Age-related late-onset disease heritability patterns and implications for genome-wide association studies

Roman Teo Oliynyk[1,2,*]

[1]**Centre for Computational Evolution, University of Auckland, Auckland 1010, New Zealand**
[2]**Department of Computer Science, University of Auckland, Auckland 1010, New Zealand**
[*]**roli573@aucklanduni.ac.nz**

Abstract

Background. Genome-wide association studies and other computational biology techniques are gradually discovering the causal gene variants that contribute to late-onset human diseases. After more than a decade of genome-wide association studies efforts, these can account for only a fraction of the heritability implied by familial studies, the so-called "missing heritability" problem.

Methods. Computer simulations of polygenic late-onset diseases in an aging population have quantified the risk allele frequency decrease at older ages, caused by individuals with higher polygenic risk scores becoming ill proportionately earlier. This effect is most prominent for diseases characterized by high cumulative incidence and high heritability, examples of which include Alzheimer's disease, coronary artery disease, cerebral stroke, and type 2 diabetes.

Results. The late-onset diseases incidence rate grows exponentially for decades after early onset ages, guaranteeing that the cohorts for genome-wide association studies overrepresent older individuals with lower polygenic risk scores, whose disease cases are disproportionately due to environmental causes such as old age itself. This mechanism explains diminishing with age clinical predictive power, lower familial studies heritability and genome-wide association studies discovery power. It also explains relatively constant with age heritability found for late-onset diseases of lower prevalence, exemplified by cancers.

Conclusions. For late-onset polygenic diseases showing high cumulative incidence together with high initial heritability, rather than using relatively old age-matched cohorts, study cohorts combining the youngest possible cases with the oldest possible controls may significantly improve the discovery power of genome-wide association studies.

Introduction

Throughout the ages, late-onset diseases (LODs) were considered a bane of the lucky few who survived to an advanced age. Over the last couple of centuries, continuous improvements in sanitation, life and work environments, vaccinations, disease prevention, and medical interventions have extended the average life expectancy by decades.

With a growing fraction of the population being of advanced age, the leading causes of mortality are now heart disease, cancer, respiratory disease, stroke, and notably Alzheimer's disease and other dementias (Murphy et al., 2017). The need—and with it, the effort being made—to determine the causes of late-onset diseases is ever increasing, and one of the targets of medicine has become combating aging in addition to specific age-related diseases (Franceschi et al., 2018).

One of the major goals of computational biology is to identify gene variants that lead to increased odds of late-onset diseases. Nevertheless, polygenic LODs remain resistant to the discovery of sufficient causal gene variants that would allow for accurate predictions of an individual's disease risk (Manolio et al., 2009; Clarke and Cooper, 2010; Kumar et al., 2016). This is despite the fact that LODs with varied symptoms and phenotypes show high heritability in twin and familial studies (Zaitlen and Kraft, 2012).

At a young age, the human organism usually functions as well as it ever will. With time, the organism's functions decline, leading to the common image of aging as one of thinning hair and a loss of pigmentation in what remains, increased wrinkling and altered pigmentation of the skin, reductions in height, muscle and bone mass, joint pain, and deficits in hearing, sight and memory (Fedarko, 2018). The combination of genetic liability, environmental factors, and the physiological decline of multiple organism systems leads to individual disease presentation. Detrimental gene variants may be either protective or exacerbating factors, compared to the average distribution of common gene variants that defines human conditions as it applies to polygenic LODs.

Genome-wide association studies (GWASs) researchers often set an unrealistic expectation that a combination of causal single nucleotide polymorphisms (SNPs)—also known as a polygenic score—will, irrespective of the patient's age, completely predict an individual's predisposition to an LOD to a degree matching the maximum heritability found in familial studies (Naj and Schellenberg, 2017; Silva et al., 2015). The lost heritability debate, in the case of LODs, often treats polygenic LODs as if they were binary hereditary phenotypic features rather than facets of failure processes that arise in the human body (Oh et al., 2014) when it is past its reproductive prime and when evolutionary selection is significantly relaxed compared to younger ages (Fedarko, 2018).

GWAS can implicate a subset of SNPs that can typically explain between 10 and 20% of the genetic heritability of a polygenic LOD (Visscher et al., 2017). There are two complementary hypotheses explaining this so-called missing heritability (Eyre-Walker, 2010; Yang et al., 2012; Thornton et al., 2013; Agarwala et al., 2013). The first is the hypothesis that LODs are caused by a combination of a large number of relatively common alleles of small effect (Goldstein, 2009). GWAS has been able to discover only a small number of moderate-effect SNPs, but a large number of SNPs remain below GWASs' statistical discovery power. The second hypothesis states that LODs are caused by a relatively small number of rare moderate- or high-effect alleles with a frequency below 1% that likely segregate in various proportions into subpopulations or families (Dickson et al., 2010; North and Beaumont, 2015) and are also under the radar of GWASs' discovery power.

Both scenarios can contribute to observational facts, but their relative weights vary depending on the genetic architecture of an LOD (Park et al., 2011). In cases of high detrimentality, rare high-effect alleles become indistinguishable in their presentation from the OMIM cataloged conditions and likely will be diagnosed as a separate disease or syndrome. The population age distribution and individual disease progression of polygenic LODs can be best understood by considering the aging process itself as an ongoing loss of function, which can be modulated by the genetic liabilities resulting from both common and rare SNP distributions combined with changing environmental and lifestyle variables. It is determined (Anderson et al., 2011; Yang et al., 2015) that rare variants are unlikely to explain the majority of heritability for most complex traits.

While GWAS findings can explain only a fraction of heritability, the systematically collected SNP correlations provide a good indication of what to expect regarding the effect sizes and allele frequency distribution of as yet undiscovered SNPs (Eyre-Walker, 2010). Many studies focus on constructing hypotheses, defining the types of gene variants that could explain the missing heritability, proving why these gene variants are difficult to discover, and identifying the evolutionary processes that led to the hypothesized and observed gene variant distributions (Manolio et al., 2009; Clarke and Cooper, 2010; So et al., 2011; Zaitlen and Kraft, 2012; Thornton et al., 2013; Wood et al., 2014). These studies explore the effect sizes and allele frequencies that GWAS would expect to find for LODs as well as the genetic architecture of the complex traits and their implications for fitness.

The age-related heritability decline of some LODs has been assumed for decades. The precise magnitude of heritability change with age is typically not known for most LODs, and the effects are not understood and often ignored or overlooked. The GWASs primarily recommend homogeneity of cohort age, where same age window should be targeted, though it was suggested (Li and Meyre, 2013) that individuals with an early age of onset are likely to be enriched for genetic susceptibility. Discussing a replication study design, (Li and Meyre, 2013) state: "Once the risk of false positive association has been ruled out by initial replication studies, the focus of the association can be extended to different age windows." Another usual approach is to "age adjust" the effect by GWASs (Zaitlen et al., 2012) with the goal of removing or

averaging out the effect of aging rather than looking into its consequences more thoroughly.

One of the first geneticists to build a conceptual foundation for susceptibility to diseases, and the pioneer of the liability threshold approach, was D. S. Falconer in his studies of inheritance estimated from the prevalence among relatives (Falconer, 1965) and his 1967 follow-up study exploring the prevalence patterns of LODs, specifically diabetes (Falconer, 1967), and their decreasing heritability with age. These concepts were not followed up by systematic research, likely due to the difficulties involved in setting up large familial studies and perhaps the perceived limited clinical use of this kind of expensive and time-consuming project.

Detailed, high-granularity data on heritability by age are rare for most diseases. The familial heritability, clinical, and epidemiological statistics were available for eight prevalent LODs: Alzheimer's disease (AD), type 2 diabetes (T2D), coronary artery disease (CAD), and cerebral stroke, and four late-onset cancers: breast, prostate, colorectal, and lung cancer. These statistics served as the basis for this study's analysis and conclusions. This study investigated the model in which the polygenic risk of an individual remains constant with age, and endeavored to establish how the higher odds of becoming ill of individuals with higher polygenic liability may lead to a change of risk allele distribution as the population ages and whether this alone may explain some of the known observational facts.

A set of computer simulations quantified the change in the risk allele representation for these LODs as the population ages and determined how and why these changes affect clinical predictive power and GWAS statistical discovery power with age, more for some LODs than for others. Consequently, this study proposes a modification to GWAS cohort selection to improve statistical discovery power.

Methods

The model definition

According to (Chatterjee et al., 2016), the conditional age-specific incidence rate of the disease, $I(t|G)$ that is defined as the probability of developing the disease at a particular age t, given that a subject has been disease-free until that age, can be modeled using Cox's proportional hazards model (Cox, 1972):

$$I(t|G) = I_0(t) \cdot exp(\sum_k b_k G_k), \tag{1}$$

where $G = (G_1, \ldots, G_k)$ is the multiplicative effect of a set of risk factors, on the baseline hazard of the disease $I_0(t)$. The set of age independent variables in G could include genetic and environmental risk factors, and their interaction terms.

The following summary from (Chatterjee et al., 2016) is particularly relevant to the methodology of this research: "logistic regression methods are preferred for the evaluation of multiplicative interactions. For case-control studies, if it can be assumed that environmental risk factors are independent of the SNPs in the underlying population, then case-only and related methods can be used to increase the power of tests for gene-environment interactions. To date, post-GWAS epidemiological studies of gene-environment interactions have generally reported multiplicative joint associations between low-penetrant SNPs and environmental risk factors, with only a few exceptions." This means that the polygenic score $G = \sum_k b_k G_k$, as the lifelong characteristic of each individual, is used multiplicatively with $I_0(t)$, which encompasses environmental and aging effects.

Two recent studies (Lin et al., 2014; Bjørnland et al., 2018) emphasize the need to explore "extreme phenotype sampling" in order to improve GWAS discovery, including using cohorts nonuniform by age and other covariates. These studies provide mathematical development and simulational validation of these modes of sampling using logistic regression model. The study (Pawitan et al., 2009), which serves as a basis for genetic architectures used here, follows mathematical foundation and simulational validation of the logistic regression model of liability developed in (Noh et al., 2006), and thus further validated by the above studies.

It is important to note that this research simulations operate on model genetic architectures of the analyzed LODs, not a complete GWAS map of their experimentally discovered SNPs, because GWAS-discovered sets can explain only a fraction of these LODs heritability. These model genetic architecture SNPs are treated as "true" causal for disease liability and heritability variants, as discussed in (Chatterjee et al., 2016), rather than GWAS linked SNPs. They are used as *a priori* known constant causal SNPs that combine into individual PRSs for an LOD, as will be described further.

Taking an aging population simulation approach allows to find individuals becoming ill and, with them, the corresponding allele distribution between cases and controls, without intermediate steps and operating directly with the odds-ratio-based polygenic scores common to GWASs and clinical studies. The core of the simulation is Algorithm 1, operating on the known yearly incidence of an LOD and the PRSs for all individuals based on a modeled LOD genetic architecture:

Algorithm 1: Sampling individuals diagnosed with a disease proportionately to their polygenic odds ratio and incidence rate.

```
for age = 1 to MaxAge do
    numberIllThisYear = I(age)·N // N is unaffected population
    for i = 1 to numberIllThisYear do
        HRsum = 0 // will recalculate sum of all HRs
        for u = 1 to N do
            HRsum = HRsum + ORtoHR(G_u) // calculate the HR total
            LOOKUP(add, HRsum, u) // add u_th individual to the lookup table
        rand = RandomNumber(0, HRsum) // pick a random number
        ill = LOOKUP(find, rand, N) // found newly diagnosed
        N = N − 1 // decrement number of healthy individuals
        ProcessAndAnalyze(ill)
```

Note: an individual makes a sampling target proportionate to hazard ratio (HR) in the *LOOKUP*() table. ORs are converted to HRs similar to (Wang et al., 2013). An individual with HR = 15 will be 150 times more likely to be sampled than an individual with HR = 0.1. *ProcessAndAnalyze*() moves newly diagnosed individual from healthy to ill population pool, accounts for allele distribution, case/control ORs, etc.

Descriptively, this algorithm works as follows. In the simulation, each next individual to be diagnosed with an LOD is chosen proportionately to the relative risk of the

PRS at birth, relative to all other individuals in the as-yet-unaffected population. The number of individuals diagnosed yearly is determined using the model incidence rate curve derived from clinical statistics. In this manner, the aging process is probabilistically reproduced using a population simulation model rather than a computational model. As simulation progresses, the risk alleles are tracked for all newly diagnosed individuals and the remaining unaffected population and their representation in the affected and remaining population are statistically analyzed.
The following sections describe the model genetic architectures, the LOD incidence models and the statistical foundations of this research.

Allele distribution models

An in-depth review by (Pawitan et al., 2009) extensively analyzed models of genetic architecture and through simulations determined the number of alleles required to achieve specific heritability and estimated GWASs' discovery power. They calculated allele distributions and heritabilities and ran simulations for six combinations of effect sizes and minor allele frequencies (MAFs). Relying on the conclusions of (Pawitan et al., 2009) in this research allows to avoid repeating the preliminary steps of evaluating the allele distributions needed to achieve required heritability levels. The (Pawitan et al., 2009) alleles represent the whole spectrum ranging from common low-frequency low-effect-size alleles to very rare high-effect high-frequency alleles. The five most relevant architectures were implemented in this study; see Table 1.

Table 1. Genetic architecture scenarios

Scenario	MAF	Odds ratio
A. Common low	0.073 - 0.499	1.05 - 1.15
B. Modest low	0.0365 - 0.2495	1.05 - 1.15
C. Rare low	0.0146 - 0.0998	1.05 - 1.15
D. Rare medium	0.0146 - 0.0998	1.28 - 2.01
E. Rare high	0.0073 - 0.0499	1.63 - 4.05

Allele distributions as modeled by (Pawitan et al., 2009).

It is also handy for repeatable allele tracking, rather than generating the continuous random spectrum of allele frequencies and effect sizes, to follow the (Pawitan et al., 2009) configuration and discretize the MAFs into five equally spaced values within the defined range, with an equal proportion of each MAF and an equal proportion of odds ratios. For example, for scenario A, the MAFs are distributed in equal proportion at 0.073, 0.180, 0.286, 0.393, and 0.500, while the OR values are 1.15, 1.125, 1.100, 1.075, and 1.05, resulting in 25 possible combinations. Having multiple well-defined alleles with the same parameters facilitated the tracking of their behaviors with age, LOD, and simulation incidence progression.
An individual polygenic risk score β can be calculated as a sum of all alleles' effect sizes, which is by definition a log(OR) (natural logarithm of odds ratio) for each allele, also following (Pawitan et al., 2009):

$$\beta = log(OR) = \sum_k a_k log(OR_k), \tag{2}$$

where a_k is the number of risk alleles (0, 1 or 2) and OR_k is the odds ratio of additional liability presented by the k-th allele.
Variance of the allele distribution is determined by:

$$var = 2\sum_k p_k(1-p_k)(log(OR_k))^2, \tag{3}$$

where p_k is the frequency of the k-th genotype (Pawitan et al., 2009).
The contribution of genetic variance to the risk of the disease is heritability:

$$h^2 = \frac{var}{var + \pi^2/3}, \tag{4}$$

where $\pi^2/3$ is the variance of the standard logistic distribution (Noh et al., 2006). For example, the number of variants needed for the Scenario A LODs is summarized in Table 2. Following (Pawitan et al., 2009), the

Table 2. Heritability of analyzed LODs and an example of required variant numbers for common low-effect variants scenario A

	Highly prevalent LODs				Cancers			
	AD	T2D	CAD	Stroke	Prostate	Colorectal	Breast	Lung
Heritability	0.795	0.69	0.55	0.41	0.57	0.40	0.31	0.095
Variants	3575	2125	1175	625	1250	600	400	100

variants are assigned to individuals with frequency proportionate to MAF p_k for SNP k, producing according to Hardy–Weinberg principle three genotypes (AA, AB or BB) for each SNP with frequencies p_k^2, $2p_k(1-p_k)$ and $(1-p_k)^2$. The mean value β_{mean} for the population distribution can be calculated using equation:

$$\beta_{mean} = 2\sum_k p_k \cdot log(OR_k) \tag{5}$$

Customarily, the individual PRSs are normalized relative to G_{mean}, resulting in zero mean initial population PRS, making it easy to compare higher and lower risk individuals.
A part of simulation functionality is to allocate the genetic architectures and to calculate variance Eq. (3) of each genetic architecture instance described above. Each genetic architecture listing is represented in Supplementary Data 1 executable folder, for example the file "CommonLow.txt" lists the variants describing Scenario A (only three columns are used for this simulation: SNP - internal use identifier, EAF - effect allele frequency, and OR). In case of Scenario A, $var = 0.09098$ for single set of SNPs listed in this file. Rearranging Eq. (4) and changing the multiple allows to find number of variants sets for each LOD seen in Table 2 and to closely match the target heritability. Each simulation run calculates population PRS variance and records heritabilities and allele distributions for the case and control populations as the simulated age progresses.

Evaluating GWAS statistical power

GWAS statistical power is the estimate of power to detect associations between DNA variants and a trait, and depends on the experimental sample size, the distribution of effect sizes and the frequency of these variants in population (Visscher et al., 2017). The statistical power calculations are very useful in a case/control study design for determining the minimum number of samples that will achieve adequate statistical power; conventionally 80% considered an acceptable power (Hong and Park, 2012). To achieve higher power, a disproportionately larger number of cases and controls may be required, which often would be unrealistic for a cohort study. A number of statistical power calculators are available, for example (Sham and Purcell, 2014). The most helpful for this study was the Online Sample Size Estimator (OSSE) (Online Sample Size Estimator).

The GWAS progress over the last decade, particularly regarding polygenic traits, was made to large extent due to the ever increasing cohort sizes. The cohort size is one of the main limiting factors of GWAS discovery power, making it an important benchmark for this study. Here the cohort size is defined as a number of cases and controls that will be needed to achieve 80% statistical discovery power when the case/control allele frequency changes with cohort age for a subset of representative alleles in the model genetic architectures. For each such allele in the simulated population, the allele frequency for cases and controls is tracked as age progresses. The difference between these MAFs gives the non-centrality parameter (NCP) λ for two genetic groups (Sham and Purcell, 2014; Luan et al., 2001):

$$\lambda = N \cdot p_1 \cdot p_2 \cdot (\beta_1 - \beta_2)^2, \tag{6}$$

where N is the overall population sample size and $p_1 and p_2$ the fractions of cases and controls, and β_1 and β_2 are the case and control mean log(OR) for an allele of interest. The values $p_1 = p_2 = 0.5$, or an equal number of cases and controls, are used throughout this publication.

Having obtained NCP λ from Eq. (6), (Luan et al., 2001) recommend using SAS or similar statistical software to calculate the statistical power, using the following SAS statement:

$$StatPower = 1 - PROBF(FINV(0.99999995, 1, N-4), 1, N-4, \lambda). \tag{7}$$

The conversion of this equation to its R equivalent, which was used to process the simulation output, is:

$$StatPower = 1 - pf(qf(PSign, 1, N-4), 1, N-4, \lambda), \tag{8}$$

where $PSign = 0.99999995$ corresponds to $5 \cdot 10^{-8}$ significance level. The outputs of this conversion was validated with the Online Sample Size Estimator (OSSE) (Online Sample Size Estimator). This equation returns statistical power based on a case/control number and the NCP as calculated above. To find the number of cases needed for 80% GWAS discovery power, having the $(\beta_1 - \beta_2)$, a rapid convergence R routine was used to iterate the values of N until the value of $StatPower$ matched to 0.8 with an accuracy better than ±0.01% for each age and allele distribution of interest.

LOD incidence functional approximation

The Chapter S3 in Supplementary File 1 describes functional approximations of the yearly incidence for Alzheimer's disease, type 2 diabetes, coronary artery disease, and cerebral stroke, and four late-onset cancers: breast, prostate, colorectal, and lung cancer. As a short summary, for all of the above LODs, the incidence rate curves can be approximated during the initial disease onset periods with a close to exponential yearly incidence growth. This exponential growth continues for decades, see Table 3 and Chapter S3 in Supplementary File 1.

Table 3. Until what age LOD incidence rate rises exponentially

	AD	T2D	CAD	Stroke	Breast C.	Prostate C.	Colorectal C.	Lung C.
Age, years	103	55	81	79	48	62	72	70

Later the growth may flatten at old age, as in the case of T2D, slow down like in cases of cerebral stroke and CAD, or continue exponentially to very advanced age as in case of Alzheimer's disease. An R script automates the determination of the best fit for logistic and exponential regression from the clinical incidence data.

Sampling based on the LOD incidence rate and individual PRS

The incidence rate functional approximations for the analyzed LODs are used to find the average number of diagnosed individuals N_d for each year of age t as a function of the incidence rate $I(t)$ and the remaining population unaffected by an LOD $N_u(t)$ in question:

$$N_d(t) = I(t) \cdot N_u(t), \tag{9}$$

In the next year of age, the unaffected population will be smaller by the number of individuals diagnosed in the previous year $N_d(t)$:

$$N_u(t+1) = N_u(t) - N_d(t) = N_u(t)(1 - I(t)). \tag{10}$$

The numbers of individuals projected to become ill per year, as well as remaining unaffected population, are applied in Algorithm 1.

For the PRS of the simulated population based on odds ratios built using the logistic regression model (Pawitan et al., 2009), if an LOD is characterized by low incidence within an age interval, and the odds ratio is close to 1, odds ratio values are practically identical to hazard ratio or relative risk. For example, (Song et al., 2014) treat OR and RR as equivalent in case of breast cancer in their simulation study. For higher values an OR usually significantly exceeds RR. An adjustment formula by (Zhang and Kai, 1998) approximating OR to hazard ratio subject to the modeled incidence for the LODs under consideration can be expressed as:

$$HR_u = \frac{OR_u}{1 + I(t) \cdot (OR_u - 1)}, \tag{11}$$

where HR_u and OR_u are the estimated hazard ratio for polygenic score OR of uth unaffected individual.

Individual values analysis and cohort simulation

It can be expected that for an LOD with higher incidence and heritability, the fraction of the highest PRS individuals will diminish faster with age. For such LODs the relatively lower PRS individuals will represent the bulk of the LOD cases at an earlier age compared to LODs with lower incidence and heritability. The LODs are characterized by a wide range of heritability and progression pattern of incidence rates with age. For example, T2D and breast cancer start their incidence rise relatively early but reach quite different levels at older age, while colon cancer and AD start later and also reach quite different maximum incidence and cumulative incidence levels; see Fig. S1 in Supplementary File 1. In the absence of mortality, both due to general frailty and other LODs, the incidence progression makes it appear that, sooner or later, depending on the incidence magnitude, the majority of the population would be diagnosed with every LOD. In reality, this does not happen because of the ongoing mortality from all causes.

Two main LOD simulation types are described next:

1. **The individual values analysis of polygenic risk scores and risk allele frequency for individuals diagnosed with disease at each specific age and the remaining population at this age.** The abbreviation “IVA" is interchangeably used with “individual values analysis" in this publication.

The IVA uses one-year age slices, and is performed as follows. Initially, the mean and variance of the PRS for the whole population are calculated. Next, based on the required incidence value for each year, individuals are picked from the unaffected population through randomly sampling the population with a probability proportionate to the individual’s PRS, as summarized in Algorithm 1. These individuals become the cases for the relevant year’s IVA, and the mean and variance of their PRS are also calculated and recorded. The mortality does not need to be applied to IVA scenario because it affects the future cases and controls in equal numbers, and accounting for mortality would only result in a smaller population being available for analysis. To track the GWAS statistical discovery power, the same nine representative variants (configurable) are tracked for all LODs simulated. The process continues this way until the maximum desired simulation age is reached.

2. **A simulated cohort study for each of these diseases.** For the sake of brevity, a word “cohort" may also be used throughout this publication.

The clinical study cohort simulation performs an analysis identical to that described above. The difference is that here the simulated GWAS clinical studies are performed with a patient age span of 10 years, which is a typical cohort age span, although any age span can be chosen as a simulation parameter. The simulation statistics are collected using the mid-cohort age, which is the arithmetic half-age of the cohort age span. In the first simulation year, a population equal to one-tenth of the complete population goes through the steps described for IVA. Each year, an additional one-tenth starts at age 0, while the previously added individuals age by one year. This continues until all 10 ages are represented. This combined cohort proceeds to age and is subject to the disease incidence rate and mortality according to each individual’s age.

Mortality is applied, with a probability appropriate to each year of age, to both accumulated cases and controls. As the population ages, both the case and control pool numbers would diminish. Take, for example, a cohort study that includes a 10-year span, say, between 50 and 59 years old. The cases for the cohort are composed from individuals who were diagnosed with an LOD at any age either younger than or including their current age, producing a cumulative disease incidence over all preceding years of age. For example, some of the individuals that are cases now, at age 59, may have been healthy at age 58. Some of the controls in our cohort at age 51 may or may not be diagnosed at an older age, which would qualify them as cases for this cohort, but they are currently younger and healthy. Therefore, it can not be known with certainty the extent to which younger controls differ from cases, except for the fact that they are not currently diagnosed—not unlike a real statistical study cohort. As a result, the corresponding GWAS discovery power can be expected not to change as dramatically as it does for the individual values analysis.

The following additional mortality scenarios were also simulated: (a) double mortality for cases compared with the unaffected population, (b) no mortality for either cases or controls, and (c) a one-year age span cohort with no mortality for either cases or controls.

The youngest age cohort for each LOD is defined as the mid-cohort age at which the cumulative incidence for a cohort first reaches 0.25% of the population. For consistency, this threshold was considered in this study the minimum cumulative incidence age allowing for the formation of well-powered cohort studies for all analyzed LODs.

The simulation design summary

Preliminary data collection and analysis is shared by all simulation runs and include: (a) preparing the genetic architecture files and calculating number of variants needed, based on each modeled LOD heritability as described in the “Allele distribution models" section above, and (b) determining the parameters of functional approximation for LOD incidence from published statistic as described in Chapter S3 in Supplementary File 1.

A simulation run for a single LOD can be logically divided into the following four steps:

1. Build the gene variants pool as outlined “Allele distribution models" section and load the incidence rate functional approximation parameters.

2. Allocate population objects, assign individual PRSs. Allocate all other simulation objects and arrays that will be used by simulation.

3. Run simulation's Algorithm 1 from age 0 to 100 for either the IVA or the cohort study scenario, described in "Individual values analysis and cohort simulation" section. Calculate and record the simulation data in comma separated value files.
4. Determine statistical power for cases and controls for each cohort based on "Evaluating GWAS statistical power" section.
The above steps were completely reinitialized and performed for each LOD analysis. The complete simulation iterated through all eight LOD analyses in two scenarios: a per-year-of-age population IVA and a simulation of a GWAS cohort study.

Validation simulations

Based on the described above model, it can be expected that the allele distribution in a population of the same age with a given initial genetic architecture will depend solely on cumulative incidence, which represents the fraction of the population that had succumbed to a disease. The purpose of validation simulations, performing runs with (a) constant, (b) linear and (c) exponential incidence rates was to validate whether or to what extent this expectation is correct. And whether the outcomes would differ between various genetic architectures. The validation simulations confirmed that PRSs for the population controls and cases, viewed in the individual values analysis at every age, depend on the cumulative incidence and the LOD heritability, and are independent from the incidence progression shape within each genetic architecture. The validation simulation's procedures are described in Chapter S2 in Supplementary File 1.

Data sources, programming and equipment

The population mortality statistics from the 2014 US Social Security "Actuarial Life Table" (US Social Security Actuarial Life Table) provided yearly death probability and survivor numbers up to 119 years of age for both men and women.
Disease incidence data from the following sources were extensively used for analysis, using the materials referenced in Chapter S1 in Supplementary File 1 for corroboration: Alzheimer's disease (Brookmeyer et al., 1998; Edland et al., 2002; Kokmen et al., 1988; Hebert et al., 1995), type 2 diabetes (Boehme et al., 2015), coronary artery disease and cerebral stroke (Rothwell et al., 2005), and cancers (Cancer Statistics for the UK; Kuchenbaecker et al., 2017).
The simulations were performed on an Intel i9-7900X CPU-based 10-core computer system with 64GB of RAM and an Intel Xeon Gold 6154 CPU-based 36-core computer system with 288GB of RAM. The simulation is written in C++, and can be found in Supplementary Data 1. The simulations used a population pool of 2 billion individuals for the LOD simulations and 300 million for validation simulations, resulting in minimal variability in the results between runs.
The cohort simulations were built sampling at least 5 million cases and 5 million controls from the living population of initial 2 billion simulated individuals; this is also equivalent to 0.25% of the initial population. This means that the cohort study would begin its analysis only when this cumulative incidence was reached. Conversely, the analysis would cease when, due to mortality, the number of available cases or controls declined below this threshold. For all LODs, this maximum mid-cohort age was at least 100 years and, depending on LOD, up to a few years higher. This confirms that, as described later in the Discussion section, in younger cases and older controls cohort recommendations, it is feasible to form control cohorts up to 100 years of age.
The simulation runs for either all validation scenarios or for a single scenario for all eight LODs took between 12 and 24 hours to complete. The final simulation data, additional plots and elucidation, source code, and the Windows executable are available in Supporting Information. Intel Parallel Studio XE was used for multi-threading support and Boost C++ library for faster statistical functions; the executable can be built and function without these two libraries, with corresponding execution slowdown. The ongoing simulation results were saved in comma separated files and further processed with R scripts during subsequent analysis, also available in Supplementary Data 1.

Statistical analysis

Large variation between simulation runs complicates population and genome models analysis. This issue was addressed in this study by using a large test population, resulting in negligible variability between runs. The statistical power estimates deviated less than 1% in two-sigma (95%) confidence interval, except for early Alzheimer's disease cohort starting ar 1.5% and going below 1% threshold within 4 years (see **TwoSDFraction.csv* files in Supplementary Data 1). In addition to operating with reliable data, this eliminated the need for the confidence intervals in the graphical display.

Results

Validation simulations for all genetic architectures

The validation simulations for all scenarios described in Methods Table 1, were performed not as a model of a specific disease, but to determine the behavior of all allele scenarios and the resulting allele frequency change under simple controlled and comparable-to-each-other incidence scenarios. It was important to characterize all genetic architectures, and to learn differences and similarities in behavior between them.
These simulations confirmed that a change in the population's mean PRS and a change in the cases' mean PRS, viewed as instantaneous values for each age, are dependent on the cumulative incidence and the magnitude of initial genetic model heritability. If mortality is not included, they are not dependent on the shape of incidence progression with age (see Fig. S2 in Supplementary File

1) and are qualitatively similar between the genetic architectures (seeFig. S3 in Supplementary File 1). This means that, when the same level of cumulative incidence is reached with any incidence pattern, the allele distribution for diagnosed cases and for the remaining unaffected population is identical.

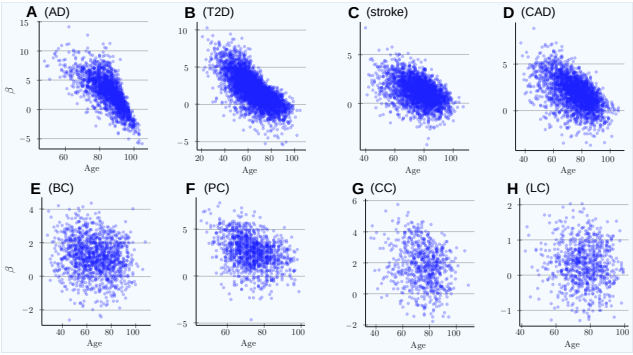


Figure 1. Polygenic risk scores of individuals diagnosed with an LOD as a function of age.
(A) Alzheimer's disease, **(B)** type 2 diabetes, **(C)** cerebral stroke, **(D)** coronary artery disease, **(E)** breast cancer, **(F)** prostate cancer, **(G)** colorectal cancer, **(H)** lung cancer.
Scatter plots show the distributions of PRS for cases diagnosed as age progresses, with ongoing mortality. $\beta = log(OddsRatio)$ visually implies the underlying heritability and incidence magnitudes. If the regression line can be easily drawn, dropping diagonally as age progresses, there is a combination of high heritability and increasing cumulative incidence. Otherwise, a plot appears as a relatively symmetrical blob.

Analysis of common low effect size genetic architecture scenario

The simulation results for the eight analyzed LODs are presented next.

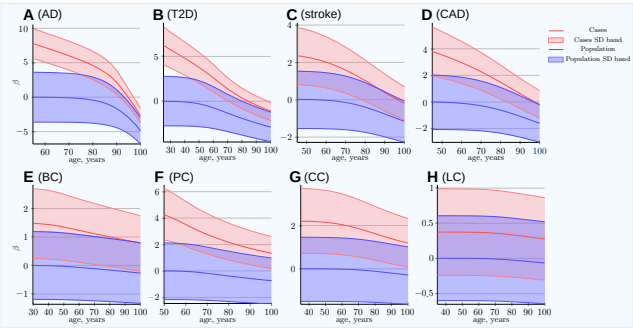


Figure 2. Polygenic risk score difference between newly diagnosed individuals and the remaining unaffected population.
(A) Alzheimer's disease, **(B)** type 2 diabetes, **(C)** cerebral stroke, **(D)** coronary artery disease, **(E)** breast cancer, **(F)** prostate cancer, **(G)** colorectal cancer, **(H)** lung cancer.
SD band is a band of one standard deviation above and below the cases and the unaffected population of the same age. For highly prevalent LODs, at very old age the mean polygenic risk of new cases crosses below the risk of an average healthy person at early onset age. *(Common low-effect-size alleles (scenario A), showing largest effect variant with MAF=0.5, OR=1.15).*

The IVA and cohort simulations were performed for all genetic architecture scenarios, from low to high effect

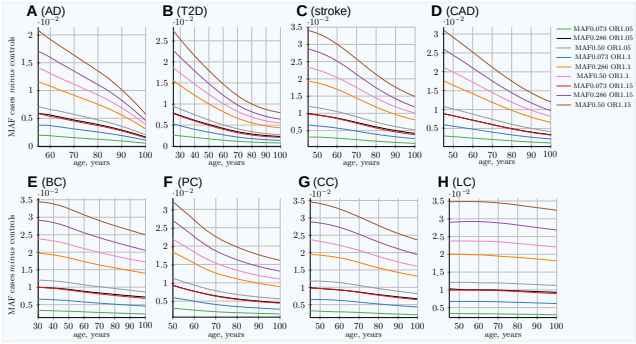


Figure 3. Allele frequency difference between newly diagnosed individuals and remaining population of the same age.
(A) Alzheimer's disease, **(B)** type 2 diabetes, **(C)** cerebral stroke, **(D)** coronary artery disease, **(E)** breast cancer, **(F)** prostate cancer, **(G)** colorectal cancer, **(H)** lung cancer.
The **MAF cases *minus* controls** value is used to determine GWAS statistical power; see Eq. (6). Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change, see also Fig. S5 in Supplementary File 1. *(Displayed here: 9 out of 25 SNPs, which are described in Methods for common low-effect-size alleles - scenario A).*

sizes, and common to low allele frequencies, and the results were found to be qualitatively consistent between all these scenarios. As a consequence, this report primarily focuses on the common low-effect-size genetic architecture scenario A, which the latest scientific consensus considers to be the genetic architecture behind the majority of polygenic LODs; the results are virtually identical for scenarios B and C, as validated in Fig. S3 in Supplementary File 1, making it unnecessary to present separate figures for these two scenarios.

The scatter plots in Fig. 1 show the distributions of PRSs for cases diagnosed as age progresses for the common low-effect-size genetic architecture scenario A. The PRS of individuals diagnosed with an LOD and the change with age of the average LOD PRS of the unaffected population are demonstrated in Fig. 2. The color bands show a one standard deviation spread for cases and controls, which, in the case of newly diagnosed cases, represents approximately two-thirds of the diagnoses at each age. This figure demonstrates how the initially high average polygenic risk of newly diagnosed cases declines as the most predisposed individuals are diagnosed with each passing year of age. The average PRS of the unaffected population decreases much more slowly. At advanced old age, the average polygenic risk of the newly diagnosed is lower than the risk for an average individual in the population at a young age; this is true for all four highly prevalent LODs: AD, T2D, CAD, and stroke.

This phenomenon is a consequence of the effect alleles frequency change, in which the highest effect alleles show the greatest difference between the diagnosed and the remaining unaffected population and also show the fastest change in frequency difference with age. Statistically, individuals possessing the higher risk alleles are more likely to succumb and to be diagnosed earlier, thus removing the allele-representative individuals from the unaffected population pool, see Fig. 3. These plots show the most

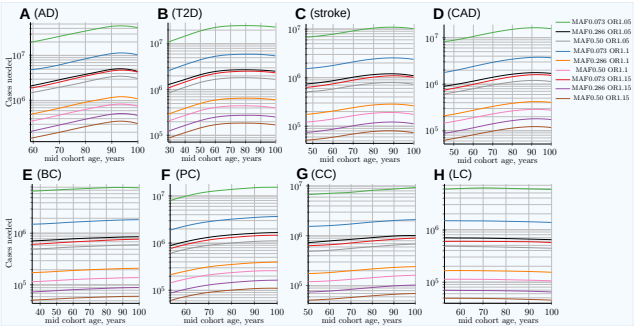


Figure 4. Change in number of cases needed to achieve 0.8 discovery power in age matched cases and controls cohort design.
(A) Alzheimer's disease, **(B)** type 2 diabetes, **(C)** cerebral stroke, **(D)** coronary artery disease, **(E)** breast cancer, **(F)** prostate cancer, **(G)** colorectal cancer, **(H)** lung cancer.
The matched age cohorts require larger number of participants to achieve the same GWAS discovery power compared to the youngest cohort age. There is a noticeable difference between cancers (with small exception of prostate cancer, see Discussion) and other LODs. *(Displayed here: 9 out of 25 SNPs, which are described in Methods for Common low-effect-size alleles - scenario A).*

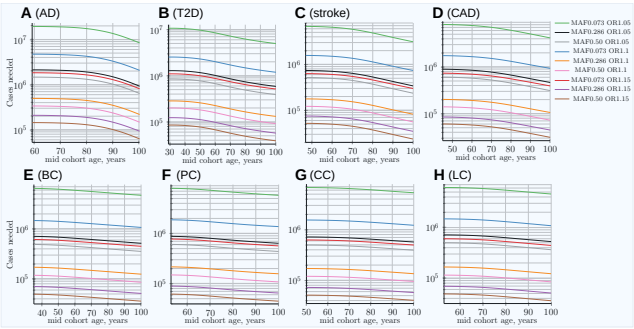


Figure 5. Change in number of cases needed for 0.8 discovery power in a cohort study when using progressively older controls compared to fixed-age young-cases.
(A) Alzheimer's disease, **(B)** type 2 diabetes, **(C)** cerebral stroke, **(D)** coronary artery disease, **(E)** breast cancer, **(F)** prostate cancer, **(G)** colorectal cancer, **(H)** lung cancer.
Cases' mid-cohort age is leftmost age (youngest plot point); control mid-cohort ages are incremental ages. The number of cases needed for 0.8 discovery power is smaller when using older age controls, particularly for LODs showing the most prominent increase in the number of cases needed for older age in matched age cohorts seen in Fig. 4. *(Displayed here: 9 out of 25 SNPs, which are described in Methods for Common low-effect-size alleles - scenario A).*

dramatic change for AD and T2D—LODs that possess the highest cumulative incidence and heritability. The smallest change corresponds to the lowest incidence and heritability LOD: lung cancer.

It is important to note that the absolute MAF values for diagnosed at particular age cases and controls do not change much with age progression for all LODs. For example, for T2D the allele with OR=1.15 and initial population MAF=0.2860, the allele frequency is: 0.2860 for controls and 0.3088 for cases at age 25, and changes to 0.2789 for controls and 0.2871 for cases at age 80 in IVA case, only a few percent change. At the same time, the relative differences change correspondingly from 0.0228 to 0.0081, a 2.7 times change, which is very significant for GWAS discovery power, as can be seen in Eq. (6). The absolute MAF change is even less prominent in cohort scenario, as seen in Fig. S13 in Supplementary File 1 that shows the same allele. The small change in the absolute value for older age groups makes it virtually impossible to analyze this effect using for example GWAS SNP database statistics for different age groups. The effect would be hidden behind interpersonal and populational genetic variability in hundreds and thousands of SNPs, changing their balance slightly with age in case of common low-effect-size genetic scenario. This effect is long established for high detrimentality variants, like for example, BRCA1/2 gene mutations in case of breast cancer (Kuchenbaecker et al., 2017), and APOE e4 allele for AD (Farrer et al., 1997), where there is known lower representation of these gene variant carriers among older undiagnosed individuals.

The cohort simulation shows a much more averaged change for these same scenarios because cohorts represent accumulative disease diagnoses from earlier ages, while mortality removes older individuals; see Fig. S5 in Supplementary File 1. While the MAF difference between cases and controls shown in the above figures is illustrative by itself, it is most important for determining the GWAS statistical discovery power using Eq. (6), Eq. (8), and from there the number of cases necessary to achieve 0.8 (80%) statistical power. From these equations, it is apparent that GWAS statistical discovery power diminishes as a complex function of a square of case/control allele frequency difference. The change with age of the number of cases needed for 0.8 GWAS discovery power for an age matched case/control cohort study is presented in Fig. 4. In the hypothetical IVA case, the number of individuals required to achieve GWAS discovery power increases fast, see Fig. S6 in Supplementary File 1. This is a quite informative instantaneous value of statistical power; however, neither GWAS nor clinical studies ever consist of individuals of the same age, due to the need to have a large number of individuals to maximize this same statistical power. The cohort scenario is correspondingly less extreme, as seen in Fig. 4. These plots show an increase in the number of participants needed to achieve GWAS statistical power between the lowest effect and frequency to the highest effect and frequency alleles; this number varies more than one-hundred fold between alleles within the genetic architecture.

This number of the required cohort participants is quite similar for the same effect alleles among all eight LODs; for example, the highest effect allele for each LOD requires $5\cdot 10^4$–$1.4\cdot 10^5$ cases for 0.8 GWAS discovery power at younger ages. The change in allele frequency with age between cases and controls varies quite widely among LODs, with the highest being AD and the lowest lung cancer; see Fig. 3. The red line in Fig. 6 summarizes the multiplier—how many times the number of participants needs to increase as the cohort is aging—comparing to the youngest possible cohort age for eight analyzed LODs. These cohort results are simulated with identical mortality for cases and controls. Mortality has an impact on the

cohort allele distribution.
Table 4 combines the heritability and incidence of the LODs with the summarized simulation results from the cohort simulation, also seen in Fig. 4 and Fig. S4 in Supplementary File 1.

Table 4. LOD statistics and age matched cohort simulation summary

	Highly prevalent LODs				Cancers			
Disease	AD	T2D	Stroke	CAD	Breast	Prostate	Colorectal	Lung
Lifetime risk %	10–20	55	25–30	32–49	12	12	< 4.5	<6.9
Heritability %	79–80	69	38–44	50–60	31	57(42)	40	8–18
Maximum yearly incidence %	> 20	2.5	4.4	3.6	<0.5	<0.8	<0.6	<0.6
	ΔMAF between cases and controls							
Youngest cohort	0.020	0.026	0.034	0.032	0.034	0.031	0.034	0.035
age 80y	0.015	0.018	0.028	0.023	0.032	0.024	0.031	0.035
age 100y	0.014	0.019	0.028	0.023	0.032	0.023	0.029	0.036
	Cases needed for 0.8 statistical power							
Youngest cohort	$1.4 \cdot 10^5$	$8.7 \cdot 10^4$	$5.3 \cdot 10^4$	$6.0 \cdot 10^4$	$5.0 \cdot 10^4$	$6.1 \cdot 10^4$	$4.9 \cdot 10^4$	$4.9 \cdot 10^4$
age 80y	$2.6 \cdot 10^5$	$1.8 \cdot 10^5$	$7.9 \cdot 10^4$	$1.1 \cdot 10^5$	$5.8 \cdot 10^4$	$1.0 \cdot 10^5$	$6.1 \cdot 10^4$	$4.7 \cdot 10^4$
age 100y	$3.0 \cdot 10^5$	$1.7 \cdot 10^5$	$7.3 \cdot 10^4$	$1.1 \cdot 10^5$	$5.9 \cdot 10^4$	$1.1 \cdot 10^5$	$6.9 \cdot 10^4$	$4.5 \cdot 10^4$
Multiple cases needed, early to 80y	1.9	2.1	1.5	1.8	1.15	1.6(1.35)	1.25	1.0

The values for MAF and cases needed for 0.8 (80%) GWAS statistical discovery power are for the common alleles, low effect size scenario A. Cohorts span 10 years. The results shown are for the allele with MAF=0.5, OR=1.15, the largest effect allele that requires the smallest number of cases/controls. “Maximum incidence %" is the largest incidence at older age. “Cases mult." is the multiple of the number of cases needed for the 80-year-old cohort to achieve the same statistical power as the early cohort. Prostate cancer heritability is 57% according to (Hjelmborg et al., 2014). Shown in braces 42% heritability (Grönberg, 2003), more in line with other three cancers.

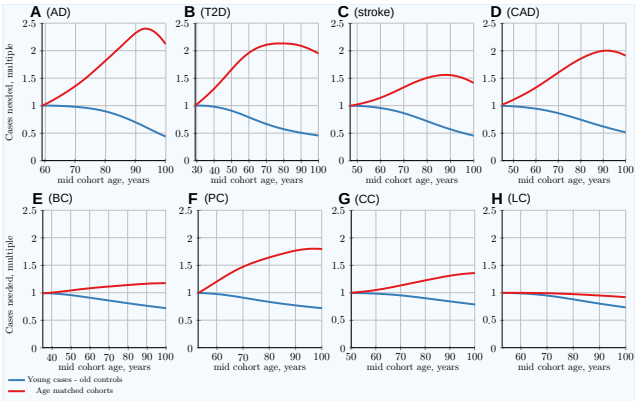


Figure 6. Per LOD comparison: youngest possible cases and increasingly older controls *vs* classical age matched cohorts.
(A) Alzheimer’s disease, **(B)** type 2 diabetes, **(C)** cerebral stroke, **(D)** coronary artery disease, **(E)** breast cancer, **(F)** prostate cancer, **(G)** colorectal cancer, **(H)** lung cancer.
The multiplier of how many times fewer cases needed in young cases / older controls scenario is represented by blue lines, strongly contrasting the increasing with age number of cases needed for the same GWAS discovery power in a classic age matched study design represented by red lines. *(Common low-effect-size alleles scenario A).*

Validating more extreme mortality scenarios

The more extreme mortality scenarios—both lower and higher—than one could expect in a real cohort study, were validated in this set of simulations. The results were relatively close to those presented for equal case/control mortality. The extreme cases of (a) no mortality for either cases and controls and (b) double the mortality of cases compared to controls produce very similar allele distributions before the age of 85, while somewhat diverging at older ages. The double mortality of cases compared to controls is higher than the clinically known mortality for the analyzed LODs. While it may have been a realistic scenario a century ago before modern health care, it is certain that patient mortality is lower these days. In addition, a one-year cohort without mortality was used as the most extreme validation case. This scenario can also be considered an individual cumulative case, which counts everyone who became ill by specified age as cases, and everyone healthy at that age as controls. These validation cohort scenarios are summarized in Fig. S7 in Supplementary File 1.
The mortality analysis was applied to one LOD at a time. This research did not attempt to estimate increased mortality for multiple disease diagnoses. (Collerton et al., 2009) followed a cohort of individuals over the age of 85 in Newcastle, England, and found that, out of the 18 common old age diseases they tracked, a man was on average diagnosed with four and a woman with five, not to mention a plethora of other less common diseases and their causal share in individual mortality.

Evaluating rare medium effect size genetic architecture scenario

Other genetic architecture scenarios produce qualitatively similar patterns, specifically differing in the number of cases needed to achieve 0.8 statistical power for medium- and large-effect genetic architecture scenarios. The rare medium-effect-size alleles (scenario D) results are presented in Fig. S8, Fig. S9, Fig. S10, and Fig. S11 in Supplementary File 1. There, at younger ages, the MAF difference between cases and controls is larger for rare medium-effect-size alleles. The number of cases and controls needed to achieve 80% GWAS statistical power for all eight LODs is approximately five times lower, a direct consequence of variants’ larger effect sizes. This result perhaps excludes the scenario of rare medium-effect-size alleles being causally associated with the LODs reviewed here, because GWAS studies would have an easier time discovering a large number of causal SNPs. From a qualitative perspective, all reviewed genetic architecture scenarios provide similar pattern of increasing number of cohort study cases needed to maintain the same discovery power with age progression.

Advantage of using youngest possible case cohorts and oldest control cohorts

The scenarios simulating the number of cases needed when the case cohort uses the youngest possible participants with increasingly older control cohorts are presented in Fig. 5 and Fig. S12 in Supplementary File 1. The multiplier of how many times fewer cases needed in this scenario is represented by blue lines in Fig. 6, strongly contrasting the increasing with age number of cases needed for the same GWAS discovery power in a classic age matched study design. The summary comparing two cohort designs is presented in Fig. 7.
Thus, cohorts composed of the youngest possible cases and the oldest available controls show improvement in

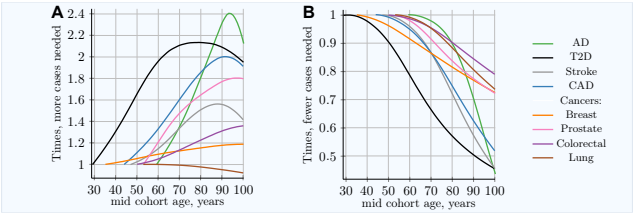


Figure 7. Advantage of using youngest possible cases and increasingly older controls, compared to classical age matched cohorts.
(A) Relative **increase in number of cases needed** for 0.8 discovery power in cohort study when using progressively older **same-age case and control cohorts**. **(B)** Relative **decrease in number of cases needed** for 0.8 discovery power in cohort study when using **progressively older control cohorts compared to fixed-age young-cases cohort**. The youngest age cohort for each LOD is defined as the mid-cohort age at which the cumulative incidence for a cohort first reaches 0.25% of the population. Therefore, the leftmost point on each LOD line is the reference (youngest) cohort, and as cohorts age, the cohort case number multiple required to achieve 0.8 statistical power is relative to this earliest cohort. While all alleles display a different magnitude of cases needed to achieve the required statistical power, the multiplier change with age is almost identical for all alleles within a genetic architecture scenario. *(Common low-effect-size alleles scenario A)*.

GWAS discovery power, or equivalently, a smaller number of participants needed to achieve the same discovery power.

Discussion

Performing a comprehensive set of validation simulations allowed to determine and generalize the change in allele distribution with cumulative incidence increase for all described in Methods genetic architectures. The simulation results show that for all genetic architectures the change in the PRS depends on the cumulative incidence and the magnitude of heritability. When the same level of cumulative incidence is reached, the allele distribution difference between diagnosed cases and a remaining unaffected population is identical. It depends on the value of the cumulative incidence, and not on the incidence pattern that led to achieving a particular value of the cumulative incidence in the validation scenarios.

Next, the simulations were performed using the incidence rate patterns and model genetic architectures of each analyzed LOD, determining allele distribution change with age, and the resulting impact on GWAS discovery power. These results compared well with findings from clinical, GWAS, and familial heritability studies, summarized below.

There are numerous reports of diminishing with age heritability, clinical predictive power, and GWAS discovery power for these LODs. (Almgren et al., 2011) observe T2D heritability equal 0.69 for the 35–60 year age of onset group, and negligible heritability for older ages. GWAS studies (de Miguel-Yanes et al., 2011; Mühlenbruch et al., 2013) segregating T2D risk SNPs by age find that the risk factor values are higher in the younger than 50 years age group, compared to the older cohorts. Regarding the variant types that are most likely associated with T2D, (Fuchsberger et al., 2016) find that, with a high degree of certainty, they can attribute T2D liability to common variants, not rare high-effect variants. A cardiovascular disease (myocardial infarction) study by (Nielsen et al., 2013) finds that predictive power of parental history decreases for ages older than 50. (Schulz et al., 2004) finds that familial history is the best predictor of ischemic stroke for individuals under the age of 60. A review based on Framingham's study (Seshadri et al., 2010) finds that the parental predictive power of stroke diminishes for ages above 65. The heritability of Alzheimer's disease is estimated at 80% from twin studies (Naj and Schellenberg, 2017); the heritability is 79% (Gatz et al., 2006) at approximately 65 years of age and diminishes with increasing age. GWAS studies (Tan et al., 2013; Shen and Jia, 2016; Naj and Schellenberg, 2017) come to similar conclusions. As a summary, the familial history predictive power for the above LODs is highest for younger ages, specifically < 65 years of age for AD, < 50 for CAD, < 60 for stroke, and < 50 for T2D.

For the above LODs, the simulations results show high PRS for the earliest diagnosed cases. The risk allele case/control difference and PRSs of newly diagnosed cases decrease rapidly with age progression. At a very old age, the individuals whose genotype would be considered a low risk at an earlier age are the ones diagnosed with the disease; Fig. 2. This also reinforces the validity of the clinical observation that the major risk factor for LODs is age itself.

The four cancers display a noticeably different pattern. The PRSs for earliest onset cases are lower than for the above LODs, and this risk changes much less with age than for the above LODs. These results explain the familial heritability studies observations: for three out of the four most prevalent cancers, the twin studies has shown relatively constant heritability with age progression (Grönberg, 2003; Möller et al., 2016; Mucci et al., 2016; Graff et al., 2017). Determining lung cancer heritability change with age has proven somewhat more elusive for researchers (Hjelmborg et al., 2016), and no definitive conclusions have been published, to a large extent due to generally low documented heritability and substantial environmental component of this disease.

Prostate cancer is the only cancer that is somewhat controversial. Its heritability is reported at 57% by (Hjelmborg et al., 2014), and prostate cancer reaches the highest maximum instance rate of the four most prevalent cancers reviewed. Therefore, the relative MAF between cases and controls is likely to be higher than for other cancers, according to the above observations and the results of the validation simulations. Yet, the same article finds that the heritability of prostate cancer remains stable with age. Possible explanations may be that either this twin study result is somehow biased and the heritability of prostate cancer is lower than stated in (Hjelmborg et al., 2014), or perhaps this is a phenomenon specific to the populations or environmental effects of Nordic countries. Perhaps the earlier familial study (Grönberg, 2003), which reported a heritability estimate of 42%, would be closer to the UK population incidence data used here. The verification simulation with 42% heritability showed the re-

sults that closer matched the patterns exhibited by other cancers, see the resulting value shown in parentheses for the cases multiple in Table 4. A more exhaustive literature investigation of the reviewed LODs is presented in Chapter S1 in Supplementary File 1.

GWAS statistical discovery power is impaired by the change in individual distribution of the PRS at older ages. A larger number of cases/controls is needed at older ages to achieve the same statistical discovery power. The first four LODs, which exhibit larger heritability and cumulative incidence compared to cancers, require an increased number of participants in a case/control study for older ages. The cancers show a small increase in the number of participants required to achieve the same statistical power.

Individual values analysis, in which the individuals diagnosed each year are compared to all not-yet-sick individuals, shows fast increase in the number of cases hypothetically needed to achieve the same statistical power, but this scenario would be not practical for a clinical study. The age matched cohort studies benefit from the fact that the diagnosed individuals are accumulated from the youngest onset to the age of being a case in the cohort study, as well as averaging over the cohort age range, resulting in more moderate increase in number of participants required, or slower decrease in GWAS discovery power for older cohorts. The age matched cohort studies would require 1.5–2.1 times larger number of participants at age 80 compared to youngest possible age matched cohorts in case of stroke, CAD, AD, and T2D.

Designing cohorts composed of the youngest possible cases and the oldest available controls improves the GWAS discovery power due to larger difference in risk allele frequency between cases and controls. This improvement leads to a lower number of participants being needed for GWAS when applied to the highest cumulative incidence and heritability LODs—so much so that about 50% fewer participants are required to achieve the same GWAS statistical power with control cohorts aged between 90 and 100 years matched to the youngest case cohorts, as opposed the opposite same magnitude effect in older age matched cohorts. In this scenario, notably (20–25%) fewer participants are also needed to achieve the same statistical power in cancer GWAS, including lung cancer.

Conclusions

This research was conducted with the goal to establish whether any of the observational phenomena, including decreasing heritability with age for some notable LODs, and the limited success of LOD GWAS discovery can be explained by allele proportions changing between cases and controls due to the higher odds of more susceptible individuals being diagnosed at an earlier age.

The simulation results reported above show that these phenomena can indeed be explained and predicted by the heritability of the LODs and their cumulative incidence progression. Simulating population age progression under the assumption of relative disease liability remaining proportionate to individual polygenic risk, confirmed that individuals with higher risk will become ill and diagnosed proportionately earlier, leading to a change in the distribution of risk alleles between new cases and the as-yet-unaffected population in every subsequent year of age. With advancing age, the mean polygenic risk of the aging population declines. The fraction of the highest risk individuals diminishes even faster.

While the number of the most susceptible individuals and the mean population susceptibility decreases, the incidence of all LODs initially grows exponentially, doubling in incidence every 5 to 8.5 years (see Methods), and remains high at older ages, leading to a high cumulative incidence for some LODs. The increasing incidence rate while polygenic risk decreases for the as-yet-unaffected population can be explained as a consequence of the aging process, which itself is the major LOD risk factor. At old age people who have low genetic or familial susceptibility are increasingly becoming sick with an LOD.

LODs with low cumulative incidence and low familial heritability produce a smaller change in the allele distribution between affected individuals and the remaining population. Most prevalent cancers have a reported stable heritability with age, and therefore these GWASs are less affected by the increasing age of the participant cohorts. For these diseases the modification of the cohort age selection favoring younger cases and older controls will also lead to noticeable improvement in statistical power, somewhat less prominent than for the higher incidence and heritability LODs.

Four of the most prevalent LODs—Alzheimer's disease, coronary artery disease, cerebral stroke, and type 2 diabetes—exhibit both a high cumulative incidence at older age and high heritability. These simulation results show that a GWAS study of any polygenic LOD that displays both high cumulative incidence at older age and high initial familial heritability will benefit from using the youngest possible participants as cases rather than age matching or statistically adjusting or compensating for age. In addition, GWAS cohort studies would benefit from using as controls participants who are as old as possible. This would allow for an additional increase in statistical discovery power due to the higher difference in risk allele frequency between cases and controls. For most LODs, there is an ample number of still-unaffected individuals at older ages who can participate in the control cohorts.

Competing interests

The author declares that there are no competing interests.

Grant information

The author declares that no grants were involved in supporting this work.

Acknowledgments

The author thanks Alexei J. Drummond at the University of Auckland for a number of helpful and challenging discussions.

Supplementary material

Supplementary File 1 Supplementary chapters and figures are attached after the main manuscript.

Data availability

Supplementary Data 1 A zip file containing the simulation executable, the source code, R scripts, batch files, and simulation results.

References

V. Agarwala, J. Flannick, S. Sunyaev, D. Altshuler, and G. Consortium. Evaluating empirical bounds on complex disease genetic architecture. *Nature genetics*, 45(12):1418–1427, 2013.

P. Almgren, M. Lehtovirta, B. Isomaa, L. Sarelin, M. Taskinen, V. Lyssenko, T. Tuomi, L. Groop, and B. S. Group. Heritability and familiality of type 2 diabetes and related quantitative traits in the botnia study. *Diabetologia*, 54(11):2811, 2011.

C. A. Anderson, N. Soranzo, E. Zeggini, and J. C. Barrett. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS biology*, 9(1): e1000580, 2011.

T. Bjørnland, A. Bye, E. Ryeng, U. Wisløff, and M. Langaas. Powerful extreme phenotype sampling designs and score tests for genetic association studies. *Statistics in medicine*, 2018.

M. W. Boehme, G. Buechele, J. Frankenhauser-Mannuss, J. Mueller, D. Lump, B. O. Boehm, and D. Rothenbacher. Prevalence, incidence and concomitant co-morbidities of type 2 diabetes mellitus in south western germany-a retrospective cohort and case control study in claims data of a large statutory health insurance. *BMC Public Health*, 15(1):855, 2015.

R. Brookmeyer, S. Gray, and C. Kawas. Projections of alzheimer's disease in the united states and the public health impact of delaying disease onset. *American journal of public health*, 88 (9):1337–1342, 1998.

Cancer Statistics for the UK. *Available at http://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk (accessed November 10, 2018)*, 2018.

N. Chatterjee, J. Shi, and M. García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17(7):392, 2016.

A. J. Clarke and D. N. Cooper. Gwas: heritability missing in action? *European Journal of Human Genetics*, 18(8):859, 2010.

J. Collerton, K. Davies, C. Jagger, A. Kingston, J. Bond, M. P. Eccles, L. A. Robinson, C. Martin-Ruiz, T. von Zglinicki, O. F. James, et al. Health and disease in 85 year olds: baseline findings from the newcastle 85+ cohort study. *Bmj*, 339:b4904, 2009.

D. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

J. M. de Miguel-Yanes, P. Shrader, M. J. Pencina, C. S. Fox, A. K. Manning, R. W. Grant, J. Dupuis, J. C. Florez, R. B. D'agostino, L. A. Cupples, et al. Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. *Diabetes care*, 34 (1):121–125, 2011.

S. P. Dickson, K. Wang, I. Krantz, H. Hakonarson, and D. B. Goldstein. Rare variants create synthetic genome-wide associations. *PLoS biology*, 8(1):e1000294, 2010.

S. D. Edland, W. A. Rocca, R. C. Petersen, R. H. Cha, and E. Kokmen. Dementia and alzheimer disease incidence rates do not vary by sex in rochester, minn. *Archives of neurology*, 59(10): 1589–1593, 2002.

A. Eyre-Walker. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107(suppl 1): 1752–1756, 2010.

D. Falconer. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of human genetics*, 31(1):1–20, 1967.

D. S. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, 29(1):51–76, 1965.

L. A. Farrer, L. A. Cupples, J. L. Haines, B. Hyman, W. A. Kukull, R. Mayeux, R. H. Myers, M. A. Pericak-Vance, N. Risch, and C. M. Van Duijn. Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease: a meta-analysis. *Jama*, 278(16):1349–1356, 1997.

N. S. Fedarko. Theories and mechanisms of aging. In *Geriatric Anesthesiology*, pages 19–25. Nature Publishing Group, 2018.

C. Franceschi, P. G. Garagnani, C. Morsiani, M. Conte, A. Santoro, A. Grignolio, D. Monti, M. Capri, and S. Salvioli. The continuum of aging and age-related diseases: common mechanisms but different rates. *Frontiers in Medicine*, 5:61, 2018.

C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, . C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, 2016.

M. Gatz, C. A. Reynolds, L. Fratiglioni, B. Johansson, J. A. Mortimer, S. Berg, A. Fiske, and N. L. Pedersen. Role of genes and environments for explaining alzheimer disease. *Archives of general psychiatry*, 63(2):168–174, 2006.

D. B. Goldstein. Common genetic variation and human traits. *New England Journal of Medicine*, 360(17):1696, 2009.

R. E. Graff, S. Möller, M. N. Passarelli, J. S. Witte, A. Skytthe, K. Christensen, Q. Tan, H.-O. Adami, K. Czene, and J. R. Harris. Familial risk and heritability of colorectal cancer in the nordic twin study of cancer. *Clinical Gastroenterology and Hepatology*, 15(8):1256–1264, 2017.

H. Grönberg. Prostate cancer epidemiology. *The Lancet*, 361 (9360):859–864, 2003.

L. E. Hebert, P. A. Scherr, L. A. Beckett, M. S. Albert, D. M. Pilgrim, M. J. Chown, H. H. Funkenstein, and D. A. Evans. Age-specific incidence of alzheimer's disease in a community population. *Jama*, 273(17):1354–1359, 1995.

J. Hjelmborg, T. Korhonen, K. Holst, A. Skytthe, E. Pukkala, J. Kutschke, J. R. Harris, L. A. Mucci, K. Christensen, K. Czene, et al. Lung cancer, genetic predisposition and smoking: the nordic twin study of cancer. *Thorax*, pages thoraxjnl–2015, 2016.

J. B. Hjelmborg, T. Scheike, K. Holst, A. Skytthe, K. L. Penney, R. E. Graff, E. Pukkala, K. Christensen, H.-O. Adami, N. V. Holm, et al. The heritability of prostate cancer in the nordic twin study of cancer. *Cancer Epidemiology and Prevention Biomarkers*, 23(11):2303–2310, 2014.

E. P. Hong and J. W. Park. Sample size and statistical power calculation in genetic association studies. *Genomics & informatics*, 10(2):117–122, 2012.

E. Kokmen, V. Chandra, and B. S. Schoenberg. Trends in incidence of dementing illness in rochester, minnesota, in three quinquennial periods, 1960–1974. *Neurology*, 38(6):975–975, 1988.

K. B. Kuchenbaecker, J. L. Hopper, D. R. Barnes, K.-A. Phillips, T. M. Mooij, M.-J. Roos-Blom, S. Jervis, F. E. Van Leeuwen, R. L. Milne, and N. Andrieu. Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers. *Jama*, 317(23):2402–2416, 2017.

S. K. Kumar, M. W. Feldman, D. H. Rehkopf, and S. Tuljapurkar. Limitations of gcta as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences*, 113 (1):E61–E70, 2016.

A. Li and D. Meyre. Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. *International journal of obesity*, 37(4):559, 2013.

D.-Y. Lin, R. Tao, W. D. Kalsbeek, D. Zeng, F. Gonzalez II, L. Fernández-Rhodes, M. Graff, G. G. Koch, K. E. North, and G. Heiss. Genetic association analysis under complex survey sampling: the hispanic community health study/study of latinos. *The American Journal of Human Genetics*, 95(6):675–688, 2014.

J. Luan, M. Wong, N. Day, and N. Wareham. Sample size determination for studies of gene-environment interaction. *International Journal of Epidemiology*, 30(5):1035–1040, 2001.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, and A. Chakravarti. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

S. Möller, L. A. Mucci, J. R. Harris, T. Scheike, K. Holst, U. Halekoh, H.-O. Adami, K. Czene, K. Christensen, N. V. Holm, et al. The heritability of breast cancer among women in the nordic twin study of cancer. *Cancer Epidemiology and Prevention Biomarkers*, 25(1):145–150, 2016.

L. A. Mucci, J. B. Hjelmborg, J. R. Harris, K. Czene, D. J. Havelick, T. Scheike, R. E. Graff, K. Holst, S. Möller, R. H. Unger, et al. Familial risk and heritability of cancer among twins in nordic countries. *Jama*, 315(1):68–76, 2016.

K. Mühlenbruch, C. Jeppesen, H.-G. Joost, H. Boeing, and M. B. Schulze. The value of genetic information for diabetes risk prediction–differences according to sex, age, family history and obesity. *PLoS One*, 8(5):e64307, 2013.

S. L. Murphy, J. Xu, K. D. Kochanek, S. C. Curtin, and E. Arias. Mortality in the united states, 2016. *NCHS Data Brief, no 293*, 293, 2017.

A. C. Naj and G. D. Schellenberg. Genomic variants, genes, and pathways of alzheimer's disease: An overview. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 174 (1):5–26, 2017.

M. Nielsen, C. Andersson, T. A. Gerds, P. K. Andersen, T. B. Jensen, L. Køber, G. Gislason, and C. Torp-Pedersen. Familial clustering of myocardial infarction in first-degree relatives: a nationwide study. *European heart journal*, 34(16):1198–1203, 2013.

M. Noh, B. Yip, Y. Lee, and Y. Pawitan. Multicomponent variance estimation for binary traits in family-based studies. *Genetic epidemiology*, 30(1):37–47, 2006.

T.-L. North and M. Beaumont. Complex trait architecture: the pleiotropic model revisited. *Scientific reports*, 5:9351, 2015.

J. Oh, Y. D. Lee, and A. J. Wagers. Stem cell aging: mechanisms, regulators and therapeutic opportunities. *Nature medicine*, 20 (8):870, 2014.

Online Sample Size Estimator. *The Bioinformatics Institute. Available at http://osse.bii.a-star.edu.sg/ (accessed November 10, 2018)*, 2018.

J.-H. Park, M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, Z. Wang, S. J. Chanock, J. F. Fraumeni, and N. Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108(44):18026–18031, 2011.

Y. Pawitan, K. C. Seng, and P. K. Magnusson. How many genetic variants remain to be discovered? *PloS one*, 4(12):e7969, 2009.

P. Rothwell, A. Coull, L. Silver, J. Fairhead, M. Giles, C. Lovelock, J. Redgrave, L. Bull, S. Welch, and F. Cuthbertson. Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (oxford vascular study). *The Lancet*, 366(9499):1773–1783, 2005.

U. Schulz, E. Flossmann, and P. Rothwell. Heritability of ischemic stroke in relation to age, vascular risk factors, and subtypes of incident stroke in population-based studies. *Stroke*, 35(4): 819–824, 2004.

S. Seshadri, A. Beiser, A. Pikula, J. J. Himali, M. Kelly-Hayes, S. Debette, A. L. DeStefano, J. R. Romero, C. S. Kase, and P. A. Wolf. Parental occurrence of stroke and risk of stroke in their children: the framingham study. *Circulation*, 121(11): 1304–1312, 2010.

P. C. Sham and S. M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335, 2014.

L. Shen and J. Jia. An overview of genome-wide association studies in alzheimer's disease. *Neuroscience bulletin*, 32(2): 183–190, 2016.

C. T. Silva, J. A. Kors, N. Amin, A. Dehghan, J. C. Witteman, R. Willemsen, B. A. Oostra, C. M. van Duijn, and A. Isaacs. Heritabilities, proportions of heritabilities explained by gwas findings, and implications of cross-phenotype effects on pr interval. *Human genetics*, 134(11-12):1211–1219, 2015.

H.-C. So, A. H. Gui, S. S. Cherny, and P. C. Sham. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genetic epidemiology*, 35(5): 310–317, 2011.

M. Song, P. Kraft, A. D. Joshi, M. Barrdahl, and N. Chatterjee. Testing calibration of risk models at extremes of disease risk. *Biostatistics*, 16(1):143–154, 2014.

L. Tan, J.-T. Yu, W. Zhang, Z.-C. Wu, Q. Zhang, Q.-Y. Liu, W. Wang, H.-F. Wang, X.-Y. Ma, and W.-Z. Cui. Association of gwas-linked loci with late-onset alzheimer's disease in a northern han chinese population. *Alzheimer's & dementia*, 9(5):546–553, 2013.

K. R. Thornton, A. J. Foran, and A. D. Long. Properties and modeling of gwas when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. *PLoS genetics*, 9(2):e1003258, 2013.

US Social Security Actuarial Life Table. *Social Security Administration (US). Available at https://www.ssa.gov/oact/STATS/table4c6.html (accessed November 10, 2018)*, 2014.

P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

Z. Wang et al. Converting odds ratio to relative risk in cohort studies with partial data information. *J Stat Softw*, 55(5): 1–11, 2013.

A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11): 1173–1186, 2014.

J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, and R. J. Loos. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375, 2012.

J. Yang, A. Bakshi, Z. Zhu, G. Hemani, A. A. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. Perry, I. M. Nolte, and J. V. van Vliet-Ostaptchouk. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*, 47(10):1114, 2015.

N. Zaitlen and P. Kraft. Heritability in the genome-wide association era. *Human genetics*, 131(10):1655–1664, 2012.

N. Zaitlen, S. Lindström, B. Pasaniuc, M. Cornelis, G. Genovese, S. Pollack, A. Barton, H. Bickeböller, D. W. Bowden, S. Eyre, et al. Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS genetics*, 8 (11):e1003032, 2012.

J. Zhang and F. Y. Kai. What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *Jama*, 280(19):1690–1691, 1998.

Age-related late-onset disease heritability patterns and implications for genome-wide association studies (SUPPLEMENTARY)

Roman Teo Oliynyk[1,2,*]

[1]**Centre for Computational Evolution, University of Auckland, Auckland 1010, New Zealand**
[2]**Department of Computer Science, University of Auckland, Auckland 1010, New Zealand**
[*]**roli573@aucklanduni.ac.nz**

Contents:

1 Supplementary Chapter S1
2 Supplementary Chapter S2
3 Supplementary Chapter S3
4 SUPPLEMENTARY FIGURES

Supplementary Chapter S1: LOD heritability patterns with age based on familial, GWAS and clinical studies

If we were to make a general statement that the heritability of LODs always decreases with age, we would not be entirely correct. A review of the clinical, GWAS and familial studies on polygenic LOD heritability within the typical age range of disease onset leads to a grouping of LODs into two broad categories: those with decreasing heritability with age and those with increasing or relatively constant heritability with age.

Next, we review these categories in detail, focusing primarily on eight of the most prevalent LODs we analyze in our simulations. We use these categories to organize the observational knowledge so we can apply this knowledge to the main article simulations and in turn verify the simulation results.

LODs with decreasing heritability with age

There is a large number of ***highly environmentally affected LODs that exhibit decreasing heritability with age***. Three are some of the highest lifetime risk diseases: coronary artery disease, cerebral stroke, and type 2 diabetes; see Table 1 summarized from [1, 2, 3], and [4].

Supplementary Table 1. Population statistics of LODs characterized by decreasing heritability with age

Statistic	Alzheimer's	CAD	Stroke	T2D
Lifetime risk, USA (%)	10m, 20w	49m, 32w	25m, 30w	55
Mortality assigned, USA (%)	4.2	23.1	5.2	2.9
Heritability (%)	79	50–60	38–44	69
Best predictability, age	< 65	< 55	< 60	< 50

Lifetime risk numbers, when marked, "w" for women, "m" for men.

As early as 1967, Falconer [5] noted that "*the increase of incidence associated with a variable age of onset can be due to either an increase of the mean liability or an increase of the variance of liability. Consideration of the changes of liability that individuals may undergo as they grow older shows that an increase of variance with increasing age is to be expected, and since the additional variance is likely to be mainly environmental, a reduction of the heritability is to be expected.*" Falconer further pointed out that "*the heritability of liability to diabetes, estimated from the sib correlation, decreases with increasing age. For people under 10, heritability is about 70 or 80%, and it drops to about 30 or 40% in people aged 50 and over. The decrease of the heritability is attributable to an increase of environmentally caused variation. The increased environmental variation is not enough to account in full for the increasing incidence, so there is probably also an increase of the mean liability with increasing age.*"

In the 1960s, the distinction between autoimmune Mendelian type 1 diabetes and late-onset polygenic type 2 diabetes (T2D) was not known, but it was suspected that there may be two distinct mechanisms. However, this conclusion of an increase in liability with age, and accordingly blurred heritability, is observed for T2D and other LODs.

The greatest heritability for T2D is observed in the 35–60 (0.69) year age of onset group [6]), and heritability dropped to only 0.31 when the upper age range was increased to 75 (making the age range 35–75). In the over-60 group, the "environmental" component is the primary cause of new T2D cases. The environmental component in this case includes systemic and tissue-specific deterioration with age and the cumulative external environmental effects with increased time duration. Just as Falconer did 60 years earlier, the authors note that T2D heritability decreases with age, and liability may be more accurately predicted in younger individuals.

One review [7] cites two studies that corroborate the authors' view. One concludes that recalculating the genetic risk for T2D by splitting a cohort by age below and above 50 years using 40 T2D risk SNPs finds that the risk factor values are higher in the younger group [8]. Meanwhile, [6] correlates the heritability and familiality of T2D with quantitative traits and finds a very significant drop in heritability over the age of 60.

The conclusion is that, for reliable GWASs, younger is better: T2D patients under the age of 60—or, even better, under the age of 50—should be chosen. Regarding the variant types that are most likely associated with T2D,

[9] finds that, with a high degree of certainty, they can attribute T2D liability to common variants, not rare high-effect variants.
Nielsen et al.'s cardiovascular disease (myocardial infarction) study [10] provides implicit confirmation of decreasing heritability with age. The predictive power of parental history is as follows: paternal RR = 3.30 for ages <50 and RR = 1.83 for ages >50; maternal RR = 3.23 for ages <50 and RR = 2.31 for ages >50.
Schulz [11] finds that familial history is the best predictor of **ischemic stroke** for individuals under the age of 60, with an overall OR of 1.73. Relative OR compared to the under-60 cohort was 0.95 for the 60–70 age band and 0.77 for individuals over the age of 75.
A review based on Framingham's study [12] supplies very useful information about parental history of stroke. Even though the grouping on the parental side is stroke under 65, on the descendant side, there are statistics showing RR both below and above 65 years of age. For descendants whose parents had a stroke before age 65, the stroke RR was determined. Overall, RR = 3.79 under 65 years and 2.21 over 65 years; ischemic stroke HR = 5.45 under 65 years and 2.47 over 65 years. Additional implicit information from this data, which supports the same conclusion, is listed in [13].
The heritability patterns for these diseases are summarized in Table 3. There is qualitative and, increasingly, quantitative knowledge about the progressively declining heritability of these diseases at ages above 50, as well as the decreasing associated familial and GWAS predictive power; see [10, 11, 12, 14, 3], and [9]. These studies find that familial history is the better predictor of next-generation disease only when the parental generation participants are relatively young; see [8, 7, 6], and Table 1.
An environmental effect on the heritability of cardiovascular disease and T2D with age is evident [5, 15], including influences such as spousal environment [16].
In addition, T2D is a major co-morbidity factor for CAD and cerebral stroke, as well as causally correlated adiposity and hypertension, which are by themselves associated with CAD and cerebral stroke and other LODs. In the presence of T2D, these diseases develop years and even decades earlier than the typical onset ages [17]. For instance, twin studies on the heritability of BMI (a co-morbidity often preceding T2D) show the highest heritability of 85% at 18 years of age, after which heritability slowly declines throughout the lifespan [18].
It must be noted that the majority of diseases are influenced to various degrees by environmental factors. The three diseases we just reviewed show incomparably higher environmental influence than Alzheimer's disease (AD). For AD, not only lifestyle but also painstakingly developed medications can barely influence the progression of the disease. In contrast, CAD, cerebral stroke and T2D are often considered by the medical community as primarily influenced by lifestyle and environment [19, 20, 17, 21].
In conclusion, the highly prevalent LODs exhibiting high environmental correlation with onset ages also show decreasing heritability with age. This is combined with an exponential incidence increase with age. In the case of CAD and cerebral stroke, the exponential incidence rate increase proceeds beyond 80 years of age.
Another type of *LOD showing heritability that declines with age can be described as a mode of failure with aging.* Alzheimer's disease begins relatively late, but from there, its incidence rises exponentially to extremely old age [22]. The heritability of Alzheimer's disease is estimated at 80% from twin studies [23]; the heritability is 79% [24] at approximately 65 years of age and diminishes with increasing age, as indicated both by familial and GWAS studies [25, 26, 23].
A clinical study documenting the association between the APOE genotype and Alzheimer's disease [27, 28] reports the change in odds ratio with age of APOE e4/e4 and APOE e3/e4 carriers, which we summarize for the Caucasian population in Table 2.

Supplementary Table 2. Alzheimer's disease odds ratio by age and APOE alleles, relative to e3/e3 allele carriers

APOE allele / Age (y)	55	60	65	70	75	80	85	90
e4/e4 OR	14.1	15.0	14.3	12.1	9.5	6.1	3.7	2.0
e4/e3 OR	3.5	3.7	3.8	3.6	3.3	2.7	2.3	1.7

Values summarized from [27].

Another review [23] concludes that the typical age at onset is 68.8 years for APOE e4/e4 carriers, 75.5 years for e3/e4 carriers, and 84.3 years for carriers without e4. Moreover, the APOE e4 effect is age dependent, giving a broad stroke assessment that the e4 allele effect is most prominent between 60 and 79 years and gradually diminishes after 80 years of age. This fits well with the assessment [27] summarized in Table 2.
In Table 3, we summarize the information in the literature about the decreasing heritability of the LODs referenced above.

Supplementary Table 3. Heritability and risk statistics for decreasing heritability with age LODs

Disease	*Heritability/risk, younger age*	*Heritability/risk, older age*
AD e3/e4 [27]	OR=3.8, 65y	OR=1.7, 90y
AD e4/e4 [27]	OR=15.0, 60y	OR=2.0, 90y
CAD paternal [10]	RR=3.30, $< 50y$	RR=1.83, $> 50y$
CAD maternal [10]	RR=3.23, $< 50y$	RR=2.31, $> 50y$
Stroke [11]	OR=1.63, $< 60y$	OR=0.77, $> 70y$
Stroke all [12]	RR=3.79, $< 65y$	RR=2.21, $> 65y$
Stroke ischemic [12]	RR=5.45, $< 65y$	RR=2.47, $> 65y$
T2D [6]	$h^2 = 0.69$, 35-60y	$h^2 = 0.31$, 35-75y

OR = odds ratio; RR = relative risk; h^2 = heritability

The model presented in [22] hypothesized that if the AD incidence curve could be delayed by five years, the overall prevalence of AD would be half the projected rate, assuming unchanged mortality from other causes. AD prevalence in this study is limited by applying a 1.4 mortality multiplier to AD patients compared with the unaffected population.
While AD progression is difficult to influence with lifestyle

changes or medications, AD incidence at comparable ages has decreased by about 30% since the 1980s in many Western countries [29, 30] due to undetermined causes. As life expectancy increases, AD lifetime incidence and prevalence are expected to regain ground.

In conclusion, AD shows an exponentially increasing incidence rate up to the most advanced ages, while also displaying heritability that declines with age.

LODs with stable heritability with age

We group LODs with relatively constant heritability with age and infrequent types of LODs with increasing heritability with age in this category. As found in the reviewed literature, the increase in heritability, when observed, is moderate. We find that the diseases showing *slightly increasing heritability with age* are those affecting the skeletal system, for instance, osteoarthritis, particularly of large joints such as the hip or lower back. One study [31] shows that both the incidence and heritability of advanced osteoarthritis of the hip and lower back increase with age.

It is evident that younger cases are more environmentally and less genetically correlated. For example, osteoarthritis at a younger age is often due to trauma rather than genetics [32, 33]. At the age of 60, the genetic and environmental components are both close to 50%, and by the age of 70, the heritability increases to 75% and stays close to this level into the 90s. Heritability is even higher and increases with advanced age for osteoarthritis of the spine at multiple locations [34].

We note that the increase in heritability for these diseases is relatively modest and extends from an initially high level. Many osteoarthritis-affected structures and corresponding diagnoses, with different ages of maximum incidence and heritability by sex and age, do not follow this pattern [35].

The osteoporosis findings are similarly varied, with studies finding no heritability of pathology for some bone structures and strong heritability for others [36]. Specifically, the osteoporosis associated with bone breaks is very heritable and shows a slight increase in heritability into older age [37]. This is explicable by the fact that, for osteoporosis, the main risk component—the shape and size of the bone—is strongly heritable. Genetics in this case determines the early developmental stages of an organism, when the structures take shape. Similar reasoning applies to osteoarthritis, which is related to defects in collagen and connective tissue formation. The malignancy happens after many decades of life, when wear, deterioration and diminishing repair capacity cross the threshold leading to pathology.

In conclusion, we find that some LODs based on early development of an organism's structures may display strong heritability late in life and even increasing diagnostic heritability as aging progresses. GWAS has found only a small set of SNPs that provide very limited risk prediction for these diseases [38, 33]. Apparently, the research cannot be impeded by the increasing heritability with age of the GWAS study cohorts.

Relatively stable heritability with advancing age is a distinguishing feature of cancers. Accurate information about heritability at different ages is not sufficiently explored for most cancers. Fortunately, during this decade, a number of studies has shed light on the age-related heritability of three out of the four most prevalent cancers, and these allow us to extrapolate the expectations to the fourth: lung cancer.

The lifetime risk of developing any type of cancer in the US is 38% for women and 40% for men [39], and the 2016 fraction of mortality death directly attributed to cancer was 21.8%, the second-highest after heart disease [40]. In the UK, the corresponding numbers are higher, at 47% and 53%, respectively [41, 42], with the higher likelihood perhaps attributable to the UK's longer life expectancy. Each specific type of cancer constitutes a small fraction of overall lifetime risk, with breast, prostate, lung, and colorectal cancer being the four most prevalent.

Next, we summarize the latest heritability and incidence research for these four cancers.

Breast cancer (BC) Breast cancer (BC) is a well-researched cancer, with studies delving into all aspects of BC. Like prostate cancer, the two largest genetic predictors for BC are mutations in the BRCA1 and BRCA2 genes. The BRCA1/2 genes are involved in the homologous repair of double-stranded DNA breaks, working in combination with at least 13 known tumor suppressor proteins [43]. Defects in BRCA1/2 proteins disable homologous double-stranded DNA break repair, and the cell falls back on the use of imprecise non-homologous repair mechanisms; this leads to the accumulation of mutations, eventually leading to cancer. BRCA1/2 mutations are the most important predictor of breast cancer. The review by Haley [43] states that the frequency of BRCA mutations varies with geographic location and ethnicity, ranging from a 0.02% mutation carrier rate in some populations to 2.6% in the Ashkenazi Jewish population due to ancient founder mutations. Other founder mutations have been reported in the Dutch, Swedish, French Canadian, Icelandic, German, and Spanish populations. In Ontario, Canada, for instance, the frequency of mutation carriers is 0.32% for BRCA1 and 0.69% for BRCA2 [44].

An early study [45] analyzing families with at least four cases of BC found that the disease was linked to BRCA1 in 52% of cases and BRCA2 in 32% of cases (with only 16% remaining for other causes). Taking into account ovarian cancer in addition to BC resulted in 81% of cases being due to BRCA1, while 76% of cases in families with both male and female BC were due to BRCA2.

The lifetime risk of BC for women both in the US and the UK is 12% [39, 42]. As Haley [43] summarized, carriers of BRCA1 have a lifetime risk of developing BC equal to 60–70%, and an additional 40% risk of developing ovarian, fallopian, or primary peritoneal cancers. For BRCA2 carriers, the risks are 45–55% for BC and 25% for ovarian cancer. These numbers closely correspond to the aforementioned study [45].

Moller et al. [46] present in-depth breast and ovarian cancer heritability by age data for BRCA1/2 carriers.

The study shows that the genetic liability, while exhibiting a slight downward trend, remains relatively constant and exceeds the common environmental component at all ages.

One of the most recent studies [47] provides further clarification, stating that BC incidences increase rapidly in early adulthood until the ages of 30 to 40 for BRCA1 carriers and until the ages of 40 to 50 for BRCA2 carriers, and thereafter remain at a relatively constant incidence rate of 2–3% per year until at least 80 years of age; see Table 4. Our calculations based on this data show that the initial incidence ramp-up is exponential before turning into the constant horizontal incidence rate approximation; a logistic approximation also fits. The exponential doubling rate, until it reaches the constant incidence level, is also consistent with all other diseases we reviewed, showing a five-year doubling incidence time for BRCA1 and eight years for BRCA2 (the BRCA1 calculation, based only on two data points, is less accurate). A much earlier review study [48] collects the same kind of statistics as [47] and arrives at similar conclusions.

Supplementary Table 4. BRCA1/2 carriers incidence rate by age, data from [47]

Gene	≤20	21-30	31-40	41-50	51-60	61-70	71-80
BRCA1 (%)	0	0.59	2.35	2.83	2.57	2.50	1.65
BRCA2 (%)	0	0.48	1.08	2.75	3.06	2.29	2.19
BRCA1 cum risk (%)	0	4	24	43	56	66	72
BRCA2 cum risk (%)	0	4	13	35	53	61	69

Moller et al.'s study [46] finds a somewhat lower lifetime BC risk of 8.1% in Nordic countries compared to 12% in the US and estimates heritability at 31%.

In addition to BRCA1/2, [49] and [43] also list a number of high penetrance gene mutations—the TP53, PTEN, STK11, and CDH-1 gene mutations—giving a lifetime probability of cancers in general of about 90% and specifically a female breast cancer probability above 50%.

Several rare gene mutations are also associated with a breast cancer relative risk in the range of 1.5–5.0: CHEK2, PALB2, ATM, BRIP1CHEK2, PALB2, ATM, and BRIP1. In aggregate, the above high-effect mutations are correlated with only approximately 10% of hereditary breast cancers [44, 43].

To date, GWAS attempts to find common polygenic variants of low effect size have had only limited success. One review study [50] outlines the history and accomplishments of breast cancer GWAS over a decade of research. The most recent high-powered consortium study [51] included 122,977 cases and 105,974 controls of European ancestry and 14,068 cases and 13,104 controls of East Asian ancestry. The study verified 102 previously reported SNPs, finding that 49 of them were reproducible. The study also found that the majority of discovered SNPs reside in non-coding areas of the genome. The discovered set of polygenic SNPs allows for the explanation of approximately 4% of heritability on top of the 14% explained by known high-penetrance SNPs, bringing the predictive power to 18%. This GWAS estimates the familial heritability of breast cancer at 41%—a possible exaggeration because it significantly exceeds the 31% estimated by [46] and the 27% estimated by [52].

Breast cancer conclusions: The familial heritability studies and BRCA1/2 clinical studies show that breast cancer heritability is relatively constant over the age of 40 for both mutations. A number of high-penetrance gene mutations can explain an additional fraction of heritability, totaling 10–14%.

The GWAS study described above [51] also finds that multiple SNPs located in non-coding areas are correlated with the candidate gene promoters and activity modifier areas. This improves the possibility that the common variant component may be able to explain a larger fraction of heritability. It appears at this time, based on Moller et al.'s statistics [46], that breast cancer heritability for the polygenic component may also be relatively constant after the age of 40 or may only slightly decline with age.

Prostate cancer (PC) The effects and risks of the BRCA1/2 genes and their mutations described in the breast cancer section apply in a very similar way to the PC incidence.

A study by Le Carpentier ([53]) found that lifetime PC risks are approximately 20% for BRCA1 mutations carriers and 40% for BRCA2 mutation carriers, while, overall, BRCA1/2 is associated with only 2% of all PC cases. In addition, BRCA1/2 accounts for 10% of male breast cancer cases. The lifetime risk of male breast cancer in mutation carriers is estimated be 5–10% for BRCA1 mutations and 1–5% for BRCA2 mutation carriers. Therefore, compared to breast cancer, BRCA1/2 mutations are associated with a smaller fraction of heritability.

The lifetime risk of PC in men is estimated at 6% for Danish cohorts and 12% for Finnish, Norwegian, and Swedish cohorts. The lifetime risk of developing PC in the US and the UK is 12% [39, 42]. PC heritability is estimated at 57% [54, 52] and 42% by an older study [55].

The Nordic twin study [54] presents strong evidence that the heritability of PC remains stable or even slightly increases between the ages of 65 and 100. The fraction of PC attributed to high detrimentality mutations is low, similar to breast cancer. Known rare high-effect-size variants such as BRCA1/2, ATM, and HOXB13 explain only 10–12% of heritability [56, 57, 58, 53]. Recently, Eeles et al. [59], using an imputed meta-analysis for 145,000 men, reported that the GWAS polygenic score they obtained explains 33% of the familial relative risk.

Wu [56] concluded that the search for the missing heritability may be better served by high-coverage WGS; however, due to the cost and complexity, it is not currently feasible to obtain this much high-quality data. In the absence of more predictive genetic data, Wu [56] notes that the best predictor for PC is age itself.

Prostate cancer conclusions: The conclusions for PC heritability are very much the same as for breast cancer. While the heritability is higher than that of BC, it appears even more constant to slightly increasing with age, notwithstanding the smaller number of known large-effect rare mutations that can be used to explain the heritability of PC.

Colorectal cancer (CRC) The lifetime risk of developing CRC in the US is 4.1% for women and 4.5% for men [39]. In the UK, the corresponding numbers are 5% and 7% [42].

The Nordic twin studies [52, 60] estimate CRC heritability at 40%. A number of studies includes a separate classification for colon cancer with a heritability of 15% and rectal cancer with a heritability of 14%, while the combined percentage is more than double the individual ones. This example may be showing that, while subdivisions exist in the medical diagnoses that may make a difference for surgical or treatment purposes, and while even the carcinogenicity manifestations may differ between subareas of the organ, from the perspective of the heritability of the liability, they are inherited as a single condition.

CRC heritability is also relatively constant between the ages of 50 and 95 in twin studies [60]. Compared to the two previously reviewed cancers, there is a larger number of identified predisposing mutations and syndromes, such as Lynch syndrome, familial adenomatous polyposis, Peutz–Jeghers syndrome, juvenile polyposis syndrome, MUTYH-associated polyposis, NTHL1-associated polyposis, and polymerase proofreading-associated polyposis syndrome [61, 62].

Graff et al.'s study [60] concludes that although a small number of genetic variants have a substantial effect on CRC, a considerable portion of its heritability is thought to result from multiple low-risk variants. De [61]) concur that penetrant high-effect gene variants are found in 5–10% of CRC cases. A GWAS review [63] finds that more than 50 SNPs have been identified as credibly associated with CRC risk, yet these only account for a small proportion of heritability. In GWAS, common genome-wide variants are able to account for 8% of heritability.

Colorectal cancer conclusions: The conclusions are much the same as for BC and PC.

Lung cancer (LC) The lifetime risk of developing LC in the US is 6.0% for women and 6.9% for men [39]. In the UK, the corresponding numbers are 5.9% and 7.6% [42].

The LC pattern of heritability is not easy to ascertain. According to Kanwal [64], approximately 8% of lung cancers are inherited or occur as a result of a genetic predisposition. The Nordic twin studies review [52] estimates the LC heritability at 18% (within a likely range of 0–42%). Heritability studies require controlling for environmental factors, particularly tobacco smoking. It is perhaps for this reason that the Nordic twin studies consortium, which was invaluable in the three other cancer analyses, primarily restricted itself to analyzing the effects of tobacco smoking on LC [65].

Factors such as asbestos, industrial smoke and pollutants, high levels of domestic radon in some areas of the world, or exposure of miners to radon or other sources of radiation may influence incidence and, if not accounted for, may affect heritability estimates [66, 67, 68]. Hereditary mutations of genes that regulate DNA repair, including BRCA1/2, TP53 and others, also increase the risk of LC, as with almost any cancer [64].

Due to the low heritability of LC, GWAS' success at identifying predictive common SNPs has been limited [69]. Some studies explain part of the LC incidence by reference to causal epigenetic effects [70]. The heritability value of 18% that Mucci et al. [52] gives has a very broad range. An earlier study [71] noted that tobacco smoking is by far the largest causal factor for LC, and the heritability of smoking may outweigh any other LC heritability.

Mucci et al. [52] also considered smoking, but the high value reported by them exceeds the previous consensus and may need further corroboration. LC perhaps belongs to the difficult-to-analyze non-additive traits of heritability noted in [72]. We will consider LC heritability to be closer to 10%.

Lung cancer conclusions: In conclusion, we lack an age-related heritability pattern for LC, and while we cannot make definitive conclusions, we can hypothesize that LC follows a similar pattern to the other three cancers we reviewed.

In summary, the heritability patterns of cancers were not systematically investigated until quite recently. A small number of familial studies [54, 46, 43, 60] and a more recent study that is especially informative about BRCA1/2 mutations' incidence with age [47] finally allowed us to determine that cancer heritability remains relatively constant with age. Table 5 summarizes the findings of the publications for breast, prostate, colorectal, and lung cancers. We have not found a study ascertaining lung cancer heritability with age; data may be difficult to collect due to the relatively low heritability of lung cancer.

Most lung cancer incidence is environmental, and lung cancer does not have specific highly detrimental mutations that may cause a noticeable fraction of heritability. We hypothesize that the mostly polygenic fraction of lung cancer heritability is similarly stable with age, as is the case with the other three cancers we reviewed.

Supplementary Table 5. Heritability by age patterns for most common cancers

Cancer	Breast	Prostate	Colorectal	Lung
Lifetime risk, USA (%)	12	12	4.5m 4.1w	6.9m 6w
Heritability (%)	31	57	40	8–18
Incidence from highlydetrimental mutations (%)	10–14	10–12	5–10	minor
Polymorphic incidence (%)	86–90	88–90	90–95	major
Heritability trend(50y–100y)	flat / slight decline [45, 48, 44, 49, 52, 43, 46, 47, 51]	flat / slight incline [54, 56, 57, 58, 53, 59]	flat [62, 63, 61, 60]	likely flat [66, 69, 67, 68, 52, 65, 64, 70, 73]

Lifetime risk numbers, when marked, "w" for women, "m" for men.

Because cancer development is primarily a consequence of mutations and epigenetic effects leading to unconstrained propagation of the clonal cell population, in the long run, cancers are inevitable for most multicellular organisms, including humans [74, 75, 76, 77, 78].

Due to cancer's constant heritability with age, the effect of age is likely to be insignificant for GWASs' discovery of cancer polygenic scores and their corresponding predictive power. This could also apply to any LOD that follows a similar heritability pattern that is relatively constant with age.

Supplementary Chapter S2: Incidence functional approximation used in preliminary validations

To determine the effect of disease incidence with age progression on allele frequencies in the population and the difference in allele frequencies between the newly affected and remaining unaffected populations, we used three incidence dependencies with age.

1) Constant incidence:

$$I(t) = a, \tag{1}$$

where a is a constant representing a horizontal line, and we chose yearly incidence values of 0.0015, 0.005, and 0.02 (0.15% to 2%).

2) Linear incidence:

$$I(t) = b\,t, \tag{2}$$

where b is a slope of the linear progression with intercept 0, and we chose slope values of 0.003, 0.01, and 0.04. This means that incidence starts at 0 and increases to an incidence equal to 0.3%, 1%, and 4%, respectively, at 100 years of age to match the cumulative incidence of 1) above.

These values were chosen to simplify the evaluation via simulation. We ran the simulation with mortality 0, and the values were chosen so the cumulative incidence is the same%)—0.44 (44%)—at 100 years of age for the highest of either the constant or linear incidence progression.

3) In addition, we used an evaluation exponential incidence progression:

$$I(t) = 3.05 \cdot 10^{-5} e^{0.1178t}, \tag{3}$$

fitted to achieve a similar cumulative incidence at the most advanced age.

In all five scenarios described in the main article, the values of the case and control means and standard deviation/variance are identical when the cumulative incidence reaches the same level.

We validated two heritability scenarios, 30.5% and 80.5%, see Table 6.

Supplementary Table 6. Linear and constant incidence validation scenarios

	A	B	C	D	E
Scenario 1. Variants:	400	625	1375	50	25
Achieved heritability:	0.3068	0.308	0.3075	0.296	0.3142
Scenario 2. Variants:	3725	5850	12775	500	225
Achieved heritability:	0.8047	0.8064	0.8049	0.8078	0.8048

The target heritability is 0.305 (30.5%) for validation scenario 1 and 0.805 (80.5%) for validation scenario 2 due to the genetic architecture model requiring multiples of 25 variants.

Validating allele distribution change in model genetic architectures using systematic incidence progressions

We ran a set of validation simulations to verify the behavior of the model genetic distributions for the three types of incidence progression described above. The validation simulations based on the constant, linear and exponential incidence rates confirmed that both of the mean polygenic scores, for the population and for the cases, viewed in the individual values analysis for each age depend on the cumulative incidence and the magnitude of heritability, with neither being dependent on the shape of incidence progression with age.

We found from the validation simulations that the cumulative incidence, regardless of the incidence progression pattern, produces virtually identical polygenic score distribution for cases and the remaining unaffected population; see the genetic common allele low effect size plotted in Supplementary Fig. 2.

Between the genetic architectures, there is also a relatively small difference in the polygenic scores of the population and the cases; see Supplementary Fig. 3. As can be seen, the low-effect-size scenarios A, B, and C, progressing in allele frequency from common to rare, are practically indistinguishable from each other.

The higher-effect-size architectures (D and E) show a slightly larger fraction of higher-polygenic-score individuals or, more precisely, a slightly larger representation of higher- and low-polygenic-score individuals. The qualitative picture is close to identical among all five scenarios.

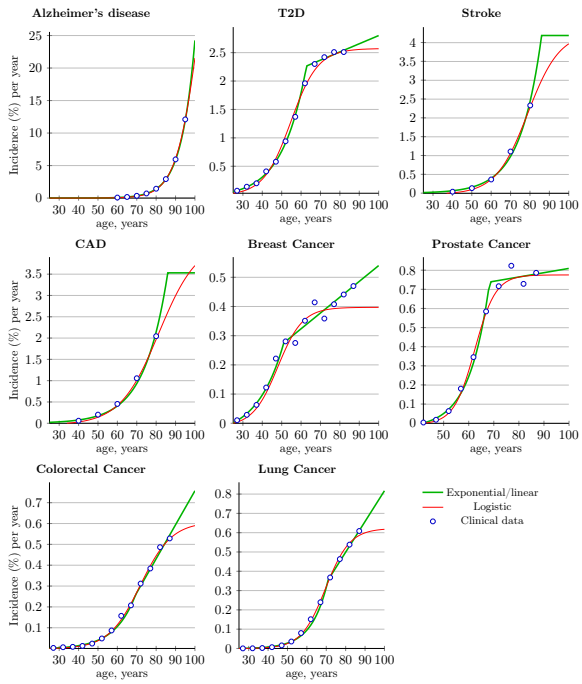
Supplementary Chapter S3: LOD incidence functional approximation

We apply the simulations to eight of the most prevalent LODs: Alzheimer's disease, type 2 diabetes, coronary artery disease, and cerebral stroke, and four late-onset cancers: breast, prostate, colorectal, and lung cancer.

First, we describe the functional approximation of the clinical incidence data we used for our simulations. The LODs incidence progression with age is presented in Supplementary Fig. 1. The initial incidence rate (fraction of population newly diagnosed per year) increases exponentially with age. This exponential growth continues for decades. Then the growth in older cohorts may flatten, as in the case of T2D [17]. In the case of cerebral stroke and CAD, the clinical studies indicate a slowdown of the incidence for individuals over the age of 85; accordingly [79], we used a constant level for the exponential approximation Eq. 4.

Alzheimer's disease, on the other hand, continues exponentially past the age of 95, reaching incidences above 20% [22]. Cancer progression reaches only a small fraction of the incidence levels of the above mentioned LODs, even for the four most prevalent cancers. Generalizing to other cancers, the incidence is much lower for more than a hundred of the less prevalent cancer types.

To evaluate each LOD's allele redistribution with age, it was necessary to approximate the yearly incidence from much rougher-grained statistics. An R script automates the determination of the best fit for logistic and exponential regression from the clinical incidence data. The script also calculated lifetime incidence from our functional approximations; it closely matched the disease clin-



Supplementary Fig. 1. LOD clinical incidence rates and functional approximations.

Two functional approximations of clinical data: exponential followed by linear and logistic. The R script automating NLM (nonlinear model) regression for both approximation curves is available in Supporting Information.

ical statistics presented in Tables 1 and 5.

The incidence approximation $I(t)$ is represented mathematically by Eq. 4. a, b, and c are exponential approximation parameters, i and s are the linear regression intercept and slope, respectively, and t is time in years.

$$I(t) = \begin{cases} ae^{bt} + c, & \text{until intersection with the line, below} \\ i + st, & \text{thereafter} \end{cases} \tag{4}$$

A logistic approximation of the clinical data is shown as a red line on Supplementary Fig. 1. It is characterized by the following equation:

$$I(t) = \frac{a}{1 + e^{(c-t)/b}} + d, \tag{5}$$

The incidence rate in the logistic curve slows faster than the incidence rate in the exponential curve and also approximates the incidence rate with age. It follows a similar pattern, with an initial exponential rise and a logistic inflection point happening at quite advanced ages. Thus, the clinical data and corresponding approximations show the higher representation of older people in the patient cohorts.

For all LODs, we observed decades-long initial exponential sections in the incidence curve. The exponent conforms to a relatively narrow range of doubling the incidence rate, fitting between 5 and 8.5 years. While the absolute incidence rate differs significantly, the exponent constant multiplier a, which is equivalent to the linear regression intercept for $log(a)$ in the $I(t)$ function, mainly controls the rise, or the initial incidence onset, of the incidence rate Supplementary Fig. 1.

From we find the logistic recursion inflection points at values shown in Table 7. Exponential incidence rise follows with high precision up to the ages shown in the table and rapid incidence rate rise continues past these ages.

Supplementary Table 7. Until what age LOD incidence rate rises exponentially

	AD	T2D	CAD	Stroke	Breast C.	Prostate C.	Colorectal C.	Lung C.
Age, years	103	55	81	79	48	62	72	70

The logistic approximation produced a good, simple fit for seven of the eight diseases. While we could also have used the logistic approximation for breast cancer, the exponential-plus-linear approximation showed a better fit and was therefore used instead.

As this paper makes extensive reference to the incidence of LODs, we should clarify some of the commonly used terms. A lifetime incidence, also called a cumulative rate, is calculated using the accepted method of summing the yearly incidences [80]:

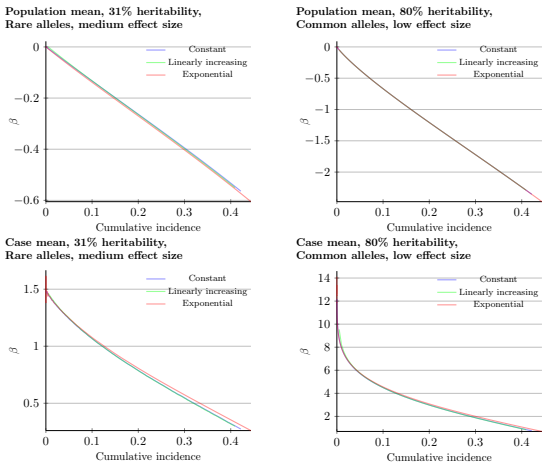
$$I_{lifetime} = \sum_{t=0}^{t_{max}} I(t), \tag{6}$$

For larger incidence values, the resulting sum produces an exaggerated result. It may become larger than 1 (100%), in which case the use of an approximate clinical statistic called cumulative risk overcomes this issue and is more meaningful. This is much like compound interest, which implicitly assumes an exact exponential incidence progression [80]:

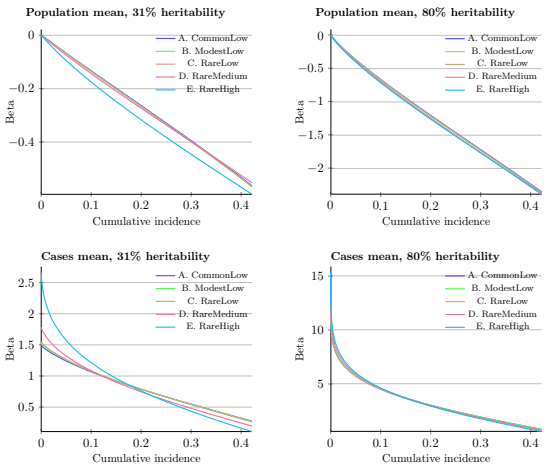
$$CumRisk = 1 - e^{-I_{lifetime}}. \tag{7}$$

Cumulative risk Eq. 7 is also an approximation because, in any practical setting, the statistic is complicated by ongoing population mortality, multiple diagnoses, and other factors. In addition, cumulative incidence and cumulative risk can be used to find values for any age of interest, not only lifetime. When needed in our simulations, we use the exact diagnosis counts to calculate the precise cumulative incidence for every age.

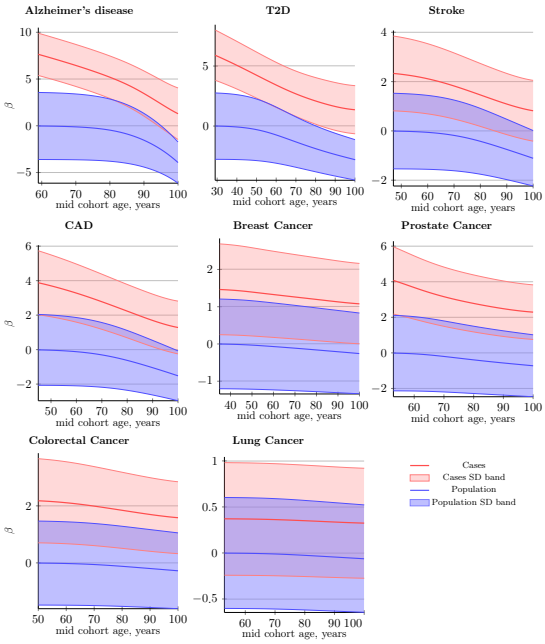
SUPPLEMENTARY FIGURES



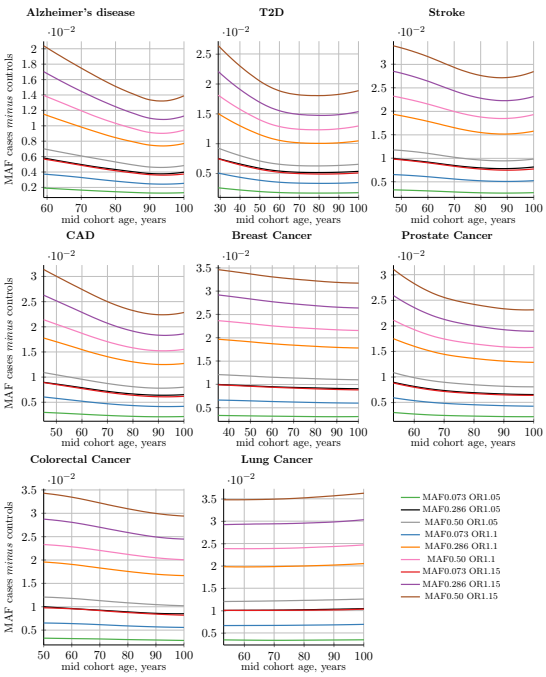
Supplementary Fig. 2. Validation simulations: constant, linear, and exponential incidence curves within the same allele architectures.
Using a constant incidence at a level of 0.5% per year, linearly increasing incidence with a slope of 0.01%, and exponentially reaching similar cumulative incidence in a 105-year age interval. Within the same allele architecture, the β is identical, subject to the simulation population stochasticity; $\beta = log(OddsRatio)$.



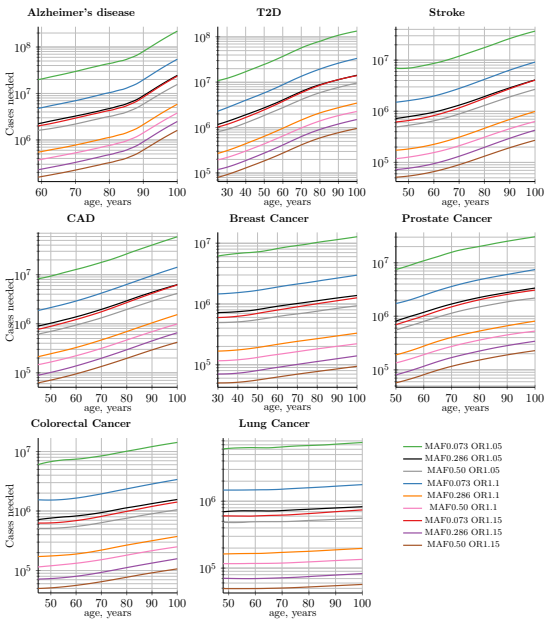
Supplementary Fig. 3. Validation simulations for five allele architectures.
The linear and constant incidence patterns give identical results for each allele architecture. The rare medium-effect-size and even rarer high-effect-size scenarios produce a fraction of higher individual betas for the same overall population variance; a relative difference is less prominent at 80% versus 31%. The three identical low-effect-size scenarios produce effectively identical β patterns; $\beta = log(OddsRatio)$.



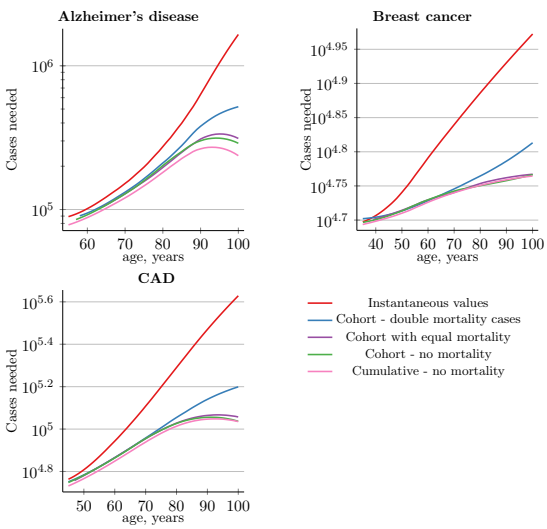
Supplementary Fig. 4. Polygenic score difference between patients and controls in a cohort simulation.
Common low-effect-size alleles (scenario A); $\beta = log(OddsRatio)$. *SD band* is a band of one standard deviation above and below the cases and the unaffected population of the same age. The cohort change and difference are less prominent than in IVA due to the accumulated diagnoses from younger cases with an averaged control polygenic risk score and mortality.



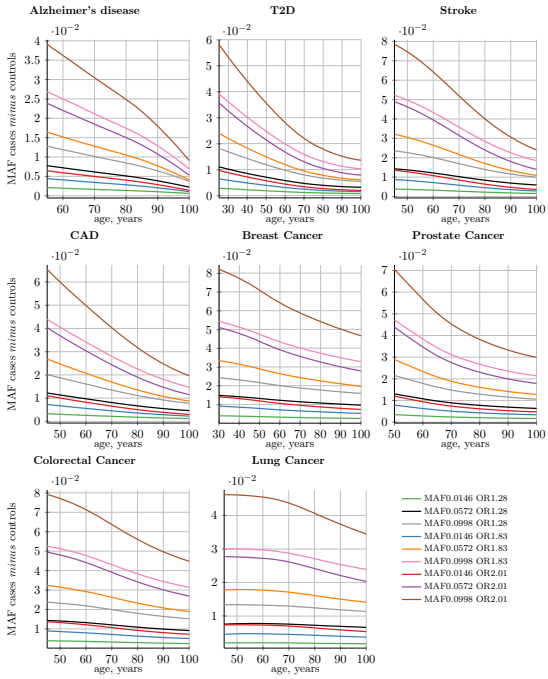
Supplementary Fig. 5. Allele frequency difference between cases and controls; cohort simulation.
Common low-effect-size alleles (scenario A). The **MAF cases *minus* controls** value is used to determine GWAS statistical power. Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change.



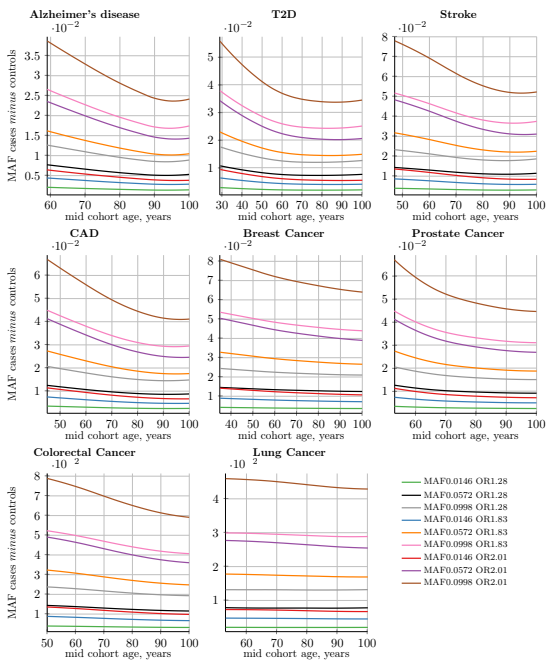
Supplementary Fig. 6. Number of cases needed to achieve 0.8 discovery power; IVA.
Common low-effect-size alleles (scenario A). The individual-diagnosed-versus-same-age-unaffected-population curve continues a steep rise in the IVA scenario. A sample of 9 out of 25 SNPs; MAF = minor (risk) allele frequency; OR = risk odds ratio.



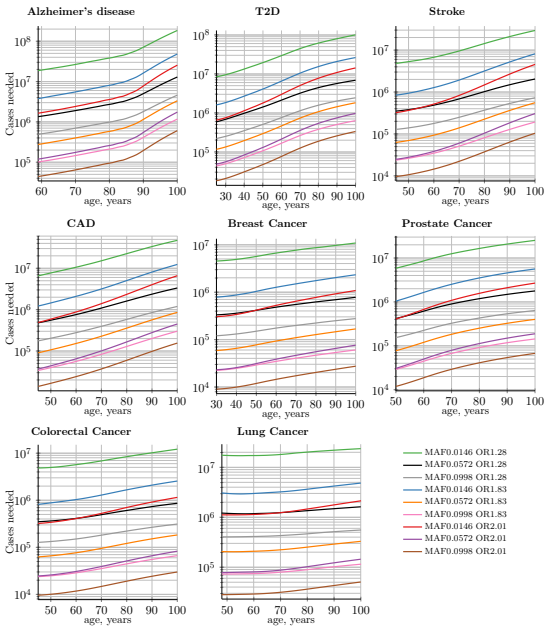
Supplementary Fig. 7. Number of cases needed for 0.8 discovery power for three LODs with representative incidence rate and initial heritability; summary of five LOD validation simulation types.
The number of cases needed for 0.8 GWAS discovery power for the clinical cohort study scenario lies between equal mortality for cases and controls and double mortality for cases; it is closer to equal mortality for the LODs we review. The divergence begins after age 85 and is even then relatively modest. "Cohort—double mortality" cases have a two-times-larger mortality than controls (doubling the value for mortality from the US "Actuarial Life Table". "Cumulative—no mortality" is the most extreme case of a one-year-span GWAS cohort; with no mortality, it requires the smallest number of cases in GWAS. Note that the logarithmic scale is very different between the three LODs.



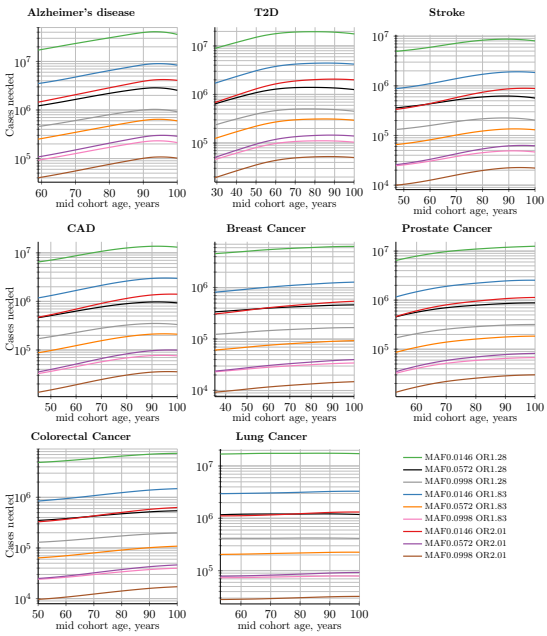
Supplementary Fig. 8. Allele frequency difference between newly diagnosed instances and the remaining unaffected population; IVA.
Rare medium-effect-size alleles (scenario D). The **MAF cases *minus* controls** value is used to determine GWAS statistical power. Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change.



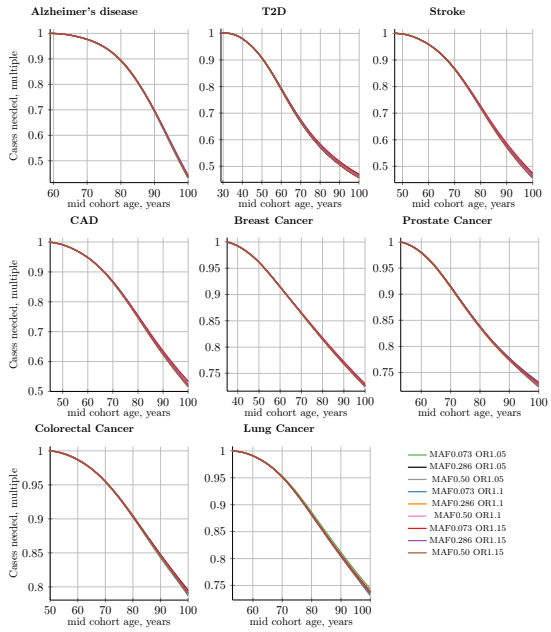
Supplementary Fig. 9. Allele frequency difference between cases and controls; cohort simulation.
Rare medium-effect-size alleles (scenario D). The **MAF cases *minus* controls** value is used to determine GWAS statistical power. Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change.



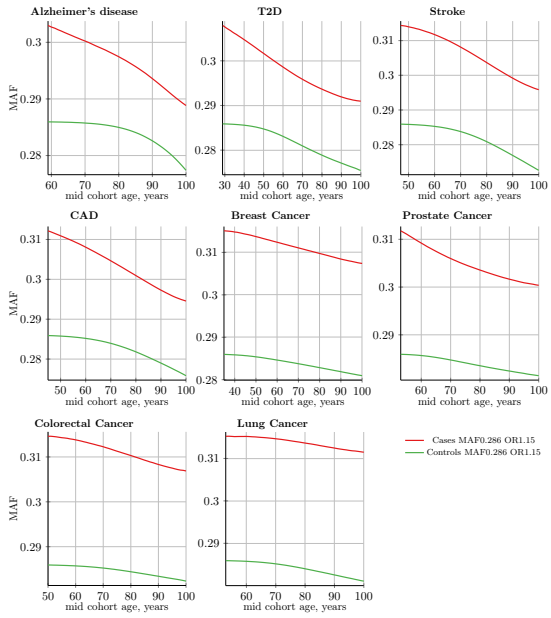
Supplementary Fig. 10. Number of cases needed to achieve 0.8 discovery power; IVA.
Rare medium-effect-size alleles (scenario D). The individual-diagnosed-versus-same-age-unaffected-population curve continues a steep rise in the IVA scenario. A sample of 9 out of 25 SNPs; MAF = minor (risk) allele frequency; OR = risk odds ratio.



Supplementary Fig. 11. Number of cases needed to achieve 0.8 discovery power; cohort simulation.
Rare medium-effect-size alleles (scenario D). The cohort curve due to the accumulative cases diagnosed at younger ages with an averaged control polygenic risk score and mortality begins at the same necessary-cases number as IVA but rises more slowly and levels out at older ages. A sample of 9 out of 25 SNPs; MAF = minor (risk) allele frequency; OR = risk odds ratio.



Supplementary Fig. 12. Times fewer the number of cases needed for 0.8 discovery power in cohort study when using progressively older control cohorts compared to fixed-age young-cases cohort.
Cases' mid-cohort age is leftmost age (youngest plot point); control mid-cohort ages are incremental ages. The number of cases needed for 0.8 discovery power is smaller when older controls are used, particularly for LODs with the highest heritability and incidence. Common low-effect-size alleles (scenario A). A sample of 9 out of 25 SNPs; MAF = minor (risk) allele frequency; OR = risk odds ratio.



Supplementary Fig. 13. MAF (minor allele frequency) absolute magnitude change with age for cases and controls; cohort simulation.
Common low-effect-size alleles (scenario A), all plots show MAF=0.286 and OR=1.15 allele. Change in absolute magnitude of each allele frequency value is relatively small with age progression. GWAS discovery power is a function of the difference in allele frequency between cases and controls. Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF absolute and relative change with cohort age progression.

Supplementary references

[1] Andreas Wienke, Niels V Holm, Axel Skytthe, and Anatoli I Yashin. The heritability of mortality due to heart diseases: a correlated frailty model applied to danish twins. *Twin Research and Human Genetics*, 4(4):266–274, 2001.

[2] S Zdravkovic, A Wienke, NL Pedersen, ME Marenberg, AI Yashin, and U De Faire. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 swedish twins. *Journal of internal medicine*, 252(3):247–254, 2002.

[3] William J Devan, Guido J Falcone, Christopher D Anderson, Jeremiasz M Jagiella, Helena Schmidt, Björn M Hansen, Jordi Jimenez-Conde, Eva Giralt-Steinhauer, Elisa Cuadrado-Godia, Carolina Soriano, et al. Heritability estimates identify a substantial genetic contribution to risk and outcome of intracerebral hemorrhage. *Stroke*, 44(6): 1578–1583, 2013.

[4] Hugo J Aparicio and Sudha Seshadri. Familial occurrence and heritability of stroke. In *Stroke Genetics*, pages 9–20. Springer, 2017.

[5] DS Falconer. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of human genetics*, 31(1):1–20, 1967.

[6] Peter Almgren, M Lehtovirta, B Isomaa, L Sarelin, MR Taskinen, Valeriya Lyssenko, Tiinamaija Tuomi, Leif Groop, and Botnia Study Group. Heritability and familiality of type 2 diabetes and related quantitative traits in the botnia study. *Diabetologia*, 54(11):2811, 2011.

[7] Philippa J Talmud, Jackie A Cooper, Richard W Morris, Frank Dudbridge, Tina Shah, Jorgen Engmann, Caroline Dale, Jon White, Stela McLachlan, Delilah Zabaneh, et al. Sixty-five common genetic variants and prediction of type 2 diabetes. *Diabetes*, page DB_141504, 2014.

[8] Jose M de Miguel-Yanes, Peter Shrader, Michael J Pencina, Caroline S Fox, Alisa K Manning, Richard W Grant, Josèe Dupuis, Jose C Florez, Ralph B D'agostino, L Adrienne Cupples, et al. Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. *Diabetes care*, 34(1):121–125, 2011.

[9] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, 2016.

[10] Mia Nielsen, Charlotte Andersson, Thomas A Gerds, Per Kragh Andersen, Thomas Bo Jensen, Lars Køber, Gunnar Gislason, and Christian Torp-Pedersen. Familial clustering of myocardial infarction in first-degree relatives: a nationwide study. *European heart journal*, 34(16):1198–1203, 2013.

[11] UGR Schulz, E Flossmann, and PM Rothwell. Heritability of ischemic stroke in relation to age, vascular risk factors, and subtypes of incident stroke in population-based studies. *Stroke*, 35(4):819–824, 2004.

[12] Sudha Seshadri, Alexa Beiser, Aleksandra Pikula, Jayandra J Himali, Margaret Kelly-Hayes, Stephanie Debette, Anita L DeStefano, Jose R Romero, Carlos S Kase, and Philip A Wolf. Parental occurrence of stroke and risk of stroke in their children: the framingham study. *Circulation*, 121(11):1304–1312, 2010.

[13] Shannon Anjelica Allport, Ngum Kikah, Nessim Abu Saif, Fonkem Ekokobe, and Folefac D Atem. Parental age of onset of cardiovascular disease as a predictor for offspring age of onset of cardiovascular disease. *PloS one*, 11(12): e0163334, 2016.

[14] Steve Bevan, Matthew Traylor, Poneh Adib-Samii, Rainer Malik, Nicola LM Paul, Caroline Jackson, Martin Farrall, Peter M Rothwell, Cathie Sudlow, Martin Dichgans, et al. Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. *Stroke*, 43(12):3161–3167, 2012.

[15] P Poulsen, K Ohm Kyvik, A Vaag, and H Beck-Nielsen. Heritability of type ii (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42(2):139–145, 1999.

[16] Sun Ha Jee, Il Suh, So Young Won, and Mi Yang Kim. Familial correlation and heritability for cardiovascular risk factors. *Yonsei medical journal*, 43(2):160–164, 2002.

[17] Michael WJ Boehme, Gisela Buechele, Julia Frankenhauser-Mannuss, Jana Mueller, Dietlinde Lump, Bernhard O Boehm, and Dietrich Rothenbacher. Prevalence, incidence and concomitant co-morbidities of type 2 diabetes mellitus in south western germany-a retrospective cohort and case control study in claims data of a large statutory health insurance. *BMC Public Health*, 15 (1):855, 2015.

[18] Cathy E Elks, Marcel Den Hoed, Jing Hua Zhao, Stephen J Sharp, Nicholas J Wareham, Ruth JF Loos, and Ken K Ong. Variability in the heritability of body mass index: a systematic review and meta-regression. *Frontiers in endocrinology*, 3:29, 2012.

[19] Donald M Lloyd-Jones, Eric P Leip, Martin G Larson, Ralph B d'Agostino, Alexa Beiser, Peter WF Wilson, Philip A Wolf, and Daniel Levy. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. *Circulation*, 113(6):791–798, 2006.

[20] Syed S Mahmood, Daniel Levy, Ramachandran S Vasan, and Thomas J Wang. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, 383(9921):999–1008, 2014.

[21] Diapedia: Epidemiology of type 2 diabetes. *Available at https://www.diapedia.org/type-2-diabetes-mellitus/3104287123/epidemiology-of-type-2-diabetes (accessed November 10, 2018)*, 2016.

[22] Ron Brookmeyer, Sarah Gray, and Claudia Kawas. Projections of alzheimer's disease in the united states and the public health impact of delaying disease onset. *American journal of public health*, 88(9):1337–1342, 1998.

[23] Adam C Naj and Gerard D Schellenberg. Genomic variants, genes, and pathways of alzheimer's disease: An overview. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 174(1):5–26, 2017.

[24] Margaret Gatz, Chandra A Reynolds, Laura Fratiglioni, Boo Johansson, James A Mortimer, Stig Berg, Amy Fiske, and Nancy L Pedersen. Role of genes and environments for explaining alzheimer disease. *Archives of general psychiatry*, 63(2):168–174, 2006.

[25] Lan Tan, Jin-Tai Yu, Wei Zhang, Zhong-Chen Wu, Qun Zhang, Qiu-Yan Liu, Wei Wang, Hui-Fu Wang, Xiao-Ying Ma, and Wei-Zhen Cui. Association of gwas-linked loci with late-onset alzheimer's disease in a northern han chinese population. *Alzheimer's & dementia*, 9(5):546–553, 2013.

[26] Luxi Shen and Jianping Jia. An overview of genome-wide association studies in alzheimer's disease. *Neuroscience bulletin*, 32(2):183–190, 2016.

[27] Lindsay A Farrer, L Adrienne Cupples, Jonathan L Haines, Bradley Hyman, Walter A Kukull, Richard Mayeux, Richard H Myers, Margaret A Pericak-Vance, Neil Risch, and Cornelia M Van Duijn. Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease: a meta-analysis. *Jama*, 278(16): 1349–1356, 1997.

[28] Yvonne Davidson, Linda Gibbons, Antonia Pritchard, Jayne Hardicre, Joanne Wren, Cheryl Stopford, Camille Julien, Jennifer Thompson, Antony Payton, Stuart M Pickering-Brown, et al. Apolipoprotein e $\varepsilon 4$ allele frequency and age at onset of alzheimer's disease. *Dementia and geriatric cognitive disorders*, 23(1):60–66, 2007.

[29] Nadine Binder and Martin Schumacher. Incidence of dementia over three decades in the framingham heart study. *The New England journal of medicine*, 375(1):92–93, 2016.

[30] Yu-Tzu Wu, Alexa S Beiser, Monique MB Breteler, Laura Fratiglioni, Catherine Helmer, Hugh C Hendrie, Hiroyuki Honda, M Arfan Ikram, Kenneth M Langa, and Antonio Lobo. The changing prevalence and incidence of dementia over time - current evidence. *Nature Reviews Neurology*, 2017.

[31] Søren Glud Skousgaard, Jacob Hjelmborg, Axel Skytthe, Lars Peter Andreas Brandt, Sören Möller, and Søren Overgaard. Probability and heritability estimates on primary osteoarthritis of the hip leading to total hip arthroplasty: a nationwide population based follow-up study in danish twins. *Arthritis research & therapy*, 17(1):336, 2015.

[32] Adae O Amoako and George Guntur A Pujalte. Osteoarthritis in young, active, and athletic individuals. *Clinical Medicine Insights: Arthritis and Musculoskeletal Disorders*, 7:CMAMD–S14386, 2014.

[33] Sophie Catriona Warner and Ana Maria Valdes. The genetics of osteoarthritis: A review. *Journal of Functional Morphology and Kinesiology*, 1(1):140–153, 2016.

[34] Tim D Spector and Alex J MacGregor. Risk factors for osteoarthritis: genetics. *Osteoarthritis and cartilage*, 12:39–44, 2004.

[35] Søren Glud Skousgaard, Axel Skytthe, Sören Möller, Søren Overgaard, and LP Brandt. Sex differences in risk and heritability estimates on primary knee osteoarthritis leading to total knee arthroplasty: a nationwide population based follow up study in danish twins. *Arthritis research & therapy*, 18:46–46, 2016.

[36] Stuart H Ralston and André G Uitterlinden. Genetics of osteoporosis. *Endocrine reviews*, 31(5):629–662, 2010.

[37] John R Shaffer, Candace M Kammerer, Jan M Bruder, Shelley A Cole, Thomas D Dyer, Laura Almasy, Jean W MacCluer, John Blangero, Richard L Bauer, and Braxton D Mitchell. Genetic influences on bone loss in the san antonio family osteoporosis study. *Osteoporosis international*, 19(12):1759–1767, 2008.

[38] John Loughlin. Genetic contribution to osteoarthritis development: current state of evidence. *Current opinion in rheumatology*, 27(3):284, 2015.

[39] Lifetime Risk of Developing or Dying From Cancer. *American Cancer Society. Available at https://www.cancer.org/cancer/cancer-basics/lifetime-probability-of-developing-or-dying-from-cancer.html (accessed November 10, 2018)*, 2018.

[40] Sherry L Murphy, Jiaquan Xu, Kenneth D Kochanek, Sally C Curtin, and Elizabeth Arias. Mortality in the united states, 2016. *NCHS Data Brief, no 293*, 293, 2017.

[41] AS Ahmad, N Ormiston-Smith, and PD Sasieni. Trends in the lifetime risk of developing cancer in great britain: comparison of risk for those born from 1930 to 1960. *British journal of cancer*, 112(5):943, 2015.

[42] Cancer Statistics for the UK. *Available at http://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk (accessed November 10, 2018)*, 2018.

[43] Barbara Haley. Hereditary breast cancer: the basics of brca and beyond. *UT Southwestern Medical Center. Internal Medicine Grand Rounds*, 2016.

[44] Harvey A Risch, John R McLaughlin, David EC Cole, Barry Rosen, Linda Bradley, Isabel Fan, James Tang, Song Li, Shiyu Zhang, and Patricia A Shaw. Population brca1 and brca2 mutation frequencies and cancer penetrances: a kin–cohort study in ontario, canada. *Journal of the National Cancer Institute*, 98(23):1694–1706, 2006.

[45] Deborah Ford, DF Easton, M Stratton, S Narod, D Goldgar, P Devilee, DT Bishop, B Weber, G Lenoir, J Chang-Claude, et al. Genetic heterogeneity and penetrance analysis of the brca1 and brca2 genes in breast cancer families. *The American Journal of Human Genetics*, 62(3):676–689, 1998.

[46] Sören Möller, Lorelei A Mucci, Jennifer R Harris, Thomas Scheike, Klaus Holst, Ulrich Halekoh, Hans-Olov Adami, Kamila Czene, Kaare Christensen, Niels V Holm, et al. The heritability of breast cancer among women in the nordic twin study of cancer. *Cancer Epidemiology and Prevention Biomarkers*, 25(1):145–150, 2016.

[47] Karoline B Kuchenbaecker, John L Hopper, Daniel R Barnes, Kelly-Anne Phillips, Thea M Mooij, Marie-José Roos-Blom, Sarah Jervis, Flora E Van Leeuwen, Roger L Milne, and Nadine Andrieu. Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers. *Jama*, 317(23):2402–2416, 2017.

[48] Anthony Antoniou, Paul DP Pharoah, Steven Narod, Harvey A Risch, Jorunn E Eyfjord, John L Hopper, Niklas Loman, Håkan Olsson, O Johannsson, Åke Borg, et al. Average risks of breast and ovarian cancer associated with

brca1 or brca2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics*, 72(5):1117–1130, 2003.

[49] Nasim Mavaddat, Antonis C Antoniou, Douglas F Easton, and Montserrat Garcia-Closas. Genetic susceptibility to breast cancer. *Molecular oncology*, 4(3):174–191, 2010.

[50] Paulo CM Lyra-Junior, Nayara G Tessarollo, Isabella S Guimarães, Taciane B Henriques, Diandra Z dos Santos, Marcele LM de Souza, Victor Hugo M Marques, Laura FRL de Oliveira, Krislayne V Siqueira, Ian V Silva, et al. Gwas in breast cancer. In *Breast Cancer-From Biology to Medicine*, pages 99–117. InTech, 2017.

[51] Kyriaki Michailidou, Sara Lindström, Joe Dennis, Jonathan Beesley, Shirley Hui, Siddhartha Kar, Audrey Lemaçon, Penny Soucy, Dylan Glubb, Asha Rostamianfar, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92, 2017.

[52] Lorelei A Mucci, Jacob B Hjelmborg, Jennifer R Harris, Kamila Czene, David J Havelick, Thomas Scheike, Rebecca E Graff, Klaus Holst, Sören Möller, Robert H Unger, et al. Familial risk and heritability of cancer among twins in nordic countries. *Jama*, 315(1):68–76, 2016.

[53] Julie Lecarpentier, Valentina Silvestri, Karoline B Kuchenbaecker, Daniel Barrowdale, Joe Dennis, Lesley McGuffog, Penny Soucy, Goska Leslie, Piera Rizzolo, and Anna Sara Navazio. Prediction of breast and prostate cancer risks in male brca1 and brca2 mutation carriers using polygenic risk scores. *Journal of Clinical Oncology*, 35(20):2240–2250, 2017.

[54] Jacob B Hjelmborg, Thomas Scheike, Klaus Holst, Axel Skytthe, Kathryn L Penney, Rebecca E Graff, Eero Pukkala, Kaare Christensen, Hans-Olov Adami, Niels V Holm, et al. The heritability of prostate cancer in the nordic twin study of cancer. *Cancer Epidemiology and Prevention Biomarkers*, 23(11):2303–2310, 2014.

[55] Henrik Grönberg. Prostate cancer epidemiology. *The Lancet*, 361(9360):859–864, 2003.

[56] Xifeng Wu and Jian Gu. Heritability of prostate cancer: a tale of rare variants and common single nucleotide polymorphisms. *Annals of translational medicine*, 4(10), 2016.

[57] Nicholas Mancuso, Nadin Rohland, Kristin A Rand, Arti Tandon, Alexander Allen, Dominique Quinque, Swapan Mallick, Heng Li, Alex Stram, and Xin Sheng. The contribution of rare variation to prostate cancer heritability. *Nature genetics*, 48(1):30, 2016.

[58] Patrick C Walsh. The search for the missing heritability of prostate cancer. *European urology*, 72(5):657–659, 2017.

[59] Rosalind Eeles, Ali Amin Al Olama, Sonja Berndt, Fredrik Wiklund, David V Conti, Mahbubl Ahmed, Sara Benlloch, Douglas Easton, Peter Kraft, Stephen J Chanock, et al. Prostate cancer meta-analysis from more than 145,000 men to identify 65 novel prostate cancer susceptibility loci. *Journal of Clinical Oncology*, 2017.

[60] Rebecca E Graff, Sören Möller, Michael N Passarelli, John S Witte, Axel Skytthe, Kaare Christensen, Qihua Tan, Hans-Olov Adami, Kamila Czene, and Jennifer R Harris. Familial risk and heritability of colorectal cancer in the nordic twin study of cancer. *Clinical Gastroenterology and Hepatology*, 15(8):1256–1264, 2017.

[61] Richarda M de Voer, Marc-Manuel Hahn, Robbert DA Weren, Arjen R Mensenkamp, Christian Gilissen, Wendy A van Zelst-Stams, Liesbeth Spruijt, C Marleen Kets, Junxiao Zhang, and Hanka Venselaar. Identification of novel candidate genes for early-onset colorectal cancer susceptibility. *PLoS genetics*, 12(2):e1005880, 2016.

[62] Shuo Jiao, Ulrike Peters, Sonja Berndt, Hermann Brenner, Katja Butterbach, Bette J Caan, Christopher S Carlson, Andrew T Chan, Jenny Chang-Claude, Stephen Chanock, et al. Estimating the heritability of colorectal cancer. *Human molecular genetics*, 23(14):3898–3905, 2014.

[63] Stephanie L Schmit, Fredrick R Schumacher, Christopher K Edlund, David V Conti, Ugonna Ihenacho, Peggy Wan, David Van Den Berg, Graham Casey, Barbara K Fortini, Heinz-Josef Lenz, et al. Genome-wide association study of colorectal cancer in hispanics. *Carcinogenesis*, 37(6): 547–556, 2016.

[64] Madiha Kanwal, Xiao-Ji Ding, and Yi Cao. Familial risk for lung cancer. *Oncology Letters*, 13(2):535–542, 2017.

[65] Jacob Hjelmborg, Tellervo Korhonen, Klaus Holst, Axel Skytthe, Eero Pukkala, Julia Kutschke, Jennifer R Harris, Lorelei A Mucci, Kaare Christensen, Kamila Czene, et al. Lung cancer, genetic predisposition and smoking: the nordic twin study of cancer. *Thorax*, pages thoraxjnl–2015, 2016.

[66] Daniel Krewski, Jay H Lubin, Jan M Zielinski, Michael Alavanja, Vanessa S Catalan, R William Field, Judith B Klotz, Ernest G Létourneau, Charles F Lynch, Joseph I Lyon, et al. Residential radon and risk of lung cancer: a combined analysis of 7 north american case-control studies. *Epidemiology*, 16(2):137–145, 2005.

[67] Shamus R Carr, Wallace Akerley, Mia Hashibe, and Lisa A Cannon-Albright. Evidence for a genetical contribution to non-smoking-related lung cancer. *Thorax*, pages thoraxjnl–2014, 2015.

[68] Jyoti Malhotra, Matteo Malvezzi, Eva Negri, Carlo La Vecchia, and Paolo Boffetta. Risk factors for lung cancer worldwide. *European Respiratory Journal*, pages ERJ–00359, 2016.

[69] Joel L Weissfeld, Yan Lin, Hui-Min Lin, Brenda F Kurland, David O Wilson, Carl R Fuhrman, Arjun Pennathur, Marjorie Romkes, Tomoko Nukui, and Jian-Min Yuan. Lung cancer risk prediction using common snps located in gwas-identified susceptibility regions. *Journal of Thoracic Oncology*, 10(11):1538–1545, 2015.

[70] Lin Shi, Minghuan Zheng, Jiayuan Hou, Bijun Zhu, and Xiangdong Wang. Regulatory roles of epigenetic modulators, modifiers and mediators in lung cancer. In *Seminars in cancer biology*, volume 42, pages 4–12. Elsevier, 2017.

[71] Ian A Yang, John W Holloway, and Kwun M Fong. Genetic susceptibility to lung cancer and co-morbidities. *Journal of thoracic disease*, 5(Suppl 5):S454, 2013.

[72] Tinca JC Polderman, Beben Benyamin, Christiaan A De Leeuw, Patrick F Sullivan, Arjen Van Bochoven, Peter M Visscher, and Danielle Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, 47(7):702, 2015.

[73] Diane C Wang and Xiangdong Wang. Tomorrow's genome medicine in lung cancer. In *Seminars in cancer biology*, volume 42, pages 39–43. Elsevier, 2017.

[74] Andriy Marusyk and James DeGregori. Declining cellular fitness with age promotes cancer initiation by selecting for adaptive oncogenic mutations. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1785(1):1–11, 2008.

[75] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

[76] Avital Guedj, Anat Geiger-Maor, Eithan Galun, Hagai Amsalem, and Jacob Rachmilewitz. Early age decline in dna repair capacity in the liver: in depth profile of differential gene expression. *Aging (Albany NY)*, 8(11):3131, 2016.

[77] Flavia Ribezzo, Yosef Shiloh, and Björn Schumacher. Systemic dna damage responses in aging and diseases. In *Seminars in cancer biology*, volume 37, pages 26–35. Elsevier, 2016.

[78] Paul Nelson and Joanna Masel. Intercellular competition and the inevitability of multicellular aging. *Proceedings of the National Academy of Sciences*, page 201618854, 2017.

[79] PM Rothwell, AJ Coull, LE Silver, JF Fairhead, MF Giles, CE Lovelock, JNE Redgrave, LM Bull, SJV Welch, and FC Cuthbertson. Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (oxford vascular study). *The Lancet*, 366(9499):1773–1783, 2005.

[80] PD Sasieni, J Shelton, N Ormiston-Smith, CS Thomson, and PB Silcocks. What is the lifetime risk of developing cancer?: the effect of adjusting for multiple primaries. *British journal of cancer*, 105(3):460, 2011.