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, New Zealand
2 Department of Computer Science, University of Auckland, Auckland, New Zealand

¤Current Address: Department of Computer Science, University of Auckland, Auckland, 1010, New Zealand
* roli573@aucklanduni.ac.nz

Abstract

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

Computer simulations of an aging population have shown that the risk allele frequency decreases at older ages because the individuals with higher polygenic risk are first to become ill. 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. The LOD incidence rate grows exponentially, doubling in incidence between 5 and 8.5 years, guaranteeing that the cohorts for GWAS studies overrepresent older individuals with lower polygenic risk scores, whose disease cases are disproportionately due to environmental causes such as old age itself. This simultaneously leads to diminished observed heritability and lower GWAS statistical power at older ages; thus, the missing heritability in GWAS is smaller than currently estimated for these LODs.

This mechanism also explains the relatively constant heritability with age reported by familial studies for the four most prevalent cancers—breast, prostate, colorectal and lung—due to a combination of their lower heritability and lower cumulative incidence; this should also be true for other LODs with similar characteristics. In addition, this mechanism explains the heritability patterns found by familial studies and in clinical practice as it relates to predicting LOD familial risks for all of these LODs.

In conclusion, for LODs 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 GWASs.

Author summary

We investigated the change in distribution of risk alleles with age progression under a model assuming that relative disease liability remains proportionate to individual polygenic risk. We found that individuals with higher polygenic risk factors become ill proportionately earlier, and the fraction of higher risk alleles diminishes for the

remaining unaffected population. This corresponds to diminishing heritability with age and lower GWAS statistical discovery power for LODs with high incidence. Even though incidence for all LODs increases exponentially with age, the effect is minimal for diseases with low prevalence and low heritability, as exemplified by cancer.

High cumulative incidence and high initial heritability are the primary determinants of risk allele relative frequency decline and consequent diminishing older age heritability and GWAS statistical power. The effect is very pronounced for a number of prevalent and highly heritable LODs, including Alzheimer's disease, coronary artery disease, cerebral stroke, and type 2 diabetes, in which the cohorts for GWAS studies over-represent older individuals whose genotype would be considered low risk earlier on and whose disease is due to old age rather than heightened genetic liability. The predictive power of GWASs is closer to the real heritability of this population than it is currently credited with.

Introduction

Throughout the ages, late-onset diseases 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 [1]. 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 [2].

One of the major goals of computational biology is to identify gene variants that lead to increased odds of late-onset diseases. The objective is to be able to predict individuals' LOD liability and, based on this knowledge, formulate preventive recommendations and treatments, with the ultimate goal of applying personalized medical interventions based on the genetic makeup of each unique individual.

With whole-genome sequencing becoming more accessible with every passing year, GWAS is being applied to all areas of genetics [3]. 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 [4–6]. This is despite the fact that LODs with varied symptoms and phenotypes show high heritability in twin and familial studies [7].

GWAS discovery is simpler in the case of Mendelian conditions, which are caused by one or a small number of large-effect mutations or SNPs [8]. For many of these diseases, causal gene variants are already well characterized. These variants can be active in a range of gene dominance, recessiveness or additivity scenarios. Mendelian conditions can be inherited or appear in a very small number of *de novo* detrimental mutations.

Natural selection efficiently removes large-effect detrimental mutations, resulting in a reasonably low genetic load in each person, equivalent to five to eight highly detrimental recessive mutations. If these were to combine and become homozygous, they could be harmful or lethal [9–13]. Still, a relatively small fraction of the population is affected by these conditions. As of May 2018, the OMIM Gene Map Statistics compendium [14] lists over 6000 genetic phenotypic conditions and close to 4000 gene mutations responsible for them.

Some LODs, such as macular degeneration [4, 15, 16] are primarily caused by a small number of high-effect variants. Macular degeneration does not exhibit a significant incidence in a form of a polygenic late-onset disease version with similar symptoms.

Therefore, even though these diseases happen late in life, they belong to the Mendelian disease category and are not classified as polygenic LODs.

The situation is much more complex in the case of polygenic LODs, in which disease liability may be associated with hundreds of small-effect gene variants or SNPs [8]. Diseases of this type are some of the most prevalent among humankind. They include cardiovascular disease—particularly coronary artery disease (CAD)—cerebral stroke, type 2 diabetes (T2D), senile dementia, Alzheimer's disease (AD), cancer and osteoarthritis (see S1 Appendix for a review of these LODs).

While computational biologists, sometimes in collaboration with specialist physicians, analyze sequencing data in an attempt to identify causal variants, physicians—especially gerontologists—have been working with aging people for centuries [2, 17]. Recently, Warner [18] stated that *"One of the criticisms raised against genetic studies is that they are far removed from clinical practice."* In this research, we focus specifically on a conceptual disconnect regarding the heritability patterns characterizing some of the most prevalent LODs and their consequences for the discovery and predictive power of GWAS.

A late-onset disease does not develop until later in life. 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 [19]. 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.

GWAS researchers often set an unrealistic expectation that a combination of causal 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 [20, 21]. 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 [22] when it is past its reproductive prime and when evolutionary selection is significantly relaxed compared to younger ages [19].

GWAS can implicate a subset of SNPs that can typically explain between 10 and 20% of the genetic heritability of a polygenic LOD [3].

There are two complementary hypotheses explaining this so-called missing heritability [23–26]. The first is the hypothesis that LODs are caused by a combination of a large number of relatively common alleles of small effect [27]. 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 [28, 29] 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 [30]. 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. We think that 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 SNPs distributions combined with changing environmental and lifestyle variables.

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 [23]. 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 [4, 5, 7, 25, 31, 32]. 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.

With larger sample sizes, GWASs are uncovering increasing numbers of causal SNPs. Nevertheless, there are no examples of polygenic LODs in which the majority of causal SNPs have been found. GWAS is a tool that has now been in use for more than a decade and continues to make progress [3]. Future tools that may be more successful in solving this puzzle are bound to appear over time. One relatively new approach is to create a number of targeted mutations in either human cell cultures or model organisms and evaluate the consequences for the phenotypes. One method is to introduce systematically designed mutations using CRISPR/Cas9 [33–35].

Another promising future approach may reverse engineer and classify causal SNPs using synthetic DNA editing, as proposed by the Genome Project-Write, which intends to develop technologies that will allow for the synthetization of complete human chromosomes, nucleotide by nucleotide [36].

One example is the research being conducted in Harvard's Church Lab, which is taking the first steps in applying this approach to the human genome [37, 38]. It may be a new method or some new variation of a method in current use, or a combination of these, that will lead to broad success. Based on the rapid rate of progress, it may take only a few decades to gain more complete, actionable and predictive LODs genetics knowledge.

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. At best, the effect may be "age adjusted" by GWASs [39] 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 [40] and his 1967 follow-up study exploring the prevalence patterns of LODs, specifically diabetes [41], 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.

We collected familial heritability, clinical, and epidemiological statistics 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 our analysis and conclusions.

The research rationale

Over the past decade, thousands of GWAS studies have been conducted, finding genetic risk variants for many polygenic diseases. The aggregate of these variants for a

population or an individual is called their polygenic risk for a particular disease. One could question what happens to individual risk and population polygenic risk as organisms age, individually or as a statistical age cohort. However, from our perspective, the following questions are the most interesting: Does individual polygenic risk remain in the same proportions between individuals when they get older, assuming we knew their polygenic score for a particular LOD when they were younger? As an organism ages, does the risk become proportionately larger for all individuals, effectively meaning a multiplier is applied to each individual polygenic score? If this is the case, is the multiplier proportional, or does the initial polygenic score ratio become more or less extreme among older individuals who are diagnosed with or unaffected by an LOD? Or is it the other way around: does aging result in decreasing genetic risk while overall frailty and LOD risk increases? Does random environmental noise rather than genetics account for the rapid age-related increase in LOD incidence that we observe?

It is obvious that individuals with higher polygenic risk have a higher likelihood of becoming ill earlier. Could this fact explain some of the observed LOD heritability known from familial studies, clinical observations and GWASs?

We chose to study a model in which the polygenic risk remains constant between individuals 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. One way to test our model is to design an aging population simulation and determine the quantitative change of allele frequency as the population ages.

More genetically predisposed individuals are more likely to be diagnosed with an LOD and become "cases" than less predisposed individuals. The allele distribution for a one-year-older population will contain fewer still-healthy high-risk individuals because such individuals fall ill with an LOD to which they are more liable in a higher proportion than lower-risk individuals. The same process continues the next year, and the next. With age progression, the frequency of higher-polygenic-risk individuals and—to an extent that we will determine—the frequency of risk alleles will decline. At the same time, a clinically significant incidence of all LODs, somewhat different for each LOD, begins at some relatively late age and increases exponentially for decades, as we will describe later.

While we may not know all underlying causes and cannot reproduce the incidence of each disease from general principles, we have clinical statistics reporting the incidence of a number of LODs in five or ten-year age bands. From these statistics, we can discover functional approximations that we can use in an aging population simulation and perform all necessary analysis as the simulation progresses.

We performed a set of computer simulations clarifying 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, we recommend a modification to GWAS cohort selection to improve statistical discovery power.

Results

First, we summarize the results of the validation simulations described in Materials and Methods. The validation simulations were performed, not as a model of a specific disease, but to determine the behavior of all allele models and the resulting allele frequency change under simple controlled and comparable-to-each-other incidence scenarios. Three validation simulation scenarios implemented constant, linear, and exponential incidence rate change.

These simulations confirmed that a change in the population's mean polygenic score and a change in the cases' mean polygenic score, 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 S1 Fig and are qualitatively similar between the genetic architectures S2 Fig.

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.

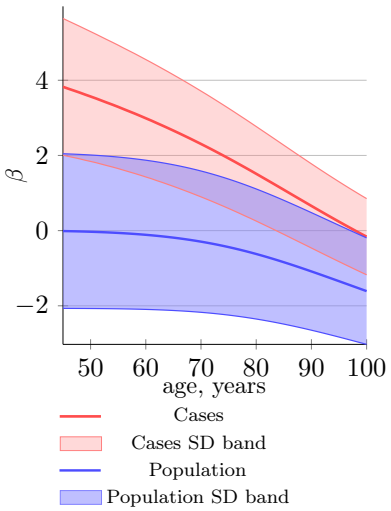

Fig 1. Polygenic score difference between newly diagnosed individuals and the remaining population: coronary artery disease IVA example

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. For all highly prevalent LODs, the mean polygenic risk of new cases crosses below the risk of an average person at early onset age. See S4 Fig, representing IVA, and S5 Fig, representing cohort study for all LODs.

Next, we present the simulation results for the eight chosen representative LODs. It is important to note that we used model genetic architectures for these disease, not a complete GWAS map of their SNPs, because GWAS-discovered sets explain only a fraction of their heritability. For this reason, we ran all the scenarios and genetic architecture models, from low to high effect size and common to low allele frequency, and found the results to be consistent for all these models. It is also known that these diseases are associated with a small fraction of high-effect variants. For example, a fraction of early cases for late-onset Alzheimer's disease is associated with the APOEe4 allele [20, 42–45]; this fraction belongs to a rare high-effect-size allele model that, as we validated, is characterized by similar allele frequency change between cases and controls. This high effect size fraction, however, is not difficult for GWAS to discover, and indeed APOEe4 was found by much less advanced methods a decade before GWAS commenced. For this reason, we focus in this report on low effect size genetic architecture models (A, B and C), which the latest consensus considers to be the likeliest hiding place for GWAS lost heritability; the results apply to all model genetic architectures described here.

Please refer to Materials and Methods for a description of individual values analysis (IVA) and cohort simulation (cohort) for interpretation of the following simulation results. We found that, similarly to the validation simulations, the average LOD polygenic score of the population changes with age.

Fig 1 shows the change in the average polygenic score for IVA cases and controls in the case of coronary artery disease; see similar plots for all LODs in S4 Fig. 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 each year. The average polygenic score 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 (S4 Fig).

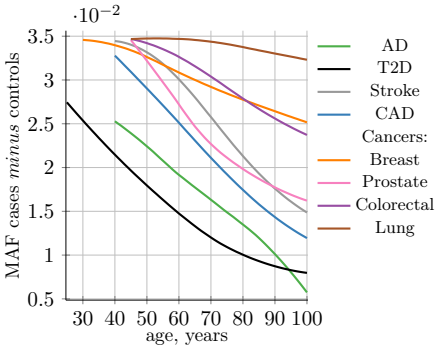

Fig 2. Allele frequency difference between newly diagnosed individuals and remaining population of the same age

Common low-effect-size alleles (scenario A); largest effect variant MAF = 0.5; OR = 1.15. The **MAF cases *minus* controls** value is used to determine GWAS statistical power; see Eq (7). Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change; see S6 Fig.

When we look at the allele level, we see that this change 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. This is because 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. Fig 2 shows a summary for the highest effect allele, MAF = 0.5, OR = 1.15, for all LODs for the individual scenario (IVA). Detailed information for multiple alleles with effect sizes ranging from highest to lowest for all LODs can be found in S6 Fig. These figures show the most 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.

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 3 and the more detailed information for all LODs presented in S7 Fig.

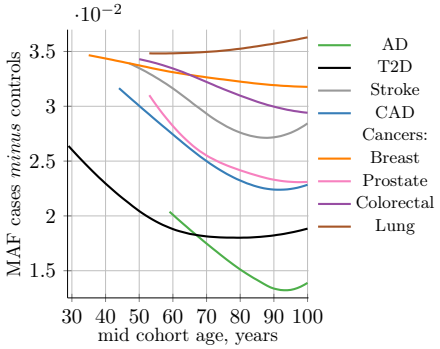

Fig 3. Allele frequency difference between cases and controls, by cohort age

Common low-effect-size alleles (scenario A); largest effect variant: MAF=0.5; OR=1.15. The **MAF cases *minus* controls** value is used to determine GWAS statistical power; see Eq (7). Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change; see S7 Fig.

While the MAF difference between cases and controls shown in the figures is illustrative by itself, it is most important for determining the GWAS statistical discovery power using Eq (7), Eq (9), 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 is exemplified by CAD in Fig 4.

In the hypothetical IVA case, the number of individuals required to achieve GWAS discovery power increases fast. 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 can be seen in Fig 4. The respective plots for all LODs can be found in S8 Fig and S9 Fig. 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 changes more than one-hundred fold.

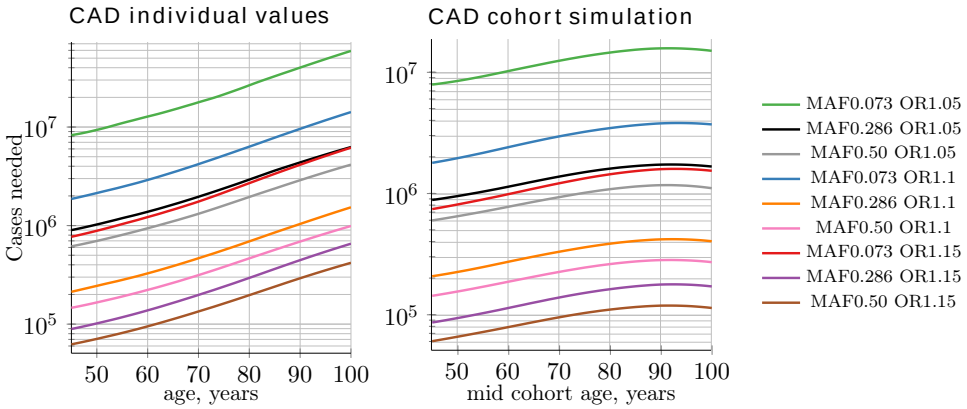

Fig 4. Number of cases needed to achieve 0.8 discovery power for IVA and a GWAS cohort: coronary artery disease example

Common low-effect-size alleles (scenario A). These two plots show the comparative number of cases needed to achieve 0.8 GWAS statistical discovery power. The curve representing the individual diagnosed versus the unaffected population of the same age continues a steep rise in the IVA scenario. 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 number of necessary cases 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).

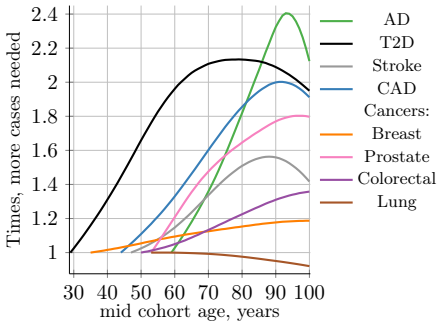

Fig 5. Relative increase in number of cases needed for 0.8 discovery power in cohort study when using progressively older same-age case and control cohorts

The youngest age cohort for each LOD is defined as the mid-cohort age at which the cumulative incidence for a cohort first reaches 0.5% of the population. We consider this the minimum cumulative incidence age allowing for the formation of well-powered cohort studies. 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.

Common low-effect-size alleles (scenario A) 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.

This value is very similar among all eight LODs, and the highest effect allele for each LOD requires $5 \cdot 10^4$–$1 \cdot 10^5$ 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 S9 Fig. Fig 5 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 our eight LODs. Additional figures detailing each LOD can be found in Supporting Information; see S4 Fig and S5 Fig. These cohort results are simulated with identical mortality for cases and controls. Mortality has an impact on the cohort allele distribution. We considered it necessary to validate more extreme mortality scenarios—both lower and higher—than one could expect in a real cohort study. We found that 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 diseases we analyze. While it may have been a

realistic scenario a century ago before modern health care, it certainly is lower these days. In addition, as the most extreme validation case, we used a one-year cohort without mortality. This scenario can also be considered an individual cumulative case. It takes as cases everyone ill at each age and everyone healthy at that age as controls. These validation cohort scenarios are summarized in S10 Fig.

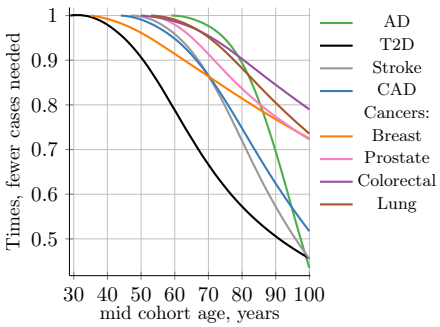

Fig 6. Relative decrease in number of cases needed for 0.8 discovery power in cohort study when using progressively older control cohorts compared to fixed-age young-cases cohort.
The youngest age cohort for each LOD is defined as the mid-cohort age at which the cumulative incidence for a cohort first reaches 0.25% of the population. We consider this the minimum cumulative incidence age allowing for the formation of well-powered cohort studies. Therefore, the leftmost point on each LOD line is the reference (youngest) cases cohort. Control mid-cohort ages are incremental ages, and as a cohort ages, the cohort case number multiple of fewer cases/controls required to achieve 0.8 statistical power is relative to this earliest cohort.
Common low-effect-size alleles (scenario A). While all alleles display different magnitudes 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; see S16 Fig.

The mortality analysis was applied to one LOD at a time. We did not attempt to estimate increased mortality for multiple disease diagnoses. Collerton et al. [46] 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. Other genetic architecture models 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 models. In Supporting Information, we present the medium-effect-size alleles (scenario D); see S11 Fig, S12 Fig, S13 Fig, and S14 Fig. There, at younger ages, the MAF difference between cases and controls is larger for 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 exclusively medium-effect-size- alleles being causally associated with the LODs we review here, because GWAS studies would have an easier time discovering a large number of SNPs. From a qualitative perspective, all reviewed genetic architecture models provide similar results.

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 6, S15 Fig, and S16 Fig.

In the next section, we will discuss and summarize these findings.

Discussion

For each of the LODs other than cancers, the earliest diagnosed individual's polygenic risk score is very high, and the higher the heritability, the larger the fraction of higher risk individuals. This is in line with Pawitan et al.'s [47] genetic architecture model, which expects higher variance in the polygenic score to achieve higher heritability.

Table 1 shows the mean polygenic odds ratios of diagnosed individuals with age. While the ages of initial onset and incidence progression, comparing odds ratio for early

onsets, and the ages of 80 and 100 correspond to Table 1 in S1 Appendix, showing that predictive power for the four LODs other than cancers is best at younger ages, specifically < 65 years of age for AD, < 55 for CAD, < 60 for stroke, and < 50 for T2D. The polygenic score of the earliest cases under the genetic architecture models we used, particularly for AD, rivals that of the Mendelian genetic diseases.

Table 1. Simulated polygenic score odds ratio of cases compared to initial population mean

	Highly prevalent LODs				Cancers			
Disease	AD	T2D	Stroke	CAD	Breast	Prostate	Colorectal	Lung
OR early	1500	320	10	41	4.3	43 (10)	8.7	1.4
OR, 80y	98	1.3	2.8	4.5	2.7	9 (4.1)	5.1	1.4
OR, 100y	0.07	0.3	0.82	0.85	2.2	3.8 (2.5)	3.3	1.3

Note: Prostate cancer heritability is 57% [48]. Shown in braces 42% heritability [49], in line with other cancers we analyze. The effective hazard ratio after the Eq (15) moderates. The early age is chosen as the earliest age achieving mid-cohort incidence of 0.25% for the corresponding LODs, as described in Methods and Materials.

Turning our attention to cancers, we find that even initial onset odds ratios are relatively low, and these ratios change much less with age than for the above LODs. Prostate cancer is the only cancer that is somewhat controversial. Its heritability is reported at 57% by [48], 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 other cancers, according to our model. Yet, the same article [48] 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 [48], or perhaps this is a phenomenon specific to the populations or environmental effects of Nordic countries.

Perhaps the earlier familial study [49], which reported a heritability estimate of 42%, would be closer to the UK population incidence data we used in our simulations. We ran a verification simulation with 42% heritability, and the values matched the patterns exhibited by other cancers.

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; see S4 Fig. This confirms the clinical observation that the major risk factor for LODs is age itself.

Our simulations also support the conclusion that, even at the most advanced age, there will remain a fraction of population whose genotype makes it very unlikely that they will be diagnosed with a particular LOD within their lifespan. This also matches clinical observations. Recent news reports revealed that the oldest known person alive, and the last person known to have been born in the 19th century, Nabi Tajima, died in Japan last year at the age of 117. At advanced age, mortality rapidly decreases an individual's chances of being diagnosed with an additional LOD, and extremely few live as long as Tajima.

We mentioned earlier a longstanding observation that the heritability of LODs decreases with age. This observation, based on prominent examples such as decreasing T2D, CAD, AD, or stroke heritability with age, could lead to a conclusion that this is a general rule. For a more thorough review of the chosen LODs, see S1 Appendix, where [1, 18, 20, 41–45, 48, 50–111], and [112] have shown that if we were to make a general statement that the heritability of LODs always decreases with age, we would not be quite correct. In S1 Appendix, we describe some notable exceptions.

The LODs with heritability that diminishes with age—Alzheimer's disease, coronary artery disease, and type 2 diabetes—exhibit a noticeable and well-known decline in

heritability in familial studies and when applied to prediction of a person's liability to a disease based on parental history. The volume of literature is somewhat smaller for cerebral stroke, but the conclusions are similar to CAD. The GWAS research shows better results if younger participants are used in cohort studies; see S1 Appendix.

For three of the four most prevalent cancers, the reported twin heritability has shown relatively constant heritability with age progression. Determining lung cancer heritability has proven somewhat more elusive for researchers, and no definitive conclusions have been published, to a large extent due to the low documented heritability and substantial environmental component of this disease. We think that our simulations and the general pattern they exposed explain the reason for these difficulties.

It is also not inconceivable that the age effect may not be exactly linear and the relative LOD odds ratios may to a small extent increase with age. Thus, prostate cancer—and all cancers, for that matter—would show more constant heritability with age than is justified by case/control MAF and odds ratio changes. If this is so, it appears that the non-linear effect is relatively minor because it does not cause more constant heritability for the four highly prevalent LODs other than cancers.

Table 2 combines the heritability and incidence of the LODs with the summarized simulation results from the cohort simulation (also S5 Fig).

Table 2. Summary of LODs' heritability and incidence and corresponding case/control ΔMAF and required cohort size change with age

	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	69	38–44	50–60	31	57(42)	40	8–18
Maximum incidence %	> 20	2.5	4.4	3.6	<0.5	<0.8	<0.6	<0.6
	ΔMAF between cases and controls							
early	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 stat power							
early	$1.4\cdot 10^5$	$8.7\cdot 10^4$	$5.3\cdot 10^4$	$6.0\cdot 10^4$	$5.0\cdot 10^4$	$6.1\cdot 10^4$	$4.9\cdot 10^4$	$4.9\cdot 10^4$
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	$3.0\cdot 10^5$	$1.7\cdot 10^5$	$7.3\cdot 10^4$	$1.1\cdot 10^5$	$5.9\cdot 10^4$	$1.1\cdot 10^5$	$6.9\cdot 10^4$	$4.5\cdot 10^4$
Cases mult., 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. For simplicity, we show the allele with MAF = 0.5, OR = 1.15, the allele that requires the smallest number of cases/controls in this genotype scenario. "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. In braces, value for 42% prostate cancer heritability.

This table shows a snapshot for a simulated GWAS study with an age span of 10 years near the earliest disease onset and centered on 80 years of age. The table summarizes the results for the largest-effect allele.

GWAS statistical discovery power is impaired by the change in individual distribution of the polygenic score. 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.

It is notable that, except for AD, which requires 1.5 times more cases at age 100, for the other three LODs, the number of cases needed does not change significantly after the age of 80. The cancers show a small increase in the number of participants required to achieve the same statistical power, with only one outlier: prostate cancer.

The cohort studies benefit from the fact that the diagnosed individuals are accumulated from the youngest onset to the age of a case in the cohort study. Comparatively, individual values analysis, in which the individuals diagnosed each year are compared to all not-yet-sick individuals, shows much faster change in the number of cases hypothetically needed to achieve the same statistical power; see S4 Fig.

To find a mitigating scenario for the GWAS' loss of discovery power with increasing age of participants, we analyzed a scenario in which the age of the cases is fixed at the lowest reasonably possible cohort age for the eight LODs and the control age is chosen from increasingly older cohorts. Discovery power improves with increased control cohort age; see summary Fig 6, S15 Fig, and S16 Fig. As expected, 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. In this scenario, notably (20–25%) fewer participants are also needed to achieve the same statistical power in cancer GWAS, including lung cancer.

Conclusions

We undertook this research 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.

We found that these phenomena can indeed be explained and predicted primarily by the heritability of the LODs and their cumulative incidence progression. By simulating population age progression under the assumption of relative disease liability remaining proportionate to individual polygenic risk, we found 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, as they become ill proportionate to their polygenic risk score or their odds ratio of becoming ill.

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, and remains high at older ages, leading to a high cumulative incidence for some LODs. We explain the increasing incidence rate while polygenic risk decreases for the as-yet-unaffected population as a consequence of the aging process itself being the major LOD risk factor.

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. On a yearly basis, the difference in the highest-effect allele frequency between newly diagnosed individuals and the remaining population changes quite rapidly for these diseases. In a typical cohort design, cohort GWAS statistical discovery power is less affected than it is in individual values analysis. This is due to

the fact that GWAS cohorts are composed of individuals whose disease accumulated from an earlier age up to the time at which they were included as cases in the study cohort. On the older age spectrum, the mortality of the population ultimately limits the increase in the number of oldest patients in LOD study cohorts.

Our 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, we conclude that 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. While finding a high number of young cases may be problematic, for most LODs, there is an ample number of still-unaffected individuals at older ages.

In addition, we find that the expectations placed on the predictive power of GWAS polygenic scores for diseases with high heritability and high cumulative incidence should be relaxed. In part, this exaggerated expectation is caused by an obvious perception bias held by readers of the published familial and twin heritability studies.

Usually, maximal heritability numbers are more likely to attract readers' notice, and often only these high numbers are published or discussed. Twin studies, in the rare cases when they report heritability change with age, show much smaller heritability at older ages, as we demonstrated in S1 Appendix.

Typically, GWAS reviews use these maximum numbers to set their expectations. These high heritability values are plainly incorrect for older ages for four of our LODs. Lower heritability at an older age means that people who have no genetic or familial susceptibility are increasingly becoming sick with an LOD.

Not all LODs are affected in this way; LODs with low cumulative incidence and low familial heritability produce a much 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 practically unaffected by the age of the participant cohorts.

As a final conclusion, for LODs like AD, T2D, stroke, and CAD, adjusting the cohort selection could lead to achieving the same GWAS discovery power with half as many participants, rather than, in the worst case scenario, requiring perhaps double the number of participants with progressively older cohorts with the traditional age matched cohort design.

Materials and methods

The model definition

Let us shortly review a formal model, which will lead to the population simulation approach we implemented. One could attempt to determine an individual probability of succumbing to an LOD based on a polygenic risk score. The environmental effect, which includes both aging and cumulative external environmental effects, acts as a multiplier over the genetic component. Therefore, the yearly LOD incidence of population u can be described by Eq 1:

$$I(t) = \frac{E(t)}{N(t)} \sum_{u=1}^{N(t)} G_u, \tag{1}$$

where the yearly incidence $I(t)$ is the fraction of individuals diagnosed with a disease out of the individuals as yet unaffected at the start of the year and $E(t)$ is the

environmental effect, a multiplier reflecting increased liability due to the ongoing aging processes.

For the purposes of this model, we consider the age/environment effect as changing with age identically for all individuals of the same age. G_u is the genotype liability distribution for as-yet-undiagnosed individuals $u \in 1, N$. We use the letter u to symbolize the undiagnosed count. $N(t)$ denotes the remaining unaffected population at age t, after accounting for previously diagnosed individuals and the accumulated mortality.

We divide by $N(t)$—the number of individuals at the beginning of each year—to obtain the incidence, rather than the number of individuals newly diagnosed with an LOD within a year. Each individual's genetic risk is considered fixed from birth. The genotype mix of the individuals changes with time, not only with the remaining number of individuals.

In this equation, G_u can be simply understood as a function of probability for each individual based on that individual's genotype liability. Then, for each year, the multiple $E(t) \cdot G_u$ can be understood as each individual's probability of becoming ill. With advancing age, the multiplier $E(t)$ increases, as does the probability of becoming ill. Then, we could try to infer the function $E(t)$.

However, GWASs and clinical studies use odds ratios rather than probabilities. GWAS also uses a polygenic score for individuals, genotypes, and SNPs. We would have to determine a representation of probability from a polygenic score. After inferring $E(t)$, we would run this function against the same population to identify individuals becoming ill every year, analyze their allele distribution, and draw conclusions based on what we find. The model function is reasonable, but this inference approach would not be very straightforward and, if successful, would be open to differences in interpretation and therefore questions of applicability of the results to the polygenic score while GWAS and clinical studies operate with odds-ratio-based polygenic scores.

Taking an aging population simulation approach allows us 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.

Knowing the yearly incidence of an LOD and the polygenic risk scores (ORs) for each individual based on modeled LOD genetic architecture, we use Algorithm 1 as the core of our simulations.

```
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
        end
        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)
    end
end
```

Note: an individual makes a sampling target proportionate to hazard ratio (HR) in the $LOOKUP()$ table. ORs are converted to HRs similar to [113]. 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.

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

Descriptively, this algorithm works as follows. We set up a simulation in which each

next individual to be diagnosed with an LOD is chosen proportionately to the relative risk of the polygenic odds ratio assigned 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 curve that we derive from clinical statistics. In this manner, we probabilistically reproduce the aging process using a population simulation model rather than a computational model. As simulation progresses, we track the risk alleles for all newly diagnosed individuals and the remaining unaffected population and statistically analyze their representation in the affected and remaining population.

In the next subsection, we will describe the model allele architecture and the incidence models we use as the parameters of this algorithm.

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 [114]:

$$I_{lifetime} = \sum_{t=0}^{t_{max}} I(t), \tag{2}$$

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 [114]:

$$CumRisk = 1 - e^{-I_{lifetime}}. \tag{3}$$

Cumulative risk Eq (3) 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.

Allele distribution models and statistical analysis

An in-depth review [47] extensively analyzed models of genetic architecture and through their simulations determined the number of alleles required to achieve specific heritability and estimate GWASs' discovery power. The authors calculated allele distributions and heritabilities and ran simulations for six combinations of effect sizes and minor allele frequencies (MAFs). By relying on the conclusions of [47], we were able to avoid repeating the preliminary step of evaluating the allele distributions needed to achieve the heritability level for the LODs assessed in this research.

These allele frequencies represent the whole spectrum ranging from common low-frequency low-effect-size alleles to very rare high-effect high-frequency alleles. In our simulations, we verify the five most relevant architectures; see Table 3.

Table 3. 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 [47].

We also found it handy for repeatable allele tracking, rather than generating the continuous random spectrum of allele frequencies and effect sizes, to follow the [47] 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 model 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. Having multiple well-defined alleles with the same parameters facilitated the tracking of their behaviors with age, LOD, and simulation incidence progression.

We calculated an individual polygenic risk score as a sum of all alleles' effect sizes, which is by definition a log(OR) (odds ratio) for each allele, also following Eq (4) Eq (5) from [47]:

$$log(OR) = \sum_k a_k log(OR_k), \tag{4}$$

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. In our publication figures, for brevity, we also use the notation β ($\beta = log(OR)$).

Variance of the allele distribution is determined by:

$$var = 2 \sum p_k(1 - p_k)(log(OR_k))^2, \tag{5}$$

where p_k is the frequency of the k-th genotype [47].

The contribution of genetic variance to the risk of the disease is heritability:

$$h^2 = \frac{var(g)}{var(g) + \pi^2/3}, \tag{6}$$

where $\pi^2/3$ is the variance of the standard logistic distribution [115].

For each allele in our simulated population, we track the allele frequency, which changes with age for cases and controls. The difference between these MAFs gives the non-centrality parameter λ for two genetic groups [8,116]:

$$\lambda = N * p_1 * p_2 * (\beta_1 - \beta_2)^2, \tag{7}$$

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

Having obtained NCP λ from Eq (7), [116] recommends using SAS or similar statistical software to calculate the statistical power, using the following SAS statement:

$$StatPower = 1 - PROBF(FINV(0.99999995, 1, N - 4), 1, N - 4, \lambda) \quad . \tag{8}$$

We converted this equation to its R equivalent, which we use to process the simulation output, as follows:

$$StatPower = 1 - pf(qf(PSign, 1, N - 4), 1, N - 4, \lambda) \quad , \tag{9}$$

where $PSign = 0.99999995$ corresponds to $5 \cdot 10^{-8}$ significance level. We validated the outputs of our conversion with the Online Sample Size Estimator (OSSE) [117]. This equation returns statistical power based on a case/control number and the NCP as calculated above.

For the purposes of our simulation, we used an equal number of cases and controls. To find the number of cases needed for 80% GWAS discovery power, having the $(\beta_1 - \beta_2)$, we iterated the value of N using a rapid convergence R routine until the value of $StatPower$ was equal to 0.8 with an accuracy better than $\pm 0.01\%$ for each age and allele distribution of interest.

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:

$$I(t) = a, \tag{10}$$

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:

$$I(t) = b\,t, \tag{11}$$

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:

$$I(t) = 3.05 \cdot 10^{-5} e^{0.1178t}, \tag{12}$$

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

In all five scenarios from Table 3, 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%, corresponding to the genetic architecture scenarios set out in Table 3; see Table 4.

Table 4. 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 allele architecture scenarios are defined in Table 3. 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 S1 Fig.

Between the genetic architectures, there is also a relatively small difference in the polygenic scores of the population and the cases; see S2 Fig. 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.

LOD incidence functional approximation

Next, 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 Fig 7. 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 [57]. 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, we used a constant level for the exponential approximation 13 [118].

Alzheimer's disease, on the other hand, continues exponentially past the age of 95, reaching incidences above 20% [58]. 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. We wrote an R script (see S1 File) to automate 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 clinical statistics presented in the Supplementary Information tables.

The incidence approximation $I(t)$ is represented mathematically by Eq (13). 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.

$$I(t) = \begin{cases} ae^{bt} + c, & \text{until intersection with the line, below} \\ i + st, & \text{thereafter} \end{cases} \tag{13}$$

A logistic approximation of the clinical data is shown as a red line on Fig 7. It is characterized by the following equation:

$$I(t) = \frac{a}{1 + e^{(c-t)/b}} + d, \tag{14}$$

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

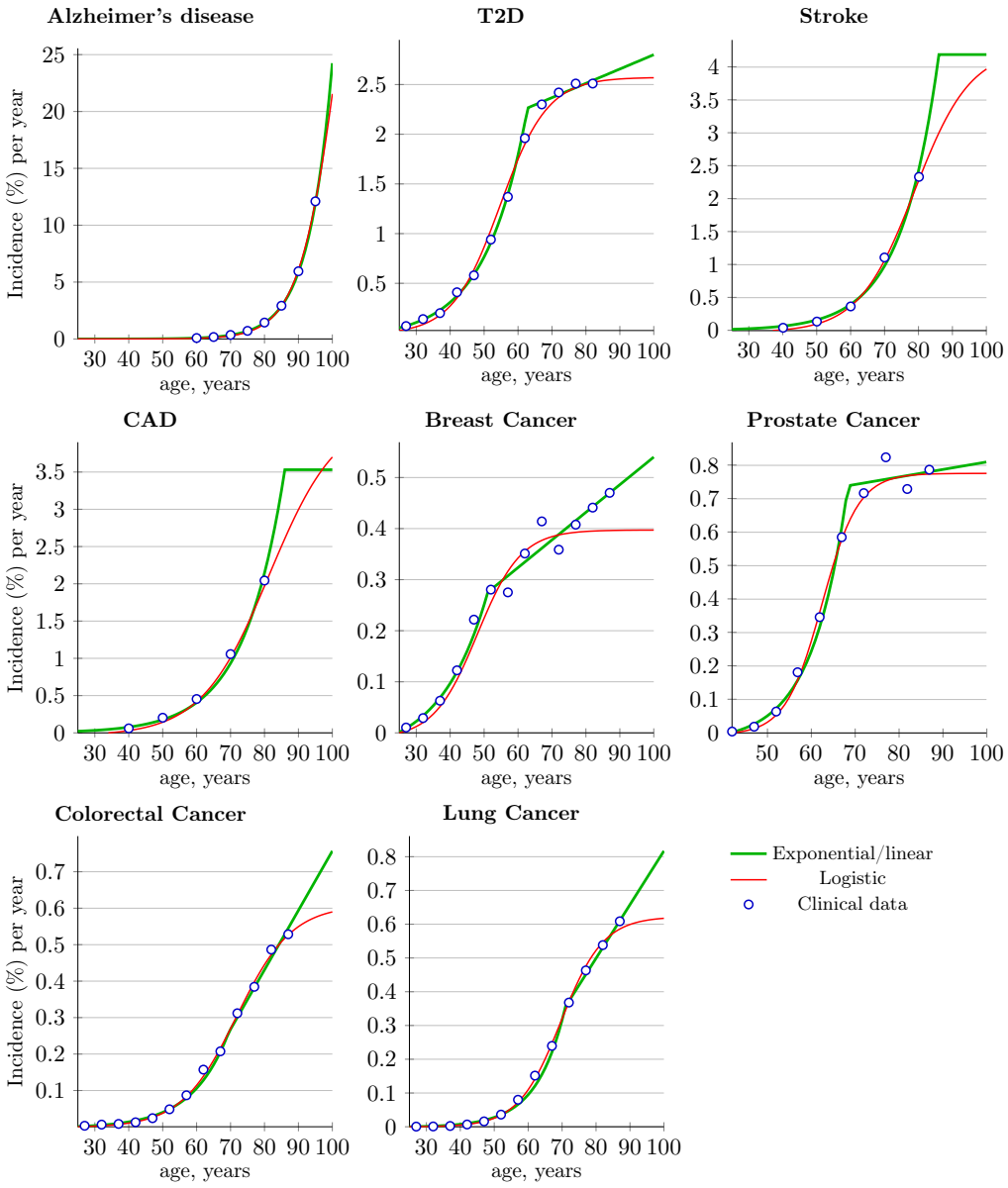

Fig 7. LOD clinical incidence rates and functional approximations
Two functional approximations of clinical data: exponential followed by linear and logistic. The R script automating NLM (nonlinear model) regression for both approximation curves is available in Supporting Information.

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 Fig 7.

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.

Using the model from [47] outlined in Table 3, scenario A, with the common low-effect variants above, the number of variants needed for the above LODs is summarized in Table 5.

Table 5. Heritability of analyzed LODs and an example required numbers of 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

Note: due to the genetic model using evenly spaced variants and MAFs, as described above, the counts of variants are necessarily in multiples of 25.

The polygenic scores of the simulated population are based on odds ratios built using the logit model [47]. 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, [119] treat OR and RR as equivalent in case of breast cancer in their simulation study. For higher values an OR usually significantly exceeds RR. We use an adjustment formula by [120] approximating OR to hazard ratio subject to our modeled incidence for the LODs under consideration using equation:

$$HR_u = \frac{OR_u}{1 + I(t) \cdot (OR_u - 1)}, \tag{15}$$

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

The simulation design

The simulation design employs several steps that are common in population genetics simulations. The gene variants pool is built as outlined above and summarized in Table 3 for an appropriate scenario from Table 4 or Table 5. The population individuals were allocated and, based on the genetic architecture model chosen, the polygenic score was generated for each individual.

To track the GWAS statistical discovery power, for the final simulation runs, the same nine representative variants were tracked for all LODs simulated.

Each simulation run was iterated through all eight LODs and analyzed two scenarios: a per-year-of-age population individual values analysis and a simulation of a GWAS clinical study with a population consisting of individuals of mixed ages.

Individual values analysis and cohort simulation

It can be expected from this observation that the higher an LOD incidence and heritability, the faster the fraction of the highest-polygenic-score individuals will diminish with age and, progressively at an earlier age, lower-polygenic-score individuals will represent the bulk of the LOD cases.

While the general principle is now determined from the validation simulations, 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 incidence and cumulative incidence levels; see Fig 7.

In the absence of mortality, both due to general frailty and other LODs, the incidence progression leads us to believe 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.

We performed two main LOD simulation types:

1. **The individual values analysis of polygenic scores and risk allele frequency for individuals diagnosed with disease at each specific age and the remaining population at this age.** For brevity, we also interchangeably use "IVA" in this publication. The individual values analysis uses one-year age slices, as follows.

First, the mean and variance of the polygenic score for the whole population were calculated. Then, for each year, the step of determining (sampling out) the individuals who had become ill that year followed. Based on the required incidence value for each year, several individuals were picked from the unaffected population through randomly sampling the population with a probability proportionate to the individual's polygenic score odds ratio, as summarized in Algorithm 1.

When completed, these individuals became the cases for the relevant year's individual values analysis, and the mean and variance of their polygenic score were also calculated and recorded. Note that mortality does not need to be applied to this simulation 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.

In addition, the specified variants of interest were also counted for cases and controls and later used to estimate GWAS statistical power.

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, we also use "cohort" throughout this publication.

The clinical study cohort simulation performs an analysis identical to that described above. The difference is that we simulate GWAS clinical studies 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. We use the mid-cohort age in our statistics, which is the arithmetic half-age of the cohort age span. In the first simulation year, a population to equal 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 be subject to the disease incidence rate and mortality according to each individual age.

Mortality was 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. For validation, we ran additional simulations with (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. 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, we do not know the extent to which younger controls differ from cases, except for the fact that they are not currently diagnosed—not unlike a real study cohort.

We can expect a quite smoothed pattern of polygenic score change if we watch this cohort aging. 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 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. We consider this the minimum cumulative incidence age allowing for the formation of well-powered cohort studies.

Data sources, programming and equipment

We used population mortality numbers from the 2014 US Social Security "Actuarial Life Table" [121], which provides 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 S1 Appendix for corroboration: Alzheimer's disease [58, 122–124], type 2 diabetes [57], coronary artery disease and cerebral stroke [118], and cancers [59, 77].

To run simulations, we used 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++ using MS Visual Studio 2015. 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; 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 we describe 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 S1 File.

Acknowledgments

My gratitude goes to Alexei J. Drummond at the University of Auckland for a number of helpful and challenging discussions.

References

1. Murphy SL, Xu J, Kochanek KD, Curtin SC, Arias E. Mortality in the United States, 2016. NCHS Data Brief, no 293. 2017;293.

2. Franceschi C, Garagnani PG, Morsiani C, Conte M, Santoro A, Grignolio A, et al. The continuum of aging and age-related diseases: common mechanisms but different rates. Frontiers in Medicine. 2018;5:61.

3. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics. 2017;101(1):5–22.

4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–753.

5. Clarke AJ, Cooper DN. GWAS: heritability missing in action? European Journal of Human Genetics. 2010;18(8):859.

6. Kumar SK, Feldman MW, Rehkopf DH, Tuljapurkar S. Limitations of GCTA as a solution to the missing heritability problem. Proceedings of the National Academy of Sciences. 2016;113(1):E61–E70.

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

8. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nature Reviews Genetics. 2014;15(5):335.

9. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. Genome biology. 2016;17(1):241.

10. Lynch M. Mutation and human exceptionalism: our future genetic load. Genetics. 2016;202(3):869–875.

11. Gao F, Keinan A. High burden of private mutations due to explosive human population growth and purifying selection. BMC genomics. 2014;15(4):S3.

12. Muller HJ. Our load of mutations. American journal of human genetics. 1950;2(2):111.

13. Morton NE, Crow JF, Muller HJ. An estimate of the mutational damage in man from data on consanguineous marriages. Proceedings of the National Academy of Sciences. 1956;42(11):855–863.

14. [Internet]. OMIM Gene Map Statistics; 2018 (accessed May 31, 2018). Available from: `http://omim.org/statistics/geneMap`.

15. Jager RD, Mieler WF, Miller JW. Age-related macular degeneration. New England Journal of Medicine. 2008;358(24):2606–2617.

16. Sobrin L, Ripke S, Yu Y, Fagerness J, Bhangale TR, Tan PL, et al. Heritability and genome-wide association study to assess genetic differences between advanced age-related macular degeneration subtypes. Ophthalmology. 2012;119(9):1874–1885.

17. Lakatta EG, Levy D. Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises: Part I: aging arteries: a "set up" for vascular disease. Circulation. 2003;107(1):139–46.

18. Warner SC, Valdes AM. The genetics of osteoarthritis: A review. Journal of Functional Morphology and Kinesiology. 2016;1(1):140–153.

19. Fedarko NS. Theories and Mechanisms of Aging. In: Geriatric Anesthesiology. Nature Publishing Group; 2018. p. 19–25.

20. Naj AC, Schellenberg GD. Genomic variants, genes, and pathways of Alzheimer's disease: An overview. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2017;174(1):5–26.

21. Silva CT, Kors JA, Amin N, Dehghan A, Witteman JC, Willemsen R, et al. Heritabilities, proportions of heritabilities explained by GWAS findings, and implications of cross-phenotype effects on PR interval. Human genetics. 2015;134(11-12):1211–1219.

22. Oh J, Lee YD, Wagers AJ. Stem cell aging: mechanisms, regulators and therapeutic opportunities. Nature medicine. 2014;20(8):870.

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

24. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature genetics. 2012;44(4):369–375.

25. Thornton KR, Foran AJ, Long AD. Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS genetics. 2013;9(2):e1003258.

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

27. Goldstein DB, et al. Common genetic variation and human traits. New England Journal of Medicine. 2009;360(17):1696.

28. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS biology. 2010;8(1):e1000294.

29. North TL, Beaumont M. Complex trait architecture: the pleiotropic model revisited. Scientific reports. 2015;5:9351.

30. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proceedings of the National Academy of Sciences. 2011;108(44):18026–18031.

31. So HC, Gui AH, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genetic epidemiology. 2011;35(5):310–317.

32. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature genetics. 2014;46(11):1173–1186.

33. Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nature biotechnology. 2014;32(4):347–355.

34. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339(6121):819–823.

35. Haney MS, Kramer NJ, Morgens DW, Jovičić A, Couthouis J, Li A, et al. CRISPR-Cas9 screens in human cells and primary neurons identify modifiers of C9orf72 dipeptide repeat protein toxicity. bioRxiv. 2017; p. 129254.

36. Boeke JD, Church G, Hessel A, Kelley NJ. Genome Project-write: A Grand Challenge Using Synthesis, Gene Editing and Other Technologies to Understand, Engineer and Test Living Systems October 31, 2016; 2018 (accessed July 8, 2018). Available from: `http://engineeringbiologycenter.org/resources/`.

37. Thompson D, Aboulhouda S, Hysolli E, Smith C, Wang S, Castanon O, et al. The future of multiplexed eukaryotic genome engineering. ACS chemical biology. 2017;13:313–325.

38. Kohman RE, Kunjapur AM, Hysolli E, Wang Y, Church GM. From Designing the Molecules of Life to Designing Life: Future Applications Derived from Advances in DNA Technologies. Angewandte Chemie. 2018;57(16):4313–4328.

39. Zaitlen N, Lindström S, Pasaniuc B, Cornelis M, Genovese G, Pollack S, et al. Informed conditioning on clinical covariates increases power in case-control association studies. PLoS genetics. 2012;8(11):e1003032.

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

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

42. Tan L, Yu JT, Zhang W, Wu ZC, Zhang Q, Liu QY, et al. Association of GWAS-linked loci with late-onset Alzheimer's disease in a northern Han Chinese population. Alzheimer's & dementia. 2013;9(5):546–553.

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

44. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis. Jama. 1997;278(16):1349–1356.

45. Davidson Y, Gibbons L, Pritchard A, Hardicre J, Wren J, Stopford C, et al. Apolipoprotein E $\varepsilon 4$ allele frequency and age at onset of Alzheimer's disease. Dementia and geriatric cognitive disorders. 2007;23(1):60–66.

46. Collerton J, Davies K, Jagger C, Kingston A, Bond J, Eccles MP, et al. Health and disease in 85 year olds: baseline findings from the Newcastle 85+ cohort study. Bmj. 2009;339:b4904.

47. Pawitan Y, Seng KC, Magnusson PK. How many genetic variants remain to be discovered? PloS one. 2009;4(12):e7969.

48. Hjelmborg JB, Scheike T, Holst K, Skytthe A, Penney KL, Graff RE, et al. The heritability of prostate cancer in the Nordic Twin Study of Cancer. Cancer Epidemiology and Prevention Biomarkers. 2014;23(11):2303–2310.

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

50. Ahmad A, Ormiston-Smith N, Sasieni P. 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. 2015;112(5):943.

51. Almgren P, Lehtovirta M, Isomaa B, Sarelin L, Taskinen M, Lyssenko V, et al. Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia. 2011;54(11):2811.

52. Amoako AO, Pujalte GGA. Osteoarthritis in young, active, and athletic individuals. Clinical Medicine Insights: Arthritis and Musculoskeletal Disorders. 2014;7:CMAMD–S14386.

53. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, 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. 2003;72(5):1117–1130.

54. Aparicio HJ, Seshadri S. Familial Occurrence and Heritability of Stroke. In: Stroke Genetics. Springer; 2017. p. 9–20.

55. Bevan S, Traylor M, Adib-Samii P, Malik R, Paul NL, Jackson C, et al. Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. Stroke. 2012;43(12):3161–3167.

56. Binder N, Schumacher M. Incidence of Dementia over Three Decades in the Framingham Heart Study. The New England journal of medicine. 2016;375(1):92–93.

57. Boehme MW, Buechele G, Frankenhauser-Mannuss J, Mueller J, Lump D, Boehm BO, et al. 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. 2015;15(1):855.

58. Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. American journal of public health. 1998;88(9):1337–1342.

59. [Internet]. Cancer Statistics for the UK; 2018 (accessed May 31, 2018). Available from: `http://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk`.

60. Carr SR, Akerley W, Hashibe M, Cannon-Albright LA. Evidence for a genetical contribution to non-smoking-related lung cancer. Thorax. 2015; p. thoraxjnl–2014.

61. de Miguel-Yanes JM, Shrader P, Pencina MJ, Fox CS, Manning AK, Grant RW, 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. 2011;34(1):121–125.

62. de Voer RM, Hahn MM, Weren RD, Mensenkamp AR, Gilissen C, van Zelst-Stams WA, et al. Identification of novel candidate genes for early-onset colorectal cancer susceptibility. PLoS genetics. 2016;12(2):e1005880.

63. Devan WJ, Falcone GJ, Anderson CD, Jagiella JM, Schmidt H, Hansen BM, et al. Heritability estimates identify a substantial genetic contribution to risk and outcome of intracerebral hemorrhage. Stroke. 2013;44(6):1578–1583.

64. Nijpels G. Diapedia: Epidemiology of type 2 diabetes; 2018 (accessed May 31, 2018). Available from: `https://www.diapedia.org/type-2-diabetes-mellitus/3104287123/epidemiology-of-type-2-diabetes`.

65. Eeles R, Al Olama AA, Berndt S, Wiklund F, Conti DV, Ahmed M, 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;.

66. Elks CE, Den Hoed M, Zhao JH, Sharp SJ, Wareham NJ, Loos RJ, et al. Variability in the heritability of body mass index: a systematic review and meta-regression. Frontiers in endocrinology. 2012;3:29.

67. Ford D, Easton D, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The American Journal of Human Genetics. 1998;62(3):676–689.

68. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. Nature. 2016;536(7614):41–47.

69. Graff RE, Möller S, Passarelli MN, Witte JS, Skytthe A, Christensen K, et al. Familial risk and heritability of colorectal cancer in the nordic twin study of cancer. Clinical Gastroenterology and Hepatology. 2017;15(8):1256–1264.

70. Guedj A, Geiger-Maor A, Galun E, Amsalem H, Rachmilewitz J. Early age decline in DNA repair capacity in the liver: in depth profile of differential gene expression. Aging (Albany NY). 2016;8(11):3131.

71. Haley B. Hereditary breast cancer: the basics of BRCA and beyond. UT Southwestern Medical Center Internal Medicine Grand Rounds. 2016;.

72. Hjelmborg J, Korhonen T, Holst K, Skytthe A, Pukkala E, Kutschke J, et al. Lung cancer, genetic predisposition and smoking: the Nordic Twin Study of Cancer. Thorax. 2016; p. thoraxjnl–2015.

73. Jee SH, Suh I, Won SY, Kim MY. Familial correlation and heritability for cardiovascular risk factors. Yonsei medical journal. 2002;43(2):160–164.

74. Jiao S, Peters U, Berndt S, Brenner H, Butterbach K, Caan BJ, et al. Estimating the heritability of colorectal cancer. Human molecular genetics. 2014;23(14):3898–3905.

75. Kanwal M, Ding XJ, Cao Y. Familial risk for lung cancer. Oncology Letters. 2017;13(2):535–542.

76. Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, et al. Residential radon and risk of lung cancer: a combined analysis of 7 North American case-control studies. Epidemiology. 2005;16(2):137–145.

77. Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. Jama. 2017;317(23):2402–2416.

78. Lecarpentier J, Silvestri V, Kuchenbaecker KB, Barrowdale D, Dennis J, McGuffog L, et al. Prediction of breast and prostate cancer risks in male BRCA1 and BRCA2 mutation carriers using polygenic risk scores. Journal of Clinical Oncology. 2017;35(20):2240–2250.

79. [Internet]. Lifetime Risk of Developing or Dying From Cancer. American Cancer Society; 2018 (accessed May 31, 2018). Available from: `https://www.cancer.org/cancer/cancer-basics/lifetime-probability-of-developing-or-dying-from-cancer.html`.

80. Lloyd-Jones DM, Leip EP, Larson MG, d'Agostino RB, Beiser A, Wilson PW, et al. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation. 2006;113(6):791–798.

81. Loughlin J. Genetic contribution to osteoarthritis development: current state of evidence. Current opinion in rheumatology. 2015;27(3):284.

82. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. The Lancet. 2014;383(9921):999–1008.

83. Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. European Respiratory Journal. 2016; p. ERJ–00359.

84. Mancuso N, Rohland N, Rand KA, Tandon A, Allen A, Quinque D, et al. The contribution of rare variation to prostate cancer heritability. Nature genetics. 2016;48(1):30.

85. Marusyk A, DeGregori J. Declining cellular fitness with age promotes cancer initiation by selecting for adaptive oncogenic mutations. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer. 2008;1785(1):1–11.

86. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M. Genetic susceptibility to breast cancer. Molecular oncology. 2010;4(3):174–191.

87. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92.

88. Möller S, Mucci LA, Harris JR, Scheike T, Holst K, Halekoh U, et al. The heritability of breast cancer among women in the Nordic Twin Study of Cancer. Cancer Epidemiology and Prevention Biomarkers. 2016;25(1):145–150.

89. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial risk and heritability of cancer among twins in Nordic countries. Jama. 2016;315(1):68–76.

90. Nelson P, Masel J. Intercellular competition and the inevitability of multicellular aging. Proceedings of the National Academy of Sciences. 2017; p. 201618854.

91. Nielsen M, Andersson C, Gerds TA, Andersen PK, Jensen TB, Køber L, et al. Familial clustering of myocardial infarction in first-degree relatives: a nationwide study. European heart journal. 2013;34(16):1198–1203.

92. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. Diabetologia. 1999;42(2):139–145.

93. Ralston SH, Uitterlinden AG. Genetics of osteoporosis. Endocrine reviews. 2010;31(5):629–662.

94. Ribezzo F, Shiloh Y, Schumacher B. Systemic DNA damage responses in aging and diseases. In: Seminars in cancer biology. vol. 37. Elsevier; 2016. p. 26–35.

95. Risch HA, McLaughlin JR, Cole DE, Rosen B, Bradley L, Fan I, et al. Population BRCA1 and BRCA2 mutation frequencies and cancer penetrances: a kin–cohort study in Ontario, Canada. Journal of the National Cancer Institute. 2006;98(23):1694–1706.

96. Schmit SL, Schumacher FR, Edlund CK, Conti DV, Ihenacho U, Wan P, et al. Genome-wide association study of colorectal cancer in Hispanics. Carcinogenesis. 2016;37(6):547–556.

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

98. Seshadri S, Beiser A, Pikula A, Himali JJ, Kelly-Hayes M, Debette S, et al. Parental occurrence of stroke and risk of stroke in their children: the Framingham study. Circulation. 2010;121(11):1304–1312.

99. Shaffer JR, Kammerer CM, Bruder JM, Cole SA, Dyer TD, Almasy L, et al. Genetic influences on bone loss in the San Antonio Family Osteoporosis study. Osteoporosis international. 2008;19(12):1759–1767.

100. Shi L, Zheng M, Hou J, Zhu B, Wang X. Regulatory roles of epigenetic modulators, modifiers and mediators in lung cancer. In: Seminars in cancer biology. vol. 42. Elsevier; 2017. p. 4–12.

101. Skousgaard SG, Hjelmborg J, Skytthe A, Brandt LPA, Möller S, Overgaard S. 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. 2015;17(1):336.

102. Skousgaard SG, Skytthe A, Möller S, Overgaard S, Brandt L. 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. 2016;18:46–46.

103. Spector TD, MacGregor AJ. Risk factors for osteoarthritis: genetics. Osteoarthritis and cartilage. 2004;12:39–44.

104. Talmud PJ, Cooper JA, Morris RW, Dudbridge F, Shah T, Engmann J, et al. Sixty-five common genetic variants and prediction of type 2 diabetes. Diabetes. 2014; p. DB_141504.

105. Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. 2015;347(6217):78–81.

106. Walsh PC. The search for the missing heritability of prostate cancer. European urology. 2017;72(5):657–659.

107. Wang DC, Wang X. Tomorrow's genome medicine in lung cancer. In: Seminars in cancer biology. vol. 42. Elsevier; 2017. p. 39–43.

108. Weissfeld JL, Lin Y, Lin HM, Kurland BF, Wilson DO, Fuhrman CR, et al. Lung cancer risk prediction using common SNPs located in GWAS-identified susceptibility regions. Journal of Thoracic Oncology. 2015;10(11):1538–1545.

109. Wienke A, Holm NV, Skytthe A, Yashin AI. The heritability of mortality due to heart diseases: a correlated frailty model applied to Danish twins. Twin Research and Human Genetics. 2001;4(4):266–274.

110. Wu X, Gu J. Heritability of prostate cancer: a tale of rare variants and common single nucleotide polymorphisms. Annals of translational medicine. 2016;4(10).

111. Wu YT, Beiser AS, Breteler MM, Fratiglioni L, Helmer C, Hendrie HC, et al. The changing prevalence and incidence of dementia over time - current evidence. Nature Reviews Neurology. 2017;.

112. Zdravkovic S, Wienke A, Pedersen N, Marenberg M, Yashin A, De Faire U. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. Journal of internal medicine. 2002;252(3):247–254.

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

114. Sasieni P, Shelton J, Ormiston-Smith N, Thomson C, Silcocks P. What is the lifetime risk of developing cancer?: the effect of adjusting for multiple primaries. British journal of cancer. 2011;105(3):460.

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

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

117. [Internet]. Online Sample Size Estimator. The Bioinformatics Institute; 2018 (accessed May 31, 2018). Available from: `http://osse.bii.a-star.edu.sg/`.

118. Rothwell P, Coull A, Silver L, Fairhead J, Giles M, Lovelock C, et al. 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. 2005;366(9499):1773–1783.

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

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

121. [Internet]. Actuarial Life Table. Social Security Administration (US); 2018 (accessed May 31, 2018). Available from: `https://www.ssa.gov/oact/STATS/table4c6.html`.

122. Edland SD, Rocca WA, Petersen RC, Cha RH, Kokmen E. Dementia and Alzheimer disease incidence rates do not vary by sex in Rochester, Minn. Archives of neurology. 2002;59(10):1589–1593.

123. Kokmen E, Chandra V, Schoenberg BS. Trends in incidence of dementing illness in Rochester, Minnesota, in three quinquennial periods, 1960–1974. Neurology. 1988;38(6):975–975.

124. Hebert LE, Scherr PA, Beckett LA, Albert MS, Pilgrim DM, Chown MJ, et al. Age-specific incidence of Alzheimer's disease in a community population. Jama. 1995;273(17):1354–1359.

125. Allport SA, Kikah N, Saif NA, Ekokobe F, Atem FD. Parental age of onset of cardiovascular disease as a predictor for offspring age of onset of cardiovascular disease. PloS one. 2016;11(12):e0163334.

126. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Archives of general psychiatry. 2006;63(2):168–174.

127. Lyra-Junior PC, Tessarollo NG, Guimarães IS, Henriques TB, dos Santos DZ, de Souza ML, et al. GWAS in Breast Cancer. In: Breast Cancer-From Biology to Medicine. InTech; 2017. p. 99–117.

128. Yang IA, Holloway JW, Fong KM. Genetic susceptibility to lung cancer and co-morbidities. Journal of thoracic disease. 2013;5(Suppl 5):S454.

129. Polderman TJ, Benyamin B, De Leeuw CA, Sullivan PF, Van Bochoven A, Visscher PM, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature genetics. 2015;47(7):702.

Supporting information

S1 Appendix. LODs heritability patterns with age based on familial, GWAS, and clinical studies A review of the clinical, GWAS, and familial studies of polygenic LOD heritability within the typical age range of disease onset leads to grouping LODs into two broad categories: decreasing heritability with age and increasing or relatively constant heritability with age. We review these categories in detail, focusing primarily on eight of the most prevalent LODs we analyze in our simulations.

S1 Appendix
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 6 summarized from [63, 109, 112], and [54].

Table 6. 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 [41] 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 [51]), 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 [104] 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 [61]. Meanwhile, [51] 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, [68] 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 [91] 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 [97] 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 [98] 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 [125].

The heritability patterns for these diseases are summarized in Table 8. 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 [55, 63, 91, 97, 98], and [68]. These studies find that familial history is the better predictor of next-generation disease only when the parental generation participants are relatively young; see [51, 61, 104], and Table 6.

An environmental effect on the heritability of cardiovascular disease and T2D with age is evident [41, 92], including influences such as spousal environment [73].

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 [57]. 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 [66].

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 [57, 64, 80, 82].

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 [58]. The heritability of Alzheimer's disease is estimated at 80% from twin studies [20]; the heritability is 79% [126] at approximately 65 years of age and diminishes with increasing age, as indicated both by familial and GWAS studies [20, 42, 43].

A clinical study documenting the association between the APOE genotype and Alzheimer's disease [44, 45] 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 7.

Table 7. 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 [44].

Another review [20] 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 [44] summarized in Table 7.

In Table 8, we summarize the information in the literature about the decreasing heritability of the LODs referenced above.

The model presented in [58] 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 [56, 111] 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.

Table 8. Heritability and risk statistics for decreasing heritability with age LODs

Disease	*Heritability/risk, younger age*	*Heritability/risk, older age*
AD e3/e4 [44]	OR=3.8, 65y	OR=1.7, 90y
AD e4/e4 [44]	OR=15.0, 60y	OR=2.0, 90y
CAD paternal [91]	RR=3.30, $< 50y$	RR=1.83, $> 50y$
CAD maternal [91]	RR=3.23, $< 50y$	RR=2.31, $> 50y$
Stroke [97]	OR=1.63, $< 60y$	OR=0.77, $> 70y$
Stroke all [98]	RR=3.79, $< 65y$	RR=2.21, $> 65y$
Stroke ischemic [98]	RR=5.45, $< 65y$	RR=2.47, $> 65y$
T2D [51]	$h^2 = 0.69$, 35-60y	$h^2 = 0.31$, 35-75y

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

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 [101] 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 [18, 52]. 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 [103].

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 [102].

The osteoporosis findings are similarly varied, with studies finding no heritability of pathology for some bone structures and strong heritability for others [93]. Specifically, the osteoporosis associated with bone breaks is very heritable and shows a slight increase in heritability into older age [99]. 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 [18, 81]. 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 [79], and the 2016 fraction of mortality death directly attributed to cancer was 21.8%, the second-highest after heart disease [1]. In the UK, the corresponding numbers are higher, at 47% and 53%, respectively [50, 59], 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 [71]. 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 [71] 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 [95].

An early study [67] 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% [59, 79]. As Haley [71] 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 [67].

Moller et al. [88] 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 [77] 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 9. 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 [53] collects the same kind of statistics as [77] and arrives at similar conclusions.

Table 9. BRCA1/2 carriers incidence rate by age, data from [77]

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 [88] 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, [86] and [71] 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 [71, 95].

To date, GWAS attempts to find common polygenic variants of low effect size have had only limited success. One review study [127] outlines the history and accomplishments of breast cancer GWAS over a decade of research. The most recent high-powered consortium study [87] 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 [88] and the 27% estimated by [89].

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 [87] 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 [88], 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 ([78]) 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% [59, 79]. PC heritability is estimated at 57% [48, 89] and 42% by an older study [49].

The Nordic twin study [48] 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 [78, 84, 106, 110]. Recently, Eeles et al. [65], 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 [110] 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 [110] 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 [79]. In the UK, the corresponding numbers are 5% and 7% [59].

The Nordic twin studies [69, 89] 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 [69]. 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 [62, 74].

Graff et al.'s study [69] 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 [62]) concur that penetrant high-effect gene variants are found in 5–10% of CRC cases. A GWAS review [96] 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 [79]. In the UK, the corresponding numbers are 5.9% and 7.6% [59].

The LC pattern of heritability is not easy to ascertain. According to Kanwal [75], approximately 8% of lung cancers are inherited or occur as a result of a genetic predisposition. The Nordic twin studies review [89] 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 [72].

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 [60, 76, 83]. 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 [75].

Due to the low heritability of LC, GWAS' success at identifying predictive common SNPs has been limited [108]. Some studies explain part of the LC incidence by reference to causal epigenetic effects [100]. The heritability value of 18% that Mucci et al. [89] gives has a very broad range. An earlier study [128] 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. [89] 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 [129]. 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 [48, 69, 71, 88] and a more recent study that is especially informative about BRCA1/2 mutations' incidence with age [77] finally allowed us to determine that cancer heritability remains relatively constant with age. Table 10 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.

Table 10. 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 highly detrimental mutations (%)	10–14	10–12	5–10	minor
Polymorphic incidence (%)	86–90	88–90	90–95	major
Heritability trend (50y–100y)	flat / slight decline [53, 67, 71, 77, 86–89, 95]	flat / slight incline [48, 65, 78, 84, 106, 110]	flat [62, 69, 74, 96]	likely flat [60, 72, 75, 76, 83, 89, 100, 107, 108]

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 [70, 85, 90, 94, 105].

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.

S1 File. Source code, executable, scripts and configurations. This zip file contains the simulation executable, the source code and the project for MS Visual Studio 2015, R scripts and corresponding batch files for producing functional approximations of clinical incidence and post analysis of simulation results.

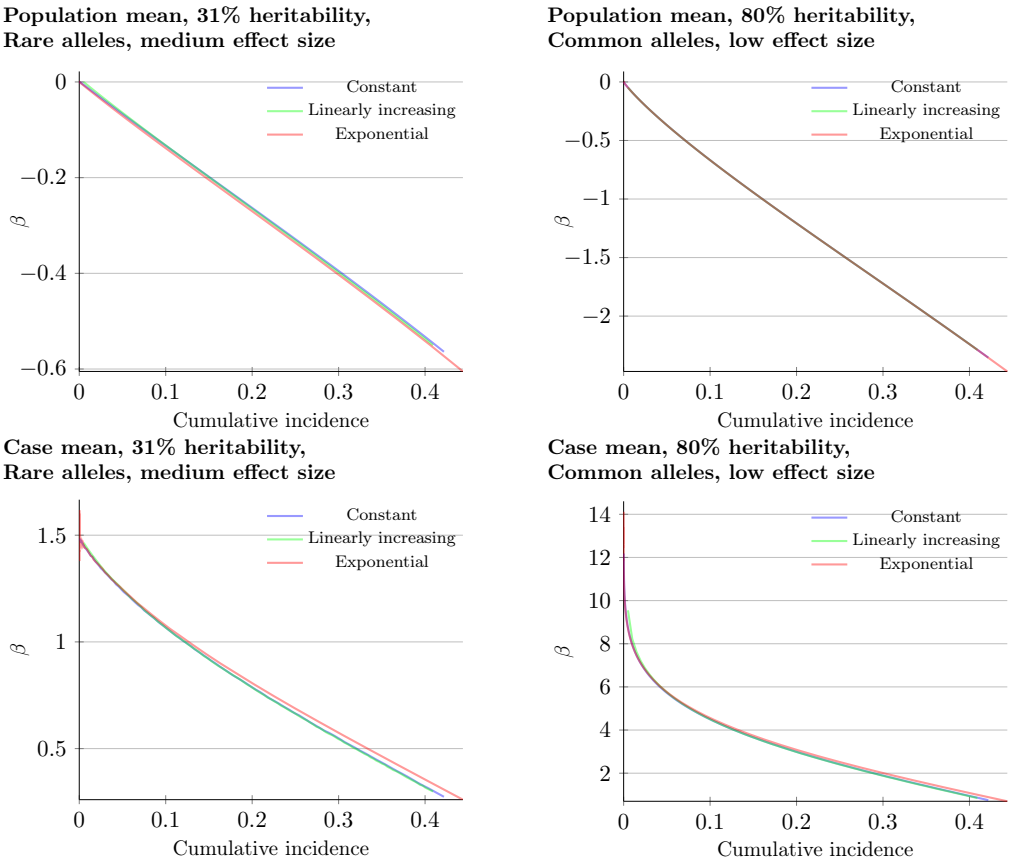

Fig 8. S1 Fig

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; $\beta = log(OddsRatio)$.

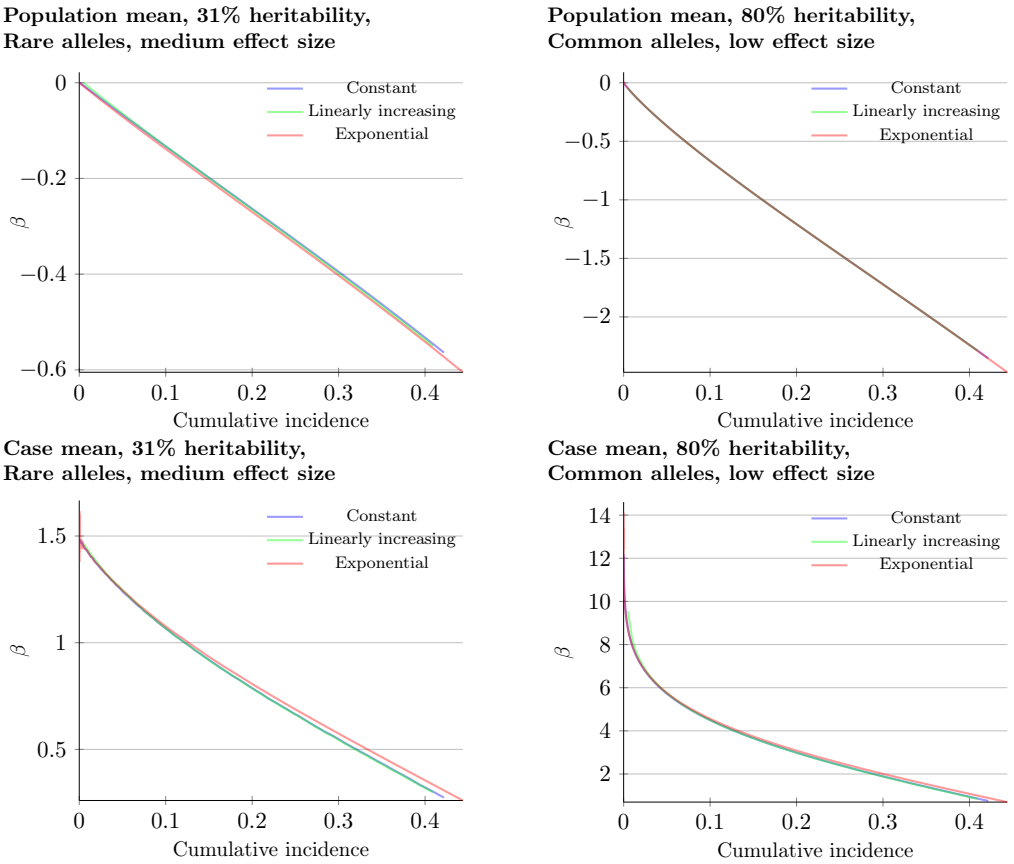

Fig 9. S2 Fig

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; $\beta = log(OddsRatio)$.

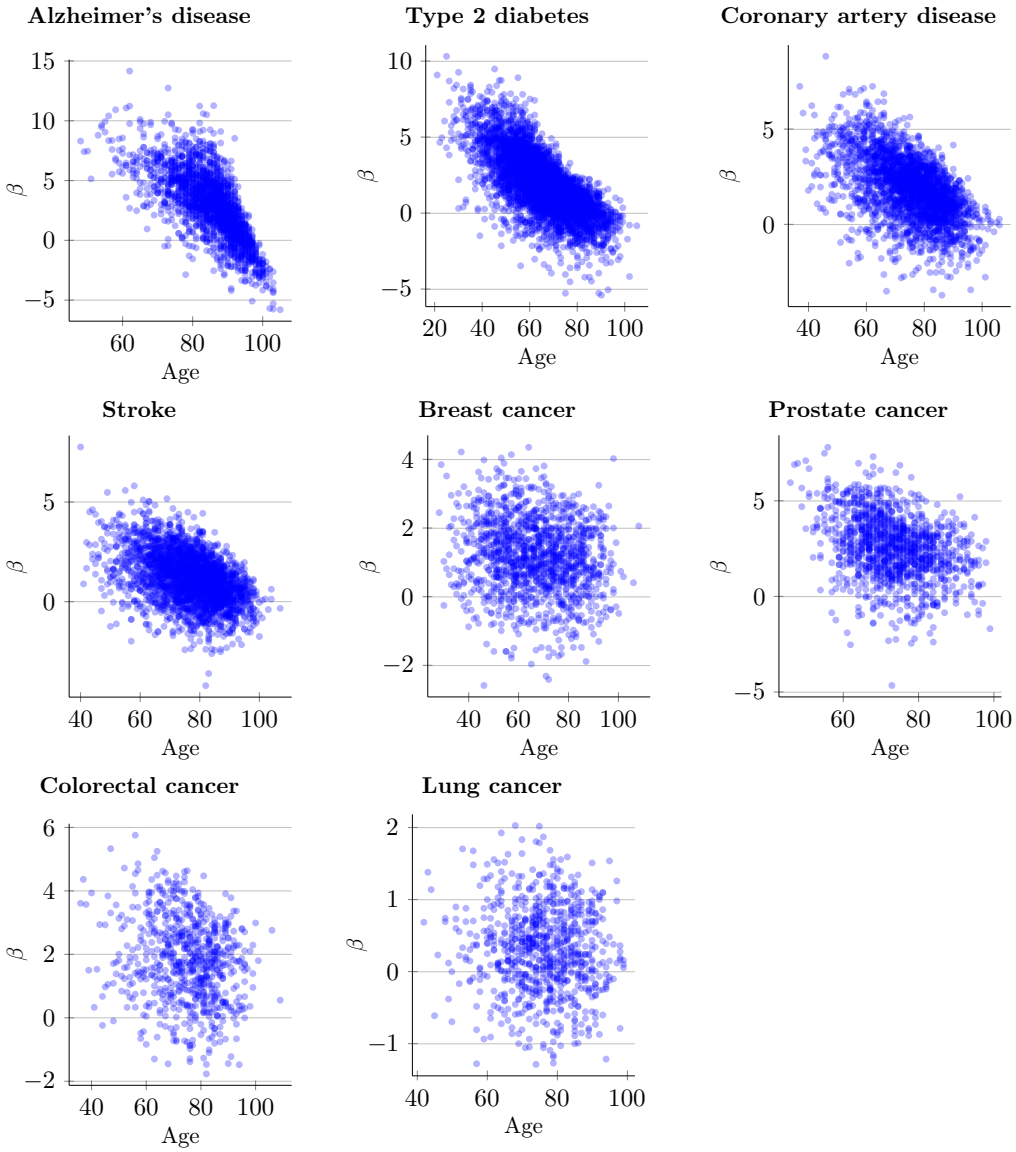

Fig 10. S3 Fig

Polygenic scores of individuals diagnosed with LOD as a function of age; IVA Scatter plots show the distributions of polygenic scores for cases diagnosed as age progresses. $\beta = log(OddsRatio)$, with mortality, visually implies the underlying heritability and incidence magnitudes. If the regression line drops diagonally, there is a combination of high heritability and increasing cumulative incidence. Otherwise, a plot appears as a relatively symmetrical blob.

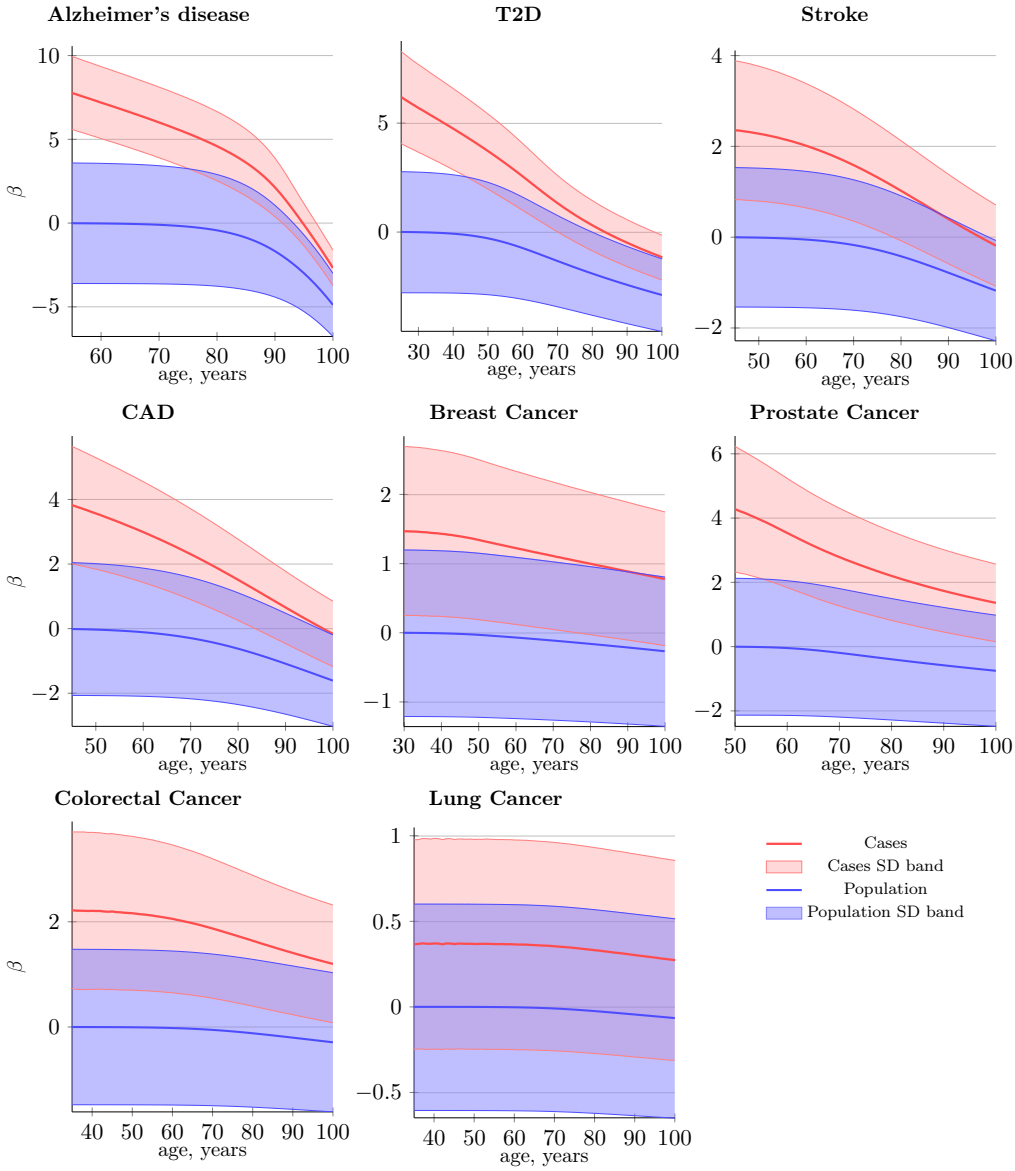

Fig 11. S4 Fig

Polygenic score difference between newly diagnosed individuals and the remaining unaffected population; IVA 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. For all highly prevalent LODs, the mean polygenic risk of new cases crosses below the risk of an average person at early onset age.

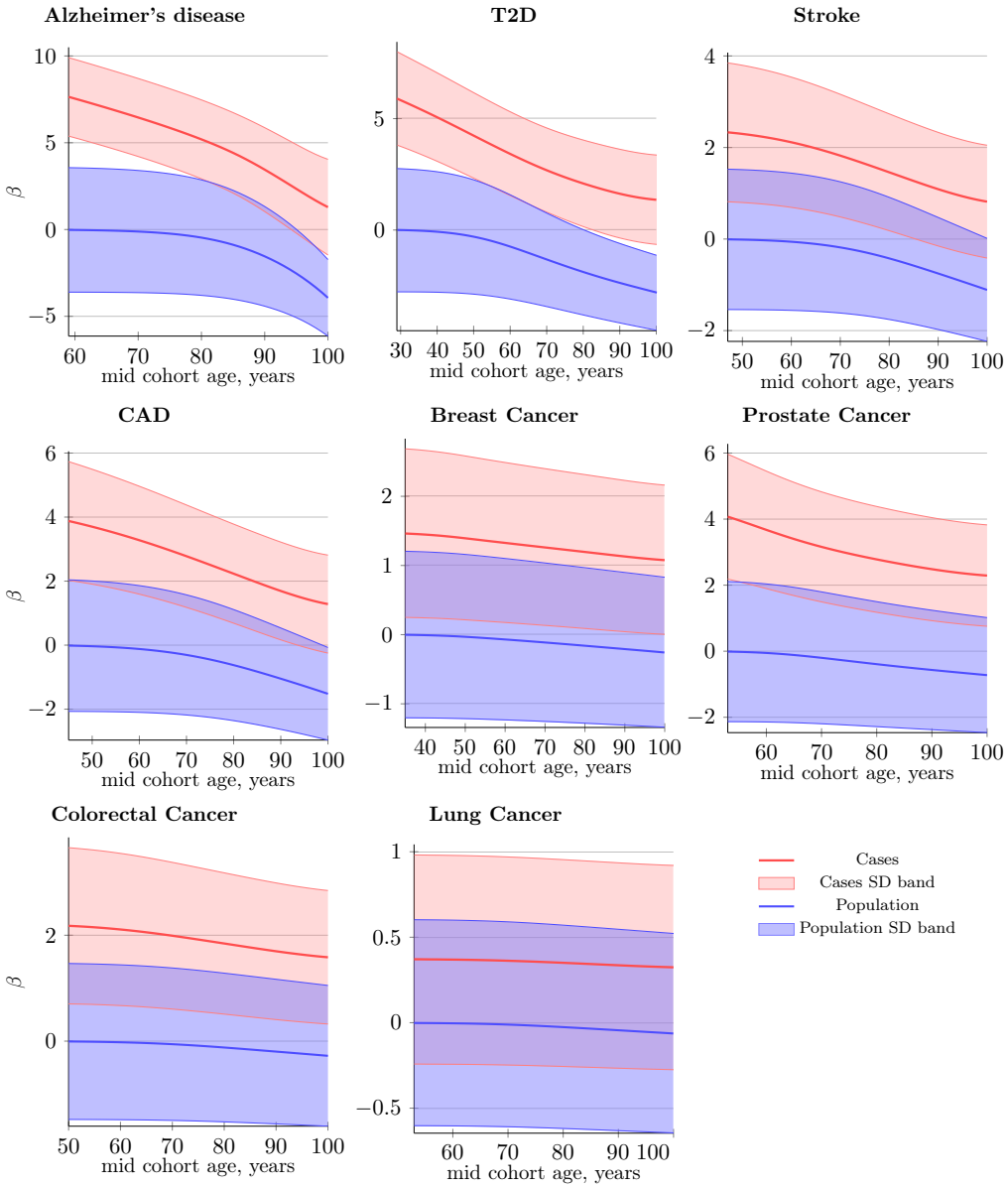

Fig 12. S5 Fig

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.

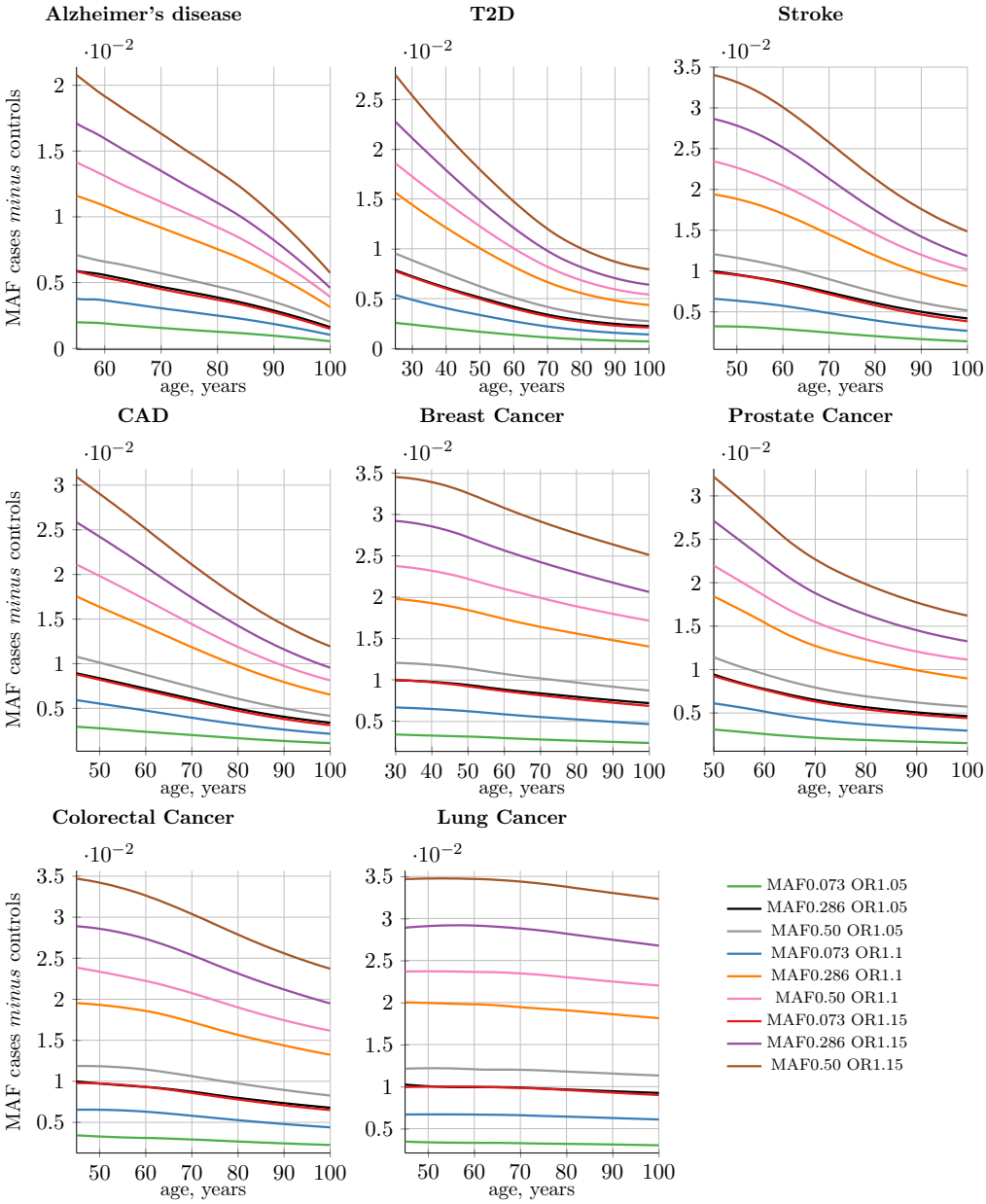

Fig 13. S6 Fig

Allele frequency difference between newly diagnosed instances and the remaining unaffected population; IVA Common low-effect-size alleles (scenario A). The **MAF cases *minus* controls** value is used to determine GWAS statistical power; see Eq (7). Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change.

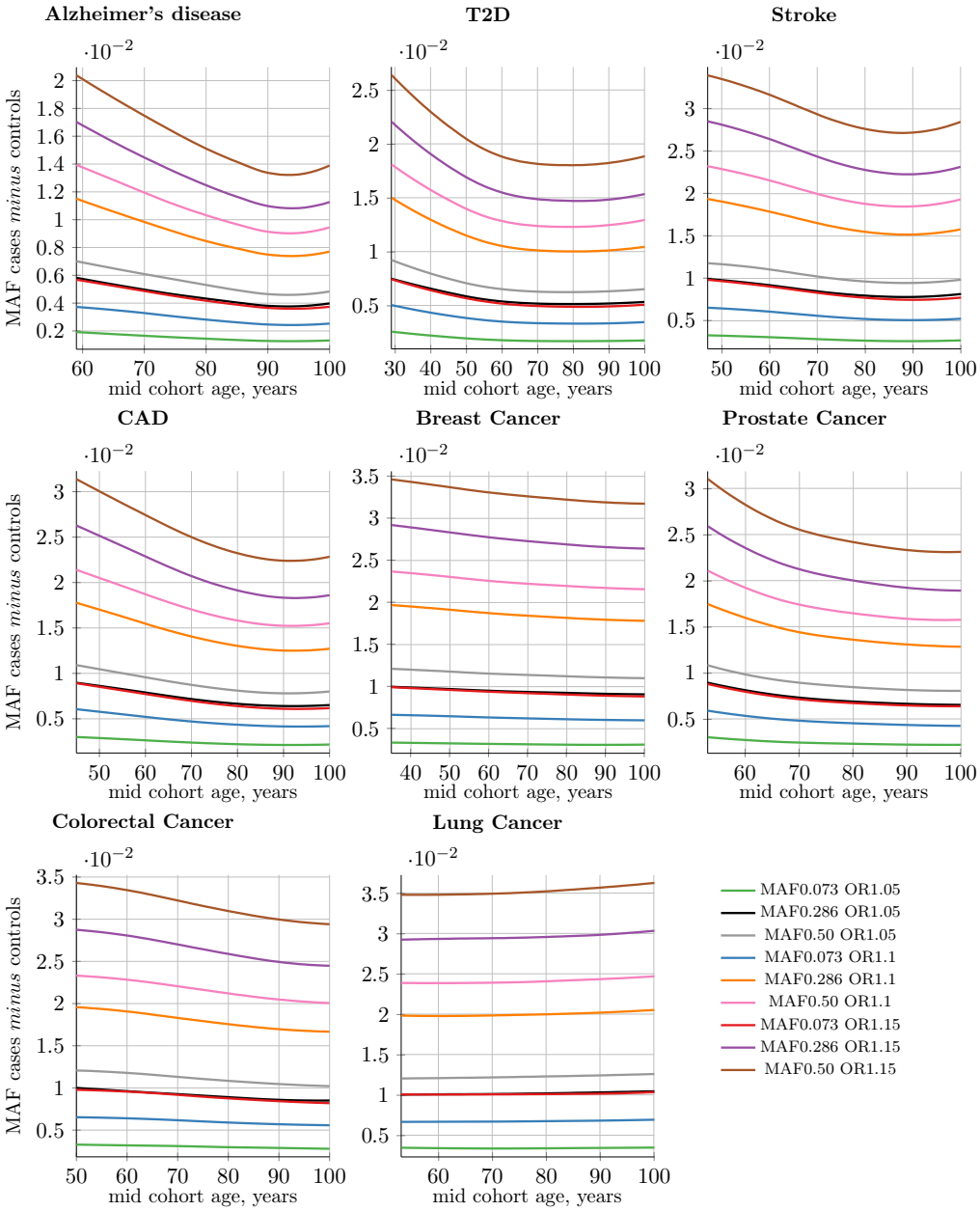

Fig 14. S7 Fig

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; see Eq (7). Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change.

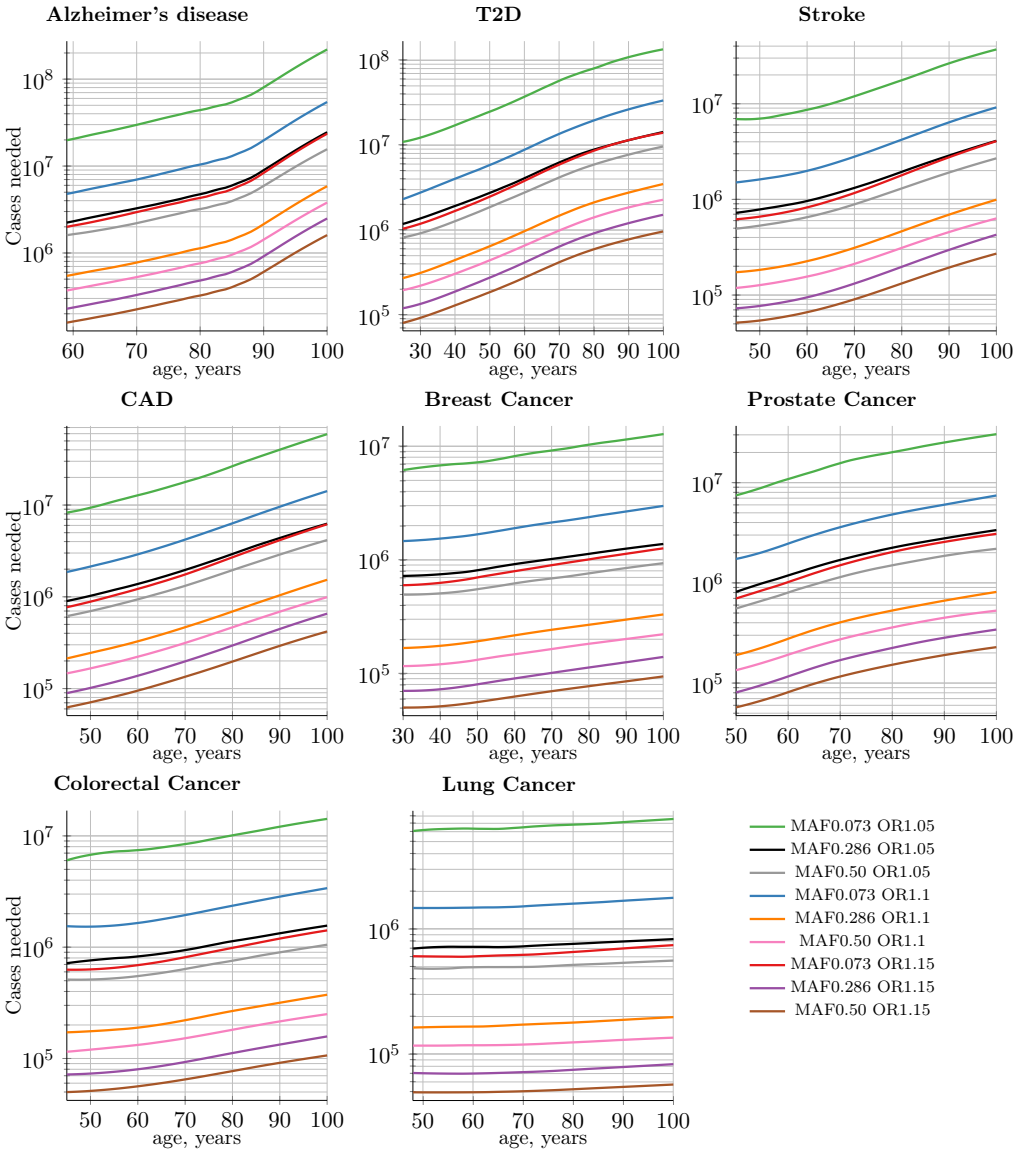

Fig 15. S8 Fig

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.

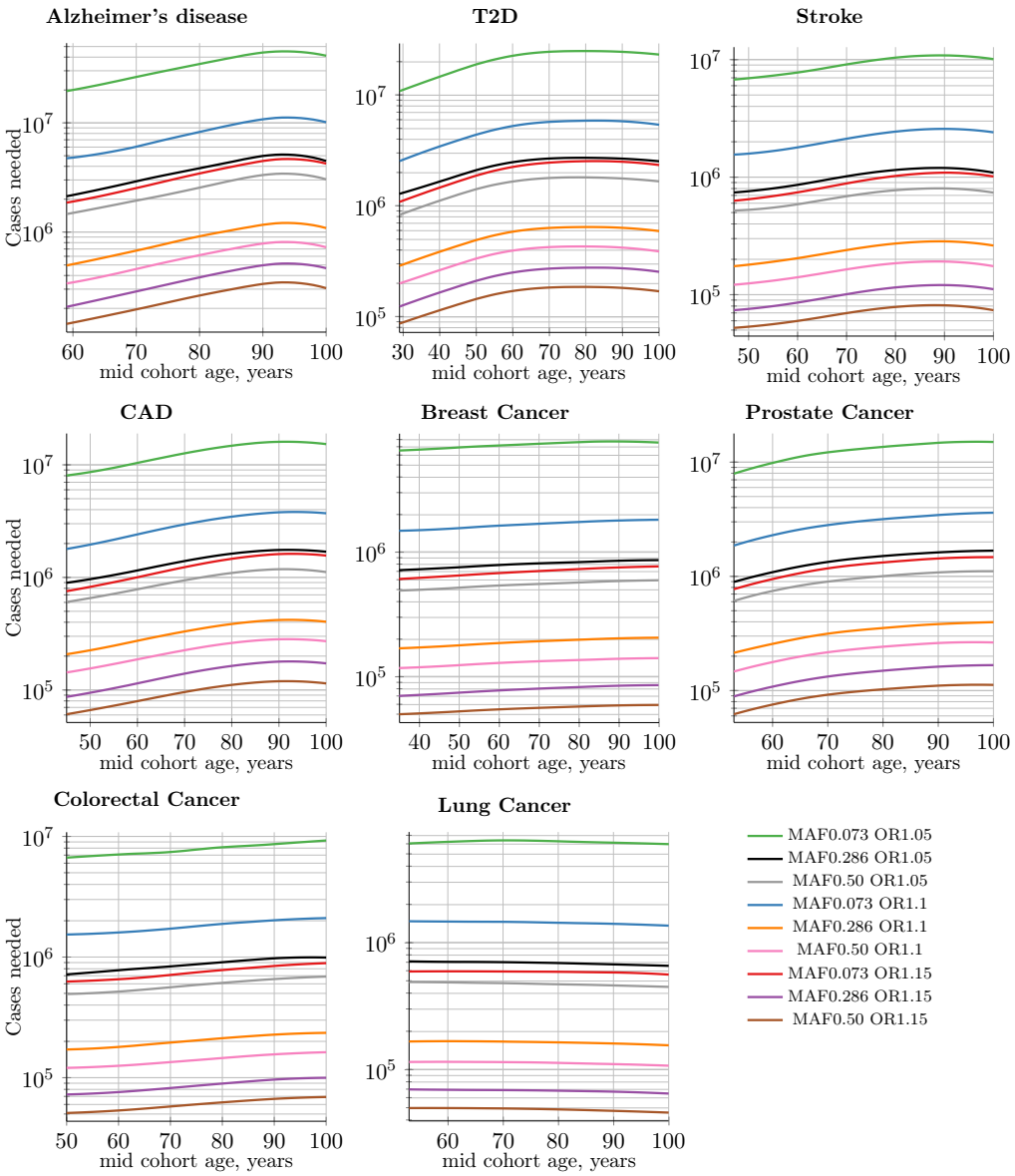

Fig 16. S9 Fig

Number of cases needed to achieve 0.8 discovery power; cohort simulation Common low-effect-size alleles (scenario A). The cohort curve due to the accumulative cases diagnosed at younger ages with an averaged control polygenic risk score and mortality starts at the same necessary cases number as the 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.

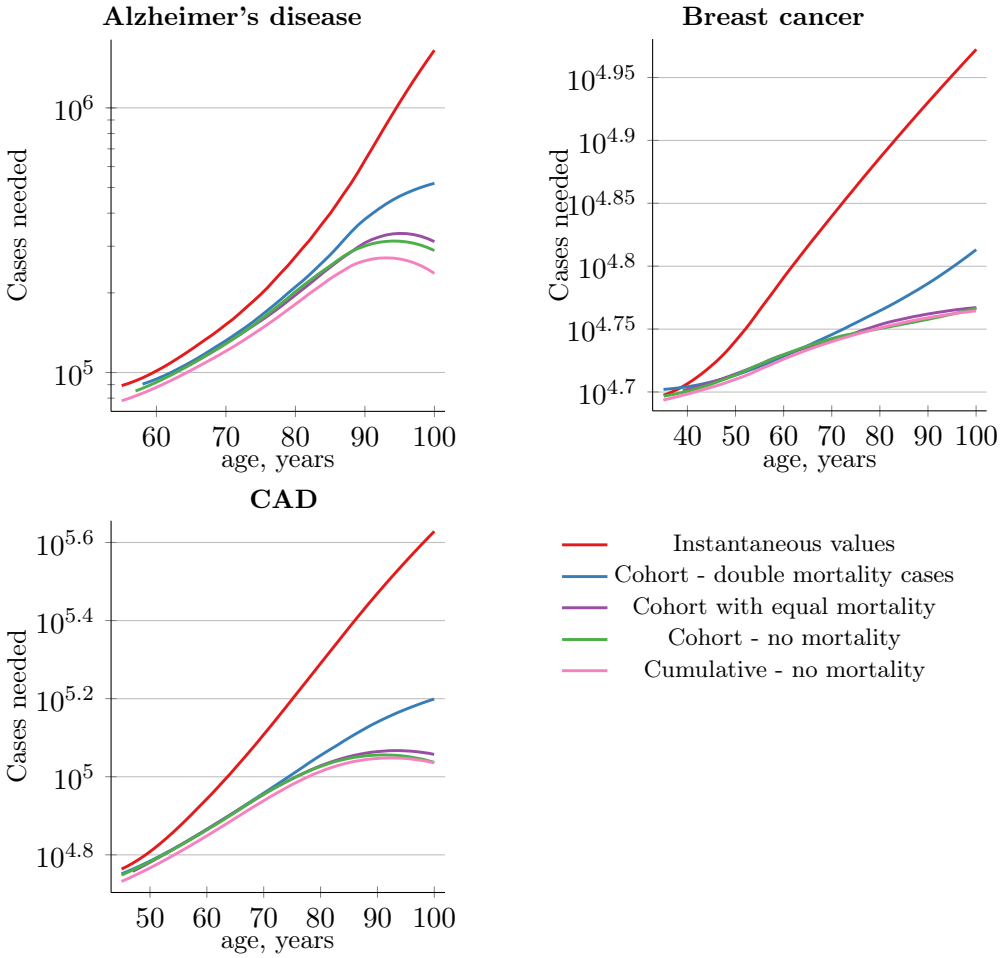

Fig 17. S10 Fig

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.

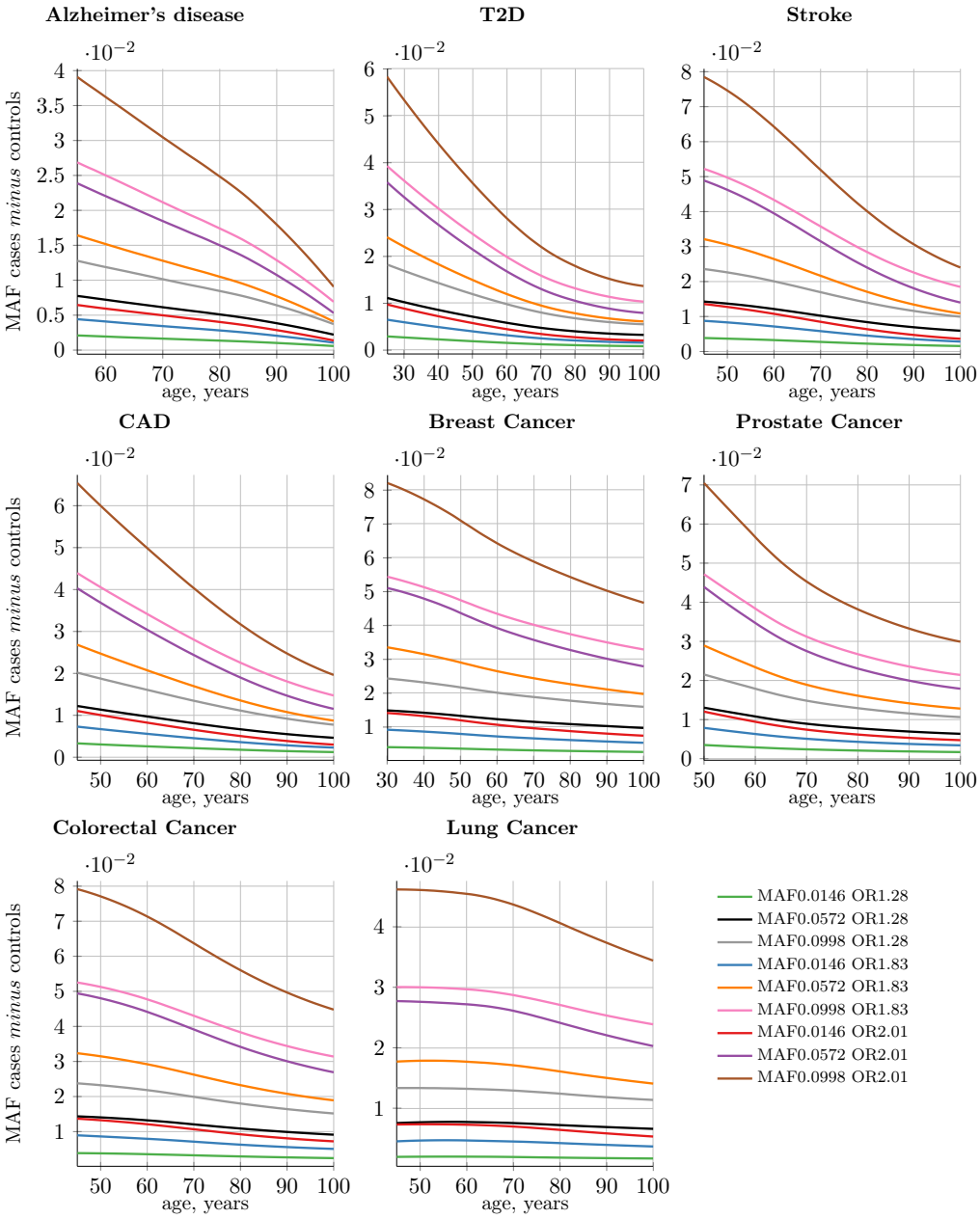

Fig 18. S11 Fig

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; see Eq (7). Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change.

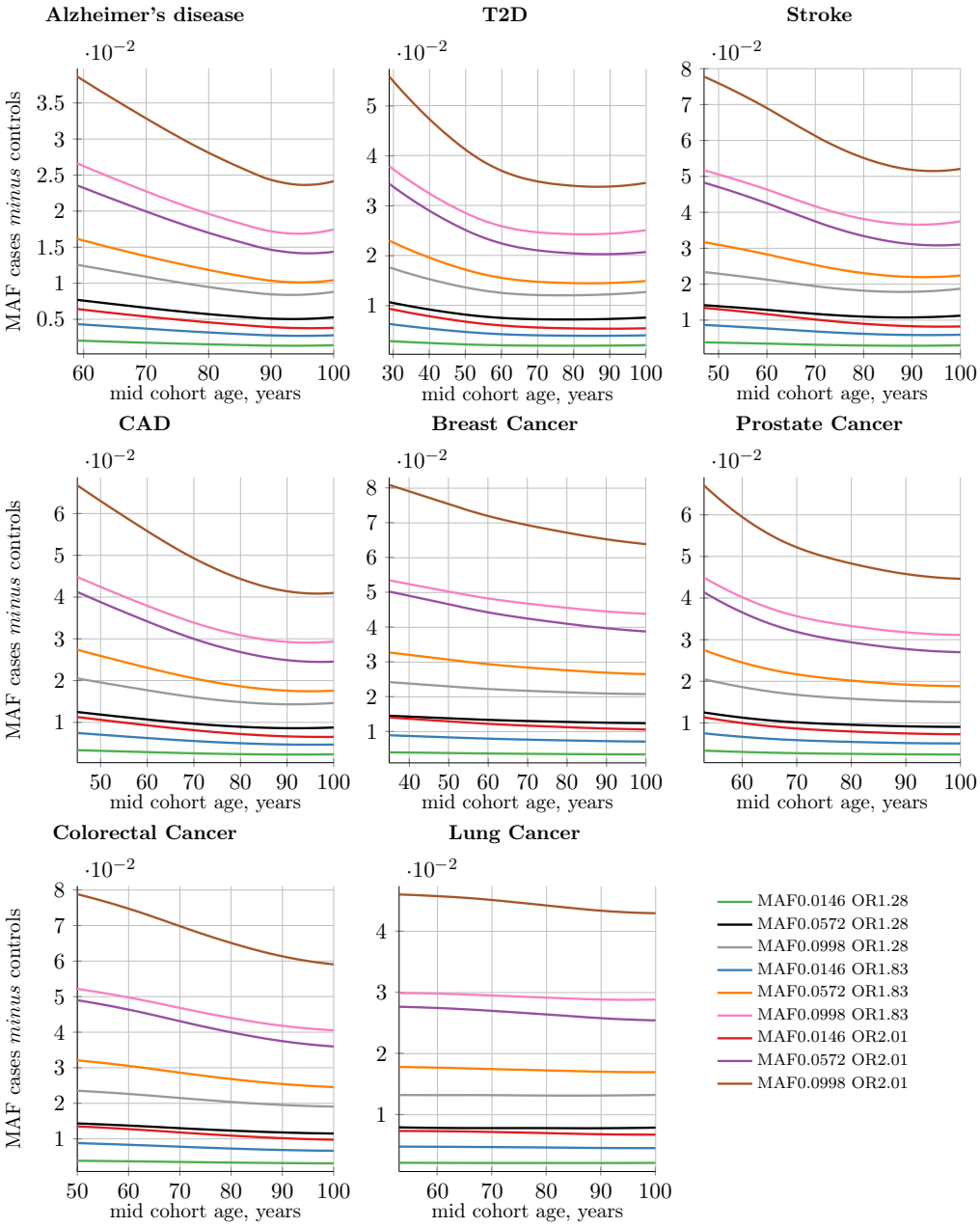

Fig 19. S12 Fig

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; see Eq (7). Rarer and lower-effect-size (OR) alleles are characterized by a lower MAF relative change.

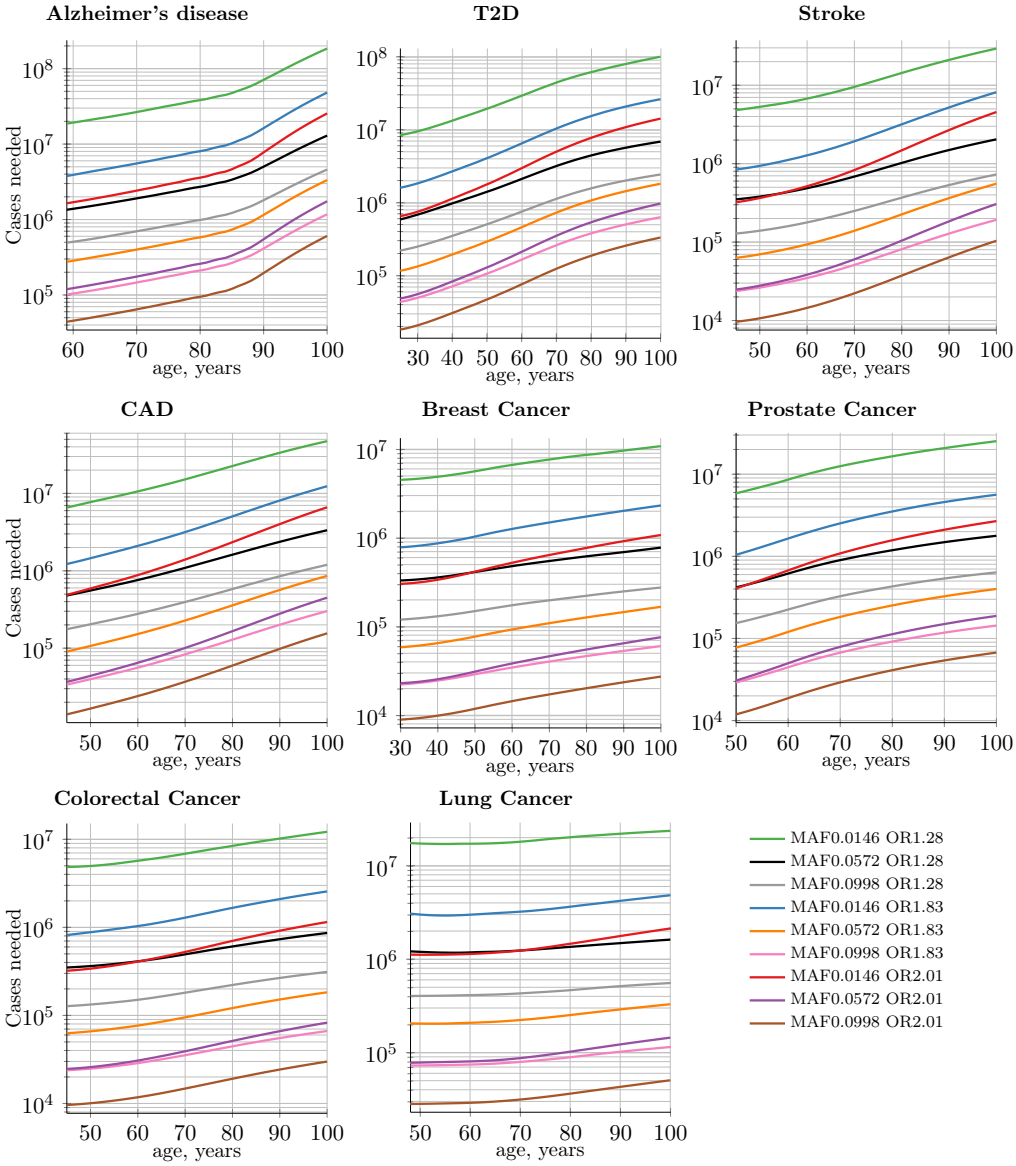

Fig 20. S13 Fig

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.

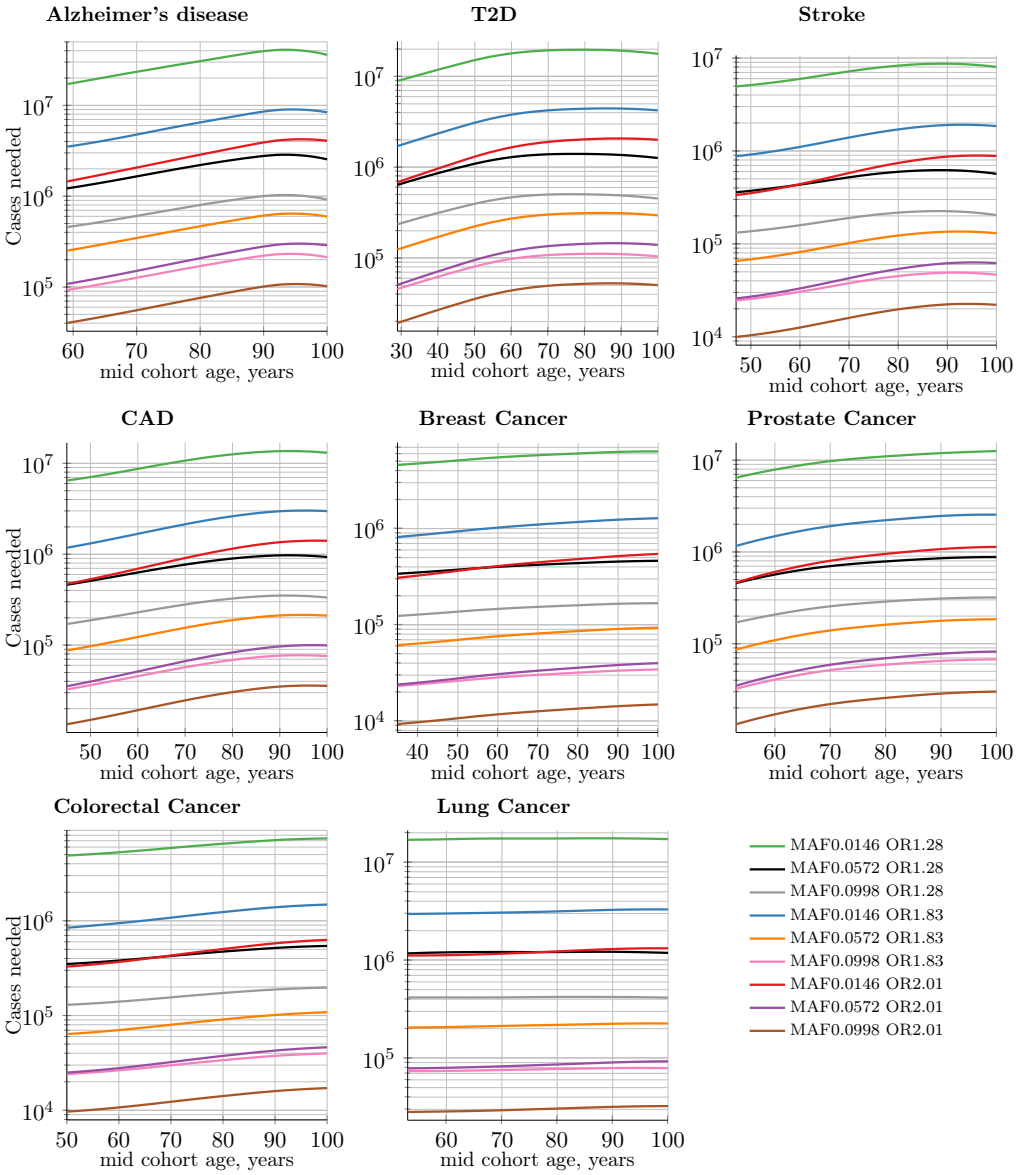

Fig 21. S14 Fig

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.

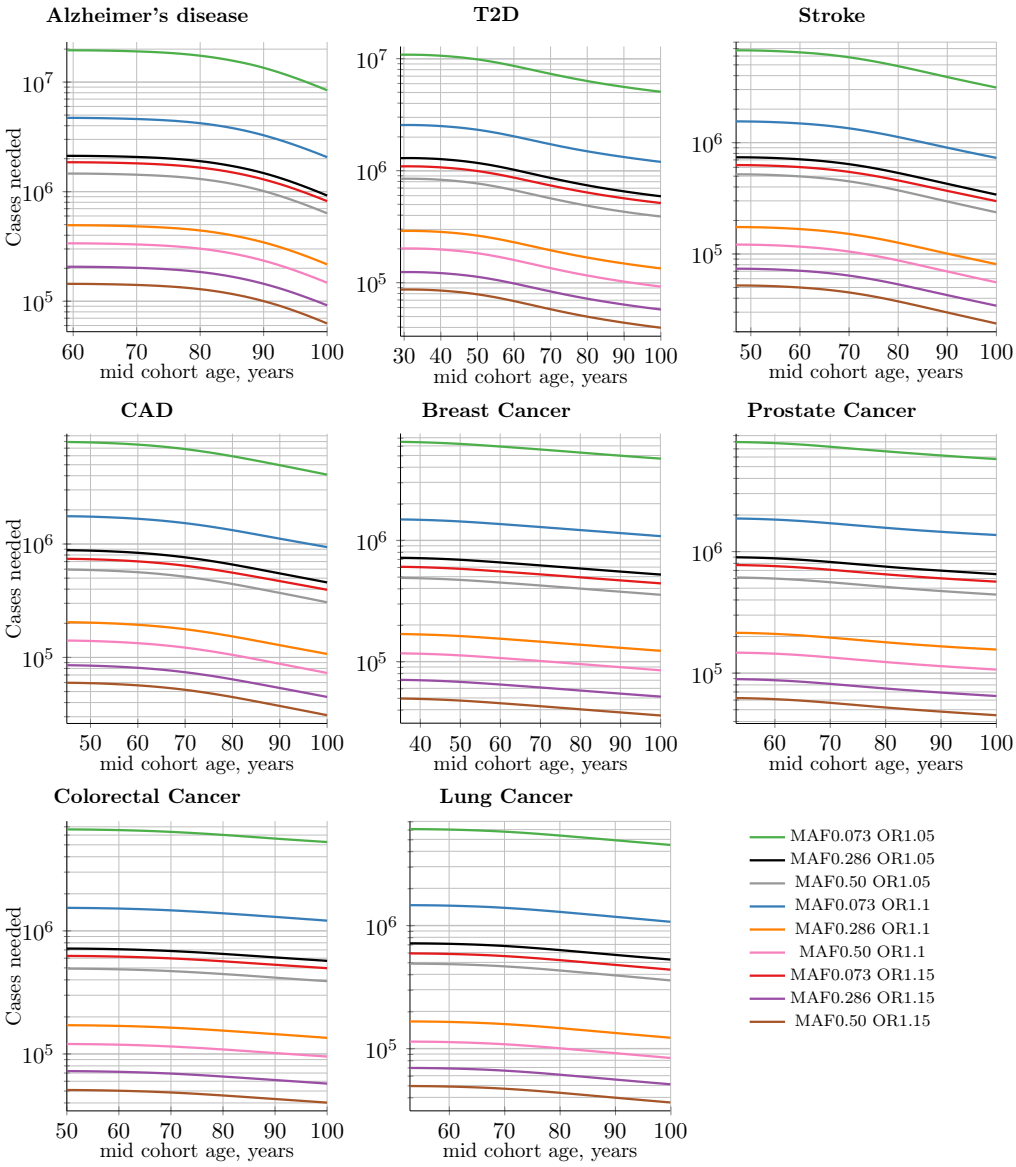

Fig 22. S15 Fig

Relative change in 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 using older age controls, particularly for LODs with the most prominent increase in the number of cases needed for older age cohorts. LODs and LOD cancers show distinct multiple groupings. Common low-effect-size alleles (scenario A). A sample of 9 out of 25 SNPs; MAF = minor (risk) allele frequency; OR = risk odds ratio.

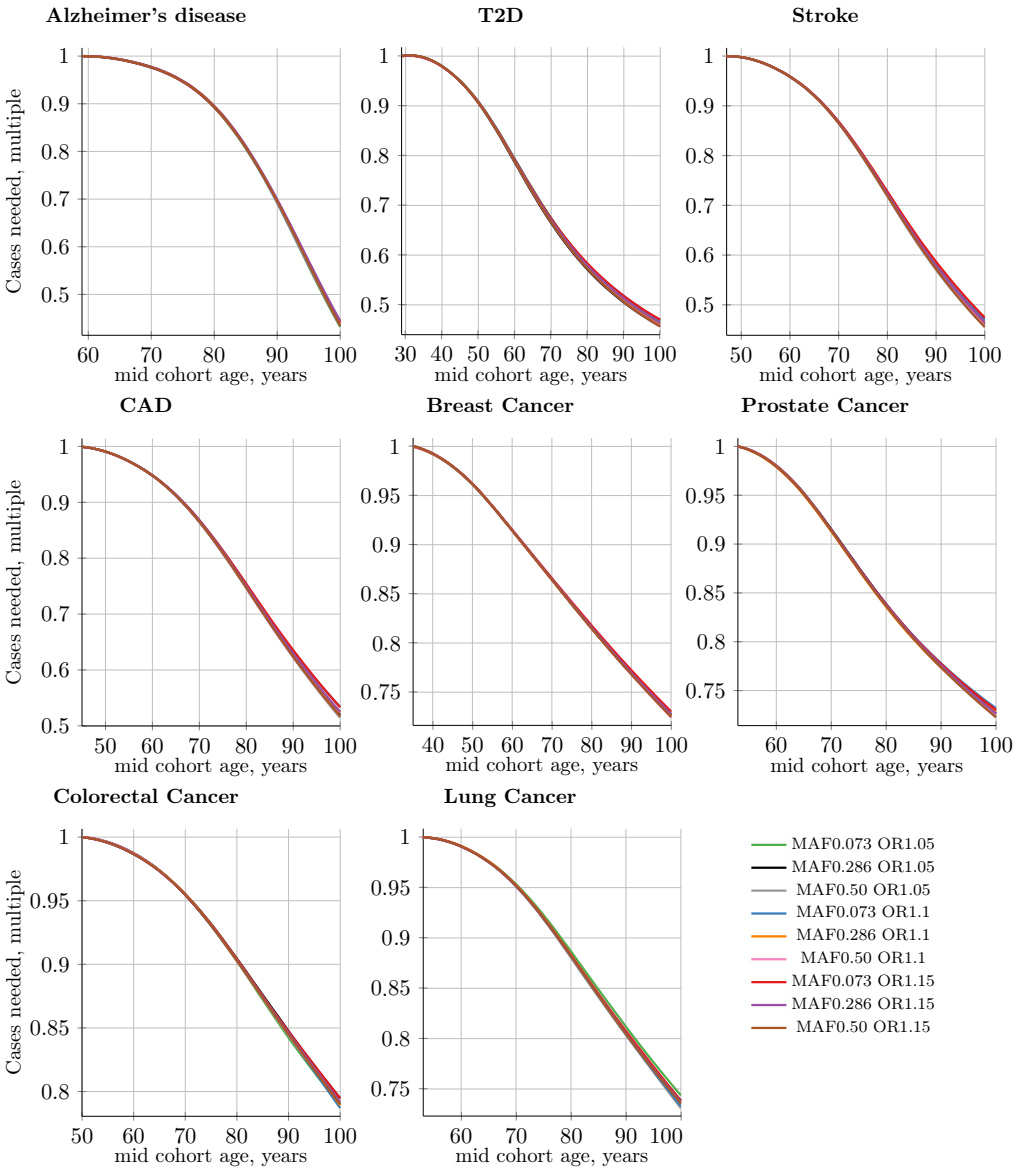

Fig 23. S16 Fig

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.