Genetic prediction and adverse selection

ONLINE APPENDIX

Eduardo Azevedo[*] Jonathan Beauchamp[†] Richard Karlsson Linnér[‡]

September 27, 2024

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

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

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

A Background on genetics, PGIs, and GWASs

A.1 DNA and single-nucleiotide polymorphisms (SNPs)

This study investigates common **complex** (or **non-Mendelian**) **diseases**—i.e., common diseases that are influenced by many genetic variants (and by the environment) (Schork, 1997). Research has established that complex diseases are typically highly polygenic—i.e., a very large number of genetic variants are involved, most or all of which have only a tiny individual effect (Visscher et al., 2021). Thus, the risk of a complex disease cannot be accurately predicted from only one or a few genetic variants. Even though each variant typically has a tiny effect, the overall genetic effect across all the variants can be considerable. The heritability—defined as the share of the variation in disease liability that is attributable to genetic factors—is often substantial, reaching ∼50% on average across complex diseases (Polderman et al., 2015).

The human genome consists of two roughly 3-billion-long sequences of base pairs encoded in DNA (NHGRI, 2023). One of the two sequences is inherited from the mother and the other is inherited from the father. Each sequence consists of four possible nucleotide bases: adenine ("A"), cytosine ("C"), guanine ("G"), and thymine ("T"). Two complementary strands—the "forward" and the "reverse" strand—make up the DNA molecule. They are connected by bonding base pairs like the rungs on a ladder; due to a property called complementarity, A always bonds with T, C with G, and vice versa. This property means that the information content of both strands can be recorded perfectly from a single strand.

The maternal and paternal sequences are each packaged into 23 chromosomes, thus yielding 23 *pairs* of chromosomes in the offspring. One of these pairs contains the sex chromosomes (Chromosome 23). We exclude the sex chromosomes from our analysis because most large genetic association studies omit these (Sun et al., 2023). This decision will not impact our results much because the sex chromosomes generally contribute little to the heritability of complex traits and disease (<5%) (Visscher et al., 2006). Each genomic position on the (non-sex) chromosome pairs consists of two base pairs, one that gets passed on by the mother (e.g., A on the forward strand and T on the reverse strand) and one by the father (e.g., G and C on the forward and reverse strand, respectively).

The nucleotides at most positions in the genome are identical across all (or nearly all) humans (NHGRI, 2023). The small fraction of nucleotides that do vary are called single-nucleotide polymorphisms (SNPs), which can be viewed as the smallest unit of genetic variation.[1] At most SNPs, only two possible nucleotide bases have ever been observed among

[1] Other types of genetic variants exist, including copy number variants (CNVs), insertions, and deletions. Genes are protein-coding sequences of nucleotides. Here, we only focus on SNPs, because they are correlated

humans on the forward (or reverse) strand. Such SNPs are called “bi-allelic”. SNPs with more than two possible nucleotide bases—called “multi-allelic”—do exist but are rare and are excluded from our analysis. To code a variable for a given SNP, geneticists arbitrarily select one of the two possible nucleotides, label it as the reference nucleotide, and then sum the number of non-reference (or “effect-coded”) nucleotides. The resulting SNP variable thus takes a value of either 0, 1, or 2. For more details, see Beauchamp et al. (2011), Benjamin et al. (2012), or Uffelmann et al. (2021).

A.2 Modeling the genetics of polygenic diseases

As indicated in the main text, we follow standard practice in epidemiological genetics and model a disease D with a liability threshold model, according to which one contracts the disease if one's liability $\mathcal{L}$ is positive. We also assume the additive genetic model (Falconer and Mackay, 1996):

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

where k is a constant; $\mathbf{X}$ is a row vector that contains the measured variants, and $\mathbf{b}$ is a column vector that contains their true effect sizes on D; $G^* = \mathbf{X}\,\mathbf{b}$ is the individual's true additive genetic factor for D; ψ is the disturbance term; and the operator $\{..\}$ is equal to 1 if its argument is true and to 0 otherwise. Thus, $\mathcal{L}$ is the weighted sum of an individual's genetic variants, where the weights are the variants' effect sizes on $\mathcal{L}$, plus a disturbance term that captures all other influences on D. It is commonly assumed that ψ follows a normal (or logistic) distribution, in which case the liability threshold model becomes a probit (or logistic) regression of the disease on the liability.

A.3 Polygenic indexes (PGIs)

As explained in the main text, in practice neither $\mathbf{b}$ nor G^* are observed. Instead, one can construct estimates $\hat{\mathbf{b}}$ using estimates from a previously published genome-wide association study (GWAS) of the disease and use them to construct a polygenic index (PGI) G.[2] A GWAS is a large-scale genetic study in which a trait of interest is regressed on each of millions of genetic variants, separately. When the trait of interest is a dichotomous disease, probit or logit regressions are used. This relatively simple approach has proven successful

with much of the variation due to these other types of variants (Sudmant et al., 2015) and capture a large fraction of the heritability of complex traits (Yengo et al., 2022).

[2] PGIs are also commonly called “polygenic risk scores” (“PRSs”).

in discovering replicable SNP associations. Large-scale GWASs are now discovering novel, replicable SNP associations at an unprecedented rate (Tam et al., 2019). To achieve sufficient sample size, the largest published GWASs are usually meta-analyses of results derived in multiple biobanks. The next section discusses the GWASs from which we obtained the regression estimates needed to construct the estimates $\hat{\mathbf{b}}$

For each individual in the data that passes our sample quality-control procedure (described below), the PGI G for a given disease is the weighted sum of the SNP variables, with each SNP variable weighted by its corresponding effect size estimate for the disease:

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

Observe that

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

where $\boldsymbol{\epsilon}$ is the measurement error in $\hat{\mathbf{b}}$ and $\epsilon = \mathbf{X}\,\boldsymbol{\epsilon}$ is the error in the PGI G. As mentioned, complex diseases are influenced by a large number of variants with tiny effects. Thus, by a simple application of the central limit theorem, both $G^{*} = \mathbf{X}\,\mathbf{b}$ and $\epsilon = \mathbf{X}\,\boldsymbol{\epsilon}$ are approximately normally distributed. Further, because $\boldsymbol{\epsilon}$ (the error in $\hat{\mathbf{b}}$) arises because of sampling variation in the GWAS, and because the GWAS is conducted in a sample that is independent from the analysis sample, $\mathbf{X} \perp \boldsymbol{\epsilon}$ and $G^{*} \perp \epsilon$.

A given SNP's GWAS estimate comes from a regression of the trait or disease on the SNP (and some baseline controls); that regression usually does not control for the SNPs that are located near the focal SNP in the genome and that are typically correlated with it. Thus, the focal SNP's GWAS estimate captures both that SNP's effect as well as those of nearby, correlated SNPs (such SNPs are said to be in "linkage disequilibrium" with the focal SNP). We computed the PGIs using the software PRS-CS (Ge et al., 2019; Choi et al., 2020), which models patterns of correlation across SNPs and use these to transform GWAS regression estimates into estimates of $\mathbf{b}$ (i.e., into $\hat{\mathbf{b}}$). PRS-CS also aims to improve the PGIs' signal-to-noise ratio by applying a continuous shrinkage prior in penalized regression to downweight the regression estimates of correlated and/or weakly associated SNPs.

PRS-CS has been shown to perform well for several of our diseases of interest, and about as well as alternative methods like LDpred2 (Privé et al., 2020). By default, PRS-CS restricts the set of SNPs to those covered by the reference map of the International HapMap 3 Consortium[3] (Altshuler et al., 2010), and then removes any SNPs with MAF $< 1\%$ in

[3] The HapMap 3 SNPs (Altshuler et al., 2010) are a standard set of SNPs used in the PGI literature (Becker et al., 2021), and have been shown to capture the bulk of the heritability attributable to common SNPs (MAF $<1\%$) in populations of European ancestry (Ge et al., 2019; Yengo et al., 2022).

the 1000 Genomes reference panel. This set of SNPs was used to compute all our PGIs except that for Alzheimer's disease. As we further discuss in Online Appendix A.5, to tag the well-known *APOE* risk types, our PGI for Alzheimer's had to be augmented by a single additional SNP—rs429358—that is not covered by the otherwise comprehensive HapMap3 set. We ran PRS-CS using its default parameters.

After PRS-CS, we used the PLINK2 software (Chang et al., 2015) to compute the PGIs in the UK Biobank data, while using the PRC-CS adjusted GWAS coefficients as the weights ($\hat{\mathbf{b}}$). Our protocol for computing PGIs is similar to those of recent studies (e.g., Becker et al. (2021) or Ge et al. (2019)) and it follows recent scientific guidelines (see, e.g., Figs 1–2. in Choi et al., 2020).

It is important to note that the GWAS regression estimates and the transformed estimates $\hat{\mathbf{b}}$ are measures of the SNPs' *associations* with the trait, rather than of their effects on the trait. The PGIs we construct do not therefore have a clear causal interpretation. Nonetheless, they are useful predictive tools, which is exactly what is needed for this study. For simplicity, we sometimes refer to $\hat{\mathbf{b}}$ as effects, even though there is no clear causal interpretation.

A.4 GWAS, quality control (QC), and meta-analysis

This section describes how we searched and gathered GWAS "summary statistics", which include the GWAS regression coefficients we use to compute the $\hat{\mathbf{b}}$. Importantly, to avoid overfitting, these summary statistics and GWAS regression coefficients must be estimated in datasets that are independent of the UK Biobank (Wray et al., 2013).

In summary, we computed PGIs using large-sample GWAS summary statistics for the seven diseases of interest. To quantify the size of a GWAS, it is useful to use 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). The effective sample sizes of the GWASs from which we obtained the summary statistics for the seven diseases are shown in Table A.1.

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

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

Disease	**Abbreviation**	$Neff$	**Primary reference**
Alzheimer's disease	ALZ	126,275	Lambert et al. (2013)
Breast cancer (F only)	BRC	236,094	Zhang et al. (2020)
Coronary artery disease	CAD	117,486	Fernandez-Rozadilla et al. (2023)
Colorectal cancer	CRC	162,973	Nikpay et al. (2015)
Prostate cancer (M only)	PRC	137,933	Schumacher et al. (2018)
Schizophrenia	SCZ	146,348	Trubetskoy et al. (2022)
Type 2 diabetes	T2D	193,440	Mahajan et al. (2018)

Notes: Additional details are reported in Supplementary Table 1. Some of the listed GWAS were meta-analyzed with publicly available results derived in the FinnGen Biobank (Kurki et al., 2023). Additional study references and acknowledgments are listed in the Acknowledgements (Online Appendix F).

analysis that excludes the UK Biobank estimates from the meta-analysis. We aimed to find such a hold-out version whenever possible. The latest search was done on January 19, 2024.

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

We applied GWAS quality control (QC) to each downloaded set of summary statistics, as is standard practice (Winkler et al., 2014). (In Online Appendix A.6, we also describe how we applied QC to the individual-level genetic data in the UKB.) Our GWAS QC-protocol was based on the often-cited protocol developed by the Social Science Genetic Association Consortium (SSGAC) (Karlsson Linnér et al., 2019; Okbay et al., 2022), which is a continuation of the older "industry-standard" protocol of the GIANT consortium (Winkler et al., 2014). The GWAS QC-protocol serves two main purposes: (1) to remove SNPs that for technical reasons are likely to worsen the signal-to-noise ratio of the PGI (e.g., rare SNPs); and (2) to ensure that SNP coordinates, reference nucleotides, and other per-SNP statistics get aligned across files (e.g., to align them all with the forward strand in the genome reference data).

The protocol was applied using the EasyQC software package (Winkler et al., 2014) and it removed (i) multi-allelic or non-SNP variants (e.g., indels), (ii) SNPs on the sex chromosomes, (iii) SNPs with MAF <0.5% (the PRS-CS software additionally removes SNPs with MAF

<1% in the reference data), (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. Next, we visually inspected a series of diagnostic plots that were produced by EasyQC, and found no conspicuous issues. Because the imputation quality metric was missing from most of the public summary statistics files, we did not filter on this metric. However, we note that most of the downloaded files were already filtered on imputation quality by the original studies, and that the vast majority of HapMap 3 SNPs are well imputed.

A.5 Alzheimer's disease and the *APOE* region

The *APOE* region on chromosome 19 (coordinates 45,384,477–45,432,606, human reference genome build 37/hg 19) has a well-established large effect on the risk of Alzheimer's disease (Scheltens et al., 2021; Lambert et al., 2013). There are six primary risk types of the gene product apolipoprotein E (*ApoE*), called $\varepsilon 2/\varepsilon 2$ through to $\varepsilon 4/\varepsilon 4$. These risk types can be assayed by measuring only two "tag-SNPs" (rs7412 and rs429358) (Faul et al., 2021). The greatest disease risk is conferred by the C nucleotide of rs429358, which has a considerably larger effect than rs7412. Genetic testing for *APOE* status has been available for years and is commonly included in consumer genetic health reports (Ryan et al., 2021). We model *APOE* status as a component of the PGI (observed by the consumer) rather than as a covariate observable to insurers because insurers typically cannot use that information in underwriting.[4]

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

[4] Genetic testing to determine *APOE* status is normally considered a predictive genetic test, so genetic information bans typically forbid insurers from observing this test result (Nabholz and Rechfeld, 2017; Dixon et al., 2024), even when that information is included in medical records that would otherwise be permissible for insurers to observe during risk classification.

GWAS coefficients, followed by merging this two-SNP PGI with the remaining non-*APOE* genome-wide PGI that fully excludes the *APOE* region, to make our final "augmented" PGI for Alzheimer's disease.

We benchmarked the augmented PGI against the standard approach of modeling the *APOE* risk types using dummy variables. Online Appendix Figure A.1 shows that the standard dummy-variables approach successfully predicts the risk of disease according to expectation. For our benchmark, we ran a series of probit regressions to evaluate the incremental McKelvey & Zavoina pseudo-R^2 from adding each of the following to a specification with only the non-genetic covariates $\boldsymbol{W}$: (i) the dummies, (ii) the dummies and the non-*APOE* genome-wide PGI, or (iii) the augmented PGI. A saturated model that included both the dummies and the augmented PGI was also evaluated. The benchmark showed that the best incremental R^2 (4.9%) was achieved by the augmented PGI (iii). The saturated model performed marginally worse. Thus, the augmented PGI was selected for use.

A.6 Genotype data, ancestry, and sample quality control (QC)

The analysis reported in the paper was conducted with the harmonized genetic data resource collected, maintained, and distributed by the UK Biobank (Data Category 100319). This resource is described in detail by Bycroft et al. (2018). Imputed genetic data are 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 rare. There are about 10.3 million imputed variants with MAF $> 1\%$ in the UK Biobank.

As indicated in the main text, we followed standard practice in applied genetic research and restricted our analysis to individuals of European ancestry. There is a lack of well-powered GWAS in non-European populations (Ruan et al., 2022), and PGIs constructed using GWAS estimates obtained in samples of European ancestry don't perform well in samples of different ancestries (Martin et al., 2017). In addition to this, to control for population stratification, we include the first 10 principal components (PCs) of the genetic relatedness matrix as control variables (Price et al., 2006; Choi et al., 2020) when estimating our econometric model.[5] We used the standard set of genetic PCs that are distributed with the UKB genetic data release (Data Category 22009).

To restrict the UK Biobank sample to participants of European ancestry, we relied on the approach described in Karlsson Linnér et al. (2019). Specifically, participants were

[5] 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 genetic *prediction* rather than *causation*, population stratification is not a major concern. We nonetheless follow standard practice in applied genetic research and control for the top 10 PCs.

considered to be of European ancestry if they (i) self-reported their ethnic background to be "White", "White British", "White Irish", or "Any other white background" (Data Field 21000), and (ii) have a value on the first genetic PC that is ≤ 0. This approach identifies 447,863 participants that form a tight cluster on the genetic PCs. That cluster is distributed distinctly from the other major population genetic groups, pointing to shared European genetic ancestry.

Next, we applied the following sample-level quality control (QC) filters, using the pre-computed quality metrics distributed by the UKB with the imputed genetic data resource (Data Category 100313). Participants were removed when (1) there was a mismatch between self-reported and genetic sex (Data Fields 31, 22001), (2) there was evidence of sex-chromosome aneuploidy (Data Field 22019), (3) they were classified as outliers on either genotype heterozygosity or missingness rates based on directly genotyped SNPs (Data Field 22027).

A.7 Inspection of PGI distributions

Online Appendix Figure A.2 plots the distribution of the PGIs for the seven diseases. We inspected the distributions for (1) departure from normality and (2) sex differences. All PGIs were found to be approximately normally distributed, except for the PGI for Alzheimer's disease. Because of the large risk conferred by the two tag-SNPs for *APOE*, the distribution of this PGI has two noticeable bumps to the right. We could not identify any meaningful differences in the PGI distributions across sexes. Nevertheless, we standardized the PGIs so that the have mean zero and unit variance separately by sex.

B Definition of the implicit tax

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

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

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

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

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

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

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

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

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

C Health data and risk factors

The UKB has collected self-reported information on a large number of health, psychological, and socioeconomic variables via a touchscreen questionnaire administered at the baseline assessment. This resource is described in detail by Sudlow et al. (2015). The baseline visit occurred in 2006–2010. The touchscreen procedure was followed by a structured follow-up interview by a trained nurse to verify and map the self-reported medical conditions and other health outcome to standardized diagnostic categories, and to run various physiological measurements and lab assays. The most recent version of the study data were refreshed and downloaded from the UK Biobank Access Management System on June 19, 2023.

Since the baseline assessment took place, the UKB has been enriched through linkage with several electronic health record data sources. Most of the healthcare in the UK is delivered by the publicly-funded National Health Service (NHS) (Kelly and Stoye, 2020), with which 98% of the population is registered and tracked (Sudlow et al., 2015). 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 the NHS (or its subsidiaries). However, the UK has no centralized system specifically for the primary care records from general practices, meaning that the primary care records has worse coverage than the other linked sources described below. For living participants, new information is still being recorded in their electronic health records. The main electronic health record sources linked to the UKB are:

i. Primary care records (Category 3000)
ii. Hospital inpatient records (Category 2000)
iii. National cancer registries (Category 100092)
iv. National death registries (Category 100093).

Some of these data are still in the process of being linked. Unfortunately, because of the UK's decentralized system for primary care records, primary care records are not yet linked for all participants; only about 45% of the participants have been linked thus far. Nevertheless, because the disease outcomes we study are fairly serious, and because the linkage of the cancer and death registries are largely complete, we consider it unlikely that we are missing a substantial number of disease cases because of missing primary care records. However, the UKB participants have been found to be healthier than the general population (Schoeler et al., 2023), so disease cases may be fewer in our data than in the general population.

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

C.1 Coding the disease outcomes and comorbidities

This section details how we coded the disease outcomes and cormorbidities. Our starting point for coding the disease outcomes is the "First occurrence of health outcomes" resource (Category 1712), which is generated and distributed by the UK Biobank. This resource is the result of a harmonization and merger of the self-reported medical data with the linked electronic health record 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 electronic health record sources (i–iv).

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

The mapping of ICD10 codes to Phecodes for our seven diseases of interest is reported in Supplementary Table 2. With the exception of coronary artery disease, our diseases of interest could all be captured by a single Phecode. To code coronary artery disease, which is defined as a collection of many underlying heart and circulatory conditions, we relied on the approach of the previous literature and considered the entire ICD10 Chapter I20–I25 "Ischaemic heart diseases" (Aragam et al., 2022), which maps to a total of six different

Phecodes.

The next section describes the selection of control variables, some of which are comorbidities, e.g., asthma or bipolar disorder. The comorbidities were coded analogously using the Phecode system.

C.2 Coding the epidemiological risk factors and comorbidities

An important contribution of this study is to control for observable information that may correlate with the PGIs. We carefully reviewed the epidemiological literature and clinical guidelines of public health organizations (e.g., the National Cancer Institute or the Centre for Disease Control) to identify the main observable risk factors that are used in clinical practice for screening or diagnostic purposes. This information is typically accessible to insurance companies during medical underwriting. Supplementary Table 3 reports the disease-specific risk factors that were identified. Our resulting list of epidemiological risk factors was verified independently by two medical doctors.

The UKB data are rich enough for us to code almost all the risk factors and comorbidities identified by the review. A handful of risk factors were eventually omitted or proxied (see Panel B in Supplementary Table 3), either because they were unspecific (e.g., diet) or because of limited data availability (e.g., substance use). Instead, the effects of diet were proxied by relevant anthropometric measures or biomarkers, such as BMI, hypertension, systolic blood pressure, and hypercholesterolemia (which are also observable risk factors in and of themselves). Similarly, general psychopathology was proxied by bipolar disorder, depression, and a neuroticism score. Only three risk factors were omitted completely: brain injury, cannabis use, and substance use. Notably, we were able to code family history (father, mother, or siblings) for six of the seven diseases. The exception is schizophrenia, for which family history data is not available and for which we instead used family history of depression as a proxy.

The data fields used to code each risk factor are reported in Supplementary Table 4. We defined a set of nine baseline covariates that were included in all the analyses: age (at the most recent observation), sex (omitted from any sex-specific analysis), Townsend's deprivation index, education (years of schooling), BMI, alcohol consumption (drinks per week), smoking (never vs. former vs. current smokers), a dummy variable indicating physical inactivity, and systolic blood pressure. These nine baseline covariates 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).

In addition to these baseline covariates, we coded risk factors and comorbidities specific to each of the seven disease. 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 the next section C.3). 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. Supplementary Table 4 lists the baseline and disease-specific covariates and Supplementary Table 5 reports sample descriptive statistics.

C.3 Adjustment of age-dependent covariate values

An objective of this study is to predict the risk of disease by a particular age (e.g., by age $a = 65$). The data is informative of the age of onset for the medical conditions and comorbidities we study, but for most participants, we only observe their covariate values once (at the baseline assessment in 2006–2010). Therefore, after estimating our econometric model, when predicting disease risk by a particular age, the values of the age-dependent covariates were adjusted for age before being inputted in the prediction model. 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. Supplementary Table 4 lists which covariates were age-adjusted.

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

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 the age polynomial to model the covariate as a function of age a: $W(a) = a_1 \times a + a_2 \times a^2 + a_3 \times a^3 + a_4 \times a^4 + e$. 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, for each covariate, 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, with the exception of "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 did not adjust the covariate values of participants with age $a_i < 65$ who had already experienced the event before age 65, nor of participants of $a_i > 65$ and 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 female participants older than age 65 who had not yet experienced the event were recoded as having experienced the event at age 65. The reason is that almost all women have either reached menopause (or had a hysterectomy) before this age.

We found that none of the age-dependent covariates in this study decreases as a function of age. Instead, in most cases, the age polynomial predicted a monotonic increase in the covariate. However, for a few covariates, the polynomial predicted slight non-monotonicity at the top and bottom of the age distribution, where there are much fewer observations (see, e.g., when age exceeds 80 in the "Breast cancer screening" or "Prostate cancer screening" panels in Online Appendix Figure A.4). Therefore, we forced monotonicity over the entire age range by fitting a cubic smoothing spline as implemented in the R package "mgcv". The smoothing had little to no effect on most age groups in the data (see the figure).

Because all covariates were observed to increases as a function of age, the age-adjustment procedure effectively increased the values of the age-dependent covariates for participants below 65, and decreased the values for those older than 65.

D Generalized econometric model for multiple-diseases contracts

D.1 Models for Multiple Disease Contracts

We now discuss how we model multiple disease CII contracts. The theoretical model in Section 3.1 of the main text accommodates this case, as the loss can be defined a 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 to write vectors in bold. We follow that convention everywhere else, but here we write matrices rather than vectors in bold.

D.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 dimension 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

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

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

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

and where p is the cardinality of the set of covariates. In accordance with our notational convention for this section only, we here bold matrices rather than vectors. G_f is a $\mathcal{D} \times 1$ vector; θ is a $\mathcal{D} \times p$ matrix; W is a $\mathcal{D} \times 1$ vector of covariates.

Main text Equation 2 becomes

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

where

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

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

Because Assumptions 1-5 hold for each dimension, there exist latent variables $\boldsymbol{\mathcal{L}}_d$, $d = 1, \dots, \mathcal{D}$, such that:

$$\boldsymbol{\mathcal{L}}_d = \beta_{g,d} G_{f,d} + \beta_{w,d} W + \eta_d, \tag{4}$$

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

We can write

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

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

where

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

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

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

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

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

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

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 $W_{w,d,k}$ is 0.

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

Assumption 1 (Multidimensional Assumptions)**.** *We have that*

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

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

D.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 $\mathrm{Cov}[\eta]$ and $\mathrm{Cov}[V]$. Estimating $\mathrm{Cov}[V]$ is simple because, by equations (2) and (3),

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

so that

$$\mathrm{Cov}[V] = \mathrm{Cov}[G_c - \boldsymbol{\theta} W] - \mathrm{Cov}[\epsilon]. \tag{5}$$

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

Online Appendix 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

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

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 here 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

$$\boldsymbol{A}g_c + \boldsymbol{B}\boldsymbol{\theta}w + \boldsymbol{C}\nu, \tag{6}$$

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

$$\begin{aligned}\mathcal{L} =& \boldsymbol{\beta}_g G_f + \boldsymbol{\beta}_w w + \eta\\ \mathcal{L} =& \boldsymbol{\beta}_g(\boldsymbol{A}g_c + \boldsymbol{B}\boldsymbol{\theta}w + \boldsymbol{C}\nu) + \boldsymbol{\beta}_{\boldsymbol{w}} w + \eta\\ \mathcal{L} =& \boldsymbol{\beta}_g \boldsymbol{A}g_c + (\boldsymbol{\beta}_{\boldsymbol{g}}\boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\beta}_w)w + (\boldsymbol{\beta}_g\boldsymbol{C}\nu + \eta).\end{aligned} \tag{7}$$

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

$$\mathrm{Cov}[\mathcal{L}|G = g_c, W = w] = \boldsymbol{\beta}_g\boldsymbol{C}\boldsymbol{C}^T\boldsymbol{\beta}_g + \mathrm{Cov}[\eta]. \tag{8}$$

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 g_c and let

$$m := \boldsymbol{\beta}_g \boldsymbol{A} g_c + (\boldsymbol{\beta_g} \boldsymbol{B} \boldsymbol{\theta} + \boldsymbol{\beta}_w) w.$$

Then, conditional on $G_c = g_c$ and $W = w$, $\boldsymbol{\mathcal{L}}$ is multivariate Gaussian with mean m and covariance given by Equation 8. 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(8) is in the first quadrant. This probability is increasing in the $\mathrm{Cov}[\eta]_{12}$.[6] Therefore, $\mathrm{Cov}[\eta]_{12}$ is identified. The same argument also implies identification of all the other off-diagonal terms. □

D.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 $\mathrm{Cov}[V]$ and $\mathrm{Cov}[\eta]$. We estimate $\mathrm{Cov}[V]$ with Equation 5, where we use the sample analogue of $\mathrm{Cov}[G_c - \boldsymbol{\theta} W]$. To estimate $\mathrm{Cov}[\eta]$, we use Equation 8. The joint distribution of the data given $\mathrm{Cov}[\eta]$ and known parameters is given by the multivariate probit model. We fit $\mathrm{Cov}[\eta]$ by maximum likelihood.

D.5 Generating the private risk distributions

To generate the private risk distribution for each of the four scenarios we consider, we generalize the approach 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 covariates W and the vector of current PGIs G_c. We then use equation (6) 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 genetic and non-genetic risks based on the covariates, current PGIs, and the simulated future PGIs.

[6] 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.

E Overview of policy responses

To complement the discussion in the main text, we now give a brief overview of the literature on the regulation of selection markets. We also discuss how the ideas relate to genetic private information. As mentioned in the main text, regulation of selection markets typically seeks to balance the goals of efficiency and redistribution. The former aims to maximize total economic surplus, while the latter tries to help particular groups. Redistribution includes solidarity towards consumers with lower wealth, worse health status, or genes that are predictive of serious illnesses. The standard regulatory playbook contains three main approaches.

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

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

The third approach is **managed competition**. Managed competition combines private provision with regulations to improve market performance. In the United States, examples include individual health insurance exchanges and insurance options sponsored by large employers. In most other developed countries, the main example is complementary health insurance. Managed competition involves four main kinds of regulations: community rating, mandates, subsidies, and risk adjustment. Community rating imposes restrictions on what variables insurers can use to set prices. Community rating is used for redistribution reasons (such as lowering premiums for the elderly) but has the cost of increasing adverse selection (Handel et al., 2015). Mandates impose penalties on consumers who do not purchase insurance, and subsidies lower the prices of policies; these are important for increasing the number of consumers who purchase policies, which otherwise tends to be inefficiently too low due to adverse selection (Geruso et al., 2021; Chade et al., 2022; Veiga and Levy, 2022). Risk adjustment is a type of subsidy to insurers who cover sicker patients. Risk adjustments and other subsidies are thought to be important to guarantee that firms don't offer contracts that provide inefficiently low coverage (Glazer and McGuire, 2000).

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

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

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

F Acknowledgments

This research was reviewed and approved by the Committee Ethics and Data of the Leiden Law School of Leiden University (application number 2023-24). The use of the UK Biobank data for this research has also been reviewed by the George Mason University Institutional Review Board (IRB) Office, which has determined it to be exempt from IRB review. The use the UK Biobank data was also reviewed by and granted exemptions by the University of Southern California University Park Institutional Review Board and the Vrije Universiteit Amsterdam Research Ethics Committee. This research has been conducted using the UK Biobank Resource under application number(s) 99086 and 11425. This publication is part of the project "The Economic Consequences of New Genetic Testing Technologies" (project number Vl.Veni.221E.080) of the NWO Talent Programme Veni – Social Sciences and Humanities, which is (partly) financed by the Dutch Research Council (NWO). This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-8891. This research was also supported by a Faculty Research and Development Award (FRDA) at George Mason University and by the National Science Foundation (NSF award nos. 2343735 and 2343736).

F.1 Additional acknowledgments and GWAS data availability

We thank the participants and investigators of the FinnGen study for providing GWAS summary statistics for these analyses (Kurki et al., 2023). The GWAS summary statistics for Alzheimer's disease and colorectal cancer derived in the FinnGen biobank can be accessed here: https://www.finngen.fi/en/access_results. We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary statistics for these analyses (Lambert et al., 2013). The GWAS summary statistics for Alzheimer's disease and an extended acknowledgment of the relevant funding sources can be accessed here: https://www.niagads.org/datasets/ng00036. We thank the participants and investigators of the Breast Cancer Association Consortium for providing the summary statistics for these analyses (Michailidou et al., 2017; Zhang et al., 2020). The GWAS summary statistics for breast cancer and an extended acknowledgment of the relevant funding sources can be accessed here: https://bcac.ccge.medschl.cam.ac.uk/bcacdata/. We thank the participants and investigators of the CARDIoGRAMplusC4D Consortium (Nikpay et al., 2015). The GWAS summary statistics for coronary artery disease can be accessed here: http://www.cardiogramplusc4d.org/. We thank the participants and investigators that contributed to producing the following GWAS summary statistics that were downloaded from the NHGRI-EBI GWAS Catalog (Sollis et al., 2022): Colorectal cancer study ID(s) GCST012876, GCST012877, GCST012878, GCST012879, GCST012880 (Huyghe et al., 2019; Fernandez-Rozadilla et al., 2023); Prostate cancer study ID(s) GCST006085 (Schumacher et al., 2018). The GWAS summary statistics on colorectal cancer and prostate cancer downloaded from the NHGRI-EBI GWAS Catalog can be accessed here: https://www.ebi.ac.uk/gwas/downloads/summary-statistics. We thank the participants and investigators of the Psychiatric Genomics Consortium (Trubetskoy et al., 2022). The GWAS summary statistics for schizophrenia can be accessed here: https://pgc.unc.edu/for-researchers/download-results/. We thank the participants and investigators of the DIAGRAM Consortium (Mahajan et al., 2018). The GWAS summary statistics for type 2 diabetes can be accessed here: http://diagram-consortium.org/downloads.html.

References

Abramowitz, M. and Stegun, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. US Government printing office, 1972 edition, 1948.

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

Anderson, M. and Mossialos, E. Are we heading for a two tier healthcare system in the UK? *BMJ*, page o618, 2022.

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

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

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

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

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

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

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

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

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

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

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

Dixon, P., Horton, R. H., Newman, W. G., McDermott, J. H., and Lucassen, A. Genomics and insurance in the United Kingdom: increasing complexity and emerging challenges. *Health Economics, Policy and Law*, pages 1–13, 2024.

Einav, L., Finkelstein, A., and Fisman, R. *Risky business: why insurance markets fail and* *what to do about it.* Yale University Press, 2023.

Falconer, D. S. and Mackay, T. F. *Introduction to quantitative genetics.* Prentice Hall, Essex, 4th edition, 1996.

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

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

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

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

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

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

Grotzinger, A. D., Fuente, J. d. l., Privé, F., Nivard, M. G., and Tucker-Drob, E. M. Pervasive downward bias in estimates of liability-scale heritability in genome-wide association study meta-analysis: a simple solution. *Biological Psychiatry*, 93:29–36, 2022.

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

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

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

Kelly, E. and Stoye, G. The impacts of private hospital entry on the public market for elective care in England. *Journal of Health Economics*, 73:102353, 2020.

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

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

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

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

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

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

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

Nabholz, C. and Rechfeld, F. Seeing the future? How genetic testing will impact life insurance. Technical report, Swiss Re Institute, 2017.

NHGRI. Fact Sheet Human Genomic Variation, 2023.

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

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

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

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

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

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

Ryan, M. M., Cox, C. G., Witbracht, M., Hoang, D., Gillen, D. L., et al. Using direct-to-consumer genetic testing results to accelerate Alzheimer disease clinical trial recruitment. *Alzheimer Disease & Associated Disorders*, 35:141–147, 2021.

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

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

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

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

Soch, J. et al. Statproofbook/statproofbook.github.io: The book of statistical proofs, 2024.

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

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

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

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

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

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

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

Veiga, A. and Levy, Y. Optimal contract regulation in selection markets. *SSRN paper no. 4029945*, 2022.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, 2:e41, 2006.

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

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

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

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

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

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

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