Measuring Selection in Insurance Markets due to Genetic Predictions

Eduardo Azevedo* Jonathan Beauchamp† Richard Karlsson Linnér‡

January 16, 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 a rich dataset with nearly 500,000 genotyped individuals, we 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, insurance, polygenic indexes, genomic risk prediction, genetic testing

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

*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

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 for various diseases based on the 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. Building on these advances, private genetic testing companies such as 23andMe and MyHeritage have begun providing consumers with PGI-based risk assessments for various conditions (23andMe, 2023; MyHeritage, 2021). By introducing risk assessments for common diseases, PGIs complement the existing class of diagnostic genetic tests for rare diseases already used in healthcare. 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).

Genotyping costs are now comparable to those of blood or cholesterol tests (and still plummeting). 23andMe charges $229 for genotyping, while Labcorp charges $169 for their comprehensive blood 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 or utilizing genetic information known to applicants, e.g., from a consumer genetic test, 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 the unraveling of these market segments and thus leave all their consumers, including the most vulnerable ones, uninsured (Akerlof, 1970; Einav et al., 2023).

This study combines large-scale genetic data, economic theory, and econometrics to inform the debate on genetics and adverse selection in insurance markets. We focus on the

[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).

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 that are 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, but our methodology and results are also relevant to risk-rated health, life, and long-term care insurance markets.

This study makes three key contributions. The first two contributions are methodological. 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 (i.e., medical diagnoses). 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 and incremental predictive power over the covariates that are typically observed and used by insurers in underwriting.

Third, we apply the method to the critical illness insurance market to measure selection with both current- and future-day PGIs. 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 all 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 share of the variation in disease risk. For instance, the current-day PGI of prostate cancer accounts for 9.2% of the variation and—based on heritability studies—we expect future-day

[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).

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 together with econometric modeling to examine this question. Given our findings of considerable potential for selection, we discuss policy remedies in the end of the paper.

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 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 this 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- or future-day PGI.

Figure 1 shows results based on UKB data. Figure 1a shows the distribution of the risk

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

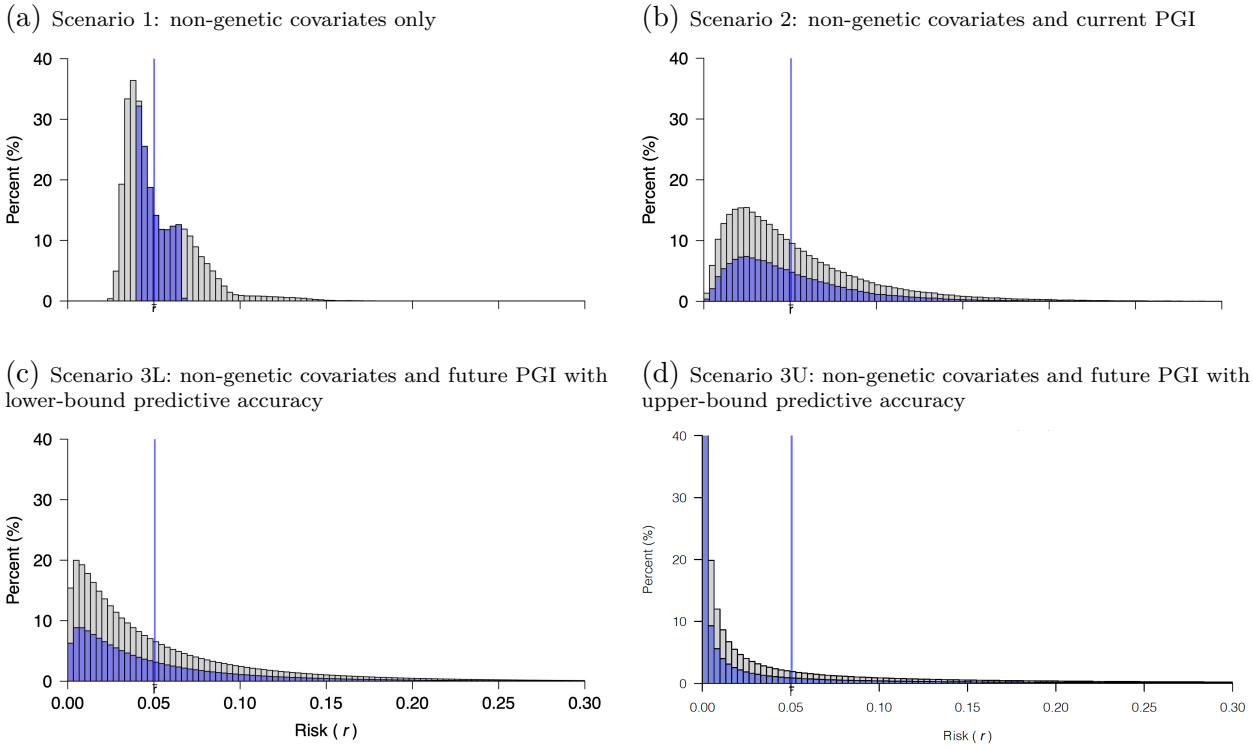


Figure 1: Distribution of the risk of contracting prostate cancer by age 65, conditional on various information sets and computed in the UKB data ($N = 181,642$). The standard risk class is shaded blue. The vertical blue line marks the average population risk ($\bar{r} = 0.053$) estimated in our data.

of contracting prostate cancer by age 65 (for males), conditional on non-genetic covariates only (Scenario 1). 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.

Figure 1b shows the distribution of risk conditional on both the non-genetic covariates and the current-day PGI of prostate cancer (Scenario 2). 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. Figures

1c and 1d show the distribution with two future-day PGIs with different assumed accuracies, both informed by existing estimates of the heritability of prostate cancer. We will consider two future-day PGIs with accuracies corresponding to what are considered reasonable lower (Scenarios 3L) and upper (Scenarios 3U) bounds. In Figure 1c, 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 Figure 1d, 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). Results for prostate cancer imply very high levels 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 diseases, 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, focusing closely on prostate cancer. Section 6 discusses our main results for all the diseases. Section 7 discusses the policy implications of our findings, and 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 are called "monogenic" disorders. 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] The proposed research focuses uniquely on polygenic diseases.

To understand how genetic data can be 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 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 $\mathcal{L}^*$:

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

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. (Throughout, we denote unobservable variables with a "*", except for error terms; random variables with uppercase and their realizations and constants with lowercase letters; and vectors with bold font.) 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 of D: $G = \mathbf{X}\,\widehat{\mathbf{b}}$. G is the **polygenic index** (**PGI**) of D.[7]

[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).

[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

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.

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 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, defined as the share of the trait variation that is attributable to additive genetic factors.

Table 1 shows, for the six 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

of correlation across variants and use these to transform the GWAS coefficients into estimates of $\mathbf{b}^*$ (i.e., into $\widehat{\mathbf{b}}$).

[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).

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)	4.9%	31%	58%
Breast cancer (F only)	8.1%	12.3%	5.3%	14%	31%
Coronary artery disease	3%	51.7% (M), 39.2% (F)	2.3%	22%	40%
Prostate cancer (M only)	0.9%	12.7%	9.2%	18%	57%
Schizophrenia	0.46%	0.7%	5.1%	24%	79%
Type 2 diabetes	3.9%	27% (M), 18.6% (F)	7.4%	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 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.,

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

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. 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). 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, and Germany. It is estimated that 20% of British workers were covered by a CII policy in 2009 (Gatzert and Maegebier, 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 blood 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 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 various disorders. 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 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, which we saw earlier, displays in blue four possible distributions of the risk of contracting prostate cancer by age 65, for consumers in the standard risk class. The distribution in Figure 1a, where consumers only have non-genetic private risk, 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 Figures 1c and 1d, where consumers observe risk conditional on 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 a 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]

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 $\mathbb{P}_{\text{data}}$.

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

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.

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* $\epsilon \perp (\mathbf{W}, V)$.

The fourth assumption is that D follows a logit 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:*

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

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 ratio of the variances of G_c and G_f equals the inverse of the known ratio of their R^2's (on the liability scale):*

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

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 genotyping data for ∼800,000 genetic variants from across the genome, 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 ∼1.1 million commonly used and well-characterized variants from the HapMap 3 Project (International HapMap Consortium, 2005). 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. 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.[12,13,14]

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.[15]

For each disease, we also 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 2 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 these risk factors in our analysis is critical, because they are typically observed and used by insurers in underwriting.

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

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

[14]As mentioned, Alzheimer's disease and breast cancer are polygenic diseases, but they can also be affected by variants with sizeable effects in the APOE and BRCA genes, respectively. We did not incorporate these specific variants in our Alzheimer's and breast cancer PGIs, because doing so would require sequencing data and because they correlate with the variants used to construct our PGIs, and so our PGIs partly capture these variants.

[15]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).

Table 2: Non-genetic risk factors for selected diseases

Disease	Key risk factors
Baseline risk factors (for all diseases)	Age, sex, Townsend Deprivation index, years of schooling, BMI, alcohol consumption, smoking (current, ex, or never), physical inactivity, systolic blood pressure, family history (did each of father, mother, brothers, or sisters ever suffer from the disease?)
Alzheimer's disease	Air pollution, [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
Coronary artery disease	Depression, diabetes, [diet], hypercholesteremia, hyperglycemia, hypertension, physical inactivity, 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], hypercholesteremia, hyperglycemia, hypertension, social isolation

Notes: The risk factors for each disease were identified from a review of the epidemiological literature. Variables capturing all identified factors were coded in the UKB, except for a few for which data was not available (indicated in square brackets).

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 Institute of Actuaries, 2012). There are also less comprehensive contracts such as cancer insurance and even single-disease CII contracts.

For our analysis, we focus on 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.

We will consider the six diseases in Table 1. These include some of the most common cancers (prostate and breast), 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.

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 interest rates.

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. In the Online Appendix, we report results for the **high risk class**, which comprises consumers with risk between 1.75 and 2.25 times the population average (Brackenridge et al., 2006).

PGIs and covariates. For each disease, we construct the current PGI G_c based on the genetic data from the UKB and prior genomic association studies, as described in Section 4.1. Table 2 lists the covariates $\mathbf{W}$ for each disease.

Scenarios and information sets. For each disease, we produce results under the

four Scenarios mentioned in the Introduction. In all four Scenarios, insurers only observe a number of discrete consumer risk classes (including the standard and high risk classes).[16] 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 ($R_f^2 = h_{SNP}^2$), and in Scenario 3U it is equal to the disease's twin heritability ($R_f^2 = h_{twin}^2$).

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 of the six diseases, our estimation is based on the identification Theorem 1. The theorem shows that all parameters of our 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 of the six diseases, our method proceeds as follows:

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

[16] 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).

cancer, for which we only use females and males, respectively. 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).[17]

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. As explained in Section 2.1, we see the former as a reasonable lower bound and the latter as an upper bound. 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).

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 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 2) 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

[17] Assumption 5 does not make sense in Scenario 1, but the model still applies if we maintain Assumptions 1-4 and let $\theta = \sigma_\epsilon^2 = \sigma_V^2 = 0$. In Scenario 2, the model applies if we let $G_f = G_c$.

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 prostrate cancer. Recall that we consider typical CII buyers in young adulthood. Each buyer can buy a one-time insurance contract that pays a lump if ever he contracts prostate cancer by age 65. For prostate cancer, we use only the male sample.

The risk distributions for prostate cancer under Scenarios 1-3, shown in Figure 1, formed the basis of our illustrative example in Section 1.1. Figure 1a plots the distribution of non-genetic risk (the expectation of a loss conditional on non-genetic covariates) in Scenario 1. The distribution is bimodal, with the mode to the right due to family history being 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, however, insurers do 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 1a corresponds to the distribution of non-genetic private risk P for the standard risk class.

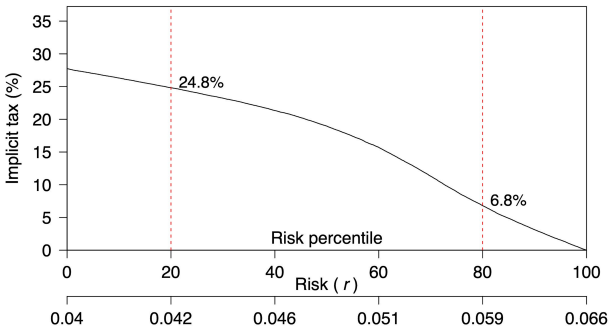
We now consider selection under the current prediction technology. Figure 1b shows in blue the distribution of private risk R (the expectation of a loss conditional on covariates and the current PGI) for the standard risk class consumers in Scenario 2. The current PGI adds considerable explanatory power. 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, there is 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.

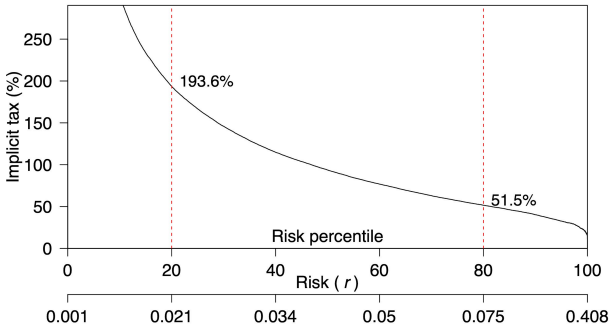
The amount of selection grows considerably with future prediction technology. Figures 1c and 1d plot, in blue, the distribution of private risk for the standard risk class based on future-day PGIs. Figure 1c uses the lower bound for the future PGI's R^2 based on prostate cancer's SNP heritability (Scenario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

We can use the implicit tax to quantify the amount of selection and compare it to that in previously studied insurance markets. Figure 2 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}.

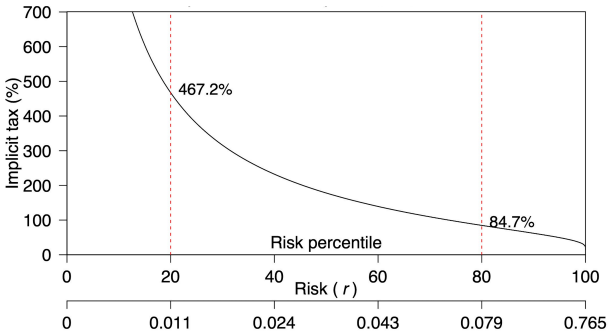
(a) Scenario 1: non-genetic covariates only



(b) Scenario 2: non-genetic covariates and current PGI



(c) Scenario 3L: non-genetic covariates and future PGI with lower-bound predictive accuracy



(d) Scenario 3U: non-genetic covariates and future PGI with upper-bound predictive accuracy

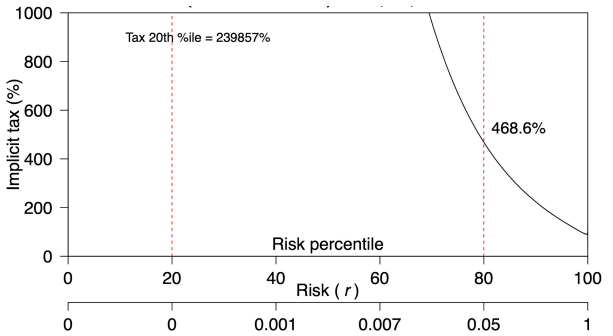


Figure 2: Implicit tax for a CII contract insuring against the onset of prostate cancer by age 65, for a consumer in the standard risk class. The sub-figures show the tax, as a function of the consumer's percentile risk of contracting prostate cancer, with the risk computed conditional on different information sets observed by the consumer.

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 actuarially fair price for the segment of consumers with risk higher 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 9.2%—see Table 1), t_{80} is now 76.9%. 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: the majority of consumers have never been genotyped and those who have been have not seen detailed predictions of their risk of contracting prostate cancer and other diseases, which we expect will become widely available in the future. Therefore, this result 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 heritability ($R^2_{f,SNP\,h^2} = 18\%$), adverse selection increases considerably, with t_{80} now reaching 118%. And in Scenario 3U, with a future-day PGI with predictive power equal to the twin heritability ($R^2_{f,twin\,h^2} = 57\%$), adverse selection increases dramatically, with t_{80} reaching 522%. 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 (let alone in functioning market segments). This suggests the market for CII for prostate cancer may unravel when future genetic prediction technology becomes widely available, if insurers are prohibited from using genetic information in underwriting.

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. These consumers know that they are basically safe. So, to cover most consumers, we would need these 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

[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.

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 six single-disease CII contracts we study. Figure 3 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 3 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 significant. t_{80} ranges from 18% for coronary artery disease to 86% for Alzheimer. The next lowest minimum implicit taxes after coronary artery disease's are the ones for breast cancer and type 2 diabetes, at are around 35%, in the upper range of market segments where Hendren (2013) found no unraveling. 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 possible, and the only question is when genetic prediction technology will become widely available.

The second main finding is that, with the expected predictive power of future PGIs, adverse selection will become crippling if access to genetic prediction becomes widespread. 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 62% for breast cancer to 900% for Alzheimers. This is considerably higher than what Hendren (2013) found even in market segments 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%

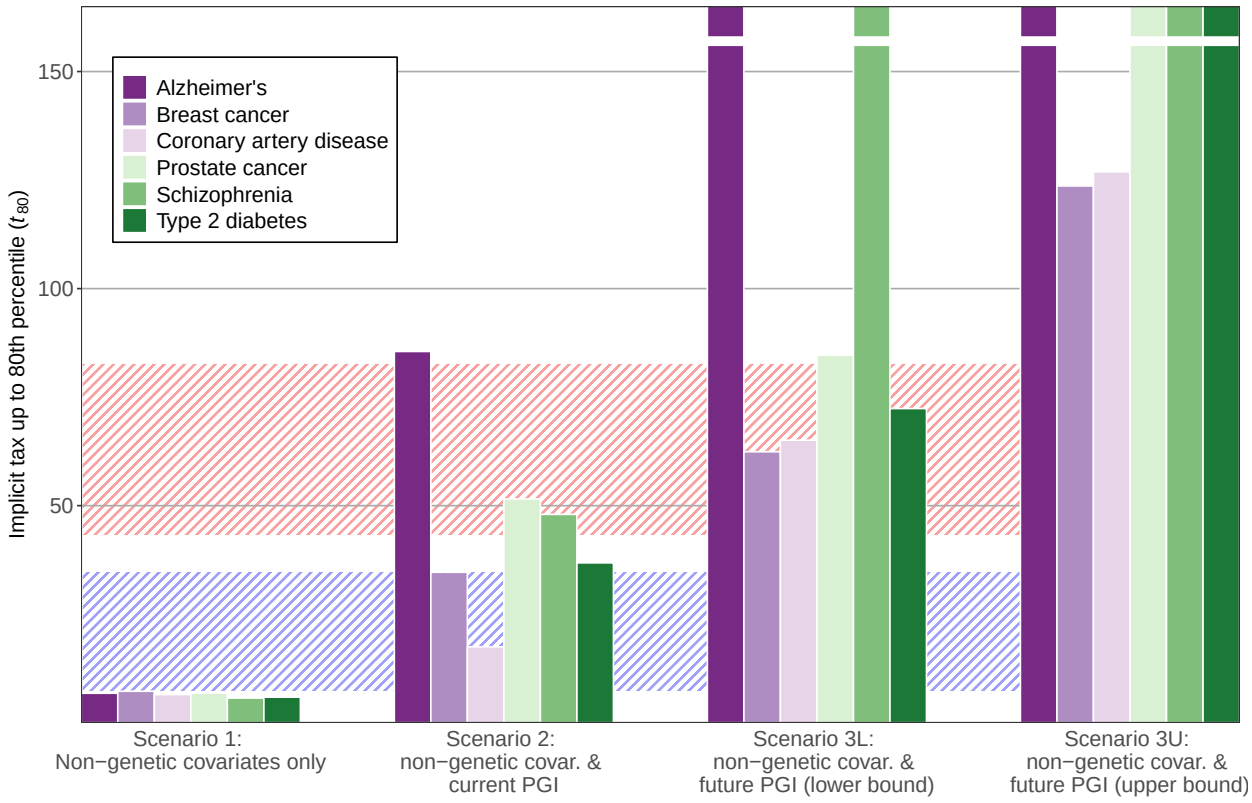


Figure 3: Minimum implicit tax up to the 80th percentile (t_{80}) for different single-disease contracts and scenarios. The figure displays t_{80} for each of the single-disease CII contracts and each of the four scenarios. The striped blue area corresponds to the range of t_{80} observed by Hendren (2013) in market segments that had not unraveled, and the striped red area corresponds to the range for market segments that had unraveled. The white horizontal bar breaks for the tallest bars indicate that these bars' heights exceed the figure's.

for all six diseases and exceeding 1,000% for three of the six diseases. This would imply extreme levels of selection. These findings suggest that, with widespread testing and future prediction technology, the markets for many single-disease CII contracts may unravel.

The third main finding is that there is substantial variation between the single-disease CII contracts. The minimum implicit taxes seem to be mainly driven by two factors. First, how good is the genetic prediction technology, as measured by the PGI's R^2. Second, how much of the information contained in the PGI is new relative to the observable covariates (Table 3). One example is coronary artery disease, which has the lowest current PGI R^2 and whose future R^2 is among the lowest based on the SNP heritability lower bound. Accordingly, coronary artery disease has the lowest implicit tax in Scenario 2, and the second lowest implicit tax in Scenario 3L. The only disease with a lower implicit tax in Scenario 3L is breast cancer.

7 Policy implications

Our results have four main policy implications. First, the current policy of genetic information bans can lead to severe market failures. Second, optimal regulation faces standard issues that have been widely studied in healthcare and insurance economics and that practical insurance regulation has long sought to address. Third, the political economy of optimal regulation is complex. Fourth, there is some potential for policies that are specially geared towards genetics.

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. For example, in life insurance and health insurance, consumers have private information about expected costs due to variables like family history. A market where consumers have private information that predicts costs is known as a selection market. There is a large amount of research and practical experience on how to regulate selection markets to address the market failures (Einav et al., 2010b, 2021, 2023) due to asymmetric information. The problem of regulating genetic information in insurance is similar. Therefore, we can draw on extensive research and practical experience from other selection markets.

We will give a brief overview of the literature and 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.

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—incluing 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 contracting 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.

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, we only consider single-disease CII contracts. We plan to include bundled contracts in the next version of this paper. Second, 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. Third, 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. Fourth, our counterfactual scenarios assume that genetic testing will be commonplace and that test results will be widely available and understood. While we strongly expect this will be the case, some in the medical community disagree. Lastly, we have assumed throughout that the predictive power of future genetic technology can be bounded by current-day heritability estimates. However, future-day heritability estimates—which is what matters for future genetic prediction—may be lower if, for instance, individuals respond to genetic tests indicating a high risk for some disease by modifying their behavior to mitigate that risk. 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. URL https://www.23andme.com/test-info/genetic-health/.

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

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

G. Akerlof. 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, October 2022. URL https://www.alliedmarketresearch.com/critical-illness-insurance-market-A19460.

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

J. P. Beauchamp, D. Cesarini, M. Johannesson, M. J. van der Loos, P. D. Koellinger, P. J. Groenen, J. H. Fowler, J. N. Rosenquist, A. R. Thurik, and N. A. Christakis. Molecular genetics and economics. *Journal of Economic Perspectives*, 25(4):57–82, 2011.

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

D. J. Benjamin, D. Cesarini, C. F. Chabris, E. L. Glaeser, D. I. Laibson, G. S.-R. S. Age, V. Guðnason, T. B. Harris, L. J. Launer, S. Purcell, et al. The promises and pitfalls of genoeconomics. *Annu. Rev. Econ.*, 4(1):627–662, 2012.

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

C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, 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.

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

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

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

J. C. Denny, L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink, J. D. Mosley, J. R. Field, J. M. Pulley, A. H. Ramirez, E. Bowton, 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.

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

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

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

L. Einav, A. Finkelstein, and R. Fisman. *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.

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

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

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

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

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

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

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

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

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

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

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

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

D. Holland, O. Frei, R. Desikan, C.-C. Fan, A. A. Shadrin, O. B. Smeland, V. S. Sundar, P. Thompson, O. A. Andreassen, and A. M. Dale. Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate gaussian mixture model. *PLOS Genetics*, 16(5):e1008612, 2020.

R. C. Howard. 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. URL www.cia-ica.ca/docs/default-source/2016/216002e.pdf.

International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437 (7063):1299–1320, 2005.

Y. Joly, H. Burton, B. M. Knoppers, I. N. Feze, T. Dent, N. Pashayan, S. Chowdhury, W. Foulkes, A. Hall, P. Hamet, N. Kirwan, A. Macdonald, J. Simard, and I. Van Hoyweghen. Life insurance: Genomic stratification and risk classification. *European Journal of Human Genetics*, 22:575–579, 2014.

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

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

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

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

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

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

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

A. S. Macdonald, H. R. Waters, and C. T. Wekwete. 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.

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

J. M. Maxwell, R. A. Russell, H. M. Wu, N. Sharapova, P. Banthorpe, P. F. O'Reilly, and C. M. Lewis. 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.

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

M. C. Mills and C. Rahal. 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. URL https://education.myheritage.com/article/understanding-your-polygenic-risk-reports/.

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

A. Okbay, Y. Wu, N. Wang, H. Jayashankar, M. Bennett, S. M. Nehzati, J. Sidorenko, H. Kweon, G. Goldman, T. Gjorgjieva, 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.

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

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

F. Privé, H. Aschard, S. Carmi, L. Folkersen, C. Hoggart, P. F. O'Reilly, and B. J. Vilhjálmsson. 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.

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

S. Saya, L. Boyd, P. Chondros, M. McNamara, M. King, S. Milton, R. D. A. Lourenco, M. Clark, G. Fishman, J. Marker, 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.

M. Simcoe, A. Valdes, F. Liu, N. A. Furlotte, D. M. Evans, G. Hemani, S. M. Ring, G. D. Smith, D. L. Duffy, G. Zhu, 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.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, 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.

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

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

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

E. Widén, N. Junna, S. Ruotsalainen, I. Surakka, N. Mars, P. Ripatti, J. J. Partanen, J. Aro, P. Mustonen, T. Tuomi, 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.

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

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

W. Zhou, M. Kanai, K.-H. H. Wu, H. Rasheed, K. Tsuo, J. B. Hirbo, Y. Wang, A. Bhattacharya, H. Zhao, S. Namba, et al. Global biobank meta-analysis initiative: Powering genetic discovery across human disease. *Cell Genomics*, 2(10), 2022.

Table 3: Summary of main results

	ALZ	BRC	CAD	PRC	SCZ	T2D
Males and/or females	M & F	F only	M & F	M only	M & F	M & F
Sample size	405,746	184,143	410,858	181,642	410,858	410,858
Cases	2,628	11,697	49,174	13,927	925	31,759

Panel A. Explanatory power of non-genetic covariates and PGIs

			Scenarios 1-3			
R^2 of covariates	36.36%	7.67%	37.09%	23.57%	19.48%	40.13%
			Scenario 2			
R^2 of G_c	4.91%	5.32%	2.32%	9.24%	5.14%	7.39%
R^2 of G_c & covar.	39.84%	12.41%	38.01%	31.25%	22.64%	43.68%
ΔR^2 of G_c	3.47%	4.74%	0.92%	7.68%	3.16%	3.56%
			Scenario 3L			
R^2 of G_f (= SNP-h^2)	33.1%	14.0%	22.0%	18.0%	24.0%	19.6%
R^2 of G_f & covar.	63.3%	20.1%	44.5%	38.5%	34.9%	47.5%
ΔR^2 of G_f	26.9%	12.4%	7.4%	14.9%	15.4%	7.4%
			Scenario 3U			
R^2 of G_f (= twin-h^2)	58.0%	31.0%	40.0%	57.0%	79.0%	72.0%
R^2 of G_f & covar.	83.9%	37.1%	60.0%	74.3%	90.0%	92.9%
ΔR^2 of G_f	47.5%	29.4%	22.9%	50.7%	70.5%	52.8%

Panel B. Risk of contracting disease (r) by age 65, conditional on non-genetic covariates only

			Full analysis sample			
Mean (SD)	0.2% (0.1 pp)	5.7% (1.7 pp)	8.2% (8.9 pp)	5.3% (2.1 pp)	0.2% (1.1 pp)	5.6% (8.1 pp)
20–80th pct	0.1–0.2%	4.5–6.6%	2.4–11.7%	3.7–6.8%	0.0–0.2%	0.6–8.7%
			Standard risk class (r = 0.75–1.25 of sample mean)			
Mean (SD)	0.1% (0.0 pp)	5.4% (0.7 pp)	8.0% (1.2 pp)	5.0% (0.8 pp)	0.2% (0.0 pp)	5.5% (0.8 pp)
20–80th pct	0.1–0.2%	4.8–6.1%	6.8–9.2%	4.2–5.9%	0.2–0.3%	4.7–6.3%
			High risk class (r = 1.75–2.25 of sample mean)			
Mean (SD)	0.3% (0.0 pp)	10.9% (0.7 pp)	16.3% (1.2 pp)	10.4% (0.8 pp)	0.4% (0.0 pp)	11.2% (0.8 pp)
20–80th pct	0.3–0.3%	10.3–11.6%	15.1–17.5%	9.6–11.3%	0.4–0.5%	10.4–12%

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	75.0%	29.3%	114.8%	27.9%	389.5%	134.6%
			Standard risk class (r = 0.75–1.25 of full sample mean)			
Scenario 1	6.8%	7.2%	6.5%	6.8%	5.7%	5.9%
Scenario 2	85.5%	34.6%	17.5%	51.5%	48.0%	36.8%
Scenario 3L	901.2%	62.4%	65.1%	84.7%	200.3%	72.4%
Scenario 3U	>10,000%	123.6%	126.9%	468.6%	>10,000%	2477.7%
			High risk class (r = 1.75–2.25 of full sample mean)			
Scenario 1	3.0%	5.2%	3.5%	3.1%	2.7%	3.0%
Scenario 2	76.4%	30.3%	14.0%	42.6%	43.3%	30.4%
Scenario 3L	639.5%	53.0%	54.3%	69.0%	167.9%	59.9%
Scenario 3U	>10,000%	98.4%	98.9%	259.3%	>10,000%	642.0%

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.

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
Assumption 3 implies that

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

Plugging in Assumption 5 and $\text{Var}[G_c] = 1$ 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

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

□

Observe that Theorem 1 and the above derivations also hold with the current-day PGI (Scenario 2), if we let $G_f = G_c$. Then, $a = 1$, $b = c = 0$, so $\beta_g = \gamma_g$ and $\beta_w = \gamma_w$—i.e., β_g and β_w are the untransformed coefficients from the probit regression (A.5) (and there is no need to obtain an estimate of θ by regressing the PGI on the covariates).

If there is no PGI (Scenario 1), then Assumption 5 no longer makes sense; if we maintain the other Assumptions 1-4 and let $\theta = \sigma_\epsilon^2 = \sigma_V^2 = 0$, then $a = b = c = 0$ and the probit regression (A.5) becomes a probit regression of the disease on the covariates only, with β_w equal to the untransformed coefficient γ_w from the regression.