predictive of mortality. In most countries, insurers are free to use these variables when underwriting policies. The laissez-faire approach minimizes adverse selection. In life insurance, it is thought that there is relatively little adverse selection, and that insurers often predict mortality better than consumers (Gottlieb and Smetters, 2021). However, the laissez-faire approach ignores redistribution. The elderly and other consumers with high mortality pay higher premiums. This is deemed acceptable in life insurance because premiums are a small part of most consumers' income, and redistribution is not a first-order issue.

The second, and opposite, approach is **government provision**. In developed countries, basic health insurance is publicly provided. One reason is that redistributive issues are central in health insurance. Insurance costs are usually considerable relative to personal income because healthcare represents a substantial share of GDP (18% in the United States). In most countries, there is widespread agreement that the destitutes should be granted some level of care, and that high-cost consumers such as the elderly should be subsidized (Einav et al., 2023). For these reasons, basic health insurance coverage is publicly provided in most developed countries. Public provision of any kind of insurance has three main potential drawbacks. First, the public sector may be less efficient. Second, innovation can be curbed, especially for innovative products like critical illness insurance. Third, the quality of public coverage might be inefficiently too high or too low.

The third approach is **managed competition**. Managed competition combines private provision with regulations to improve market performance. In the United States, examples include individual health insurance exchanges and insurance options sponsored by large employers. In most other developed countries, the main example is complementary health insurance. Managed competition involves four main kinds of regulations: community rating, mandates, subsidies, and risk adjustment. Community rating imposes restrictions on what variables insurers can use to set prices. Community rating is used for redistribution reasons (such as lowering premiums for the elderly) but has the cost of increasing adverse selection (Handel et al., 2015). Mandates impose penalties on consumers who do not purchase insurance, and subsidies lower the prices of policies; these are important for increasing the number of consumers who purchase policies, which otherwise tends to be inefficiently too low due to adverse selection (Geruso et al., 2021; Chade et al., 2022; Veiga and Levy, 2022). Risk adjustment is a type of subsidy to insurers who cover sicker patients.

Risk adjustments and other subsidies are thought to be important to guarantee that firms don't offer contracts that provide inefficiently low coverage (Glazer and McGuire, 2000).

Consider how these standard approaches would work in the case of genetic information and CII. The laissez-faire approach would be to allow companies to underwrite based on genetic information. It is likely that companies would require genetic testing, much like they require medical exams today. This would lead to little adverse selection. The downside of this policy is that critical illness insurance would be very expensive for consumers with high risk. For example, consider a consumer with a 25% risk of developing prostate cancer. This consumer would have to pay $25,000 in lifetime premiums for a $100,000 CII contract for prostate cancer. Our results suggest that, with improved prediction technology, many consumers would be in this situation, which might create pushback against the laissez-faire approach.

The public provision approach would be for governments to publicly provide CII to all consumers. This strikes us as being unlikely to be popular. Most people do not hold CII, and it seems like providing it to a broad segment of the population might be unpopular.

The managed competition approach includes a broad menu of options. In fact, the current policy of genetic information bans falls within managed competition. A genetic information ban is simply a community rating regulation. Our results suggest it is likely that this type of community rating will lead to high levels of selection in the future. In that case, industry participants and regulators may want to address selection with policies like risk adjustment and subsidies. The design of efficient policies would depend on empirical work, along the lines of existing research on current selection markets.

G Acknowledgments

This research was reviewed and approved by the Committee Ethics and Data of the Leiden Law School of Leiden University (application number 2023-24). The use of the UK Biobank data for this research has also been reviewed by the George Mason University Institutional Review Board (IRB) Office, which has determined it to be exempt from IRB review. This research has been conducted using the UK Biobank Resource under application number(s) 99086 and 217716 (pending application). Methods development and preliminary analyses were conducted using the UK Biobank Resource under application number 11425. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-8891. This publication is part of the project "The Economic Consequences of New Genetic Testing Technologies" (project number Vl.Veni.221E.080) of the NWO Talent Programme Veni – Social Sciences and Humanities, which is (partly) financed by the Dutch Research Council (NWO). This research was also supported by a Faculty Research and Development Award (FRDA) at George Mason University.

G.1 Acknowledgments and data availability GWAS summary statistics

We thank the participants and investigators of the FinnGen study for providing GWAS summary statistics for these analyses (Kurki et al., 2023). The GWAS summary statistics on Alzheimer's disease, colorectal cancer, and lung cancer derived in the FinnGen biobank can be accessed here: https://www.finngen.fi/en/access_results. We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary statistics for these analyses (Lambert et al., 2013). The GWAS summary statistics on Alzheimer's disease and an extended acknowledgment of the relevant funding sources can be accessed here: https://www.niagads.org/datasets/ng00036. We thank the participants and investigators of the Breast Cancer Association Consortium for providing the summary statistics for these analyses (Michailidou et al., 2017; Zhang et al., 2020). The GWAS summary statistics on breast cancer and an extended acknowledgment of the relevant funding sources can be accessed here: https://bcac.ccge.medschl.cam.ac.uk/bcacdata/. We thank the participants and investigators of the CARDIoGRAMplusC4D Consortium (Nikpay et al., 2015). The GWAS summary statistics on coronary artery disease can be accessed here: http://www.cardiogramplusc4d.org/. We thank the participants and investigators that contributed to the following GWAS summary statistics that were downloaded from the NHGRI-EBI GWAS Catalog (Sollis et al., 2022): Colorectal cancer study ID(s) GCST012876, GCST012877, GCST012878, GCST012879, GCST012880 (Huyghe et al., 2019; Fernandez-Rozadilla et al., 2023); Lung cancer study ID(s) GCST004748 (McKay et al., 2017); Prostate cancer study ID(s) GCST006085 (Schumacher et al., 2018). The GWAS summary statistics on colorectal cancer, lung, cancer, and prostate cancer downloaded from the NHGRI-EBI GWAS Catalog can be accessed here: https://www.ebi.ac.uk/gwas/downloads/summary-statistics. We thank the participants and investigators of the Psychiatric Genomics Consortium (Trubetskoy et al., 2022). The GWAS summary statistics on schizophrenia can be accessed here: https://pgc.unc.edu/for-researchers/download-results/. We thank the participants and investigators of

the DIAGRAM Consortium (Mahajan et al., 2018). The GWAS summary statistics on type 2 diabetes can be accessed here: http://diagram-consortium.org/downloads.html

References

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–58, 2010.

Aragam, K. G., Jiang, T., Goel, A., Kanoni, S., Wolford, B. N., et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nature Genetics*, 54:1803–1815, 2022.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., et al. A global reference for human genetic variation. *Nature*, 526, 2015.

Bastarache, L. Using phecodes for research with the electronic health record: From PheWAS to PheRS. *Annual Review of Biomedical Data Science*, pages 1–19, 2021.

Beauchamp, J. P., Cesarini, D., Johannesson, M., van der Loos, M. J., Koellinger, P. D., et al. Molecular genetics and economics. *Journal of Economic Perspectives*, 25(4):57–82, 2011.

Becker, J., Burik, C. A., Goldman, G., Wang, N., Jayashankar, H., et al. Resource profile and user guide of the polygenic index repository. *Nature human behaviour*, 5(12):1744–1758, 2021.

Benjamin, D. J., Cesarini, D., Chabris, C. F., Glaeser, E. L., Laibson, D. I., et al. The promises and pitfalls of genoeconomics. *Annu. Rev. Econ.*, 4(1):627–662, 2012.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47:D1005–D1012, 2018.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Chade, H., Marone, V. R., Starc, A., and Swinkels, J. Multidimensional screening and menu design in health insurance markets. National Bureau of Economic Research Working Paper No. 30542, 2022.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4:7, 2015.

Chatterjee, N., Shi, J., and García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17:392–406, 2016.

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15:2759–2772, 2020.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, pages 1102–1111, 2013.

Einav, L., Finkelstein, A., and Fisman, R. *Risky Business: Why Insurance Markets Fail and What to Do About It*. Yale University Press, 2023.

Faul, J., Collins, S., Smith, J., Zhao, W., Kardia, S., et al. Health and Retirement Study: APOE and Serotonin Transporter Alleles – Early Release. Technical report, University of Michigan, 2021.

Fernandez-Rozadilla, C., Timofeeva, M., Chen, Z., Law, P., Thomas, M., et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nature Genetics*, 55:89–99, 2023.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10:1776, 2019.

Geruso, M., Layton, T. J., McCormack, G., and Shepard, M. The two margin problem in insurance markets. *Review of Economics and Statistics*, pages 1–46, 2021.

Glazer, J. and McGuire, T. G. Optimal risk adjustment in markets with adverse selection: an application to managed care. *American Economic Review*, 90(4):1055–1071, 2000.

Gottlieb, D. and Smetters, K. Lapse-based insurance. *American Economic Review*, 111(8):2377–2416, 2021.

Grotzinger, A. D., Fuente, J. d. l., Privé, F., Nivard, M. G., and Tucker-Drob, E. M. Pervasive Downward Bias in Estimates of Liability-Scale Heritability in Genome-wide Association Study Meta-analysis: A Simple Solution. *Biological Psychiatry*, 93:29–36, 2022.

Handel, B., Hendel, I., and Whinston, M. D. Equilibria in health exchanges: Adverse selection versus reclassification risk. *Econometrica*, 83(4):1261–1313, 2015.

Huyghe, J. R., Bien, S. A., Harrison, T. A., Kang, H. M., Chen, S., et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nature Genetics*, 51:76–87, 2019.

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

Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613:508–518, 2023.

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45:1452–1458, 2013.

Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*, 53:420–425, 2021.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, pages 1505–1513, 2018.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48, 2016.

McKay, J. D., Hung, R. J., Han, Y., Zong, X., Carreras-Torres, R., et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature Genetics*, 49:1126–1132, 2017.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551:92, 2017.

NHGRI. Fact Sheet Human Genomic Variation, 2023.

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, 47:1121–1130, 2015.

Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., et al. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*, 54(4):437–449, 2022.

Polderman, T. J. C., Benyamin, B., de Leeuw, C., Sullivan, P. F., van Bochoven, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7):702–709, 2015.

Privé, F., Arbel, J., and Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics*, 36: 5424–5431, 2020.

Ruan, Y., Lin, Y.-F., Feng, Y.-C. A., Chen, C.-Y., Lam, M., et al. Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics*, 54(5):573–580, 2022.

Ryan, M. M., Cox, C. G., Witbracht, M., Hoang, D., Gillen, D. L., et al. Using Direct-to-Consumer Genetic Testing Results to Accelerate Alzheimer Disease Clinical Trial Recruitment. *Alzheimer Disease & Associated Disorders*, 35:141–147, 2021.

Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., et al. Alzheimer's disease. *The Lancet*, 397:1577–1590, 2021.

Schoeler, T., Speed, D., Porcu, E., Pirastu, N., Pingault, J.-B., et al. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour*, pages 1216—1227, 2023.

Schork, N. J. Genetics of Complex Disease. *American Journal of Respiratory and Critical Care Medicine*, 156:S103–S109, 1997.

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature Genetics*, 50:928–936, 2018.

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51:D977–D985, 2022.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526:75–81, 2015.

Sun, L., Wang, Z., Lu, T., Manolio, T. A., and Paterson, A. D. eXclusionarY: 10 years later, where are the sex chromosomes in GWASs? *The American Journal of Human Genetics*, 110:903–912, 2023.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., et al. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, pages 467–484, 2019.

Torkamani, A., Wineinger, N. E., and Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19:581—590, 2018.

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, pages 502–508, 2022.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., et al. Genome-wide association studies. *Nature Reviews Methods Primers*, 1:59, 2021.

Veiga, A. and Levy, Y. Optimal contract regulation in selection markets. *Available at SSRN 4029945*, 2022.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., et al. Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. *PLoS Genetics*, 2:e41, 2006.

Visscher, P. M., Yengo, L., Cox, N. J., and Wray, N. R. Discovery and implications of polygenicity of common diseases. *Science*, 373:1468–1473, 2021.

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51:1349–1348, 2019.

Willer, C. J., Li, Y., and Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26:2190–2191, 2010.

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E., et al. Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, 9:1192–212, 2014.

Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., et al. A saturated map of common genetic variants associated with human height. *Nature*, 610:704–712, 2022.

Zhang, Y. D., Hurson, A. N., Zhang, H., Choudhury, P. P., Easton, D. F., et al. Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nature Communications*, 11:3353, 2020.

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

H Online Appendix Figures

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

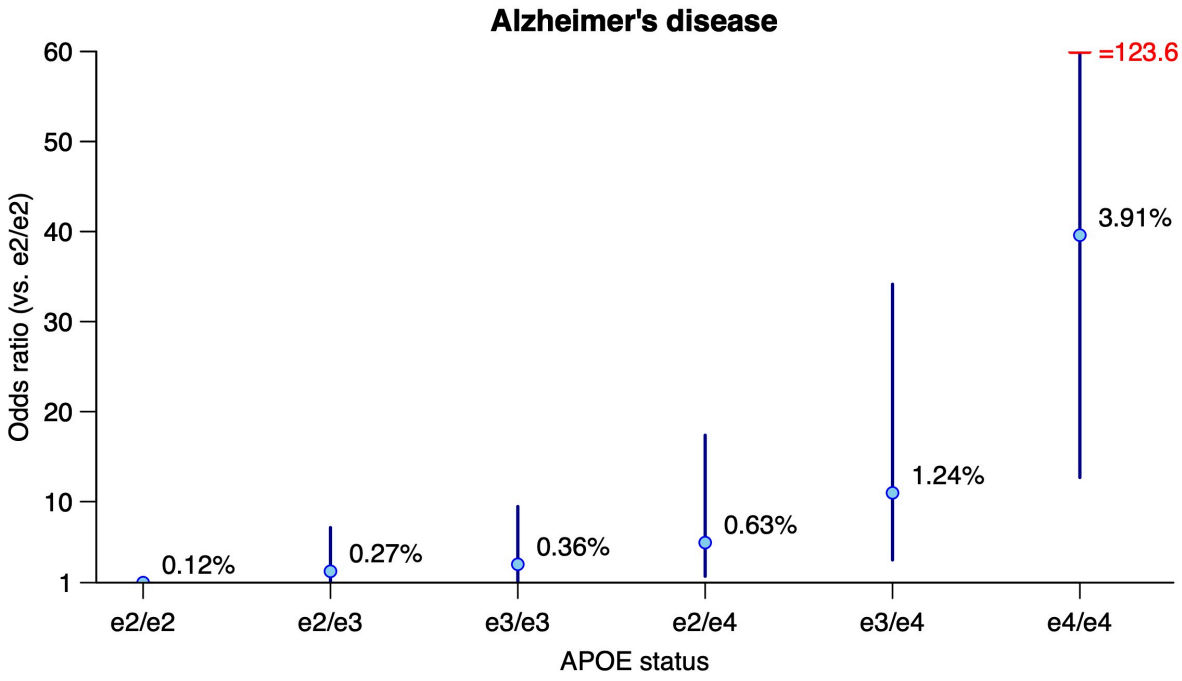


Figure 1: The effect of *APOE* risk type status on the risk of Alzheimer's disease. The error bars represent 95% confidence intervals. The odds ratios were estimated among 405,729 UK Biobank participants with logistic regression controlling for age, sex, and the first ten genetic PCs. The percentages report the observed proportion of cases by each *APOE* risk type. The *APOE* status variable could explain about 4.2% of the variance on the latent disease scale.

PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

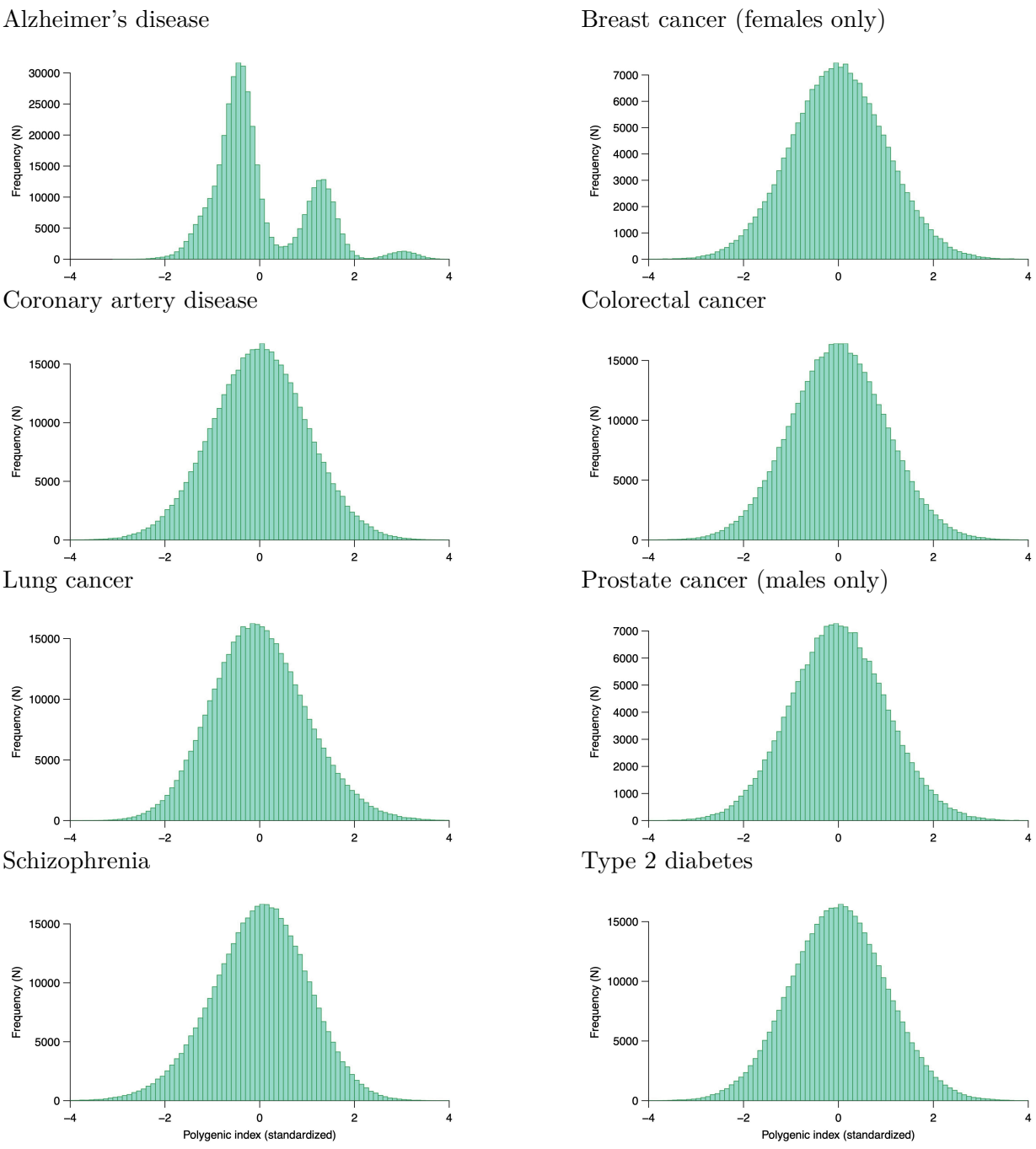


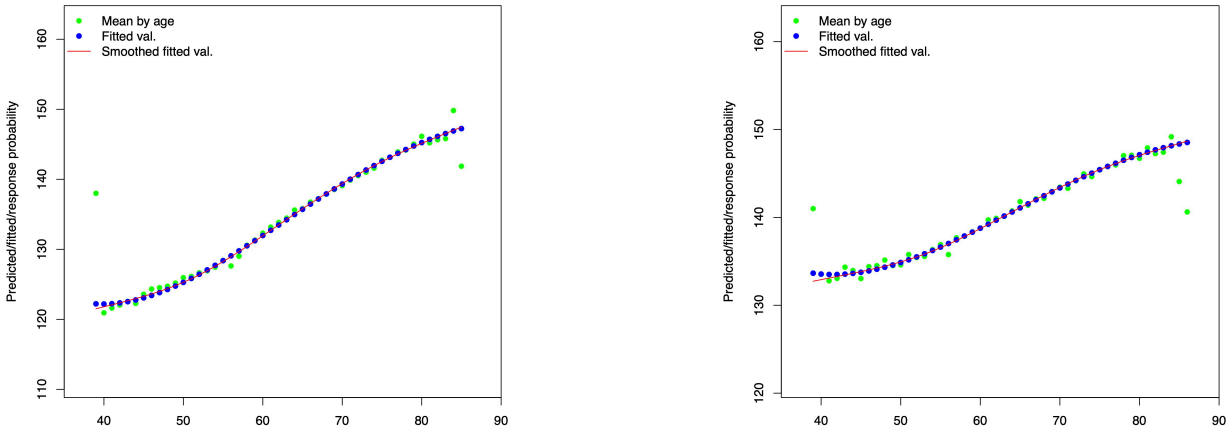
Figure 2: Histograms of the PGI distributions

Notes: The histograms display the distribution of the PGIs for our eight diseases of interest. The PGIs were standardized to mean zero and unit variance separately by sex. The PGI for Alzheimer's disease has three modes because this PGI captures the *APOE* risk types, for more information please see section B.3.

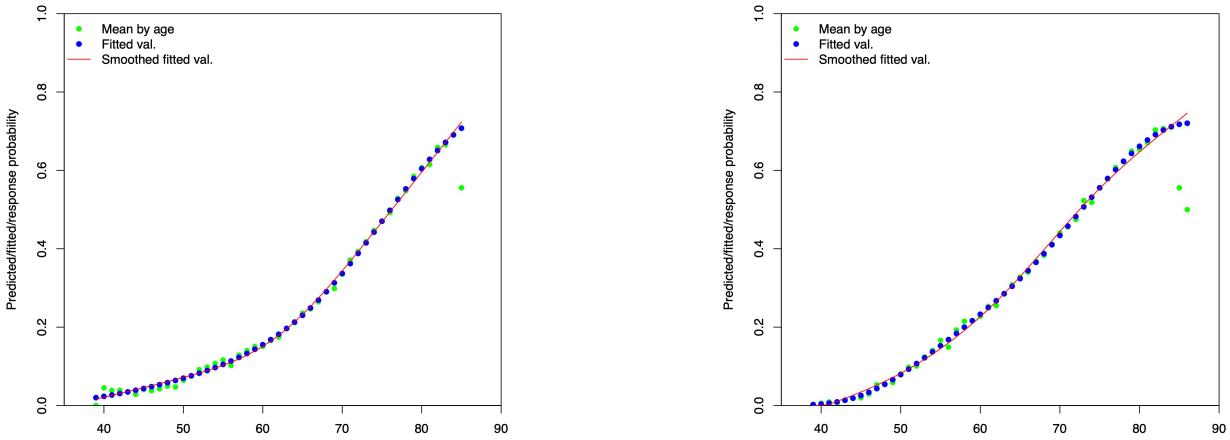
PRELIMINARY DRAFT: DO NOT CIRCULATE OR CITE

Females **Males**

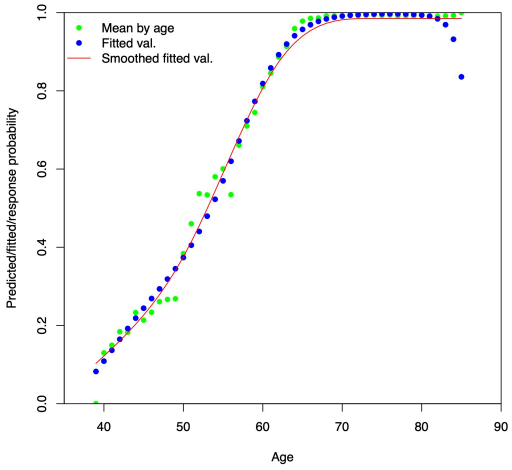
Systolic blood pressure



Hypertension



Breast cancer screening



Prostate cancer screening

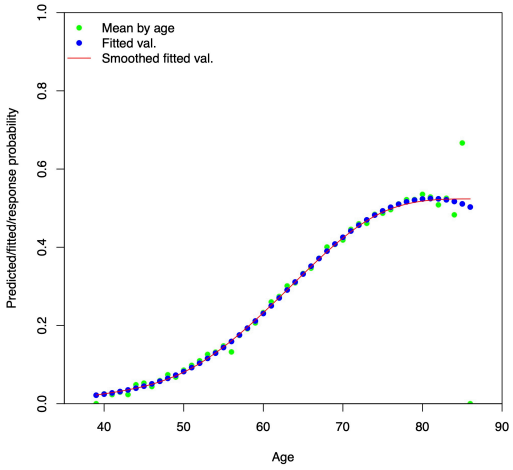


Figure 3: Interpolation of covariate values (selected examples)

Notes: The figure shows the predicted values (blue dots) from a regression of the covariate on a polynomial of age to the fourth degree. For binary covariates, the predicted values are probabilities. Green dots show the sample means by year of age. The red line shows a cubic smoothing spline applied to the predicted values to force monotonicity at the youngest and oldest ages. The interpolation procedure is described in section D.3.

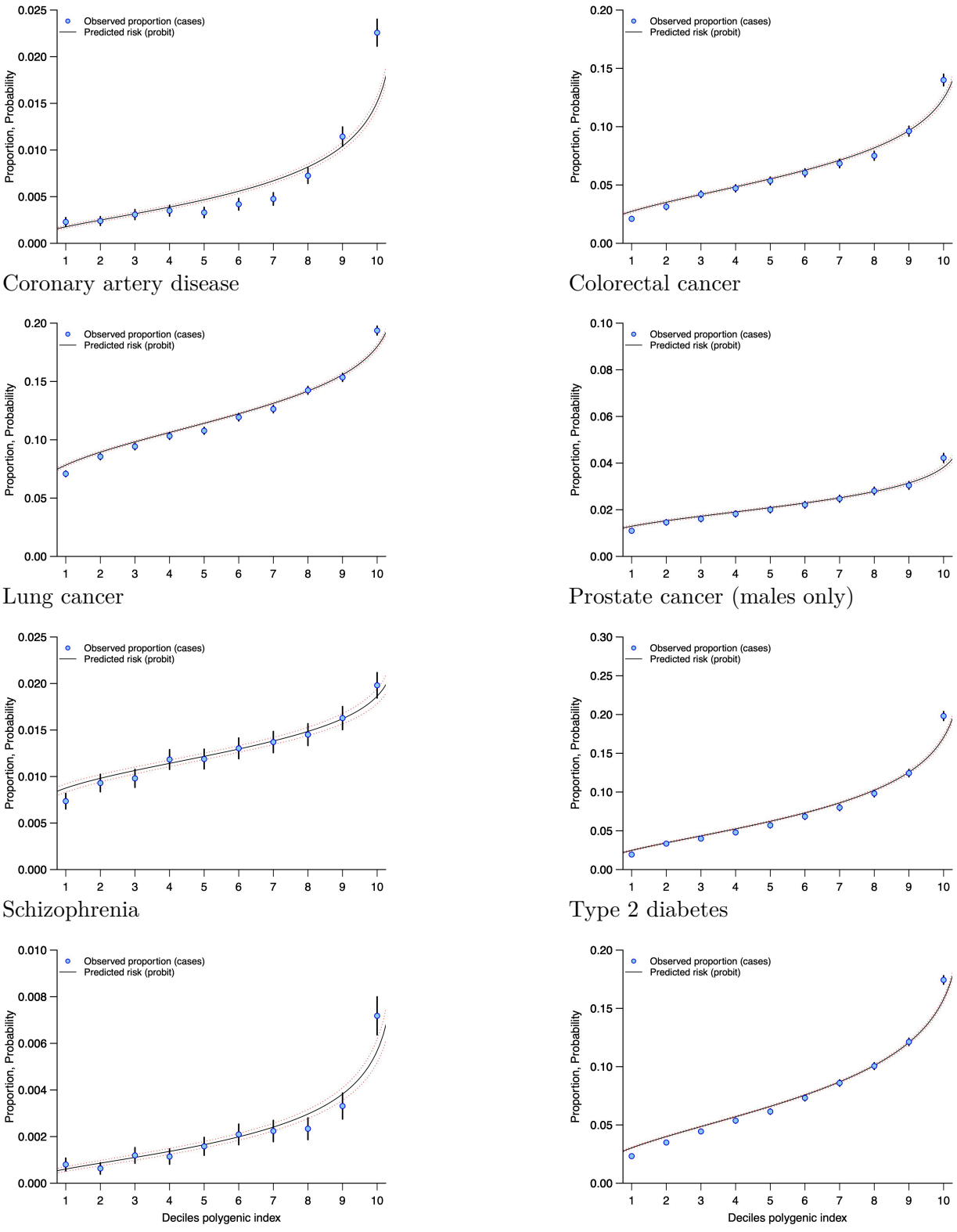


Figure 4: Observed and predicted disease risk by deciles of the PGIs

Notes: The figures display the proportion of (observed) disease cases by PGI decile, as well as the predicted disease risk (i.e., a probability) as estimated with an ad hoc probit regression of each disease on the the PGI together with age, sex, the first ten genetic PCs, a dummy for genotyping array.