ONLINE APPENDIX

Genetic prediction and adverse selection*

Eduardo Azevedo† Jonathan Beauchamp‡
Richard Karlsson Linnér§

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

1 Protocol for polygenic indexes (PGIs)

Our protocol for computing PGIs aligns with recent studies on genetic prediction, such as Becker et al. (2021) or Ge et al. (2019), and follows recent best practices and community guidelines (see, e.g., Figs 1-2 in Choi, Mak, and O'Reilly, 2020).

In summary, we computed PGIs using PRS-CS (Ge et al., 2019), which models the covariance between SNPs to transform their GWAS regression estimates into $\hat{\mathbf{b}}$, as if the SNPs were a fitted together in a multiple regression. The motivation for this transformation is that a GWAS tests millions of SNPs for association one-by-one in repeated regressions. Also, PRS-CS improves the PGIs' signal-to-noise ratio by applying a continuous shrinkage prior to downweight the effects of weakly associated SNPs that are expected to add more noise than signal to the PGI.

*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, r.karlssonlinner@vu.nl

We ran PRS-CS using its default parameters, which restricts the set of SNPs to a high-quality reference set that is frequently used in the PGI literature (i.e., the HapMap 3 SNPs) (Altshuler et al., 2010). This set has satisfactory coverage to capture the bulk of the heritability due to common SNPs in populations of European-like ancestry (Yengo et al., 2022). This set of SNPs was used to compute all our PGIs. After PRS-CS, we used PLINK2 (Chang et al., 2015) to compute the PGIs in the UKB data by weighting the genotypic data ($\mathbf{X}$) by the PRS-CS adjusted GWAS coefficients ($\hat{\mathbf{b}}$).

For the PGI of Alzheimer's disease, we augmented the standard HapMap 3 set of SNPs with two tag-SNPs that are known to capture the risk types of the *APOE* gene. These two tag-SNPs are rs7412 and rs429358, which are located in the *APOE* region on chromosome 19. We used these two SNPs to create an "APOE-only" score, and combined that score with a genome-wide PGI that excludes a 5Mb region surrounding APOE to avoid redundancy due to linkage disequilibrium.

1.1 GWAS, quality control (QC), and meta-analysis

To compute the PGIs, we need GWAS "summary statistics" for the diseases we study. Table 1 shows the studies from which we sourced our GWAS summary statistics. In particular, we need the GWAS regression coefficients of each SNP on the disease. To avoid overfitting, these regression coefficients must be estimated in data independent of the UK Biobank (Wray et al., 2013). Further documentation on the GWAS data sources is provided in the replication package.

To quantify the sample size of a GWAS, we rely on the effective sample size metric, "N_{eff}", which penalizes the total sample size of an imbalanced case-control regression so that it matches the expected statistical power of a balanced analysis with 50% cases/controls (Grotzinger et al., 2022).

We applied GWAS quality control (QC) to each downloaded set of summary statistics, as is standard practice (Winkler et al., 2014). Our GWAS QC-protocol was based on the protocol developed by the Social Science Genetic Association Consortium (SSGAC) (Karlsson Linnér, Biroli, et al., 2019), which is a continuation of an older "industry-standard" protocol (Winkler et al., 2014). The GWAS QC has two main aims: (1) to remove SNPs that for technical reasons are likely to worsen the signal-to-noise ratio of the PGI (e.g., poorly measured SNPs); and (2) to ensure

Table 1: Summary of the GWAS of the seven diseases of interest

Disease	Label	N_{cases}	$N_{controls}$	$Neff$	Reference
Alzheimer's disease	ALZ	41,197	445,030	126,275	Lambert et al. (2013)
Breast cancer	BRC	133,384	113,789	236,094	Zhang et al. (2020)
Coronary artery disease	CAD	56,424	357,721	117,486	Nikpay et al. (2015)
Colorectal cancer	CRC	60,801	123,504	162,973	Fernandez-Rozadilla et al. (2023)
Prostate cancer	PRC	79,148	61,106	137,933	Schumacher et al. (2018)
Schizophrenia	SCZ	67,390	94,015	157,013	Trubetskoy et al. (2022)
Type 2 diabetes	T2D	55,005	400,308	193,440	Mahajan et al. (2018)

Notes: the listed GWASs of ALZ and CRC were each meta-analyzed with a GWAS conducted in the FinnGen biobank (Kurki et al., 2023). Further documentation is provided in the replication package.

that SNP locations, reference nucleotide, and other per-SNP statistics get aligned across files (e.g., alignment with the forward strand in the genome reference data).

The protocol was applied using the software EasyQC (Winkler et al., 2014). It removed (i) multi-allelic or non-SNP variants, (ii) SNPs on the sex chromosomes, (iii) SNPs with MAF $<$0.5% (PRS-CS also removes SNPs with MAF $<$1%), (iv) SNPs that could not be matched with the genome reference data from the Haplotype Reference Consortium (HRC) (McCarthy et al., 2016), and (v) SNPs whose MAF differs by more than 0.2 from the reference frequency in the HRC data. Diagnostic plots were inspected, and no conspicuous issues were found. Imputation quality metric was not filtered on, because it was not available in all GWAS data. Nonetheless, the downloaded files were already filtered on imputation quality by the original studies, and HapMap 3 SNPs are, on average, well imputed.

1.2 Genotype data, ancestry, and sample quality control (QC)

Our analyses were conducted with the harmonized genetic data resource collected, maintained, and distributed by the UK Biobank (Data Category 100319), described at length in Bycroft et al. (2018). Imputed genetic data were available for about 487,000 participants. The total number of imputed genetic variants is about 97 million, but the vast majority of these are typically not tested in GWASs because they are very rare. There are about 10.3 million imputed variants with MAF $>$ 1%.

We followed standard practice and restricted our analysis to individuals of European-like ancestry. Well-powered PGIs are not readily available for other ancestral groups (Martin et al., 2017). Following best practices in applied genetic

research (Price et al., 2006), we included the first 10 genetic principal components (PCs) as control variables when estimating our econometric model, to control for population stratification.[1] We used the standard set of PCs distributed by the UKB (Data Category 22009).

To restrict the data to European-like ancestry, we relied on the approach described in Karlsson Linnér, Biroli, et al. (2019). Participants were kept for analysis if (i) they self-reported their ethnic background as "White", "White British", "White Irish", or "Any other white background" (Data Field 21000), and (ii) their value on the first genetic PC is ≤ 0. $N \sim 448,000$ individuals passed these two filters.

We applied sample-level quality control (QC) filters, using quality metrics distributed by the UKB (Data Category 100313). Participants were removed when (1) there was a mismatch between surveyed and genetic sex, (2) there was evidence of sex-chromosome aneuploidy, (3) they were classified as outliers on either genotype heterozygosity or missingness rates based on directly genotyped SNPs. $N \sim 446,500$ remained after the sample-level QC filters.

2 Health data and risk factors

The UKB has collected extensive survey and interview data on clinical and environmental disease risk factors, sociodemographics, as well as biomedical assays from blood, saliva, and urine (Sudlow et al., 2015). A touchscreen survey was followed by a structured interview by a trained nurse to verify and map any self-reported medical conditions to standard disease categories. The most recent version of the study data was pulled on the UK Biobank Research Analysis Platform on October 22, 2024.

The UKB has since been enriched through linkage with several electronic health record (EHR) data sources. Most of the healthcare in the UK is delivered by the publicly-funded National Health Service (NHS) (Kelly and Stoye, 2020). Only 7% of the population is covered by private health insurance (Anderson and Mossialos, 2022). Thus, the vast majority of healthcare in the UK is provided and tracked by

[1] Population stratification can occur when cultural or environmental differences across populations that impact a disease correlate with genetic differences that are typically non-causal for the disease. This can introduce bias in GWASs and other applied genetic research. For this study, since we are interested in *prediction* rather than *causation*, population stratification is not a concern. We nonetheless follow standard practice in genetic research.

the NHS (or its subsidiaries). However, the UK has no centralized system dedicated to maintaining primary care records from general practices, so primary care records have worse coverage than the other linked sources described below. For living participants, new information is still being recorded in their EHR. The main EHR sources linked to the UKB are:

1. Primary care records (Category 3000)
2. Hospital inpatient records (Category 2000)
3. National cancer registries (Category 100092)
4. National death registries (Category 100093).

Some of these data are still in the process of being linked. Primary care records have thus far only been linked to about 45% of the participants. Nevertheless, because the diseases we study are fairly serious, and because the cancer and death registries are virtually complete, it is unlikely that we are missing a substantial number of disease cases because of missing primary care records. Overall, the combination of self-reported and interviewed medical history with several distinct EHR sources provide good overall coverage of the disease and medical histories of the UKB participants.

2.1 Coding the disease outcomes and comorbidities

Our starting point to code the disease outcomes for statistical analysis is the "First occurrence of health outcomes" resource (Category 1712), which is maintained and distributed by the UK Biobank. This resource is the result of a harmonization of the self-reported medical data with the linked EHR sources (i–iv). However, because the "First occurrence" resource excludes cancer codes, and because it is updated at periodic intervals that differ from the periodic updates of the underlying sources (i–iv), we merged the "First occurrence" resource with the most recent data from all EHR sources (i–iv).

The disease outcomes are recorded in the data using the International Classification of Diseases (ICD) version 10, censored to first three characters (e.g., C50 "Malignant neoplasm of breast"). To code the unstructured EHR-data for statistical analysis, we relied on the "Phecode" mapping system (Denny et al., 2013). This system was designed specifically to define meaningful case-control variables from complex EHR sources (Bastarache, 2021). A key feature is to condense some 90,000

detailed codes from the International Classification of Diseases (e.g., C50.211: "malignant neoplasm of upper-inner quadrant of female breast") into about 1,900 major case-control status codes (174.1: "Breast cancer [female]"). Secondly, it combines redundant or alternative codes spread across ICD chapters. For example, although the official chapter for breast cancer is C50, there are alternative codes that some doctors may use, such as "D05 Carcinoma in situ of breast". In the latest version, the Phecode mapping system condenses some 90,000 ICD-10 codes into about 1,900 Phecodes (Bastarache, 2021).

For the exact mapping of ICD-codes to Phecodes, see the replication package. With the exception of coronary artery disease, our diseases of interest could all be captured by a single Phecode. To code coronary artery disease, which is defined as a collection of many underlying heart and circulatory conditions, we relied on the approach of the previous literature and considered the entire ICD10 Chapter I20–I25 "Ischaemic heart diseases", which maps to a total of six different Phecodes.

In addition to the diseases, any comorbidities sourced from the EHR data (described next) were also coded using the Phecode system.

2.2 Coding the epidemiological risk factors and comorbidities

It is important that our model captures the information typically observed by insurers. Therefore, we reviewed clinical guidelines and the epidemiological literature to identify the main disease-specific non-genetic risk factors that are used either by medical practitioners or by insurers to assess a person's disease risk. Our resulting list of epidemiological risk factors is shown in Table 3 in the main text and was verified independently by two medical doctors. A complete reference list is provided in the replication package.

The UKB data are rich enough for us to code most of the risk factors identified in the review. Notably, we were able to code family history (father, mother, or siblings) for six of the seven diseases (family history of schizophrenia was proxied by family history of severe depression). Only a handful of factors were eventually omitted, either because they were unspecific (e.g., diet) or because of limited data availability (e.g., substance use). Only three risk factors were omitted completely: brain injury, cannabis use, and substance use.

The exact data fields for coding the risk factors are listed in the replication pack-

age. We defined a set of nine disease-general risk factors that were included in all the analyses: age (at the most recent observation), sex (omitted from any sex-specific analysis), Townsend's deprivation index, education (years of schooling), BMI, alcohol consumption (drinks per week), smoking (current, ex, or never), a dummy variable indicating physical inactivity, and systolic blood pressure. These nine risk factors were always accompanied by the top 10 genetic PCs and by a genotyping array dummy (to control for the fact that part of the sample was genotyped on a different genotyping array), as well as by the family history variables mentioned above.

In addition to these disease-general covariates, we coded risk factors and comorbidities specific to each of the seven diseases. These disease-specific covariates were coded either as continuous variables or as dummies, with the exception of BMI and alcohol consumption (drinks per week), which were coded as percentile ranks. The reason is that percentile ranks remain more stable as people age (for more information, see Section 2.3 below). Also, because not all female participants have yet undergone menopause, we coded the variable indicating age at menopause (or hysterectomy) as zero for women who had not yet undergone menopause, and then also included a dummy indicating menopause status.

2.3 Adjustment of age-dependent covariate values

An objective of this study is to predict the risk of disease by a given age a (here, a = 65). The data is informative of the age of onset for the medical conditions and comorbidities we study, but for most participants, we only observe their covariate values once (at the baseline assessment in 2006–2010). Therefore, after estimating our econometric model, when predicting disease risk, the values of age-dependent covariates were adjusted to reflect age a before projecting the disease risk.

We did not adjust any time-fixed covariates (e.g., genotyping array dummy) or covariates that are mostly stable in older populations, such as education (years of schooling) or age at first menstruation, nor the handful of geographical variables based on the home address at the time of recruitment, such as the Townsend's deprivation index or air pollution.

To determine whether a covariate needed adjustment in the disease-risk prediction step, we regressed the covariates one-by-one on a fourth-degree polynomial

of age. Whenever the polynomial explained more than 1% of the variation, the covariate was adjusted. Otherwise, we used the observed value in both the estimation and disease-risk prediction steps. Also, to be consistent across our seven diseases of interest, the family history variables were always age-adjusted. Because all age-dependent covariates were observed to increase as a function of age, the age-adjustment procedure effectively increased the values of these covariates for participants with age $a_i < 65$, and decreased the values for those with $a_i > 65$.

The age-adjustment procedure was done separately by sex. For continuous covariates, with the exception of "age at menopause (or hysterectomy)", we first ran a linear regression of each covariate W on a fourth-degree polynomial of age: $W(a) = a_1 \times a + a_2 \times a^2 + a_3 \times a^3 + a_4 \times a^4 + e$. The estimated regression coefficients were then used to predict the values of the covariate $\bar{w}(a)$ for each year of age a observed in the data (i.e., 39–86 years). Then, we adjusted the value at age 65 for participant i observed at age a_i by adding the predicted difference between $\bar{w}(65)$ and $\bar{w}(a_i)$:

$$w_i(65) := w_i(a) + [\bar{w}(65) - \bar{w}(a_i)].$$

For binary covariates, except for "ever menopause (or hysterectomy)", we proceeded analogously, but estimated probit rather than linear regressions and projected probabilities. We only adjusted probabilities among participants with age $a_i < 65$ who had not experienced the event of the covariate, as well as among individuals with age $a_i > 65$ who had experienced the event but for which we could not observe the age of the event. We used the observed covariate values of participants with age $a_i < 65$ who had already experienced the event before age 65, and of participants with age $a_i > 65$ who had not (yet) experienced the event.

To illustrate the procedure for binary covariates, consider a woman of age $a_i = 40$ who had not yet experienced hypertension. For that woman, the observed covariate value is $x_i(40) = 0$. This was adjusted to $w_i(65) = 0 + [0.20 - 0.02]$, where $[0.20 - 0.02] = 0.18$ is the difference between the probabilities of having hypertension at ages 65 and 40, which we assume to be the probability of developing hypertension between ages 40 and 65 conditional on not having hypertension at age 40. Similarly, the covariate for a woman of age 80 and who had experienced hypertension was adjusted to $w_i(80) = 1 + [0.2 - 0.6] = 0.6$, to reflect the fact that the woman may not have had hypertension at age 65.

For the covariates "age at menopause (or hysterectomy)" and "ever menopause

(or hysterectomy)", all women were coded as having experienced the event by age 65 in the prediction step, reflecting that virtually all women reach menopause (or undergo hysterectomy) by this age. For women older than 65 with no recorded event, we recoded them as having experienced the event at age 65. For women younger than 65 with no event, we imputed an expected age at event using a Kaplan-Meier estimate of $E[\text{age at event} \mid \text{no event by age } a_i]$, after forcing the cumulative event probability to reach 100% by age 65. Women who had already experienced the event before age 65 retained their observed age at event, while those who experienced it after age 65 were recoded to event age 65. This procedure ensures that all women are treated as having experienced menopause (or hysterectomy) by age 65 in a manner consistent with observed timing patterns in the data.

3 Generalized econometric model for multiple-disease contracts

3.1 Model for multiple-disease contracts

We now discuss how we model multiple-disease CII contracts. Our simple economic framework in Section 3.1 of the main text accommodates this case, as the loss can be defined as the occurrence of any of the diseases. There are two basic ways of implementing this model empirically. The first is to use a PGI for the bundle and proceed as in the single-disease case. The second is to formally model the co-occurrence of the multiple diseases. Here, we pursue this second route, by extending the econometric model to multiple diseases.

For ease of exposition, in this section only, we modify our convention of writing vectors in bold: in this section we write matrices rather than vectors in bold.

3.2 The model

We now generalize our model to the case where there are $\mathcal{D} > 1$ diseases. Each agent in now characterized by a tuple

$$(D, G_c, G_f, W).$$

The only difference is that D, G_c and G_f are now column vectors with coordinates indexed by disease $d = 1, \ldots, \mathcal{D}$. The model is fully specified by the joint distribution $\mathbb{P}$ of all variables. We assume that, for each disease d, Assumptions 1-5 from the main text hold.

We begin by defining the key equations of the model in matrix notation. All vectors are column vectors. Main text Equation 1 becomes

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

where

$$G_f = (G_{f,1}, \ldots, G_{f,\mathcal{D}})^T,$$

$$\boldsymbol{\theta} = \begin{bmatrix} - & \theta_{w,1}^\top & - \\ - & \theta_{w,2}^\top & - \\ & \vdots & \\ - & \theta_{w,\mathcal{D}}^\top & - \end{bmatrix},$$

$$W = (W_1, \ldots, W_p)^T,$$

and

$$V = (V_1, \ldots, V_{\mathcal{D}})^T,$$

and where p is the cardinality of the set of covariates. In accordance with our notational convention for this section only, we here bold the matrix $\boldsymbol{\theta}$ but do not bold the vectors. G_f is a $\mathcal{D} \times 1$ vector; $\boldsymbol{\theta}$ is a $\mathcal{D} \times p$ matrix; W is a $p \times 1$ vector of covariates.

Main text Equation 2 becomes

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

where

$$G_c = (G_{c,1}, \ldots, G_{c,\mathcal{D}})^T$$

$$\epsilon = (\epsilon_1, \ldots, \epsilon_{\mathcal{D}})^T.$$

Because Assumptions 1-5 hold for each dimension, there exist latent variables

$\mathcal{L}_d$, $d = 1, \ldots, \mathcal{D}$, such that:

$$\mathcal{L}_d = \beta_{g,d} G_{f,d} + \beta_{w,d} W + \eta_d, \tag{3}$$

$$D_d = \{\mathcal{L}_d > 0\}.$$

We can write

$$\mathcal{L} = \beta_g G_f + \beta_w W + \eta,$$

$$D = \{\mathcal{L} > 0\},$$

where

$$\mathcal{L} = (\mathcal{L}_1, \ldots, \mathcal{L}_{\mathcal{D}})^T,$$

$$D = (D_1, \ldots, D_{\mathcal{D}}),$$

$$\boldsymbol{\beta}_g = \text{diag}(\beta_{g,1}, \ldots, \beta_{g,\mathcal{D}}),$$

$$\beta_{w,d} = (\beta_{w,d,1}, \ldots, \beta_{w,d,p})^T,$$

$$\boldsymbol{\beta}_w = \begin{bmatrix} - & \beta_{w,1}^\top & - \\ - & \beta_{w,2}^\top & - \\ & \vdots & \\ - & \beta_{w,\mathcal{D}}^\top & - \end{bmatrix},$$

$$\eta = (\eta_1, \ldots, \eta_{\mathcal{D}})^T.$$

Note that the set of relevant covariates differs across disease; when a covariate W_k is not used for a disease d, the corresponding coefficient $\beta_{w,d,k}$ is 0.

We need the following natural extension of the normality and independence assumptions from the one-disease case.

Assumption 6. *[Multidimensional assumptions] We have that*

- (η, ϵ, V) *are multivariate normal.*
- $\text{Cov}[\epsilon]$ *is diagonal.*
- η, ϵ, *and* V *are orthogonal from each other.*

We allow the covariance matrices $\text{Cov}[\eta]$ and $\text{Cov}[V]$ to be non-diagonal, so that the genetic shocks V and non-genetic shocks η can have correlations across dis-

eases. Since the PGIs for the different diseases were constructed using summary statistics from GWASs that were conducted in mostly independent datasets, it is reasonable to assume that $\text{Cov}[\epsilon]$ is diagonal.

3.3 Identification Theorem

We now prove that the model is identified. Theorem 1 in the main text implies that we can identify almost all parameters in the model by considering each disease separately. The only parameters for which identification still needs to be established are the off-diagonal terms in $\text{Cov}[\eta]$ and $\text{Cov}[V]$. Estimating $\text{Cov}[V]$ is simple because, by equations (1) and (2),

$$G_c = \boldsymbol{\theta} W + V + \epsilon$$

so that

$$\text{Cov}[V] = \text{Cov}[G_c - \boldsymbol{\theta} W] - \text{Cov}[\epsilon]. \tag{4}$$

To see how $\text{Cov}[\eta]$ is identified, we use the multivariate version of the Bayesian updating lemma:

Supplementary Information Lemma 1. *Conditional on $G_c = g_c$ and $W = w$, G_f is normally distributed with mean*

$$\boldsymbol{A} g_c + \boldsymbol{B}\boldsymbol{\theta} w$$

and variance

$$\boldsymbol{C}\boldsymbol{C}^T.$$

The constants are given by the precision matrices

$$\begin{aligned} \boldsymbol{\Lambda}_\epsilon &:= \text{Cov}[\epsilon]^{-1} \\ \boldsymbol{\Lambda}_V &:= \text{Cov}[V]^{-1} \\ \boldsymbol{\Lambda} &:= \boldsymbol{\Lambda}_\epsilon + \boldsymbol{\Lambda}_V. \end{aligned}$$

as

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{\epsilon} \\ \boldsymbol{B} &= \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_{V} \\ \boldsymbol{C}\boldsymbol{C}^{T} &= \boldsymbol{\Lambda}^{-1}. \end{aligned}$$

Proof. A proof is in Section 1.7.2 of Soch et al. (2024). Our formula corresponds to the particular case where their X is a $\mathcal{D} \times \mathcal{D}$ identity matrix. The covariance matrix ($\boldsymbol{C}\boldsymbol{C}^{T}$) can be decomposed in this form by the Cholesky decomposition. □

Therefore, conditional on $G = g_c$ and $W = w$, G_f is distributed as

$$\boldsymbol{A}g_c + \boldsymbol{B}\boldsymbol{\theta}w + \boldsymbol{C}\nu, \tag{5}$$

where ν is a standard normal $\mathcal{D} \times 1$ vector. Therefore, the conditional distribution of the latent variable $\mathcal{L}$ is

$$\begin{aligned} \mathcal{L} =& \boldsymbol{\beta}_{g} G_f + \boldsymbol{\beta}_{w} w + \eta \\ \mathcal{L} =& \boldsymbol{\beta}_{g}(\boldsymbol{A}g_c + \boldsymbol{B}\boldsymbol{\theta}w + \boldsymbol{C}\nu) + \boldsymbol{\beta}_{\boldsymbol{w}} w + \eta \\ \mathcal{L} =& \boldsymbol{\beta}_{g}\boldsymbol{A}g_c + (\boldsymbol{\beta}_{\boldsymbol{g}}\boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\beta}_{w})w + (\boldsymbol{\beta}_{g}\boldsymbol{C}\nu + \eta). \end{aligned} \tag{6}$$

Therefore, conditional on g_c and w, the covariance matrix of $\mathcal{L}$ is

$$\mathrm{Cov}[\mathcal{L}|G = g_c, W = w] = \boldsymbol{\beta}_{g}\boldsymbol{C}\boldsymbol{C}^{T}\boldsymbol{\beta}_{g} + \mathrm{Cov}[\eta]. \tag{7}$$

With this observation, we can extend the identification theorem to the multiple-disease model. The definition of identification is identical to that for the one-disease case.

Theorem 1. *Under Assumptions 1-6, the multiple diseases model is identified.*

Proof. The argument above shows identification of all parameters except for the off-diagonal terms of $\mathrm{Cov}[\eta]$. Choose w in the support of $\mathbb{P}_W$ and choose g_c and let

$$m := \boldsymbol{\beta}_{g}\boldsymbol{A}g_c + (\boldsymbol{\beta}_{\boldsymbol{g}}\boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\beta}_{w})w.$$

Then, conditional on $G_c = g_c$ and $W = w$, $\mathcal{L}$ is multivariate Gaussian with mean m and covariance given by Equation 7. Therefore, the probability that $D_1 = D_2 = 1$ is the probability that the projection to the first two coordinates of a bivariate Gaussian with mean m and covariance matrix given by Equation 7 is in the first quadrant. This probability is increasing in the $\text{Cov}[\eta]_{12}$.[2] Therefore, $\text{Cov}[\eta]_{12}$ is identified. The same argument also implies identification of all the other off-diagonal terms. □

3.4 Estimation of the econometric model

The estimation of the multiple-disease econometric model proceeds as follows. First, we estimate the single-disease model for each disease. This yields estimates for all parameters except for the off-diagonal terms in $\text{Cov}[V]$ and $\text{Cov}[\eta]$. We estimate $\text{Cov}[V]$ with Equation 4, where we use the sample analogue of $\text{Cov}[G_c - \boldsymbol{\theta} W]$. To estimate $\text{Cov}[\eta]$, we use Equation 7. The joint distribution of the data given $\text{Cov}[\eta]$ and known parameters is given by the multivariate probit model. We fit $\text{Cov}[\eta]$ by maximum likelihood.

3.5 Generating the private risk distributions

To generate the private risk distribution for each of the four scenarios we consider, we generalize the analysis for the single-disease contracts (see Section 4.3 of the main text) to the case of a multiple-disease contract. We consider a contract that pays out in case any of the diseases occurs, and calculate risks accordingly.

We start from a dataset that includes the diseases D, the covariates W, and the vector of current PGIs G_c, and estimate the econometric model. For Scenarios 3L and 3U, we use Equation 5 with the estimated parameters to draw values of G_f according to its conditional expectation given observables; we draw 10 simulated observations per original observation. We then calculate the risk of contracting any of the diseases in the contract based on the covariates and the current PGIs (for Scenarios 1 and 2) or the simulated future PGIs (for Scenarios 3L and 3U).

[2] We can show that this probability is increasing as follows. Formula 26.3.19 of Abramowitz and Stegun (1948) shows that the probability of the first quadrant for a standard bivariate Gaussian is $\frac{1}{4} + \frac{\arcsin \rho}{2\pi}$. This is increasing in the correlation coefficient ρ.

4 Robustness analysis with the Health and Retirement Study

Our main analysis uses the UKB, which is an ideal dataset that combines large sample size, genetic information, and detailed health information from exams, surveys, and the NHS health records. In this section, we reproduce our main results using the Health and Retirement Study (HRS) dataset. The HRS is a longitudinal study of older adults in the United States. The HRS is a well-known dataset that has been used in many studies of insurance and health. The goal is to assess the robustness of our findings in a different setting and with a widely used dataset. While any interested researcher can apply to use the UK Biobank, the process is more involved than for the HRS. Thus, the HRS analysis makes it easier for other researchers to replicate our results.

In our analysis, the main difficulty in using the HRS is the lack of electronic health records information. This makes it more difficult to reliably measure insurance losses that are relevant for critical illness insurance. We searched the HRS for proxy variables for the relevant insurance losses and performed basic quality checks. The most relevant proxies are built from questions about whether a doctor had told the respondent that they have a certain disease. Out of these variables, the one about heart problems was the most reliable in our quality checks. This variable is `RwHEART` in the RAND longitudinal files, where `w` is the wave number. The documentation for the RAND longitudinal file 2020v2 indicates (at p. 431) that `RwHEART` captures the following conditions: "heart attack, coronary heart disease, angina, congestive heart failure, or other heart problems". This definition is broader than CAD, but it is the closest available high-quality proxy in the HRS.

4.1 The HRS CAD contract and data construction

We use the `RwHEART` variable to define the loss for a critical illness contract that we term the "HRS CAD" contract. Because the HRS's `RwHEART` definition is broader than just CAD, we expect it to have a higher probability of occurrence and to be broader than the definition in real critical illness policies.

The genetic predictor we use is the CAD polygenic index from Karlsson Linnér and Koellinger (2022). We estimate its R^2, controlling for gender and age, to be

3.5% in the HRS. We also experimented with using the most up-to-date CAD PGI available in the HRS, `E5_MI_CARDIOGRAM15` (CARDIoGRAM 2015). This is available as part of the HRS's sensitive health information. `E5_MI_CARDIOGRAM15` is, however, a considerably older and less powerful PGI. We estimate its R^2 to be about half that of the newer PGI we use. Nevertheless, using `E5_MI_CARDIOGRAM15` instead of our newer PGI yields similar results for the amount of selection with the future PGI (although less selection with the current PGI, as expected). The noticeable difference in performance between `E5_MI_CARDIOGRAM15` and the newer PGI we use from Karlsson Linnér and Koellinger (2022) illustrates the rapid speed of advances in the quality of available PGIs.

For non-genetic covariates, we use the HRS variables that correspond as closely as possible to our UKB covariates. We use age, gender, BMI, smoking status, alcohol consumption, physical activity, cholesterol, and earnings income. See the replication code for the corresponding HRS variable names. For the SNP and twin heritability assumptions for Scenarios 3L and 3U, we used the same values as in the UKB analysis.

The final estimation sample includes 10,460 individuals and 39,137 individual-wave combinations. We conduct the same analysis as in the main text. After estimating the model, we generate the disease risk distributions under the same informational assumptions as in the main text, using only a subsample of the 7,183 individuals who had also answered questions on risk preferences. We use the risk preference questions for our equilibrium model in Section 5, and we here generate risk distributions for this subsample to facilitate comparison across sections.

Figure 1 displays the risk distribution and implicit taxes for the HRS CAD contract under the different information scenarios. The results are qualitatively similar to the UKB results for CAD.

5 Robustness analysis in a calibrated equilibrium model

Equilibrium in adverse selection models depends on the distributions of both risk and risk preferences. However, the main text considers only the distribution of (disease) risk. This section checks whether the main results are robust to considering risk preferences in a calibrated equilibrium model. We verify the robustness of the main qualitative conclusions: that selection due to genetics would be notice-

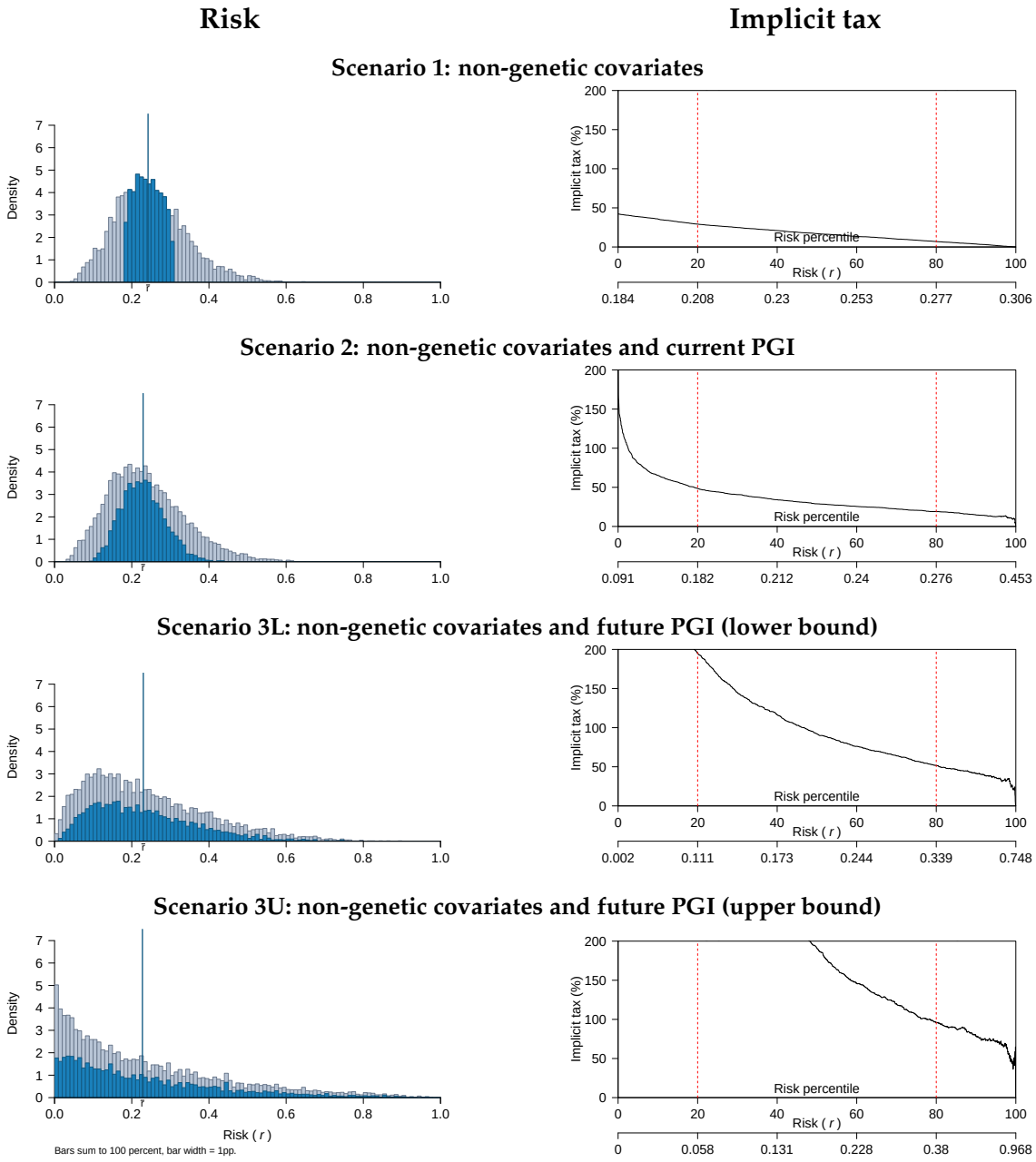

Figure 1: HRS CAD contract: risk and implicit tax

Notes: Left panel: distribution of the risk of a loss based on the HRS CAD variable. Right panel: implicit tax for consumers in the standard risk class as a function of their percentile private risk for each scenario.

able with the current PGI and would be high with future PGIs.

To reduce researcher degrees of freedom, we use a parsimonious calibrated model, similar in spirit to the exercise in Brown and Finkelstein (2008). We report the results for the HRS CAD contract for two reasons. First, the HRS data includes questions eliciting relative risk aversion. Thus, we can use the data to inform the joint distribution of risk and risk preferences, which is known to impact the degree of selection problems (Fang, Keane, and Silverman, 2008). Second, the HRS CAD contract is one of the least adversely selected in our analysis, so this is a more stringent robustness check.

5.1 Model

We consider a binary loss insurance model. We use the standard competitive equilibrium concept from Akerlof (1970) and Einav, Finkelstein, and Cullen (2010). The model uses its own notation, independent of the other sections.

Consumers i in I are heterogeneous in their **probability of loss** π_i and **relative risk aversion** γ_i. The consumer's income equals Y if no loss is incurred, and $(1 - \Delta c)Y$ if a loss is incurred. i is uniformly distributed over the unit interval. The consumer can purchase an insurance contract that pays \$1 if a loss is incurred, and \$0 otherwise. The **price** of the contract is p.

Demand. We assume that the policy payout is considerably smaller than the consumer's income. So the consumer's net gain from purchasing the contract is given by the difference in marginal utility between the two states. That is, the net gain is

$$\pi_i \ \mathrm{MU}_{\mathrm{low},i} \cdot (1 - p) + (1 - \pi_i) \ \mathrm{MU}_{\mathrm{high},i} \cdot (-p), \tag{8}$$

where $\mathrm{MU}_{\mathrm{low},i} := [(1 - \Delta c)Y]^{-\gamma_i}$ and $\mathrm{MU}_{\mathrm{high},i} := Y^{-\gamma_i}$ denote the marginal utilities of wealth in the low and high states, respectively. The consumer's **willingness to pay** for the contract, u_i, is defined as the price p that sets this net gain to zero.

Supply. The **cost** to a firm of selling the contract is the average payout plus a fixed cost: $c_i := \pi_i + F$. We include the fixed cost as in Meza and Webb (2001). The fixed cost is both realistic and a parsimonious way to calibrate the model with less than 100% of consumers buying insurance in the case without selection. We could have instead assumed other frictions such as monopoly power or moral hazard.

Equilibrium. Given the joint distribution of the willingness to pay u_i and cost c_i, we use the standard competitive selection model of Akerlof (1970) and Einav, Finkelstein, and Cullen (2010). Specifically, Einav, Finkelstein, and Cullen (2010) define the **demand curve** as

$$D(p) := \int_{u_i > p} 1 \, di.$$

That is, the demand at a price p is the share of consumers who are willing to pay at least p.

Einav, Finkelstein, and Cullen (2010) define the **average cost curve** as

$$AC(p) := \int_{u_i > p} c_i \, di / D(p).$$

The AC curve is the inverse supply curve. $AC(p)$ is the average cost of selling insurance when the quantity of consumers covered is $q = D(p)$. The defining feature of selection models is that the AC curve depends on quantity, because due to selection the probability of loss depends on which consumers are buying insurance. Einav, Finkelstein, and Cullen (2010) define an **equilibrium** as a price and quantity pair (p^*, q^*) in the intersection of the demand and supply curves.

5.2 Willingness to pay

We can gain some intuition for the demand curve by solving for the willingness to pay. Setting Equation 8 to zero and substituting u_i for p yields

$$\frac{u_i}{1-u_i} = \frac{\pi_i}{1-\pi_i} \cdot \frac{\text{MU}_{\text{low},i}}{\text{MU}_{\text{high},i}}.$$

Under the CRRA assumption,

$$\frac{u_i}{1-u_i} = \frac{\pi_i}{1-\pi_i} \cdot (1-\Delta c)^{-\gamma_i}.$$

Or, solving for u_i,

$$u_i = \underbrace{\pi_i}_{\text{probability of loss}} \cdot \underbrace{\frac{(1-\Delta c)^{-\gamma_i}}{(1-\pi_i)+\pi_i(1-\Delta c)^{-\gamma_i}}}_{\text{risk premium}}.$$

That is, willingness to pay equals the probability of loss times the risk premium, with the latter being equal to the ratio of marginal utility in the loss state to average marginal utility. This clarifies that demand and equilibrium depend on the joint distribution of risk (π_i) and risk preferences (γ_i). Our main analysis in the main text measures selection solely based on the distribution of risk, without considering risk preferences. We now proceed to calibrate the model, using the HRS data to inform the joint distribution of risk and risk preferences.

5.2.1 Calibration

We calibrate the model to the HRS CAD contract from Section 4 in this Online Appendix. We focus on consumers in the standard risk class. The probability of loss π_i is the probability of having a heart problem as defined by the CAD HRS variable. We consider the same four information scenarios as in the main text for predicting π_i: the non-genetic covariates only (Scenario 1), the non-genetic covariates and the current PGI (Scenario 2), and the non-genetic covariates and the future PGI (lower and upper bounds; Scenarios 3L and 3U). All results are reported for the standard risk class, so this is a set of consumers that firms cannot price discriminate against.

To measure γ_i, we follow Kimball, Sahm, and Shapiro (2008). They developed methods to impute a relative risk aversion coefficient from the HRS questions. The estimated coefficient is based on questions about how respondents would choose between a job with certain earnings and a job with uncertain earnings. Their method yields, for each subject, an estimated coefficient of relative risk aversion γ_i. We note that the Kimball, Sahm, and Shapiro (2008) risk aversion coefficient estimates are relatively high, with the bulk of the distribution between $\gamma = 6$ and $\gamma = 10$. Naturally, the level of demand depends on both γ and Δc. Thus, we can still accommodate a realistic overall level of demand, as long as we calibrate the model with a modest value of Δc.

Thus, we have a joint distribution of π_i and γ_i. This leaves two parameters to be

calibrated: the fixed cost F and the loss in income in the low state, Δc. We calibrate them to match the loss ratio and market size in the critical illness insurance market. We base the numbers on the UK, which is one of the countries with the most developed critical illness insurance markets. Swiss Re Institute (2022) reports the number of policies sold in the UK in 2022. There were 484,110 critical illness policies sold as additional benefits to term life insurance policies. These correspond to the bulk of critical illness insurance sold, with an additional 94,426 standalone critical illness policies. The total number of term life insurance policies sold was 1,698,301.

Given these figures, we perform the following back-of-the-envelope calculation. First, we assume that the potential market comprises the 1,698,301 consumers who purchase a term life insurance policy. This follows from the details of how these products are marketed, and from the fact they are overwhelmingly sold as additional benefits to term life insurance buyers. If we consider only the 484,110 add-on critical illness policies sold, this implies a demand of 28.5% of the market. If we consider also the 94,426 standalone critical illness policies, this implies a demand of 34%. Thus, we set our calibration target to 30% of the market. Loss ratios in the industry are widely known to be high (the loss ratio equals claims paid divided by premiums). For critical illness, loss ratios in the ballpark of 50% are often cited, indicating a substantial amount of frictions (Reuters, 2020). We set 50% as our calibration target for the loss ratio.

The calibration matches the target loss ratio and market size exactly. The calibrated parameter values are a consumption loss Δc of 12.5% and fixed cost F of \$ 0.26 per dollar of coverage.[3] Results are reported in the main text.

6 Acknowledgments

Acknowledgments are available at: https://osf.io/9ndw6/files/osfstorage.

[3]Note that the calibrated fixed costs are relatively high, and thus the first-best welfare gain from insurance and the deadweight losses from selection are relatively small. This is because this simple model has relatively little heterogeneity in preferences, and thus the only way to match the relatively low equilibrium quantity at baseline is to impose a high fixed cost. The relatively flat demand curve then makes first-best surplus from insurance relatively small.

References

Abramowitz, Milton and Irene A Stegun (1948). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 1972nd ed. US Government printing office.

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

Altshuler, David M. et al. (2010). "Integrating Common and Rare Genetic Variation in Diverse Human Populations". In: *Nature* 467, pp. 52–58.

Anderson, Michael and Elias Mossialos (2022). "Are We Heading for a Two Tier Healthcare System in the UK?" In: *BMJ*, o618.

Bastarache, Lisa (2021). "Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS". In: *Annual Review of Biomedical Data Science*, pp. 1–19.

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

Brown, Jeffrey R and Amy Finkelstein (2008). "The Interaction of Public and Private Insurance: Medicaid and the Long-Term Care Insurance Market". In: *American Economic Review* 98.3, pp. 1083–1102.

Bycroft, Clare et al. (2018). "The UK Biobank Resource with Deep Phenotyping and Genomic Data". In: *Nature* 562.7726, pp. 203–209.

Chang, Christopher C et al. (2015). "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets". In: *GigaScience* 4, p. 7.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F. O'Reilly (2020). "Tutorial: A Guide to Performing Polygenic Risk Score Analyses". In: *Nature Protocols* 15, pp. 2759–2772.

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.

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.

Fang, Hanming, Michael P Keane, and Dan Silverman (2008). "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market". In: *Journal of political Economy* 116.2, pp. 303–350.

Fernandez-Rozadilla, Ceres et al. (2023). "Deciphering Colorectal Cancer Genetics Through Multi-omic Analysis of 100,204 Cases and 154,587 Controls of European and east Asian ancestries". In: *Nature Genetics* 55, pp. 89–99.

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.

Grotzinger, Andrew D., Javier de la Fuente, Florian Privé, Michel G. Nivard, and Elliot M. Tucker-Drob (2022). "Pervasive Downward Bias in Estimates of Liability-Scale Heritability in Genome-Wide Association Study Meta-Analysis: A Simple Solution". In: *Biological Psychiatry* 93, pp. 29–36.

Karlsson Linnér, Richard, Pietro Biroli, et al. (2019). "Genome-Wide Association Analyses of Risk Tolerance and Risky Behaviors in Over 1 Million Individuals Identify Hundreds of Loci and Shared Genetic Influences". In: *Nature Genetics* 51, pp. 245–257.

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

Kelly, Elaine and George Stoye (2020). "The Impacts of Private Hospital Entry on the Public Market for Elective Care in England". In: *Journal of Health Economics* 73, p. 102353.

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.

Kurki, Mitja I. et al. (2023). "FinnGen Provides Genetic Insights from a Well-Phenotyped Isolated Population". In: *Nature* 613, pp. 508–518.

Lambert, Jean-Charles et al. (2013). "Meta-analysis of 74,046 Individuals Identifies 11 New Susceptibility Loci for Alzheimer's disease". In: *Nature Genetics* 45, pp. 1452–1458.

Mahajan, Anubha et al. (2018). "Fine-mapping Type 2 Diabetes Loci to Single-variant Resolution Using High-density Imputation and Islet-specific Epigenome Maps". In: *Nature Genetics*, pp. 1505–1513.

Martin, Alicia R et al. (2017). "Human Demographic History Impacts Genetic Risk Prediction Across Diverse Populations". In: *The American Journal of Human Genetics* 100.4, pp. 635–649.

McCarthy, Shane et al. (2016). "A Reference Panel of 64,976 Haplotypes for Genotype Imputation". In: *Nature Genetics* 48.

Meza, David de and David C. Webb (2001). "Advantageous Selection in Insurance Markets". In: *RAND Journal of Economics* 32.2, pp. 249–262.

Nikpay, Majid et al. (2015). "A Comprehensive 1,000 Genomes-Based Genome-Wide Association Meta-Analysis of Coronary Artery Disease." In: *Nature genetics* 47, pp. 1121–1130.

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.

Reuters (2020). *Are Critical Illness Insurance Policies Right for You?* Accessed: 2025-05-13. Reuters. URL: https://www.reuters.com/article/world/are-critical-illness-insurance-policies-right-for-you-idUSKBN27K1F5/.

Schumacher, Fredrick R et al. (2018). "Association Analyses of More Than 140,000 Men Identify 63 New Prostate Cancer Susceptibility Loci". In: *Nature Genetics* 50, pp. 928–936.

Soch, Joram et al. (2024). *StatProofBook/StatProofBook.github.io: The Book of Statistical Proofs*. Version 2023. DOI: 10.5281/ZENODO.4305949. URL: https://statproofbook.github.io/.

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 (2022). *Term and Health Watch 2022*. Tech. rep. Analysis of Individual Life and Health Protection Sales in the UK, Including Critical Illness Insurance. Swiss Re Institute.

Trubetskoy, Vassily et al. (2022). "Mapping Genomic Loci Implicates Genes and Synaptic Biology in Schizophrenia". In: *Nature*, pp. 502–508.

Winkler, Thomas W et al. (2014). "Quality Control and Conduct of Genome-Wide Association Meta-Analyses". In: *Nature Protocols* 9, pp. 1192–212.

Wray, Naomi R et al. (2013). "Pitfalls of Predicting Complex Traits from SNPs". In: *Nature Reviews Genetics* 14.7, pp. 507–515.

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

Zhang, Yan Dora et al. (2020). "Assessment of Polygenic Architecture and Risk Prediction Based on Common Variants Across Fourteen Cancers". In: *Nature Communications* 11, p. 3353.