Genetic prediction and adverse selection*

Eduardo Azevedo† Jonathan Beauchamp‡
Richard Karlsson Linnér§

First version: October, 2024
This version: May 14, 2026

Abstract

Recent advances have substantially improved our ability to predict disease risk with genetic data. In response, many countries have banned insurers from using genetic data, despite concerns about adverse selection. This paper measures adverse selection in a market where insurers can underwrite only on non-genetic information while consumers also have access to genetic information. We do so by combining methods from quantitative genetics, selection measures from economic theory, and genetic and electronic health record data on nearly 500,000 individuals in the UK Biobank. We focus on the critical illness insurance market and consider scenarios in which consumers have access to current or expected future genetic prediction technology. We find noticeable levels of selection with current prediction technology, and potentially crippling selection with expected future technology.

*For helpful feedback, we thank Martin Henriksson, Ralph Koijen, Patrick Turley, Peter Visscher, as well as conference and seminar participants at Stanford University, the University of Wisconsin-Madison, Uppsala University, Virginia Commonwealth University, Washington University in St. Louis, the Yale School of Management, the American Enterprise Institute, The Advances in Social Genomics Conference, the European Social Science Genetics Network conference, the Integrating Genetics in the Social Sciences conference, and the NBER Insurance Working Group meeting. For outstanding research assistance, we thank Gareth Markel. This research has been conducted using the UK Biobank resource under application numbers 217716, 99086, and 11425. This work uses data provided by patients and collected by the NHS as part of their care and support. This research was supported by the Dutch Research Council (NWO Veni project number VI.Veni.221E.080), by a Small Compute grant (NWO/SURF, no. EINF-8891), by a Faculty Research and Development Award (FRDA) at George Mason University, and by the National Science Foundation (NSF award nos. 2343735 and 2343736).

Replication materials will be archived in the journal repository at: https://zenodo.org

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

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

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

1 Introduction

Technological advancements have led to substantial increases in the predictive power of genetic data (Abdellaoui et al., 2023). Scientists can now create **"polygenic indexes"** (**PGIs**) from a person's genetic sequence that partially predict the risk of developing various common diseases, such as breast cancer or heart disease. Because common diseases are influenced by many genes, state-of-the-art PGIs sum the effects of a million or more genetic variants (Visscher et al., 2021). PGIs are complementary to the long-established use of genetic testing for rare disorders driven by a single gene, such as Huntington's disease.

The predictive power of PGIs has increased dramatically over the last ten years and has recently reached levels of clinical and personal utility (Kullo, 2025). The first clinical trials are now underway to evaluate whether PGIs can improve routine healthcare and disease screening. Large healthcare providers, such as the UK's National Health Service (NHS), have initiated trials to inform millions of patients about their PGIs. Thanks to the continued growth of massive biobanks, experts predict that PGIs' predictive power will continue to improve and to approach its theoretical upper limits in the coming decades (Abdellaoui et al., 2023).

While PGIs are not yet common in healthcare, consumers can purchase prototypes in the consumer genetics market. Tens of millions of people worldwide have purchased an at-home genetic test to explore their genealogy (Regalado, 2019). The market leaders now bundle genealogy services with PGI-based disease risk assessments, communicating absolute risk probabilities alongside comparisons to covariate-matched reference groups and the population average (23andMe, 2023; MyHeritage, 2021). The price of these assessments is declining fast, and they now often cost less than $150. These trends suggest that both patients and ordinary consumers may soon learn their PGI-based risks for various diseases.

Meanwhile, many jurisdictions have banned or restricted the use of genetic information by insurance companies (Golinghorst et al., 2022).[1] Such bans typically prohibit insurers from requesting the results of genetic tests known to applicants

[1]For a recent global overview, see Swiss Re Institute (2024a). In the US, the Genetic Information Nondiscrimination Act (GINA) bans the use of genetic information by health insurers. Some states have also legislated bans for other insurance types. Canada's Genetic Non-Discrimination Act (GNDA) and France's Penal Code both prohibit insurers from using genetic information. Some countries—including the UK, the Netherlands, and Sweden—generally ban the use of genetic information but permit it for large insured amounts or certain disorders, e.g., Huntington's disease.

or from considering any genetic information. Scholars, policymakers, and industry stakeholders have long voiced concerns that this information asymmetry may lead to adverse selection (Joly et al., 2014; Maxwell et al., 2021; Swiss Re Institute, 2024a). Selection could then lead to spiraling premiums and to the unraveling of market segments, thus leaving consumers uninsured (Akerlof, 1970; Einav, Finkelstein, and Fisman, 2023). Despite decades of research on the impact of genetic testing on insurance markets, debate remains over whether PGIs could eventually create much selection (Golinghorst et al., 2022; Dixon et al., 2024).

This study measures the potential for adverse selection due to PGIs in the market for **critical illness insurance** (**CII**). Although CII is less well known than life and health insurance, the US and global CII markets have matured into core segments of the insurance industry, with many millions of consumers relying on CII for protection (Section 2.2 provides more details). The typical CII policy provides a lump-sum payment upon diagnosis of any covered condition—typically cancers, heart disease, stroke, multiple sclerosis, paralysis, blindness, and terminal illnesses. CII is particularly well-suited for our analysis because it directly covers diseases that have been studied extensively in the genetics literature.

Our main data source is the UK Biobank (UKB) (Sudlow et al., 2015), a large dataset of 500,000 individuals with genotypic and rich health-related data, including linked patient records collected by the National Health Service (NHS). We first consider **single-disease contracts** that pay upon the onset of the covered disease. We model such contracts for seven diseases: Alzheimer's disease, breast cancer, coronary artery disease, colorectal cancer, prostate cancer, schizophrenia, and type 2 diabetes. Next, we analyze **multiple-disease contracts** that bundle these diseases into one contract that pays a lump sum upon the onset of any of the diseases.

For each contract, we measure the degree of selection that would arise if consumers had access to their current PGIs, while insurers are not allowed to use them. We consider a market segment defined by consumers with similar risk conditional on non-genetic risk factors. Because insurers underwrite on the non-genetic factors alone, they treat these consumers identically. We then measure the dispersion of risk within this segment when taking into account both non-genetic risk factors and the PGI. This dispersion captures the additional information available only to consumers, who observe both their genetic predictions and the non-genetic fac-

tors used by insurers.[2] Moreover, given the rapid improvements in PGI predictive power, we also model the distribution of risk assuming that consumers have access to future PGIs, with predictive power calibrated using heritability estimates.[3]

We report three main findings. First, there are noticeable levels of selection with the current genetic prediction technology, if we assume that all consumers observe their current PGIs. For most of the contracts, the level of selection is non-negligible, but lower than in previously studied market segments that have unraveled. Since most people do not get genetic testing today, these results are consistent with the fact that CII markets are growing and do not currently appear to be plagued by selection issues (Golinghorst et al., 2022; Gen Re, 2023b).

Second, with the expected future PGIs, the degree of selection becomes extremely high. For all contracts, we find degrees of selection that are within or above the range previously observed for market segments that had unraveled (Hendren, 2013). This is the case even for the multiple-disease contracts, which diversify risk across several uncorrelated diseases. Even in a scenario where only 50% of consumers have private knowledge about their PGIs, we find high levels of selection. These results suggest that the policy to simply ban insurers from considering genetic data is sometimes Pareto dominated. Some market segments may collapse, leaving consumers uninsured.

Third, we find that there is subtle variation in the amount of selection across contracts. For single-disease contracts, the amount of selection is mainly driven by the predictive power of the PGI that is incremental to the non-genetic risk factors.

We consider robustness checks, limitations of our analysis, and policy implications. In particular, we reproduce some of our main findings using data from the Health and Retirement Study (HRS). We also calibrate an equilibrium model of adverse selection using the HRS data.

A key explanation for the high levels of selection we find 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.

[2]We assume consumers rationally combine their non-genetic and genetic information to form an integrated risk prediction. Consumers can already access such integrated predictions via several channels, including genetic-testing companies and online risk calculators such as CanRisk; integrated predictions are also being evaluated in clinical trials and may become widely available in healthcare settings.

[3]Heritability refers to the proportion of variation in disease risk attributable to genetic factors.

For instance, the current PGI of prostate cancer accounts for 9.9% of the variation and—based on heritability studies—we expect the future PGI to account for 18.0 to 57.0%. By contrast, popular discussion of genetic prediction often focuses on rare mutations that individually have large effects on the risk of certain diseases, but that only explain a small proportion of the overall disease variation. Salient examples include mutations in the *BRCA* genes for breast cancer and in the *LDLR* gene for familial hypercholesterolemia and heart disease. Interestingly, these mutations are unlikely to create strong selection because they are rare and can be assessed from medical records or family history (Swiss Re Institute, 2024a).

1.1 Illustrative example: CII for prostate cancer

The gist of our paper can be understood with a simple example. Consider a single-disease contract for prostate cancer that pays $100,000 upon diagnosis of prostate cancer by age 65. The contract is purchased at age 35, many years before the usual age of onset of prostate cancer. The insurer's cost of selling this contract is roughly proportional to the buyer's risk of prostate cancer by age 65. We estimate the distribution of risk in a population conditional on various subsets of information (Scenarios 1, 2, 3L, and 3U, described below).

The left panel of Figure 1 shows results based on UKB data. Scenario 1 shows the distribution of the risk of prostate cancer by age 65 in an all-male population of typical consumers, conditional on non-genetic covariates only. This information is observed by both consumers and insurers and is used by insurers to set prices. There would be a non-trivial amount of selection if insurers were not allowed to use this information. But, in practice, insurers do underwrite to categorize consumers into risk classes that are priced differently.[4] Following industry practice (Brackenridge, Croxson, and Mackenzie, 2006), we define the **standard risk class** as the set of consumers whose risk of prostate cancer conditional on the non-genetic covariates is between 0.75 and 1.25 times the average population risk ($\bar{r} = 0.0547$). These consumers are shown in blue in the figure and face the same price. By definition, these consumers have almost the same risk conditional on the

[4]Underwriting is essential for the functioning of life and CII markets (see Section 2.2). The importance of underwriting can be seen from the fairly large dispersion of the population-wide risk distribution in Scenario 1 (colored gray), which reflects the substantial predictive power of the non-genetic covariates.

non-genetic covariates, and so there is almost no selection within the standard risk class in Scenario 1.

Scenario 2 shows the distribution of risk conditional on both the non-genetic covariates and the current PGI of prostate cancer. The risk distribution for the standard risk class now has a wider spread. Consumers can see their PGIs and so privately observe this risk. However, because insurers are banned from using genetic information, the standard-risk-class consumers all look like identical risks to them and are therefore charged the same price.

Scenarios 3L and 3U show the distribution of risk with two future PGIs with different assumed accuracies. These assumed accuracies are informed by existing estimates of the heritability of prostate cancer and correspond to what are considered reasonable lower (Scenario 3L) and upper (Scenario 3U) bounds for the accuracy of the future PGI.[5] In Scenario 3L, the standard risk class now has an even wider spread. Consumers at the 5th and 95th percentiles face risks of 0.4% and 16.5% of contracting prostate cancer, respectively. And in Scenario 3U, consumers at the 5th percentile face a negligible risk (evident from the spike at 0) vs. a 31.4% risk at the 95th percentile. Thus, consumers who face the same price of insurance vary widely with respect to their privately known risk. This suggests significant potential for adverse selection.

To quantify the degree of asymmetric information and adverse selection that would result from this situation, we employ the **implicit tax**, a measure of how much riskier average buyers are compared to a marginal buyer (Hendren, 2013). The right panel of Figure 1 shows the estimated implicit tax as a function of a given consumer's risk percentile for prostate cancer, for the consumers in the standard risk class. In Scenario 1, underwriting based on non-genetic risk factors reduces private information to tolerable levels. In Scenarios 2, 3L, and 3U, the implicit taxes from PGIs imply elevated levels of selection that exceed those that have been estimated for market segments that have unraveled. This, along with the results we present below for the other CII contracts we study, suggests that selection due to future genetic prediction technology is a first-order concern.

[5] These bounds correspond to different assumptions about the accuracy of the future PGI, and are not statistical bounds.

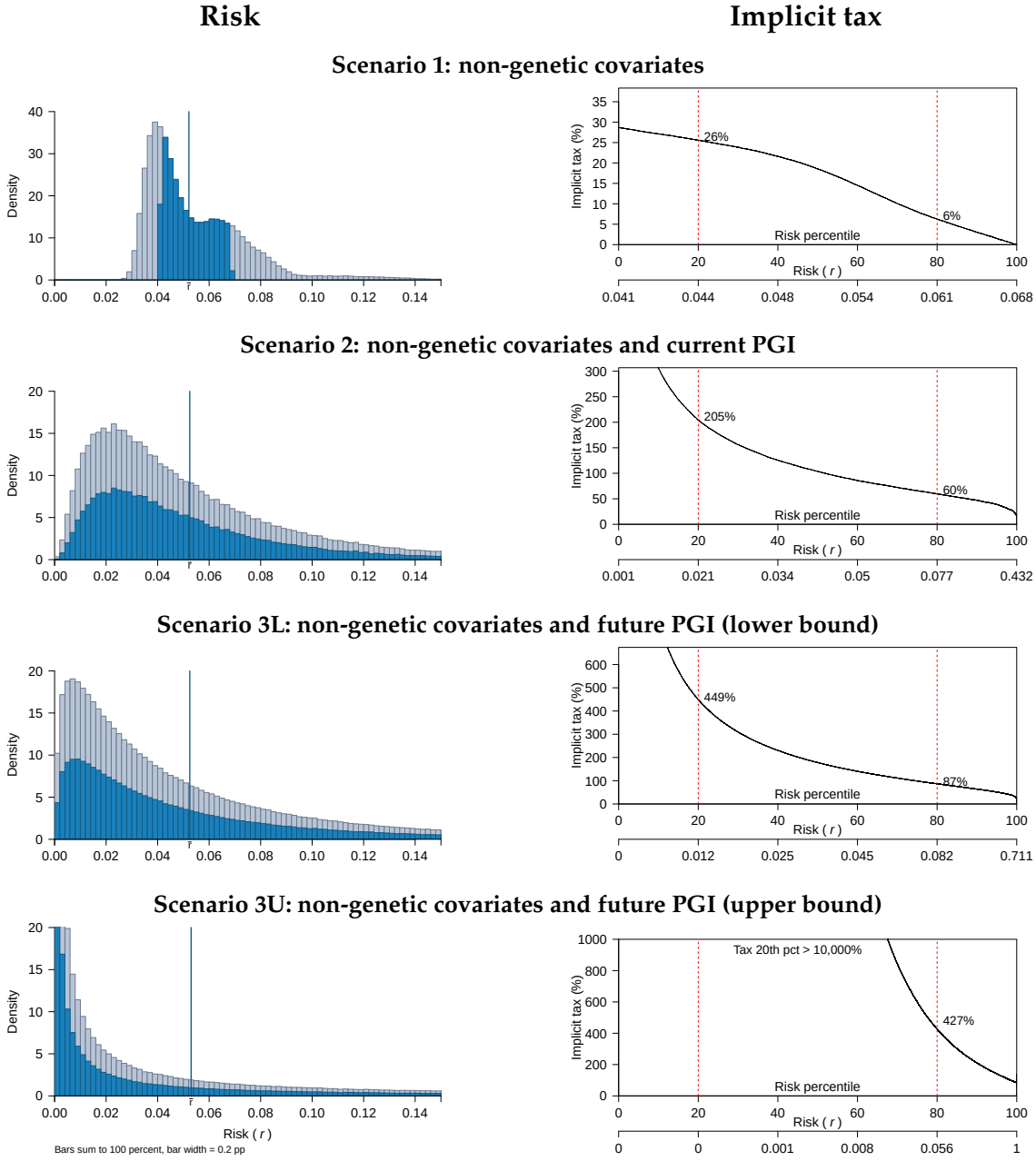

Figure 1: Single-disease CII contract for prostate: risk and implicit tax

Notes: Left panel: distribution of the risk of contracting prostate cancer by age 65, conditional on various information scenarios and computed in the UKB data ($N = 181,902$). The vertical blue line marks the average risk in the standard risk class ($\bar{r} = 0.0521$). The standard risk class individuals are shown in blue; insurers treat these individuals identically, so the blue distribution corresponds to the distribution of private risk for these individuals. Right panel: implicit tax for consumers in the standard risk class as a function of their percentile private risk of prostate cancer for each scenario.

1.2 Related literature

There is an extensive literature on asymmetric information and insurance markets (Akerlof, 1970; Rothschild and Stiglitz, 1976). As genetic research advanced in the 1990s, these canonical models were adapted to study adverse selection from genetic tests (Tabarrok, 1994; Doherty and Thistle, 1996). The literature on insurance and genetic testing expanded to consider health and life insurance (e.g., Strohmenger and Wambach, 2000; Hoy and Polborn, 2000; Hoel and Iversen, 2002) as well as policy interventions, including public insurance schemes and genetic information bans (e.g., Hindriks and De Donder, 2003; Polborn, Hoy, and Sadanand, 2006). Some work then continued on the welfare effects of policy (e.g., Barigozzi and Henriet, 2011; Bardey and De Donder, 2013). More recently, there has been renewed interest in examining policy in relation to health insurance and in studying aspects like preventive medicine (e.g., Peter, Richter, and Thistle, 2017; Posey and Thistle, 2021), as well as how the revelation of risk type interacts across insurance markets (Boyer and Glenzer, 2021).

Separately from this mostly theoretical literature in economics, a series of papers in the actuarial literature have used life table modeling and simulated data to examine the potential impact of genetic testing on insurance markets, including CII (e.g., Macdonald, Pritchard, and Tapadar, 2006; Macdonald and McIvor, 2009; Macdonald and Tapadar, 2010). These papers mostly focused on single-gene mutations or predictors based on a handful of genetic variants—consistent with the technology at the time—and often concluded that genetic tests are unlikely to cause much adverse selection. In contrast, Howard (2016) considers known rare mutations for six conditions and finds that barring insurers from using genetic data would lead to high levels of adverse selection.

More recently, a few papers have studied PGIs and insurance. Karlsson Linnér and Koellinger (2022) benchmarked PGIs for life underwriting and found that they are on par with classic underwriting factors, but did not address adverse selection, the CII market, or future predictive power. Maxwell et al. (2021) use PGIs to predict the incidence of breast cancer and coronary artery disease and show the possibility of adverse selection in different insurance markets. Zhao, Salter-Townshend, and O'Hagan (2023) use simulated data to assess how a future PGI of heart disease might affect adverse selection in CII. They find that selection increases with PGI accuracy and can be substantial.

Our study builds on this literature. It is the first (to our knowledge) to combine large-scale genetic and individual-level health data with an econometric model of single- and multiple-disease contracts, and to account for the incremental predictive power of current and future PGIs over and above that of already observable risk factors.

Our study complements a literature that shows that people respond to genetic information when making insurance decisions. Oster et al. (2010) show that carriers of the variant causing Huntington's disease (affecting less than 0.01% of the population) were 5 times more likely to own long-term care insurance. Zick et al. (2005) and 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 up to 5.76 times more likely to buy or plan to buy long-term care insurance.

1.3 Paper structure

The paper is structured as follows. Section 2 provides background on genetics and the CII market. Section 3 describes the theory and estimation procedure. Section 4 describes the data. Section 5 presents the main results. Section 6 discusses limitations, robustness, and extensions. Section 7 concludes.

2 Background

2.1 Genetics background

Some conditions, such as Huntington's disease, are caused by a single gene. Such conditions are called monogenic (Lynch and Walsh, 1999). Some traits, like eye color, are mainly caused by a small number of variants. However, the vast majority of traits are **complex traits**: they are influenced by numerous genetic variants (as well as by the environment), with each causal variant having a tiny effect (Visscher et al., 2021). Height, BMI, and cognitive ability are quintessential complex traits. The seven diseases we study here are all examples of complex diseases.[6]

[6] Alzheimer's disease and breast cancer are also complex diseases, but in addition to being affected by numerous variants with tiny effects, they are affected by a handful of common variants (in the *APOE* gene, for Alzheimer's) or by rare variants (in the *BRCA* genes, for breast cancer) with large effects. Still, the bulk of their genetic variation is due to many variants with small effects.

Consider how genetic data are used to predict a continuous complex trait Y. The entire genome contains more than 3 billion locations.[7] But most of the variation among humans is concentrated at well under 5% of these locations. For that reason, typical genotyping datasets contain data on millions instead of billions of genomic locations—those where there are common genetic variants.[8]

As is common practice in genetic epidemiology, we use a simple predictive model. We model Y as the weighted sum of an individual's genetic variants plus a disturbance term ψ:

$$Y = k + \mathbf{X}\,\mathbf{b} + \psi,$$

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 (population) regression coefficients. $G^* = \mathbf{X}\,\mathbf{b}$ is the individual's **true genetic predictor** for Y. (Throughout, we denote random variables with uppercase and their realizations and constants with lowercase letters, and vectors with bold font.)

Our predictive model with linear genetic effects and no interactions is identical to the **additive genetic model** from quantitative genetics (Falconer and Mackay, 1996), except that the latter has a causal interpretation while our model needs not have one.[9] In addition to capturing causal genetic effects, our model may also capture factors that correlate with both the variants and Y, such as culture and socioeconomic determinants of health.

In practice, neither $\mathbf{b}$ nor G^* are observed. Instead, a **polygenic index** (**PGI**) can be constructed by using estimates $\hat{\mathbf{b}}$ from a previously published **genome-wide association study** (**GWAS**)[10] of Y:

$$G = \mathbf{X}\,\hat{\mathbf{b}}.$$

The observed PGI G is a noisy measure of the trait's true genetic predictor G^*. Let

[7] Each location has a pair of elements—one from the mother and one from the father—with each element associated with the letter "A", "C", "G", or "T".

[8] Genotyping is the process of measuring an informative subset of genetic variants; sequencing is the costlier process of identifying every single element of a person's DNA.

[9] Theory and evidence agree that an additive genetic model generally captures the bulk of the genetic variation of complex traits (Visscher et al., 2021).

[10] A GWAS is a large-scale genetic study in which a trait is regressed on millions of genetic variants, separately. The method we use to create PGIs, PRS-CS (Ge et al., 2019), adjusts the GWAS estimates of $\hat{\mathbf{b}}$ for the effects of nearby, correlated SNPs as if they were regressed jointly.

ϵ be the error: $G = G^* + \epsilon$. ϵ stems from measurement error in $\hat{\mathbf{b}}$, which is due to sampling variation in the GWAS. To avoid overfitting, the GWAS is conducted in a sample that is independent of the analysis dataset in which the PGI is constructed, so ϵ and G^* are independent.

A complex trait is associated with a large number of genetic variants, each with a tiny regression coefficient. This implies, by the central limit theorem, that $G^* = \mathbf{X}\,\mathbf{b}$ is approximately normally distributed. By the same reasoning, so is ϵ. As a result, so is G—i.e., PGIs are approximately normally distributed.

This model for a continuous trait can easily be adapted to a binary **disease** D. It is common practice to use a liability threshold model such as the probit, in which case the liability $\mathcal{L}$ replaces Y, ψ is normally distributed, and $D := \{\mathcal{L} \geq 0\}$.

Given each variant's tiny regression coefficient, very large samples are needed to estimate them precisely. As GWASs have rapidly grown in sample size over the past decade (Abdellaoui et al., 2023), estimates of $\mathbf{b}$ have become more precise; as a result, the predictive power of PGIs has also increased dramatically, albeit from a very low base and with much progress still to be made (Visscher et al., 2021). The GWASs from which we obtained the $\mathbf{b}$ estimates to construct PGIs all involved over 100,000 participants (see Online Appendix Table 1). Despite this, these $\mathbf{b}$ estimates, and thus also the current PGIs, remain noisy. Table 1 shows our estimates of the current PGI R^2's for the seven diseases. Since the diseases are binary, we report their $\boldsymbol{R^2}$ **on the liability scale**, also known as the McKelvey & Zavoina pseudo-R^2 (Lee, Goddard, et al., 2012; McKelvey and Zavoina, 1975).[11]

As GWAS sample sizes keep increasing and as the set of measured variants is augmented to include rarer variants, PGI predictive power should increase further (Visscher et al., 2021). For instance, for height—arguably the most studied trait in quantitative genetics—PGI R^2's have increased as subsequent GWASs increased in size, reaching 45% (among European ancestry populations) in the latest GWAS in 5.4M individuals (Yengo et al., 2022). In theory, the upper bound for a PGI's predictive power for a given disease is the disease's heritability (Dudbridge, 2013),

[11]The liability scale R^2 is defined as $\text{Var}[\mathbf{Z}\boldsymbol{\lambda}]/\text{Var}[\mathcal{L}_r]$, where $\boldsymbol{\lambda}$ is the coefficient from a probit regression of D on $\mathbf{Z}$ (with $\mathbf{Z} = G$, here) and $\text{Var}[\mathcal{L}_r] = \text{Var}[\mathbf{Z}\boldsymbol{\lambda}] + 1$ is the variance of the liability in that regression. This R^2 is intuitively analogous to the usual R^2, but indicates the share of the variance of the liability—instead of a continuous dependent variable—that is accounted for by the covariates.

which is the share of variation in disease liability attributable to genetic factors.[12]

The last two columns of Table 1 report previous studies' estimates of the diseases' **SNP heritabilities** and **twin heritabilities**. A disease's **twin heritability** is computed by comparing the disease resemblance of monozygotic twins—who share 100% of their genomes—to that of dizygotic twins—who share 50% on average. Under some assumptions, it is a consistent estimate of the true heritability.[13] A disease's **SNP heritability** is the share of variation in its liability that is attributable to a set of genetic variants. Because not all variants are used, it underestimates the true heritability.[14] For an in-depth discussion of heritability concepts, see Markel et al. (2025) and Benjamin et al. (2024).

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)	7.1%	33.1%	58%
Breast cancer	8.1%	12.3%	6.7%	14%	31%
Coronary artery disease	3.0%	51.7% (M), 39.2% (F)	2.5%	22%	40%
Colorectal cancer	2.9%	4.6%	2.2%	9%	40.2%
Prostate cancer	0.9%	12.7%	9.9%	18%	57%
Schizophrenia	0.46%	0.7%	4.9%	24%	79%
Type 2 diabetes	3.9%	27% (M), 18.6% (F)	7.4%	19.6%	72%

Notes: Current PGI R^2's are on the liability scale and were estimated in the UKB. All other statistics come from earlier studies and were not produced with a uniform methodology. Prevalence and lifetime risk include various age-corrected measures (e.g., the prevalence of Alzheimer's disease is reported for ages above 65). Further documentation is in the replication package.

While it is impossible to know exactly how predictive a disease's future PGI will be, it is likely that it will explain at least as much as today's estimates of the disease's SNP heritability. This is because future PGIs will benefit from more pre-

[12]Under our predictive model, a PGI could in fact predict more than the heritability if it also captures non-genetic correlates of the variants and Y. Treating heritability as the upper bound, as we do below, thus errs on the conservative side for the purpose of measuring selection.

[13]One important such assumption is that the greater resemblance of monozygotic vs. dizygotic twins is entirely due to the former's greater genetic resemblance.

[14]The expression *SNP* heritability comes from the fact that the variants used are typically single nucleotide polymorphisms (SNPs). The basic idea is to compute the genetic resemblance ("relatedness") between pairs of unrelated individuals based on their SNPs, and then to quantify how much of the pairs' disease resemblance is explained by their genetic resemblance (Yang et al., 2010). Because only a subset of common SNPs are used, the computed genetic resemblance is a noisy measure of the true genetic resemblance, which leads to an underestimate of the true heritability. The heritability estimates in Table 1 are from previous studies using similar methods and variants. They can be viewed as ballpark figures.

cise estimates $\hat{\mathbf{b}}$ and from including rarer variants. At the upper bound, future PGIs are unlikely to predict more than the twin heritability, because it is estimated from familial resemblance and thus already reflects the total genetic contribution of all genetic variants. Below, we report our results under two assumptions: that future PGIs' accuracies will equal their SNP heritabilities (as lower bounds) and that they will equal their twin heritabilities (as upper bounds).

We make no attempt to pinpoint the time it will take for PGI predictive power to reach either heritability. There is debate about the relationship between GWAS sample size and PGI predictive power. Even greater uncertainty surrounds future methodological advances and the growth trajectory of GWAS sample sizes.

2.2 Critical illness insurance

Contracts and Underwriting. CII pays a lump sum upon diagnosis of any of the medical conditions listed on the policy. Typical policies cover 20 to 50 conditions, including various cancers, coronary artery disease, multiple sclerosis, stroke, and other serious conditions (Gatzert and Maegebier, 2015). Enhanced plans exist that cover additional conditions (Gen Re, 2024), while some policies (called cancer insurance) only cover cancers. Cancers and cardiovascular diseases account for ∼80% of CII claims, with cancers accounting for the bulk of these.[15]

Most CII policies terminate upon payout; the payout, which is not earmarked for healthcare costs, is often used for non-health-related expenses (PartnerRe, 2015). Policies are often renewable term plans, with guaranteed premiums that are fixed or increasing with age and that are set upfront through underwriting (Brackenridge, Croxson, and Mackenzie, 2006); they are typically sold as stand-alone products, as add-ons (often to life insurance), or as workplace benefits.

The industry relies on sophisticated medical underwriting, with the world's largest reinsurance companies playing a central role in model development (PartnerRe, 2015; Gen Re, 2023b; Swiss Re Institute, 2024a). Insurers request detailed information on health, lifestyle, and family history, and may ask for medical records, blood tests, or a medical exam; they also use contract features such as waiting

[15] In Canada, between 2006 and 2011, 72% and 20% of claims were due to cancers and cardiovascular diseases, respectively (Canadian Institute of Actuaries, 2014). Cancers and cardiovascular diseases together accounted for between 70% and 87% of claims in the US in 2022 (Gen Re, 2023b) and account for over 80% of claims in the UK (Association of British Insurers, 2022).

periods to safeguard against selection. Current underwriting procedures appear effective in controlling claims in the US market, with all companies reporting that claims were at or below expected levels in a recent survey (Gen Re, 2023b). Yet industry reports identify improved genetic prediction technology as a key emerging challenge (Swiss Re Institute, 2024a), along with improved diagnostic techniques (RGA, 2016; Swiss Re Institute, 2016).

Market Overview. CII was introduced in South Africa in the 1980s and has matured into a core supplemental health insurance product in many countries, with as many as 100 million people covered worldwide (Gen Re, 2023b; Swiss Re Institute, 2024b; Gen Re, 2023a). Market research valued the global market at over $100 billion in 2021 and projected it to grow to over $350 billion by 2031 (Allied Market Research, 2022). Market disruptions could thus have far-reaching consequences for millions of consumers and for insurers alike.

CII is widespread even in countries with near-universal healthcare systems, such as the UK (Gatzert and Maegebier, 2015), Australia (Financial Services Council, 2020), Canada (Dorse, 2024), and Hong Kong, Malaysia, and Singapore (Gen Re, 2023a). In the US, there were at least 6.3 million in-force CII policies in 2022 (Gen Re, 2023b)—now outnumbering long-term care policies (Gallagher Re, 2025)—making CII one of the country's fastest growing supplemental health products. The US market is served by numerous firms, including the market leaders in health and life insurance. In Asia, CII has been a top-selling insurance product (Gen Re, 2023a). In China, the CII market has been described as "a core pillar of the health insurance market," with 5% to 18% of working-age adults covered in 2018 (Swiss Re Institute, 2022).

3 Theory

We now introduce our economic and econometric models. We focus on single-disease contracts to facilitate exposition. Section 5.3 and Online Appendix 3 generalize these models to multiple-disease contracts.

3.1 Economic model notation

The model. Consider a population of consumers who may incur a binary loss L and can purchase an insurance contract for it. In a single-disease CII contract, the loss corresponds to contracting the disease: $L = D$.

Each consumer is characterized by four random variables $(L, G_c, G_f, \mathbf{W})$. The **loss** (or **disease**) L is a binary variable. G_c is the current polygenic index (PGI), which is a sufficient statistic of the consumer's DNA for the purpose of estimating disease risk probabilities, conditional on the current technology. G_f is the future PGI, which is a sufficient statistic of the consumer's DNA conditional on the future genetic prediction technology. G_c and G_f take real values. $\mathbf{W}$ is a row vector of non-genetic covariates that takes values in Euclidean space. The vector $(L, G_c, G_f, \mathbf{W})$ is i.i.d. across consumers with joint distribution $\mathbb{P}$.

Throughout this section, we consider groups treated identically by insurers, such as consumers in the same risk class. We assume that consumers possess additional private information, so that there may be selection.

Private information about risk. Each consumer has private information about her non-genetic covariates $\mathbf{W}$ and her PGI G, and rationally combines this information to form an integrated prediction of her disease risk.[16]

In our implementation, the population of consumers treated equally by insurers is a risk class. The insurers observe the non-genetic covariates $\mathbf{W}$ but use them only to define risk classes. Thus, any remaining variation in $\mathbf{W}$ within a risk class is ignored by insurers and is effectively private information.

Define consumers' **private risk function** as the probability $\rho(g, \mathbf{w})$ of the loss conditional on the PGI $G = g$ and the non-genetic factors $\mathbf{W} = \mathbf{w}$:

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

Define the random variable for **private risk** as $R := \rho(G, \mathbf{W})$. The realizations of losses are independent across consumers, and the same holds for R.

Information scenarios. For each contract, we consider four alternative scenar-

[16] Consumers need not perform the integration themselves. As mentioned, integrated risk predictions are already offered by genetic-testing companies and online tools such as CanRisk, and may become widely available in clinical care.

ios that differ only in consumers' private information. In all scenarios, insurers observe $\mathbf{W}$ perfectly but only use this information to classify consumers into discrete risk classes. Here, the standard risk class is defined as consumers whose risk is between 0.75–1.25 times the average population risk.

In Scenario 1, consumers observe only their non-genetic covariates (or, equivalently, they also observe a degenerate, uninformative PGI). In Scenario 2, they also observe the current PGI G_c. In Scenarios 3L and 3U, they observe a future PGI G_f, with predictive power equal to the disease's SNP heritability ($R^2_f = h^2_{\text{SNP}}$, Scenario 3L) or twin heritability ($R^2_f = h^2_{\text{twin}}$, Scenario 3U), respectively.

Measures of adverse selection. We now measure selection in market segment $\mathcal{M}$ where we assume insurers do not price discriminate among consumers with $\mathbf{W} \in \mathcal{M}$. The distribution of private risk R is informative about the degree of adverse selection when consumers have genetic information. In the prostate cancer example, Figure 1 shows the distribution of R under the different information scenarios. Scenarios with a wider distribution of R are more likely to be characterized by adverse selection.

Adverse selection depends on the entire distribution of private risk R within the market segment. To facilitate comparison, we report an economic measure of selection, the **implicit tax** (Hendren, 2013). The implicit tax $t(r)$ is defined as the extra price a consumer with private risk r would have to pay if she had to buy insurance at the actuarially fair price for all consumers with private risk r or greater, relative to her own actuarially fair price:[17]

$$t(r) := \frac{\mathbb{E}[R|R \geq r, \mathbf{W} \in \mathcal{M}]}{1 - \mathbb{E}[R|R \geq r, \mathbf{W} \in \mathcal{M}]} \Big/ \frac{r}{1-r} - 1.$$

Among selection measures that compare marginal and average types, the implicit tax has two advantages. First, Hendren (2013) proposed a no-trade theorem that shows that sufficiently high values cause insurance market disruption under

[17] Hendren (2013) gives the following motivation. Consider an insurance contract that pays \$1 if a loss happens, and otherwise charges a premium. For the contract to break even with with private risk r, the premium has to be $r/1-r$. Consider the case where this policy is purchased by all consumers with private risk r or greater. For the policy to break even the premium would have to be $\frac{\mathbb{E}[R|R\geq r, \mathbf{W}\in\mathcal{M}]}{1-\mathbb{E}[R|R\geq r, \mathbf{W}\in\mathcal{M}]}$. The consumer with private risk r would have to pay $T(r)$ times her actuarially fair price, where $T(r)$ is the second premium divided by the first. Hendren (2013) calls $T(r)$ the "pooled price ratio". We define the implicit tax $t(r)$ as $1 + t(r) := T(r)$.

certain assumptions. Second, it simplifies comparison with levels of selection reported previously. Hendren (2013) estimates the implicit tax in market segments that had unraveled (in the sense of not being served by insurers) and in functioning market segments. He reports the minimum implicit tax up to the 80th percentile of risk, which we denote t_{80}. The interpretation is that this is the tax a marginal consumer would have to be willing to pay to get a market to function with the top 20% riskiest consumers participating. Hendren estimated t_{80} to be 7–35% for functioning markets and 43–83% for markets that had unraveled.

3.2 Econometric model and identification

Our model is fully specified by $\mathbb{P}$, the joint distribution of $(L, G_c, G_f, \mathbf{W})$.[18] While G_c can be observed in the UKB data, G_f cannot. Define the distribution of the observed data as a joint distribution $\mathbb{P}_{\text{data}}$ over $(L, G_c, \mathbf{W})$. We say that the model is identified if there is a unique value of $\mathbb{P}$ consistent with $\mathbb{P}_{\text{data}}$.

The model is identified under reasonable assumptions. Our assumptions are based on two empirical regularities from quantitative genetics. First, PGIs are approximately normally distributed. And second, existing heritability estimates are informative about the likely predictive power of future PGIs.

We now state the assumptions. The first assumption is a regularity condition.

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 G_c and G_f. They are motivated by the first regularity mentioned above and by the fact that what distinguishes G_c from G_f is the size of the independent and approximately normally distributed error. Further, we take G_c to be standardized.[19]

Assumption 2. *[Gaussian future PGI] Conditional on observables, the future PGI has a Gaussian distribution*

$$G_f := \mathbf{W}\boldsymbol{\theta} + V, \tag{1}$$

[18] Our model is similar in spirit to that developed by Becker et al. (2021) for continuous outcomes.

[19] Standardizing G_c is without loss of generality because the model's parameters can be rescaled to achieve this without changing the risk distributions. We standardize G_c to save on notation and to be consistent with standard practice in the quantitative genetics literature.

with $V \sim N(0, \sigma_V^2)$ *and* $V \perp \mathbf{W}$.[20]

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

$$G_c := G_f + \epsilon, \tag{2}$$

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

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

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

$$\Pr[L = 1 | G_f = g_f, G_c = g_c, \mathbf{W} = \mathbf{w}] = \Phi(g_f \beta_g + \mathbf{w}\boldsymbol{\beta}_w). \tag{3}$$

Finally, we assume that we know the predictive power of the disease's future PGI G_f, which we denote R_f^2, from heritability studies.

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

Under these assumptions, the model is identified:

Theorem 1. *Under Assumptions 1–5, the model is identified.*

Appendix A provides the proof. The intuition is that the heritability estimates tell us how predictive the future PGI is on its own. The correlations in the current data tell us how much added value the future PGI brings over the covariates, and the normality assumptions take us from correlations to the full distribution.[21]

The proof is based on two observations. The first observation is that the posterior distribution of G_f given the data is normal and given by a standard formula

[20] Due to the large effects of variants in the *APOE* gene on the risk of Alzheimer's disease, Assumption 2 may be violated for that disease. We verified that our results for the multiple-disease contracts (discussed below) are robust to excluding Alzheimer's disease.

[21] While the non-genetic covariates do not impact the heritability bounds, they play a central role in our estimation. As discussed below, what drives adverse selection within the standard risk class is not the future PGI's total predictive power, but its predictive power incremental to $\mathbf{W}$. Heritability bounds the former; the joint distribution of G_c, $\mathbf{W}$, and disease status in the data identifies the latter.

from Bayesian statistics. The second observation is that this implies that the disease model is also probit in G_c, and that the model parameters are functions of feasible regressions that use G_c instead of G_f. This observation also yields a statistically and computationally efficient estimator.

4 Data and empirical specification

4.1 Data

We analyze data from the UK Biobank (UKB), which contains genotypic and rich health-relevant data on about 500,000 people (Sudlow et al., 2015). Participants were between 40 and 69 years old at recruitment (2006–2010). Their data is linked to the UK's National Health Service (NHS), which maintains electronic health records (EHR) on almost the entire UK population (Sudlow et al., 2015).

UKB participants provided extensive survey and interview data on clinical and environmental disease risk factors, sociodemographics, as well as biomedical assays from blood, saliva, and urine. Table 2 reports sample descriptive statistics for the seven diseases and the multiple-disease contracts (Panel A) and for a set of disease-general risk factors (Panel B).

Our sample was restricted to 446,570 participants of European ancestry, whose genotypic data passed a number of quality control filters. Well-powered PGIs are not yet readily available for non-European ancestries. To compute the current PGI G_c for each disease, we combine the participants' genotypic data with GWAS estimates sourced from the largest available GWAS for the disease that does not contain UKB data.[22] Online Appendix 1 provides additional details.

To code the unstructured EHR-data for statistical analysis, we relied on the "Phecode" mapping system (Denny et al., 2013). This system was created to define meaningful case-control variables from complex EHR sources, e.g., by condensing some 90,000 detailed codes from the International Classification of Diseases (e.g., C50.211: "malignant neoplasm of upper-inner quadrant [...]") into about 1,900 major case-control status codes (174.1: "Breast cancer [female]").

We reviewed clinical guidelines and the epidemiological literature to identify the main disease-specific non-genetic risk factors that are used either by medical

[22] To ensure that $G^* \perp \epsilon$, the GWAS sample must not overlap with the UKB.

Table 2: Sample descriptive statistics

Panel A. Descriptive statistics for the diseases

			Age of onset			
Disease	**N**	**% Cases**	**Mean**	**SD**	**5th pct**	**95th pct**
Alzheimer's disease	405,573	0.9%	75.3	5.3	65.4	82.2
Breast cancer	211,575	9.1%	57.8	10.1	41.0	74.0
Coronary artery disease	410,686	13.4%	62.7	10.5	44.5	78.0
Colorectal cancer	410,562	2.6%	65.1	9.7	47.7	78.7
Prostate cancer	181,902	8.8%	67.3	6.9	55.3	77.9
Schizophrenia	410,275	0.3%	46.2	17.6	20.2	75.2
Type 2 diabetes	410,686	8.6%	64.1	9.3	47.8	77.8
Any disease (Males only)	175,466	36.6%	62.6	10.1	45.0	77.2
Any disease (Females only)	204,672	24.7%	61.3	10.7	43.1	77.5

Panel B. Descriptive statistics for the general risk factors (N = 415,847)

Variable	**Mean (s.e.)**	**SD**	**Min**	**Max**
Age (at last observation)	66.0 (0.015)	9.89	39.0	88.0
Sex (Male = 1)	0.458 (<0.001)	0.498	0	1
Deprivation index (Townsend)	-1.52 (0.005)	2.94	-6.3	10.9
Education (years of schooling)	13.9 (0.008)	4.94	7	20
BMI	27.4 (0.007)	4.73	12.1	74.7
Drinks per week	8.34 (0.014)	9.24	0.0	52.0
Current smoker	0.097 (<0.001)	0.295	0	1
Previous smoker	0.359 (<0.001)	0.48	0	1
Never smoker	0.545 (<0.001)	0.498	0	1
Physical inactivity	0.016 (<0.001)	0.125	0	1
Systolic blood pressure (SBP)	138.5 (0.028)	18.4	72.0	253.5

Notes: For each disease, the sample sizes in Panel A reflect the number of participants in UKB with no missing values for the disease nor for the disease-general or disease-specific risk factors (listed in Table 3). "Any disease" is a dummy for having contracted any of the diseases in the multiple-disease contract.

practitioners or by insurers to assess a person's disease risk. It is important that our model captures the information typically observed by insurers. The findings of the review are listed in Table 3. We were able to code most of the risk factors identified in the search in the UKB data. All risk factors that could be coded are included in $\mathbf{W}$, together with a genotyping array dummy and the top ten genetic principal components (PCs). Adjusting for genetic PCs is standard practice (Price et al., 2006). Further documentation is provided in Online Appendix 2.

4.2 Contracts and specification

We consider standard contracts that pay a lump sum upon a loss, defined as the occurrence of a covered disease. Real-world contracts vary across many dimensions

Table 3: Non-genetic risk factors for the studied diseases

Disease	Key risk factors
General risk factors	Age, sex, Townsend Deprivation index, education (years of schooling), body mass index (BMI), alcohol consumption, smoking (current, ex, or never), physical inactivity, systolic blood pressure, family history* (did each of father, mother, or siblings ever suffer from the disease?)
Alzheimer's disease	Air pollution, [*APOE* risk type], [brain injury], [coronary artery disease], depression, diabetes type 1, [diabetes type 2], hearing loss, hypertension, physical inactivity, social isolation
Breast cancer	Diabetes type 1, [diabetes type 2], ever had breast cancer screening / mammogram, ever had hormonal replacement therapy, hormonal/oral contraceptives, age at menarche, age at menopause, age at first birth, number of live births
Coronary artery disease	Crohn's disease, diabetes type 1, [diabetes type 2], [diet], ever had bowel cancer screening, irritable bowel syndrome (IBS), polyp of colon, ulcerative colitis (and other non-infective gastro-enteritis and colitis)
Colorectal cancer	Depression, diabetes type 1, [diabetes type 2], [diet], hypercholesterolemia, hyperglycemia, hypertension, social isolation
Prostate cancer	Ever had prostate specific antigen (PSA) test
Schizophrenia	Anxiety, bipolar disorder, [cannabis use], depression, neuroticism, psychopathology**, [substance use]
Type 2 diabetes	[Diet], hypercholesterolemia, hyperglycemia, hypertension, social isolation

Notes: Risk factors were identified in a review of clinical guidelines and of the epidemiological literature, and were verified by two medical doctors (for references, see the replication package). All risk factors were coded as variables in the UKB, except those (i) that were not well-defined (e.g., diet), (ii) for which data were lacking (e.g., brain injury, cannabis use), (iii) that were a disease in the multiple-disease contract (e.g., type 2 diabetes), or (iv) that were captured by PGIs (i.e., *APOE*). These exceptions are shown in square brackets.
* For schizophrenia, family history of "severe depression" was used as a proxy. ** Psychopathology was proxied by neuroticism score, bipolar disorder, and depression.

(see Section 2.2), so we limit our analysis to particular types of contracts, especially in two key contract dimensions: the set of covered diseases and the coverage period. We also specify a population of interest (i.e., a risk class). We discuss the current PGI G_c and the covariates $\mathbf{W}$ for each disease in Section 4.1, and the four information scenarios for which we produce results in Section 3.1. With this, we have a full specification to take the model (the loss L, the current PGI G_c, and the covariates $\mathbf{W}$) to the data.

Covered diseases. We consider the seven diseases listed in Table 1. These include three of the four most common cancers in Western countries—prostate, breast, and colorectal cancer—which by themselves account for a substantial share of CII claims.[23] As a proxy for cardiovascular disease—the next most common source of claims—we include coronary artery disease.[24] We include Alzheimer's

[23]These three cancers account for a little less than half of all cancers in Western countries, and they accounted for 40% of all cancers in the UK between 2016 and 2018 (Cancer Research UK, 2020). As mentioned, cancers account for a large share of CII claims.

[24]Most vasculatory-disease-related CII claims are due to heart attacks, followed by strokes (cerebrovascular). Heart attacks are fully captured by our coronary artery disease variable.

disease because it is the most common cause of early-onset dementia, which some insurers cover. Schizophrenia is included as an example of a highly heritable but rare condition. We include type 2 diabetes because it is of public health concern and is sometimes covered by enhanced CII plans. Most other conditions covered by real CII contracts, but not by our analysis, are rare and account for a small share of claims; our multiple-disease contracts thus approximate real CII contracts.

Coverage period. As mentioned, typical CII contracts are guaranteed renewable at a predetermined premium 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 all consumers buy contracts at the same age and renew until age 65. Here, we will assume that all consumers buy contracts at age 35, a typical age of first CII purchase that is lower than the usual age of disease onset. Therefore, we will effectively consider one-time 30-year contracts. Ignoring interest rates, this means that insurers' costs are proportional to the probability that the consumer will incur a covered disease by age 65. In the terminology of our model, we set the loss $L = 1$ if the consumer ever incurs a covered disease by age 65.

We stress that this is an approximation. First, in practice, consumers do let policies lapse. Lapsation is a secondary but important driver of insurance premiums (Gottlieb and Smetters, 2021). Second, the timing of losses also influences insurers' costs, due to time discounting.

Population of interest. We focus on the population of 35 year old consumers who may purchase CII contracts and fall in the standard risk class, as defined above. Insurers typically charge the same rate to these consumers, who constitute the market segment $\mathcal{M}$ in the calculation of the implicit tax.

4.3 Estimation for single-disease contracts

For each single-disease contract, estimation is based on the identification Theorem 1. We proceed as follows.

Step 1. Estimating the econometric model. We begin by estimating our econometric model for the contract, using the observed $\mathbb{P}_{\text{data}}$ for our study sample. We pool females and males, except for breast cancer and prostate cancer. Estimation

follows the two-step estimator in Appendix B, which is based on Theorem 1.[25]

Step 2. Generating the private risk distributions. For all four scenarios, we use Equation 3 to generate the distribution of risk for the disease for all the individuals in our sample. For Scenarios 3L and 3U, we do not observe G_f, so for each individual we draw 10 values from its probability distribution conditional on G_c and $\mathbf{W}$ (given by Lemma A.1 in the Appendix), and use these to generate the risk distribution.

Individuals in the data range in age from 39 to 86. Our set of covariates $\mathbf{W}$ (see Table 3) includes age as well as a number of age-dependent covariates, such as BMI, blood pressure, and hypertension. Because age and a number of these covariates are important predictors of disease risk, we set age to 65 and adjust the most age-dependent covariates to their expected values at age 65 when using Equation 3 to estimate the risk of the disease by age 65. See Online Appendix 2.

Step 3. Identifying the risk classes. To assign the individuals to different risk classes, we use the risk distribution we generated for Scenario 1 using only the non-genetic covariates (which is what insurers use). Individuals whose risk is between 0.75–1.25 times the average population-wide risk are assigned to the standard risk class.

Step 4. Estimating the implicit tax. For each scenario, we use the scenario's risk distribution from Step 2, limited to those in the standard risk class. We compute Hendren's implicit tax measure for all percentiles of the distribution. Our main summary statistic of the amount of selection is the minimum implicit tax up to the 80th percentile, t_{80}. To obtain 95% confidence intervals, we use the bootstrap method, with 200 draws.

[25] Ideally, we would estimate the model off a population that we would observe both at age 35 and then again at age 65. The non-genetic covariates would be observed at the former and disease diagnoses at the latter age. In practice, we do not have such data; instead, we use data from a population of different ages to estimate the risk of developing each disease by age 65.

5 Results

5.1 Single-disease contracts: case study for prostate cancer

We first return to the case study of the single-disease CII contract for prostate cancer from the introduction. The left panel of Figure 1 shows the risk distributions for prostate cancer under Scenarios 1–3U.

Scenario 1 predicts risk using only the non-genetic covariates $\mathbf{W}$. The distribution is bimodal because family history variables are strong risk factors. There is wide dispersion in non-genetic risk. Without underwriting, this would result in a noticeable amount of selection; for that reason, insurers do underwrite and classify consumers into the aforementioned risk classes based on the non-genetic risk factors. The consumers in the standard risk class are shown in blue in Figure 1. Since insurers treat these equally, any variation in risk within the risk class effectively becomes private information. Thus, the blue area in Figure 1, Scenario 1, corresponds to the distribution of private risk R for the standard risk class. Scenarios 2–3U show how the distribution of private risk R changes when consumers observe their genetic predictions. There is noticeable private information with the current PGI, and a considerable amount with future PGIs.

We can use the implicit tax to quantify the amount of selection. The right panel plots the implicit taxes in the four scenarios for consumers in the standard risk class. In Scenario 1, with the non-genetic covariates only, t_{80} is equal to 6.3%. Thus, after underwriting, there is a negligible amount of selection within the standard risk class, as expected. The implicit tax t_{80} increases to 59.8% with the current PGI (Scenario 2), and to between 86.8% and 426.9% with the future PGI (Scenarios 3L and 3U). Thus, there would already be noticeable selection with today's genetic prediction technology, if all consumers had adopted this technology. And if genetic predictions approach the heritability estimates and are widely available, a single-disease CII contract for prostate cancer may not be viable.

5.2 Single-disease contracts: main results

We now extend these results to all seven single-disease CII contracts. Though there are nuances across the diseases, the big picture is consistent with the prostate cancer case study. Figure 2 reports the minimum implicit tax up to the 80th percentile

of risk (t_{80}) for each disease and under each scenario for the standard risk class. Table 4 reports detailed results.[26] There are three main findings.

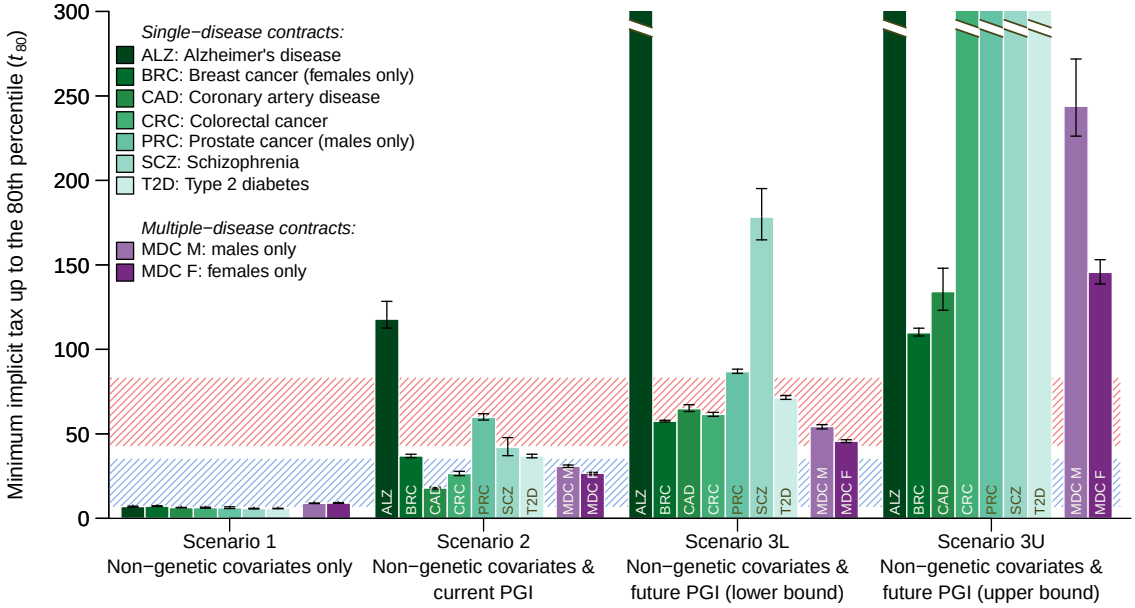

Figure 2: Minimum implicit taxes up to the 80th percentile (t_{80})

Notes: The figure displays t_{80} within the standard risk class for the single-disease and multiple-disease CII contracts, in the four scenarios. The striped blue area corresponds to the range of t_{80} observed by Hendren (2013) in market segments that had not unraveled, and the striped red area corresponds to the range for market segments that had unraveled. The breaks at the top of the tallest bars indicate that these bars' heights exceed the figure's.

First, there would be noticeable selection if the current genetic prediction technology were widely adopted: in Scenario 2, t_{80} ranges from 17.9% for coronary artery disease to 117.9% for Alzheimer's disease. In comparison with Hendren (2013), t_{80} for coronary artery disease (17.9%) and colorectal cancer (26.5%) falls in the middle of the range found for market segments that had not unraveled (shown in blue stripes in Figure 2). t_{80} for breast cancer (36.9%), schizophrenia (42.1%), and type 2 diabetes (37.0%) falls in between the ranges of no unraveling and unraveling. t_{80} for Alzheimer's disease and prostate cancer is within or above the range observed for segments that had unraveled (shown in red stripes). The Scenario 2 results suggest that current prediction technology is already sufficiently powerful to cause noticeable selection, and that the main question is when the technology will be adopted widely.

[26] The low t_{80}'s in Scenario 1 are expected, since by definition the standard risk class leaves little private risk in that scenario. These estimates suggest that our risk classification was successful.

The second main finding is that, with the expected predictive power of future PGIs, the amount of selection could become crippling. In Scenario 3L, where the future R^2's are assumed to be equal to the diseases' SNP heritabilities, t_{80} ranges from 57.5% for breast to >1,000% for Alzheimer's. Some contracts have t_{80}'s that are considerably higher than what Hendren (2013) found in market segments that had unraveled. In Scenario 3U—in which future PGIs explain as much as the diseases' twin heritabilities—t_{80} is even higher, exceeding 100% for all seven single-disease contracts, and exceeding 1,000% for three. These findings suggest that, with widespread adoption of genetic prediction, single-disease CII contracts would not be a viable product.

The third main finding is that there is variation in t_{80} across the diseases. A key driver of this variation is the incremental predictive power of the PGI over and above the non-genetic covariates (reported in Panel B of Table 4)[27]. This makes sense, since the incremental predictive power of a PGI captures the extent to which it adds private information beyond the non-genetic covariates. To see this, consider the following examples. In Scenario 3L, Alzheimer's and schizophrenia's PGIs have the highest incremental predictive power and, accordingly, the highest t_{80}'s.[28]

5.3 Multiple-disease contracts

We now consider multiple-disease contracts. We consider separate contracts for males and females that each bundle all the diseases, except breast cancer for males and prostate cancer for females. The contracts pay a lump sum upon contraction of any of the covered diseases.

Model and empirical implementation. The economic model generalizes to multiple diseases if we define the **loss** L as a random variable equal to zero or one depending on whether a consumer contracts any of the diseases listed in the contract. The econometric model also extends by specifying and estimating the

[27] Panel B reports this quantity as the PGI's ΔR^2, defined as the difference between the R^2 of a regression of the disease on the PGI and covariates and that of the same regression but on the covariates only (in the full sample). Panel B also reports the R^2's of the covariates, the PGI, and the covariates and PGI together. The R^2 of the covariates is high for coronary artery disease and type 2 diabetes, consistent with the fact that lifestyle factors play an important role for these diseases.

[28] t_{80} also varies with some features of the disease risk distribution. For instance, Alzheimer's and schizophrenia are rare within the contracts' age span, such that their 80th-percentile risk remains very low. This makes the denominator in the implicit-tax formula small, causing t_{80} to explode.

Table 4: Summary of main results (standard risk class)

	Single-disease contracts							Multiple-disease contracts	
	ALZ	BRC	CAD	CRC	PRC	SCZ	T2D		
Sex	M & F	F only	M & F	M & F	M only	M & F	M & F	M only	F only
Sample size	405,573	211,575	410,686	410,562	181,902	410,275	410,686	175,466	204,672
Cases	3,529	18,133	54,395	10,453	15,470	1,085	34,667	58,441	45,427
Panel A. Risk of contracting disease (r) by age 65, conditional on non-genetic covariates (**W**) only									
					Full analysis sample				
Mean	0.2%	7.9%	8.9%	1.9%	5.5%	0.3%	6.1%	25.6%	17.5%
SD	0.1 pp	2.1 pp	8.9 pp	1.2 pp	2.0 pp	1.2 pp	8.3 pp	13.1 pp	8.9 pp
					Standard risk class (r = 0.75–1.25 of sample mean)				
Mean	0.1%	7.6%	8.6%	1.8%	5.2%	0.2%	6.0%	24.7%	16.5%
SD	0.02 pp	1.0 pp	1.3 pp	0.3 pp	0.8 pp	0.04 pp	0.9 pp	3.6 pp	2.4 pp
Panel B. Explanatory power of non-genetic covariates (**W**) and PGIs (*full analysis sample*)									
					Scenarios 1-3				
R^2 of **W**	38.5%	6.0%	36.8%	11.5%	22.9%	20.1%	39.7%	–	–
					Scenario 2				
R^2 of G_c	7.1%	6.7%	2.5%	2.2%	9.9%	4.9%	7.4%	–	–
R^2 of G_c & **W**	43.7%	12.1%	37.8%	13.5%	31.2%	22.9%	43.4%	–	–
ΔR^2 of G_c	5.3 pp	6.1 pp	1.0 pp	2.0 pp	8.3 pp	2.8 pp	3.6 pp	–	–
					Scenario 3L				
R^2 of G_f (SNP h^2)	33.1%	14.0%	22.0%	9.0%	18.0%	24.0%	19.6%	–	–
R^2 of G_f & **W**	64.6%	18.9%	47.6%	20.1%	38.2%	35.9%	50.1%	–	–
ΔR^2 of G_f	26.2 pp	12.9 pp	10.8 pp	8.6 pp	15.3 pp	15.8 pp	10.4 pp	–	–
					Scenario 3U				
R^2 of G_f (twin h^2)	58.0%	31.0%	40.0%	40.2%	57.0%	79.0%	72.0%	–	–
R^2 of G_f & **W**	86.8%	35.3%	61.5%	55.7%	72.9%	85.4%	90.7%	–	–
ΔR^2 of G_f	48.4 pp	29.3 pp	24.7 pp	44.2 pp	50.0 pp	65.3 pp	51.0 pp	–	–
Panel C. Minimum implicit tax up to the 80th percentile of the disease risk distribution (t_{80})									
					No individual underwriting by insurers (same premium for all)				
Scenario 1	65.8%	25.4%	98.3%	41.9%	27.6%	379.4%	120.5%	69.1%	65.4%
					Standard risk class (r = 0.75–1.25 of full sample mean)				
Scenario 1	7.0%	7.2%	6.3%	6.3%	6.3%	5.8%	5.8%	8.9%	9.0%
Scenario 2	117.9%	36.9%	17.9%	26.5%	59.8%	42.1%	37.0%	30.8%	26.7%
Scenario 3L	>1,000%	57.5%	64.9%	61.5%	86.8%	178.3%	71.4%	54.4%	45.6%
Scenario 3U	>1,000%	109.8%	134.2%	410.1%	426.9%	>1,000%	>1,000%	243.9%	145.6%

Notes: In Panel B, the reported R^2's are pseudo-R^2's on the liability scale (defined in the text). The R^2 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 over the covariates for the disease, defined as the difference between the R^2 of the PGI and the covariates jointly and the R^2 of the covariates only. The four scenarios are defined in the text.

multivariate distributions across the diseases of the genetic disturbance terms V, of the current PGI error terms ϵ, and of the probit error. Online Appendix 3 generalizes the model, identification theorem, and estimation procedure.

Results. Figure 3 displays the risk distributions and implicit taxes for the multiple-disease contract for males in the standard risk class. The average population risk is 25.6%, but there is a lot of variation in individual risks, even when conditioning only on the non-genetic covariates (Scenario 1). In Scenario 2, when individuals observe their current PGIs, the spread of the distribution of private risk R in the standard risk class increases, with t_{80} = 30.8%. This implies moderate levels of selection similar to those found by Hendren (2013) in markets that had not unraveled.

In Scenarios 3L and 3U, when individuals observe their future PGIs, the distribution of private risk R becomes very spread out. In Scenario 3L, with future PGIs that explain as much as the diseases' SNP heritabilities, consumers at the 5th and 95th percentiles of R face risks of 7.8% and 50.4% of contracting a covered disease; in Scenario 3U, with future PGIs explaining as much as the twin heritabilities, the corresponding figures are 0.7% and 91.3%. Thus, with future PGIs, the standard risk class contains a range of individuals—all of whom face the same price of insurance—some with very low private risk and others with very high private risk, suggesting ample opportunity for selection. Consistent with this, $t_{80} = 54.4\%$ in Scenario 3L and 243.9% in Scenario 3U, implying elevated levels of selection that have not been observed in well-functioning market segments.

The results for the multiple-disease contract for females in the standard risk class are qualitatively similar (Table 4). Overall, the broad patterns observed for the single-disease CII contracts hold for the multiple-disease contracts. However, the implicit taxes are often lower in the bundled contracts.

6 Robustness and extensions

6.1 Robustness analysis in the Health and Retirement Study

The Health and Retirement Study (HRS) is a longitudinal dataset of ∼20,000 individuals with genotypic, health, lifestyle, and socioeconomic data. To examine the robustness of our results and to facilitate replication, Online Appendix 4 imple-

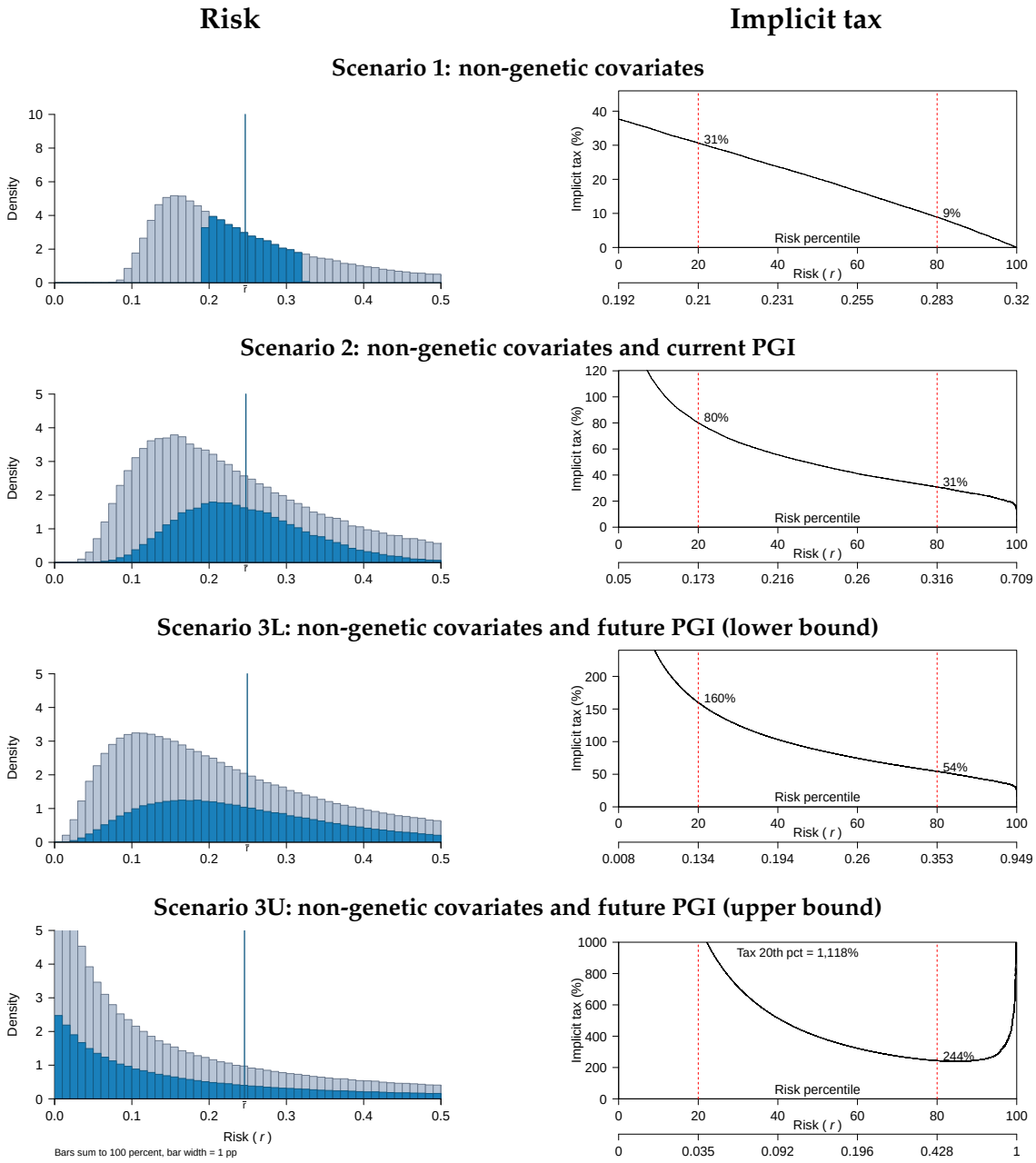

Figure 3: Multiple-disease contract for males: risk and implicit tax

Notes: Left panel: distribution of the risk of contracting any of the six diseases in the male multiple-disease contract by age 65, conditional on various information scenarios and computed in the UKB data ($N = 175,466$). The vertical blue line marks the average risk in the standard risk class ($\bar{r} = 0.2467$). The standard risk class individuals are shown in blue; insurers treat these individuals identically, so the blue distribution corresponds to the distribution of private risk for these individuals. Right panel: implicit tax for consumers in the standard risk class as a function of their percentile private risk of any of the six diseases in the male multiple-disease contract for each scenario.

ments our model in the HRS. The main limitation of the HRS is that its health and medical data are less detailed. The HRS relies largely on self-reports, whereas the UKB incorporates EHR data and a comprehensive nurse-administered assessment.

We searched the HRS for high-quality data on critical illnesses. The highest quality variable is a question about diagnoses for heart conditions. The conditions include coronary artery disease (CAD), but also some less severe conditions. For the purposes of our analysis, we define the "HRS CAD" contract as a CII contract for the HRS heart conditions. We repeated our estimation procedure in this sample.

The HRS CAD results are qualitatively similar to the UKB results. We find noticeable selection with the current PGI (t_{80} = 18.9%), and large amounts of selection with the future PGI (t_{80} = 51.8% in Scenario 3L and 96.0% in Scenario 3U). See Online Appendix 4.

6.2 Robustness analysis in a calibrated equilibrium model

Our analysis ignores risk preferences and market equilibrium. To check the robustness of our results, we consider a parsimonious calibrated equilibrium model of adverse selection in the CII market. We use the standard supply and demand model of Akerlof (1970) and Einav, Finkelstein, and Cullen (2010).

We model insurance demand using a binary loss framework. A loss reduces consumption by a fraction Δc. Consumers vary along two dimensions of heterogeneity: their probability of loss (disease risk) and their relative risk aversion. The cost for firms to provide coverage equals the expected payout plus a fixed cost F. We apply the model to the HRS CAD contract. A key advantage of using the HRS is that Kimball, Sahm, and Shapiro (2008) estimated individual-level relative risk aversion based on HRS survey questions. We use their estimate of each consumer's relative risk aversion. For disease risk, we use our econometric model.

This leaves two parameters to calibrate: Δc and F. We calibrate these to match basic CII market statistics from the UK. Based on industry estimates, we set the equilibrium quantity (i.e., the fraction of individual who buy coverage) to 30% and choose fixed costs F that imply a loss ratio of 50%. We always restrict attention to consumers in the standard risk class, and perform the calibration in Scenario 1 where the only information is from non-genetic covariates.

Results. Figure 4 displays the supply and demand graphs for each scenario.

We refer readers to Einav, Finkelstein, and Cullen (2010) for a detailed discussion of these curves. In summary, the demand curve represents the percentage of consumers who would like to purchase insurance as a function of the price. The average cost curve AC represents the cost of selling coverage to an average buyer, as a function of the fraction of consumers buying insurance. If there is no selection, the cost of selling insurance is independent of the set of consumers buying insurance, and the AC curve is flat. With adverse selection, the average cost is downward sloping because it is cheaper to sell to the average consumer in the entire market than to sell to the average consumer in a segment with higher willingness to pay (which correlates with disease risk). Einav, Finkelstein, and Cullen (2010) define the marginal cost curve MC as the cost of selling to a marginal consumer.

In Scenario 1, where consumers only observe non-genetic covariates, the AC curve is nearly flat because there is almost no selection. This is consistent with the definition of the standard risk class. The demand curve is also relatively flat because variation in disease risk is limited while variation in risk aversion is not sufficient to create large differences in willingness to pay. The equilibrium quantity is nearly identical to the calibration target of 30% of consumers buying insurance.

In Scenario 2, consumers have access to the current PGI. The AC and demand curves rotate and become steeper. The consumers who are most willing to buy insurance are on average more likely to incur a loss. These changes affect the market equilibrium. The equilibrium quantity goes down to 21.4%. Thus, giving consumers access to their current PGI creates a noticeable amount of selection.

In Scenarios 3L and 3U, consumers have access to future PGIs with predictive power given by the heritability estimates. In both cases, we see a dramatic steepening of the demand curve. This creates a large amount of adverse selection, consistent with our measures purely based on disease risk, such as the implicit tax. Accordingly, in both cases the equilibrium quantity goes to zero. The market suffers from a complete Akerlof (1970) death spiral, and there is no trade. So consumers accessing their future PGIs would result in a large, potentially crippling, amount of adverse selection.

We emphasize that these results are based on a parsimonious model calibrated from the data, but different equilibrium models may yield different results. Online Appendix 5 gives additional details.

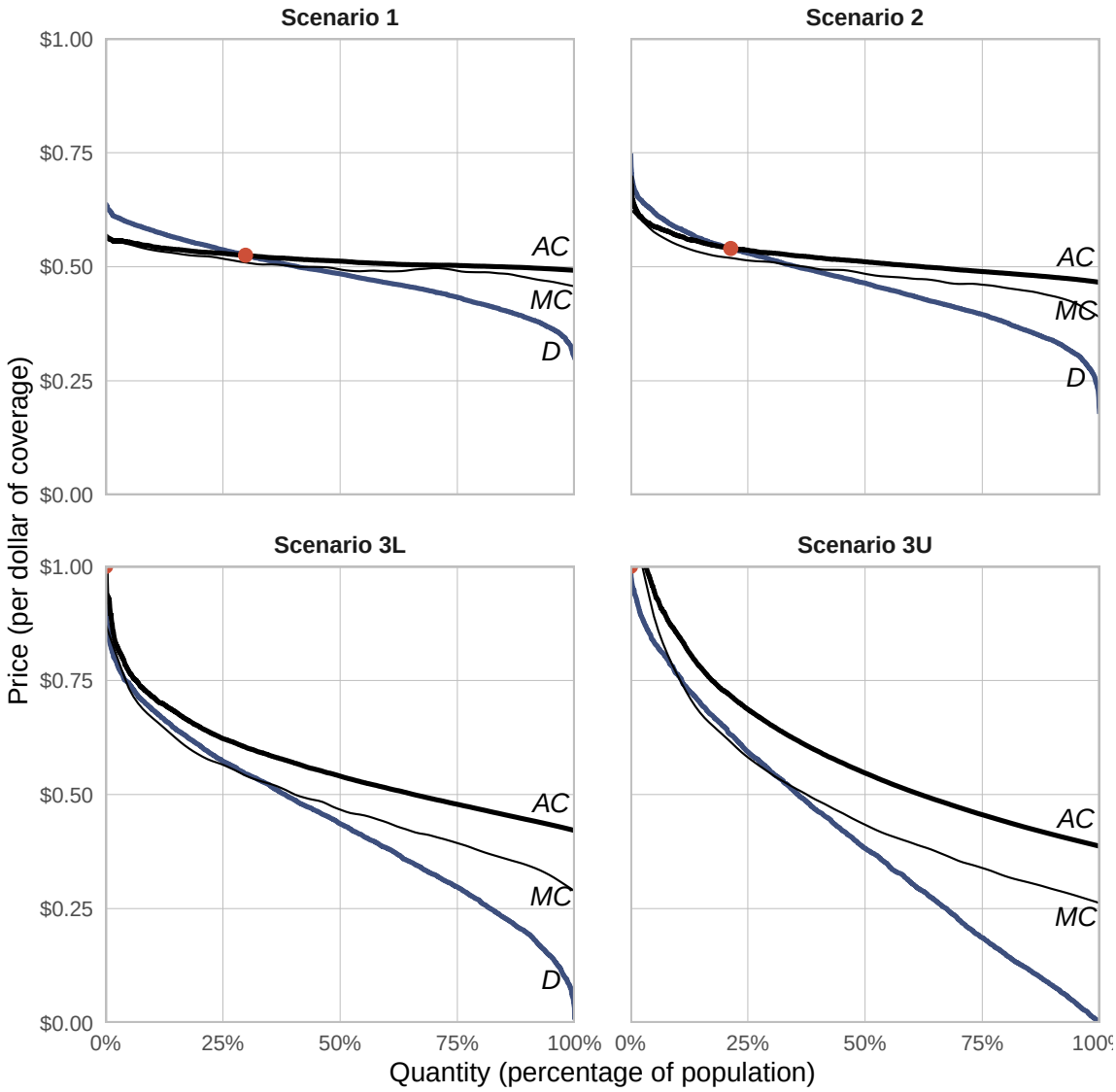

Figure 4: Supply, demand, and equilibria

Notes: This figure shows the Einav, Finkelstein, and Cullen (2010) supply (*AC*, black line), demand (*D*, blue line), and marginal cost (*MC*, thin black line) curves as well as the equilibrium point (in red) for CII under the four scenarios. The figure is based on the calibrated equilibrium model for the HRS CAD contract. The four scenarios are defined in the text.

6.3 Partial take-up of genetic testing

Scenarios 2-3U assume that 100% of consumers take up genetic prediction technology. However, there is uncertainty about the future take-up rate. The modal expert opinion is that it will be high (e.g., Meyer et al., 2024). Consistent with this, as men-

tioned, clinical trials integrating PGIs are underway, consumer genetics companies have started offering disease risk prediction, and large healthcare providers plan to inform millions of patients of their genetically predicted disease risk.

Nonetheless, to examine the robustness of our results to assuming lower take-up of genetic testing, we compute our results under two alternative take-up scenarios: 10% in the low take-up scenario and 50% in the medium scenario. As expected, the amount of selection goes down with lower take-up. In the low take-up scenario, there would not be much selection regardless of the available prediction technology. In the moderate take-up scenario, t_{80} remains problematically high for all the single-disease contracts in Scenarios 3L and 3U. For the multiple-disease contracts, t_{80} is just below Hendren's unraveling region in Scenario 3L, and is in the unraveling range in Scenario 3U. These results suggest that selection could become problematic once the predictive power exceeds the future lower bound and the take-up rate exceeds 50%.

6.4 Improvements in non-genetic prediction

Our analysis has considered expected future improvements in genetic prediction technology while treating the predictive power of non-genetic data as fixed. However, if the predictive power of non-genetic covariates correlated with PGIs also improves over time, the extent of adverse selection may be lower than what our estimates suggest. This concern, however, may be muted in light of the historically slow progress in non-genetic risk prediction.

For example, for coronary artery disease (CAD), risk prediction models have evolved—from the Framingham Risk Score (1998) to the Pooled Cohort Equations (PCE; 2013) and the AHA PREVENT equations (2023)—yet recent comparisons find broadly similar predictive performance, with only modest gains in some subgroups (Zhou et al., 2025). For other conditions, such as prostate cancer, prediction models rely on biomarkers that become informative only once the disease process has begun, and are thus ill-suited for long-run risk prediction; additional epidemiological variables yield only marginal gains in long-run prediction (Louie et al., 2015). Similarly, omics biomarkers—such as metabolomic and transcriptomic profiles—tend to capture current biological state rather than stable long-run risk, limiting their utility for long-run prediction; epigenetic measures, while sometimes

predictive of near-term outcomes, tend to integrate accumulated exposures rather than identify constitutional predisposition (Yousefi et al., 2022). This stands in stark contrast to genetic prediction: thanks to recent advances, PGIs of height and of many diseases now explain roughly 45% and 5–10% of the variation, respectively, up from zero in the early 2000s (see Section 2.1 and Table 1). Further, being fixed at conception, PGIs predict disease risk decades before onset.

6.5 Policy implications

Many developed countries currently ban insurers from using genetic information (Swiss Re Institute, 2017). Our findings suggest that, with future genetic prediction technology and widespread genetic testing, these bans would lead to high levels of adverse selection in CII, with the risk of market unravelling for most of the contracts we study.

There is a large amount of research and practical experience on how to regulate selection markets to address market failures due to asymmetric information (e.g., Einav, Finkelstein, and Levin, 2010; Einav, Finkelstein, and Fisman, 2023). The problem of regulating genetic information in insurance is similar. In a nutshell, regulation of selection markets typically seeks to balance two goals: efficiency and redistribution. Efficiency aims to maximize total economic surplus. Redistribution aims to help particular groups, in particular high risk groups who face a higher price of insurance (but also consumers with lower wealth or worse health status). The standard regulatory playbook can be divided in three approaches: laissez-faire, government provision, and managed competition. Empirical work similar to existing research on other selection markets will be necessary to design optimal policies for the CII market.

6.6 Other insurance markets

Our analysis extends in principle to health, life, and long-term care insurance, but these markets are harder to study. Unlike CII, which pays a lump sum upon diagnosis of a covered disease, health and long-term care insurance cover aggregate expenditures shaped by many conditions simultaneously, and heritability estimates and well-powered PGIs are less developed for these composite outcomes. Of the three, life insurance is closest to CII because claims are likewise triggered by a

discrete event and because this market is more similar across countries. For each insurance type, the key requirements are a current PGI with sufficient predictive power to credibly project its future power for the insured outcomes, and reliable heritability estimates; both demand large samples with genetic data matched to insured outcomes or proxies for them.

For life insurance, a PGI of lifespan could in principle predict death, but existing PGIs still have limited predictive power (e.g., Timmers et al., 2019), which may make estimation imprecise. Twin studies typically estimate the heritability of adult lifespan at ∼25% (Christensen, Johnson, and Vaupel, 2006; Timmers et al., 2019), though estimates are sensitive to the location and time period of the sample and to methodological assumptions. For health insurance, Lakhani et al. (2019) analyze claims data from 56,396 twin pairs from a large US health insurer and estimate the heritability of claim costs at 29%, and Zeeuw et al. (2021) obtain similar estimates using data from the Netherlands Twin Register linked to administrative health records. However, current PGIs of claim costs still have limited predictive power (e.g., Zeeuw et al., 2021; Lee, Jukarainen, et al., 2023). The GenCOST consortium is currently conducting a large GWAS of healthcare costs in a sample of over 2 million individuals, which should yield a more predictive PGI (May-Wilson, Nakanishi, Lee, et al., 2024). For long-term care insurance, to our knowledge, no GWAS has yet been conducted on the most relevant outcomes, such as care needs or costs.

7 Conclusion

We measure the potential for adverse selection in the CII market due to genetic prediction. We find that if genetic testing becomes widespread, there would be noticeable selection with the current prediction technology. The amount of selection would be potentially crippling with the expected future technology.

Genetic prediction is a powerful new technology that has the potential to bring about important benefits, such as personalized medicine and preventive treatments. Understanding the unintended consequences of this technology can help society mitigate its negative effects while still harvesting its benefits. This study aims to improve our understanding of these possible unintended consequences for insurance markets. Future work should explore optimal policies for the CII

market and extend the analysis to other important markets such as life, health, and long-term care insurance.

References

23andMe (2023). *Information about Genetic Health Risk Reports—23andMe*. URL: https://www.23andme.com/test-info/genetic-health/.

Abdellaoui, Abdel, Loic Yengo, Karin JH Verweij, and Peter M Visscher (2023). "15 Years of GWAS Discovery: Realizing the Promise". In: *The American Journal of Human Genetics* 110, pp. 179–194.

Akerlof, George (1970). "The Market for "lemons": Quality Uncertainty and the Market Mechanism". In: *Quarterly Journal of Economics* 84.3, pp. 488–500.

Allied Market Research (2022). *Critical Illness Insurance Market*. URL: https://www.alliedmarketresearch.com/critical-illness-insurance-market-A19460.

Association of British Insurers (2022). *ABI Guide to Minimum Standards for Criticall Illness Cover*. Tech. rep.

Bardey, David and Philippe De Donder (2013). "Genetic Testing with Primary Prevention and Moral Hazard". In: *Journal of Health Economics* 32, pp. 768–779.

Barigozzi, Francesca and Dominique Henriet (2011). "Genetic Information: Comparing Alternative Regulatory Approaches When Prevention Matters". In: *Journal of Public Economic Theory* 13, pp. 23–46.

Becker, Joel et al. (2021). "Resource Profile and User Guide of the Polygenic Index Repository". In: *Nature human behaviour* 5.12, pp. 1744–1758.

Benjamin, Daniel J, David Cesarini, Patrick Turley, and Alexander Strudwick Young (2024). "Social-science genomics: Progress, challenges, and future directions". In.

Boyer, M Martin and Franca Glenzer (2021). "Pensions, annuities, and long-term care insurance: On the impact of risk screening". In: *The Geneva Risk and Insurance Review* 46.2, pp. 133–174.

Brackenridge, R D C, Richard S Croxson, and Ross Mackenzie, eds. (2006). *Medical Selection of Life Risks*. Fifth. New York: Palgrave Macmillan.

Canadian Institute of Actuaries (2014). "Canadian Individual Critical Illness Insurance Morbidity Experience Study". In: *Mimeo*.

Cancer Research UK (2020). *Cancer Incidence for Common Cancers*. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared#heading-Zero.

Christensen, Kaare, Thomas E Johnson, and James W Vaupel (2006). "The quest for genetic determinants of human longevity: challenges and insights". In: *Nature Reviews Genetics* 7.6, pp. 436–448.

Denny, Joshua C et al. (2013). "Systematic Comparison of Phenome-wide Association Study of Electronic Medical Record Data and Genome-wide Association Study Data". In: *Nature Biotechnology*, pp. 1102–1111.

Dixon, Padraig, Rachel H. Horton, William G. Newman, John H. McDermott, and Anneke Lucassen (2024). "Genomics and Insurance in the United Kingdom: Increasing Complexity and Emerging Challenges". In: *Health Economics, Policy and Law*, pp. 1–13.

Doherty, Neil A and Paul D Thistle (1996). "Adverse Selection with Endogenous Information in Insurance Markets". In: *Journal of Public Economics* 63, pp. 83–102.

Dorse, Kevin (Feb. 2024). "Got Questions About Critical Illness Insurance? CLHIA's latest consumer guide has answers to help you". en. In: *Canadian Life & Health Insurance Association*.

Dudbridge, Frank (2013). "Power and Predictive Accuracy of Polygenic Risk Scores". In: *PLoS Genetics* 9, e1003348.

Einav, Liran, Amy Finkelstein, and Mark R Cullen (2010). "Estimating Welfare in Insurance Markets Using Variation in Prices". In: *The Quarterly Journal of Economics* 125.3, pp. 877–921.

Einav, Liran, Amy Finkelstein, and Ray Fisman (2023). *Risky Business: Why Insurance Markets Fail and What to Do About It*. Yale University Press.

Einav, Liran, Amy Finkelstein, and Jonathan Levin (2010). "Beyond Testing: Empirical Models of Insurance Markets". In: *Annu. Rev. Econ.* 2.1, pp. 311–336.

Falconer, Douglas Scott and Trudy FC Mackay (1996). *Introduction to Quantitative Genetics*. 4th. Prentice Hall, Essex.

Financial Services Council (July 2020). *Detailed Data Reveals Top Causes of Claim for the Industry*. Tech. rep. Financial Services Council.

Gallagher Re (2025). *Demographics and Long-Term Care: How insurers can turn a coming social crisis into an opportunity to help*. Tech. rep. London: Gallagher Re.

Gatzert, Nadine and Alexander Maegebier (2015). "Critical Illness Insurances: Challenges and Opportunities for Insurers". In: *Risk Management and Insurance Review* 18.2, pp. 255–272.

Ge, Tian, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller (2019). "Polygenic Prediction via Bayesian Regression and Continuous Shrinkage Priors". In: *Nature Communications* 10, p. 1776.

Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin (2013). *Bayesian Data Analysis*. 3rd. Chapman and Hall/CRC.

Gen Re (Dec. 2023a). *Insights from Gen Re's Dread Disease Survey for the Hong Kong, Malaysia, and Singapore Markets*. Tech. rep.

— (2023b). *U.S. Critical Illness and Accident Insurance Market Survey: Highlights of 2022/2023 Results*. Tech. rep. South Portland: Gen Re. URL: https://www.genre.com/content/dam/generalreinsuranceprogram/documents/surveylhci23-en.pdf.

— (Oct. 2024). *The Key Elements of Critical Illness Definitions for Mental Health Disorders*. URL: https://www.genre.com/us/knowledge/publications/2024/october/the-key-elements-of-critical-illness-definitions-for-mental-health-disorders-en.

Golinghorst, Dexter et al. (2022). "Anti-Selection & Genetic Testing in Insurance: An Interdisciplinary Perspective". In: *Journal of Law, Medicine & Ethics* 50.1, pp. 139–154.

Gottlieb, Daniel and Kent Smetters (2021). "Lapse-based Insurance". In: *American Economic Review* 111.8, pp. 2377–2416.

Hendren, Nathaniel (2013). "Private Information and Insurance Rejections". In: *Econometrica* 81.5, pp. 1713–1762.

Hindriks, Jean and Philippe De Donder (2003). "The Politics of Redistributive Social Insurance". In: *Journal of Public Economics* 87, pp. 2639–2660.

Hoel, Michael and Tor Iversen (2002). "Genetic Testing When There Is a Mix of Compulsory and Voluntary Health Insurance". In: *Journal of Health Economics* 21, pp. 253–270.

Howard, Robert CW (2016). "Genetic Testing Model for CI: If Underwriters of Individual Critical Illness Insurance Had No Access to Known Results of Genetic Tests". In: *Canadian Institute of Actuaries Report to CIA Research Committee*.

Hoy, Michael and Mattias Polborn (2000). "The Value of Genetic Information in the Life Insurance Market". In: *Journal of Public Economics* 78, pp. 235–252.

Joly, Yann et al. (2014). "Life Insurance: Genomic Stratification and Risk Classification". In: *European Journal of Human Genetics* 22, pp. 575–579.

Karlsson Linnér, Richard and Philipp D Koellinger (2022). "Genetic Risk Scores in Life Insurance Underwriting". In: *Journal of Health Economics* 81, p. 102556.

Kimball, Miles S, Claudia R Sahm, and Matthew D Shapiro (2008). "Imputing Risk Tolerance from Survey Responses". In: *Journal of the American statistical Association* 103.483, pp. 1028–1038.

Kullo, Iftikhar J. (Oct. 2025). "Clinical Use of Polygenic Risk Scores: Current Status, Barriers and Future Directions". In: *Nature Reviews Genetics*.

Lakhani, Chirag M. et al. (2019). "Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes". In: *Nature Genetics* 51.2, pp. 327–334.

Lee, Jiwoo, Sakari Jukarainen, et al. (2023). "Quantifying the causal impact of biological risk factors on healthcare costs". In: *Nature Communications* 14, p. 5672.

Lee, Sang Hong, Michael E Goddard, Naomi R Wray, and Peter M Visscher (2012). "A Better Coefficient of Determination for Genetic Profile Analysis". In: *Genetic Epidemiology* 36.3, pp. 214–224.

Louie, K.S., A. Seigneurin, P. Cathcart, and P. Sasieni (2015). "Do prostate cancer risk models improve the predictive accuracy of PSA screening? A meta-analysis". In: *Annals of Oncology* 26.5, pp. 848–864.

Lynch, M and B Walsh (1999). *Genetics and Analysis of Quantitative Traits*. Sunderland: Sinauer.

Macdonald, Angus, Delme Pritchard, and Pradip Tapadar (2006). "The Impact of Multifactorial Genetic Disorders on Critical Illness Insurance: A Simulation Study Based on UK Biobank". In: *ASTIN Bulletin* 36.02, pp. 311–346.

Macdonald, Angus and Pradip Tapadar (2010). "Multifactorial Genetic Disorders and Adverse Selection: Epidemiology Meets Economics". In: *Journal of Risk and Insurance* 77.1, pp. 155–182.

Macdonald, Angus S. and Kenneth R. McIvor (2009). "Modelling Adverse Selection in the Presence of a Common Genetic Disorder: The Breast Cancer Polygene". In: *ASTIN Bulletin* 39.2, pp. 373–402.

Markel, Gareth et al. (2025). "Nature, Nurture, and Socioeconomic Outcomes: New Evidence from Sib Pairs and Molecular Genetic Data". In: *Available at SSRN*.

Maxwell, Jessye M et al. (2021). "Multifactorial Disorders and Polygenic Risk Scores: Predicting Common Diseases and the Possibility of Adverse Selection in Life and Protection Insurance". In: *Annals of Actuarial Science* 15.3, pp. 488–503.

May-Wilson, Sebastian, Tomoko Nakanishi, Jiwoo Lee, et al. (2024). *Quantifying the impact of genetic variation on healthcare cost across 1.5 million individuals from 13 studies and 8 countries: the GenCost consortium*. American Society of Human Genetics Annual Meeting. Conference abstract.

McKelvey, Richard D and William Zavoina (1975). "A Statistical Model for the Analysis of Ordinal Level Dependent Variables". In: *Journal of Mathematical Sociology* 4.1, pp. 103–120.

Meyer, Michelle N. et al. (2024). "Potential Corporate Uses of Polygenic Indexes: Starting a Conversation About the Associated Ethics and Policy Issues". In: *The American Journal of Human Genetics* 111, pp. 833–840.

MyHeritage (2021). *Understanding Your Polygenic Risk Reports*. URL: https://education.myheritage.com/article/understanding-your-polygenic-risk-reports/.

Oster, Emily, Ira Shoulson, Kimberly Quaid, and E Ray Dorsey (2010). "Genetic Adverse Selection: Evidence from Long-Term Care Insurance and Huntington Disease". In: *Journal of Public Economics* 94.11-12, pp. 1041–1050.

PartnerRe (Sept. 2015). *Growth Potential of Critical Illness Insurance in Asia*. Tech. rep. PartnerRe.

Peter, Richard, Andreas Richter, and Paul Thistle (2017). "Endogenous Information, Adverse Selection, and Prevention: Implications for Genetic Testing Policy". In: *Journal of Health Economics* 55, pp. 95–107.

Polborn, Mattias K, Michael Hoy, and Asha Sadanand (2006). "Advantageous Effects of Regulatory Adverse Selection in the Life Insurance Market". In: *The Economic Journal* 116, pp. 327–354.

Posey, Lisa L. and Paul D. Thistle (2021). "Genetic Testing and Genetic Discrimination: Public Policy When Insurance Becomes "too Expensive"". In: *Journal of Health Economics*, p. 102441.

Price, Alkes L et al. (2006). "Principal Components Analysis Corrects for Stratification in Genome-wide Association Studies". In: *Nature genetics* 38.8, pp. 904–909.

Regalado, Antionio (2019). *More Than 26 Million People Have Taken an At-home Ancestry Test*. URL: https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/.

RGA (2016). *Critical Illness Insurance A Medical Perspective*. Tech. rep. RGA: Reinsurance Group of America. URL: https://www.rgare.com/knowledge-center/article/reflections-critical-illness-insurance-a-medical-perspective.

Rothschild, Michael and Joseph Stiglitz (1976). "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information". In: *The Quarterly Journal of Economics* 90, pp. 629–649.

Strohmenger, Rainer and Achim Wambach (2000). "Adverse Selection and Categorical Discrimination in the Health Insurance Markets: The Effects of Genetic Tests". In: *Journal of Health Economics* 19.2, pp. 197–218.

Sudlow, Cathie et al. (2015). "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". In: *PLoS Medicine* 12.3, e1001779.

Swiss Re Institute (2016). *Expert Forum on Cancer Diagnostics Conference report*. Tech. rep. Zurich: Swiss Re Institute.

— (2017). *Seeing the Future? How genetic testing will impact life insurance*. Tech. rep. Swiss Re Institute.

— (2022). *Critical Illness Insurance in China*. Tech. rep. Swiss Re Institute. URL: https://www.swissre.com/institute/research/topics-and-risk-dialogues/china/critical-illness-insurance-in-china.html.

— (2024a). *Genetic Testing in the Life & Health insurance industry*. Tech. rep. Swiss Re Institute.

— (2024b). *Protection for China's ageing society*. Tech. rep. Swiss Re Institute. URL: https://www.swissre.com/institute/research/topics-and-risk-dialogues/china/protection-china-ageing-society.html.

Tabarrok, Alexander (1994). "Genetic Testing: An Economic and Contractarian Analysis". In: *Journal of Health Economics* 13.1, pp. 75–91.

Taylor Jr, Donald H et al. (2010). "Genetic Testing for Alzheimer's and long-term care insurance". In: *Health Affairs* 29.1, pp. 102–108.

Timmers, Paul RHJ et al. (2019). "Genomics of 1 Million Parent Lifespans Implicates Novel Pathways and Common Diseases and Distinguishes Survival Chances". In: *elife* 8, e39856.

Visscher, Peter M, Loic Yengo, Nancy J Cox, and Naomi R Wray (2021). "Discovery and Implications of Polygenicity of Common Diseases". In: *Science* 373, pp. 1468–1473.

Yang, Jian, Beben Benyamin, M S Lund, Scott Gordon, Anjali K Henders, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height". In: *Nature Genetics* 42, pp. 565–569.

Yengo, Loïc et al. (2022). "A Saturated Map of Common Genetic Variants Associated with Human Height". In: *Nature* 610, pp. 704–712.

Yousefi, Paul D. et al. (2022). "DNA methylation-based predictors of health: applications and statistical considerations". In: *Nature Reviews Genetics* 23, pp. 369–383.

Zeeuw, Eveline L. de et al. (2021). "Safe Linkage of Cohort and Population-Based Register Data in a Genomewide Association Study on Health Care Expenditure". In: *Twin Research and Human Genetics* 24, pp. 103–109.

Zhao, Jinbo, Michael Salter-Townshend, and Adrian O'Hagan (2023). "A Simulation Study for Multifactorial Genetic Disorders to Quantify the Impact of Polygenic Risk Scores on Critical Illness Insurance". In: *European Actuarial Journal* 13, pp. 775–813.

Zhou, Hui et al. (2025). "Evaluation and Comparison of the PREVENT and Pooled Cohort Equations for 10-Year Atherosclerotic Cardiovascular Risk Prediction". In: *Journal of the American Heart Association* 14.4.

Zick, Cathleen D et al. (2005). "Genetic testing for Alzheimer's disease and its impact on insurance purchasing behavior." In: *Health Affairs* 24, pp. 483–490.

Appendix

A Proof of Theorem 1

Lemma A.1. *Conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$, G_f is normally distributed with mean $ag_c + b\mathbf{w}\boldsymbol{\theta}$ and variance c^2, where*

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

Proof. By Assumption 2, conditional on $\mathbf{W} = \mathbf{w}$, $G_f = \mathbf{w}\boldsymbol{\theta} + V$, 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 A.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 $\Gamma = (\mathbb{P}_{\mathbf{W}}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_V^2, \sigma_\epsilon^2)$ 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: $\boldsymbol{\theta}$. Assumption 4 implies that $\mathbb{E}[G_c|\mathbf{W} = \mathbf{w}] = \mathbf{w}\boldsymbol{\theta}$. 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 $\boldsymbol{\theta}$ is the vector of coefficients of a regression of G_c on $\mathbf{W}$.

Part 2: σ_ϵ^2. By Assumption 4, there exists a latent variable $\mathcal{L}$ for the disease such that $\mathcal{L} := G\beta_g + \boldsymbol{W}\boldsymbol{\beta}_w + \eta$, where $\eta \sim N(0,1)$, $\eta \perp (\boldsymbol{W}, V, \epsilon)$, and $L = \{\mathcal{L} > 0\}$, and where the operator $\{..\}$ is equal to 1 if its argument is true and to 0 otherwise.

Let R_c^2 be the pseudo-R^2 of a probit regression of L on the current PGI, and let R_f^2 be the pseudo-R^2 of a probit regression of L on the future PGI. We have

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

Therefore, using the fact that $\text{Cov}[\mathcal{L}, G_f] = \text{Cov}[\mathcal{L}, G_c]$ and $\text{Var}[G_c] = 1$, we have $R_c^2/R_f^2 = \text{Var}[G_f]/\text{Var}[G_c] = \text{Var}[G_f]$. R_c^2 is a function of the data, Assumption 5 says that we know R_f^2, and Assumption 3 implies that $1 = \text{Var}[G_c] = \text{Var}[G_f] + \sigma_\epsilon^2$. Plugging in the formula above, we have

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

Part 3: σ_V^2. Assumptions 3 and 2 yield

$$G_c = \mathbf{W}\boldsymbol{\theta} + V + \epsilon. \tag{A.2}$$

Taking the variance, we find $1 = \text{Var}[\mathbf{W}\boldsymbol{\theta}] + \sigma_V^2 + \sigma_\epsilon^2$, so that

$$\sigma_V^2 = 1 - \text{Var}[\mathbf{W}\boldsymbol{\theta}] - \sigma_\epsilon^2. \tag{A.3}$$

Part 4: β_g**.** As mentioned in Part 2, by Assumption 4, we have $L = \{G_f\beta_g + \mathbf{W}\boldsymbol{\beta}_w + \eta > 0\}$. Consider the distribution of L conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$. Lemma A.1 implies that $G_f = ag_c + b\mathbf{w}\boldsymbol{\theta} + c\nu$, where ν is a standard normal random variable independent of η. Therefore, conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$,

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

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

$$\Pr(L = 1 | G_c = g_c, \mathbf{W} = \mathbf{w}) = \Phi\left[g_c \frac{a\beta_g}{\sqrt{1 + c^2\beta_g^2}} + \mathbf{w}\frac{b\boldsymbol{\theta}\beta_g + \boldsymbol{\beta}_w}{\sqrt{1 + c^2\beta_g^2}} \right]. \tag{A.4}$$

Let γ be the coefficient vector of a probit regression of L on G_c and $\mathbf{W}$. Then $\gamma_g = \frac{a\beta_g}{\sqrt{1+c^2\beta_g^2}}$. This equation implies that β_g has the same sign as γ_g. Taking the square of both sides and solving, we find

$$\beta_g = \frac{\gamma_g}{\sqrt{a^2 - \gamma_g^2 c^2}}. \tag{A.5}$$

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

Part 5: $\boldsymbol{\beta}_w$**.** By Equation A.4 and using the fact that the support of $\mathbf{W}$ includes a set that spans the entire space (Assumption 1), $\boldsymbol{\gamma}_w = \frac{b\boldsymbol{\theta}\beta_g + \boldsymbol{\beta}_w}{\sqrt{1+c^2\beta_g^2}}$. Then

$$\boldsymbol{\beta}_w = \boldsymbol{\gamma}_w\sqrt{1 + c^2\beta_g^2} - b\boldsymbol{\theta}\beta_g. \tag{A.6}$$

□

B Estimation of the single-disease model

For Scenarios 1 and 2, estimation 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 PGI (Scenario 2).

For Scenarios 3L and 3U, we first estimate R_c^2, σ_ϵ^2, and $\mathbb{P}_W$ (see Appendix A). For $(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_V^2)$, we use maximum likelihood estimation. To simplify the analysis, write the vector of parameters as $(\sigma_T^2, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_V^2)$, where $\sigma_T^2 = \sigma_\epsilon^2 + \sigma_V^2$ and $\boldsymbol{\gamma}$ are the coefficients of the probit regression of L on G_c and $\mathbf{W}$ (Equation A.4). Written this way, $\boldsymbol{\beta}$ is a function of the other parameters, as given by Equations A.5 and A.6, and σ_V^2 is a function of σ_T^2.

We can decompose the likelihood function:

$$\begin{aligned} &f(L = l, G_c = g_c, \mathbf{W} = \mathbf{w} | \sigma_T^2, \boldsymbol{\gamma}, \boldsymbol{\theta}) \\ =&f_w(\mathbf{w}) \cdot f_{LM}(g_c | \mathbf{w}; \sigma_T^2, \boldsymbol{\gamma}, \boldsymbol{\theta}) \cdot f_{PM}(l | g_c, \mathbf{w}; \sigma_T^2, \boldsymbol{\gamma}, \boldsymbol{\theta}). \end{aligned}$$

The distribution of G_c conditional on $\mathbf{W} = \mathbf{w}$ is given by the linear model (LM) in Equation A.2 and the distribution of L conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$ is given by the probit model (PM) in Equation A.4. Therefore, the likelihood can be simplified to

$$f_w(\mathbf{w}) \cdot f_{LM}(g_c | \mathbf{w}; \sigma_T^2, \boldsymbol{\theta}) \cdot f_{PM}(l | g_c, \mathbf{w}; \boldsymbol{\gamma}).$$

Therefore, the log-likelihood function is a constant plus

$$\log f_{LM}(g_c | \mathbf{w}; \sigma_T^2, \boldsymbol{\theta}) + \log f_{PM}(l | g_c, \mathbf{w}; \boldsymbol{\gamma}).$$

Thus, we can fit the model by maximum likelihood separately for $(\sigma_T^2, \boldsymbol{\theta})$ with a linear model and for $\boldsymbol{\gamma}$ with a probit model. $\boldsymbol{\beta}$ and σ_V^2 can then be calculated.

For inference, the log likelihood has no interaction terms between $(\sigma_T^2, \boldsymbol{\theta})$ and $\boldsymbol{\gamma}$. Therefore, the Fisher information matrix for these parameters is block diagonal, so $(\sigma_T^2, \boldsymbol{\theta})$ and $\boldsymbol{\gamma}$ are asymptotically normal and independent, with the standard errors from their separate estimation. Because $\boldsymbol{\beta}$ and σ_V^2 are smooth functions of the other parameters, they are also asymptotically normal. We estimate the full variance-covariance matrix for the parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_V^2)$ with the delta method. Because we are using maximum likelihood, the estimator is asymptotically efficient.