Quantifying the potential for future gene therapy to lower lifetime risk of polygenic late-onset diseases

Roman Teo Oliynyk[1,2,*]

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

ABSTRACT

Within the next few decades, gene therapy techniques and genetic knowledge will sufficiently advance to support prophylactic gene therapy to prevent late-onset diseases.
Our simulations quantified the correlation between the aging process, polygenic risk score and hazard ratio change with age based on clinical incidence rate and familial heritability and estimated the outcomes of hypothetical future prophylactic gene therapy for eight highly prevalent late-onset diseases.
We confirm that prophylactic gene therapy would be beneficial in both delaying the onset age and lowering the lifetime risk of late-onset diseases. With the expected increase in lifespan, under our model, type 2 diabetes, stroke, and coronary artery disease lifetime risk will regain the pre-treatment baseline within 10-15 years of longer life. AD proves most resistant to therapeutic delay and is most affected by aging, surpassing the baseline lifetime risk within five years of a longer lifespan. The characteristics of cancers may bring longer-lasting benefits.

Introduction

In the past two decades, the human genome has been successfully sequenced. Whole genome sequencing (WGS) and genome-wide association studies (GWASs) of human and other organisms' genomes have become an everyday occurrence[1]. Our knowledge of genetic variants, particularly the single nucleotide polymorphisms (SNPs) associated with susceptibility to diseases, has become deeper and more extensive.

The experimental gene therapy techniques aimed at diseases caused by a single defective gene or a single SNP—the so-called Mendelian conditions—are being refined. Mendelian conditions cause high mortality and morbidity, but each of these conditions affects only a tiny fraction of the population. However, a large number of these conditions are known. As of May 2018, the OMIM Gene Map Statistics compendium[2] lists over 6000 phenotypic genetic conditions caused by close to 4000 gene mutations. Their frequency is limited by natural selection, which efficiently removes large-effect detrimental mutations, resulting in a reasonably low genetic load in each person, equivalent to between five and eight highly detrimental recessive mutations[3,4]. If these were to combine and become homozygous, they could be harmful or even lethal[5–7]. Many of these conditions are well characterized. In some cases, individualized genetic diagnoses, where an SNP that needs to be edited can be specified precisely, are possible. One could say that these conditions are a well-defined target waiting for the appropriate medical techniques to address it. Over the last two decades, 287 monogenic disease clinical trials have been conducted worldwide[8]. When the medical technology becomes available, individuals who receive treatment will be effectively cured and will have no need for concern about the single, specific cause of their disease. For this reason, we did not analyze Mendelian conditions as part of this study.

Instead, we focused on polygenic or complex late-onset diseases (LODs), which pose a more nuanced problem. Such diseases do not develop until later in life. There are thousands of estimated gene variants or SNPs of a typically small effect that, in combination, constitute the polygenic LOD risk of each individual[9,10]; for a graphical demonstration of the model genetic architectures used in this research, see Supplementary Fig. 1. These diseases include the old-age diseases that eventually affect most individuals and are exemplified by cardiovascular disease—particularly coronary artery disease (CAD)—cerebral stroke, type 2 diabetes (T2D), senile dementia, Alzheimer's disease (AD), cancer, and osteoarthritis.

What makes LODs different from infectious diseases or from Mendelian genetic conditions is difficulty with the concept of *cure*. The diseases of aging are primarily a consequence of an organism's decline over time, leading to increased susceptibility to many LODs[11–13]. The combination of genetic liability, environmental factors, and the physiological decline of multiple organ systems leads to individual disease presentations[14]. Detrimental gene variants are exacerbating factors[15], compared to the average distribution of common gene variants that define human conditions as they apply to polygenic LODs. The time of onset for each individual is modulated by genotype and environment[16]. While some individuals will be diagnosed at a relatively

young age, others will not be diagnosed with a particular LOD during their lifetime[17]. A low polygenic risk score (PRS) for a specific LOD makes it unlikely that an individual will become ill from that LOD but does not affect the individual's likelihood of contracting a different LOD. We all have a level of susceptibility to the major polygenic LODs, and the difference between high-risk and low-risk individuals may lie in a slightly higher- or lower-than-average fraction of detrimental gene variants.

It is important to note that many diseases, including AD and diabetes mellitus, also possess highly detrimental mutations that singularly can cause high disease liability. Usually, these are associated with a special medical diagnosis. Diabetes mellitus caused by single defects in HLA-DQA1, HLA-DQB1, or HLA-DRB1 genes[18]—called type 1 diabetes to distinguish it from the more common T2D—develops in early childhood and affects a small percentage of overall diabetes mellitus sufferers. Similarly, with AD, so-called early-onset AD (EOAD) is caused primarily by APP, PSEN1, or PSEN2 gene mutations and affects a relatively small proportion of the population, starting in their thirties, with the majority of mutation carriers being affected by the age of 65[19]. Macular degeneration[20–22] is primarily caused by a small number of high-effect variants and manifests at a relatively old age. Unlike the previous two diseases, macular degeneration does not exhibit a significant incidence in the form of a polygenic late-onset disease version with similar symptoms. Even though EOAD and macular degeneration develop late in life, they belong to the Mendelian disease category and are not classified as polygenic LODs or analyzed here.

The best cure is prevention, and the time may be nearing when prophylactic gene therapy will be attempted for the prevention of complex polygenic diseases.

Computational techniques attempting to evaluate the effects of mutations or gene variants are being developed, although their accuracy needs to improve dramatically to be applicable to personalized human genetic evaluation or treatment[23]. Similarly, while there are extensive libraries of human SNPs, including dbSNP, HapMap, SNPedia and aggregating sites[24], the information is far from actionable as far as modifying multiple personal SNPs is concerned. Locating or being able to computationally estimate a complete set of the low effect causal SNPs is knowledge that may take decades to gain. For the purposes of this research, we assumed that we know the "true" causal gene variants (see Chatterjee at al.[25]) distinct from GWAS SNPs that are, in the vast majority of representative SNPs, located relatively near in the genome and indistinguishable in GWAS analysis from the real causal variants due to a strong linkage disequilibrium.

Gene editing technologies also may be a few decades away from the time when they can be used as routinely and with the same low risk as applying an influenza vaccination to modify a large number of gene variants distributed over the human genome. The latest gene editing technique, CRISPR-Cas9[26], has supplemented and mostly replaced older technologies such as zinc-finger nucleases[27] and TALEN[28], though, for some applications, these older techniques continue to be more appropriate. While its selectivity and on-target precision are improving, CRISPR is still most effective in gene knockdown operations. For modification and repair, only a small fraction of CRISPR operations, using homologous repair with a template or a sister chromatid sequence, succeed. Successes such as that announced in the study "Enhanced proofreading governs CRISPR-Cas9 targeting accuracy"[29] are often followed by reversals, such as those detailed in "Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements"[30]. Still, CRISPR is only five years old and is a rapidly developing technology with great promise.

A potential future technology, currently in its early exploratory stages, could be synthetic genomics, currently under study in the Harvard Church laboratory[31,32]. These techniques could also help in the precise mapping of gene variant effects on disease phenotypes.

If none of the approaches mentioned above ultimately succeed in becoming reliable enough for the purposes of gene therapy, it is almost certain that a new, more suitable technique will be invented.

Notwithstanding the periodic media furors about designer babies, many questions of scientific knowledge, technical expertise, and ethics will need to be settled before this level of prophylactic gene therapy will become possible. For the purposes of this hypothetical treatise, we assumed that it is possible to precisely identify individual gene variants and their detrimental or beneficial effects and then use gene editing to modify a large number of detrimental variants. We applied this hypothetical gene therapy model to estimate what would happen to LOD progression as the population ages. We did not consider additions of artificially designed genomic sequences, but rather only corrections made to typically low-effect heterozygous in-population gene variants—a correction of a detrimental variant to a naturally occurring neutral state. For simplicity, our model uses SNPs, though the same would apply, albeit with a higher degree of complexity, to gene therapy with other gene variant types.

Many changes in lifestyle and medical care, including the prevention and treatment of infectious diseases, have extended longevity over the last century, and this trend is projected to continue; preventive gene therapies may also be one of the future factors.

For the sake of brevity, in the remainder of this study we use the terms "gene therapy," "therapy," and "treatment" interchangeably, with the understanding that we are talking about *hypothetical future preventative or prophylactic gene therapy*. Similarly, it appears most practical for a person to be born with all cells already treated, because all developmental stages are affected by the genome, which implies the importance of the germline and likely heritability of therapy. Yet, we do not know what technological possibilities and ethical considerations the future will bring; therefore, we assumed that, at birth (or age

zero), the person's genome appears with the modifications required by this study.

Modeling takes advantage of the fact that a complex process can often be probabilistically analyzed with a judiciously designed simulation, which is particularly applicable to population genetics analysis. The first models originated a century ago in the form of analytical estimates[33,34] and later, with the advent of computers, as simulations. The phenomena behind aging and LOD progression are complex. For example, individuals with a high PRS may become ill late in life or not at all during their lifespan, and vice versa for individuals with low genetic predispositions[17], as can be seen in Fig. 1**b**. Statistically, either scenario will happen with low frequency, and clinically, physicians are likely to find specific or probable causes, particularly in cases with a positive diagnosis. The real causes may be environmental, infectious, epigenetic, or pleiotropic[35], stochastically and methodologically explained by the concept of penetrance[36], etcetera. This is one of the reasons for prediction difficulties using PRS on an individual basis, where the concomitant conditions and present physiological risk factors are much better predictors. These clinical observations work on the basis of accumulated knowledge of physiology, disease patterns, and co-morbidities that allows for the elimination of a large degree of uncertainty inherent in purely genomic predictions. A good example is the Framingham General Cardiovascular Risk Score[37], which includes as major risk factors T2D, current tobacco use, age, gender, cholesterol, systolic blood pressure, and hypertension. A clinical prediction in this setting is a part of diagnosis. With simulations operating on a large population, overall trends can be observed, generalized, and projected beyond the initial input parameters. In this study, we also used a simulation approach to a complex problem, and when the model representation was found that described the baseline conditions well, we made moderate changes to the simulation input parameters and evaluated the resulting projections. We modeled the age extension by a duration long enough to make projections but not so long as to render the model results questionable.

We reviewed and analyzed four of the most common diseases—AD, T2D, cerebral stroke, and CAD—and four of the most prevalent cancers: breast, prostate, colorectal, and lung.

We aimed to quantify the effects of hypothetical future gene therapy specifically in the case of LODs in an aging population. We have previously shown[38] that the average PRS of unaffected individuals diminishes with age while the risk of the disease increases with age, as reflected in rising incidence rates. This indicates that environmental effects and aging make an increasing contribution to LOD incidence with age progression, particularly in relation to AD, T2D, and cardiovascular diseases. These are diseases that can arise in most individuals because they represent the natural modes of failure for the human body due to the particular details of its complex design[14]. They are failures of operation that may start at the molecular, cellular, or organ level and affect vital subsystems such as metabolism, circulation, and cognition. The failure does not happen immediately prior to the age of diagnosis. For example, AD deterioration begins decades before symptoms first become noticeable[39]. A similar situation holds for cardiovascular disease[40,41] and cancers[42].

While all the factors involved in LOD causality and age of onset are complex and vary between diseases and individuals, our premise is that, statistically, a PRS value will correspond to a risk of succumbing to an LOD at each year of age, and that—given each LOD's clinical incidence rate statistics and heritability from familial studies—it is possible to find, for each year of age, a correlation coefficient mapping PRS to this hazard ratio. With a model in which individual risk remains proportionate to PRS[25], we determined what we call the *aging coefficient* by correlating PRS to the age-related hazard ratio. This coefficient allows us to exactly reproduce the LOD's incidence rate for all analyzed LODs. Changing the individual PRS distribution and applying these aging coefficients allow us to emulate the prophylactic gene therapy aimed at LOD prevention and to quantify the populational change in a given LOD's incidence and lifetime risk.

We considered a number of scenarios of prophylactic gene therapy. In the most extreme scenario, we could make thousands of edits for AD, and hundreds for most of the rest of the LODs, reversing all variants constituting individual PRSs to a neutral state; this would show a resulting zero disease risk and zero incidence. We are not confident the model would be realistic with so drastic an intervention, and many less extreme scenarios are possible.

We decided on two natural scenarios, further explained in the Methods section:

Scenario 1 is where all individuals at risk undergo a preventive treatment, making their risk equal to the baseline average of the population. This way, the number of edits, while still large, is of limited magnitude. This approach also corresponds to the established medical practice of primarily treating at-risk individuals.

Scenario 2, in which every individual PRS is decreased fourfold, generally produces a more extensive effect than scenario 1. Scenario 2 could also be considered a consequence of a situation in which only part of the at-risk population is treated but in which, over a number of generations, the average PRS drops due to the average risk allele frequency in the population gradually declining.

Applying this novel approach, we quantified the correlation between the environment and aging process, the polygenic score and the hazard ratio of selected LODs at all ages, based on clinical incidence rates and familial heritability, and applied the resulting aging coefficient to simulated populations in the two scenarios of hypothetical future gene therapy. The results follow.

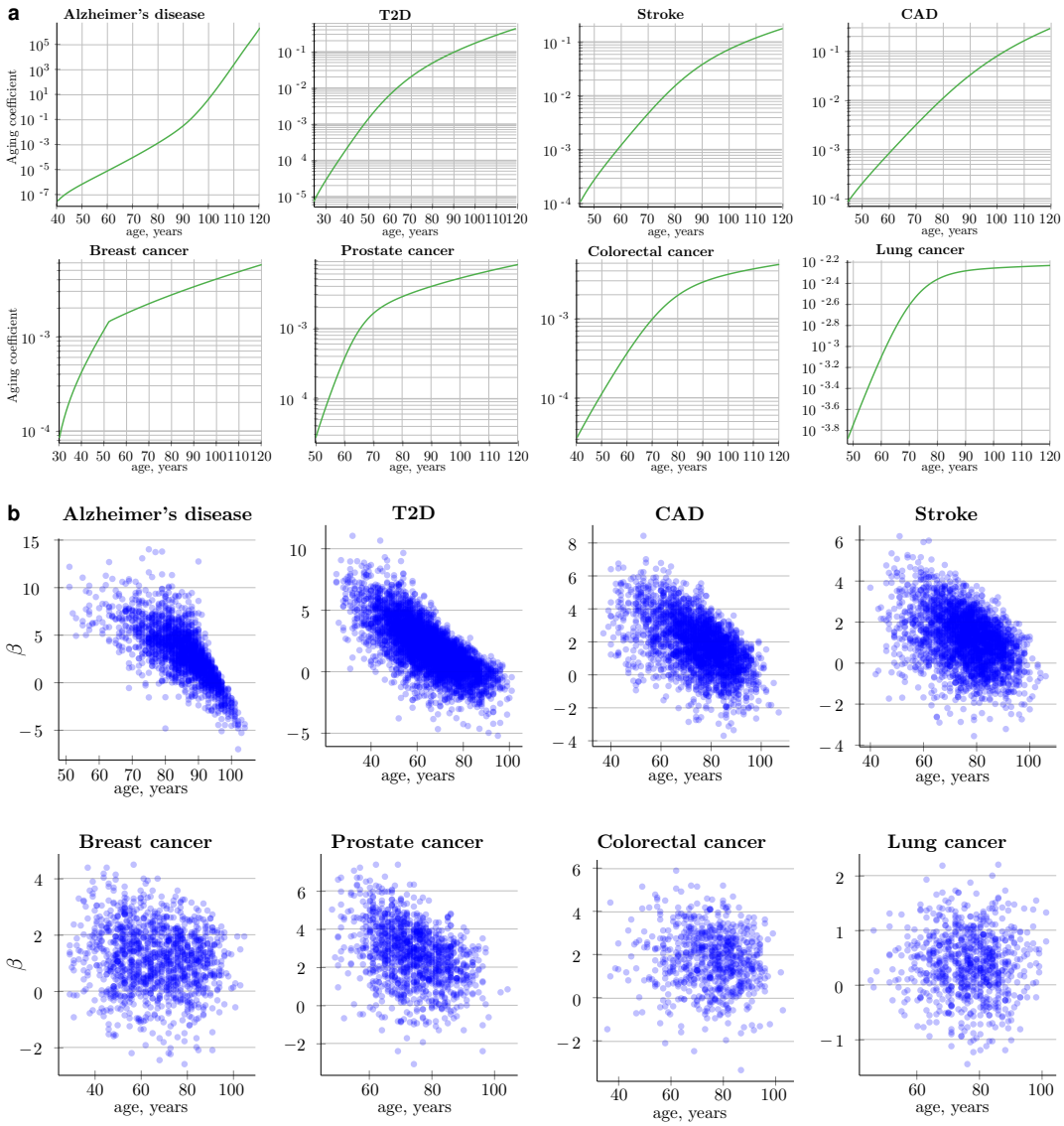


Fig. 1. Aging coefficients applied to reproduce the LOD incidence distribution by age based on individual PRS
a Aging coefficients log_{10} scale. This is the parameter that reflects increases in LOD liability with age based on the clinical incidence rate and model genetic architecture PRS. **b** Polygenic scores of individuals diagnosed with an LOD as a function of age. Scatter plots show the distributions of polygenic scores for cases diagnosed as age progresses. PRS $\beta = log(OddsRatio)$.

Results

We used the clinical incidence rate pattern to map the hazard ratio from the PRS of individuals diagnosed with an LOD at each year of age for eight LODs: AD, T2D, cerebral stroke, CAD, and four late-onset cancers: breast, prostate, colorectal, and lung.

The core of the simulation is the iterative discovery and application of the aging coefficient (see Methods Eq. 1), mapping individual PRSs to the hazard ratio specific to each LOD on a yearly basis, depicted in Fig. 1**a**. The aging coefficient incorporates combined aging and environmental effects, and the rising pattern indicates the increasing magnitude of these effects with age. It is interesting to note how the magnitude of the aging coefficient changes with age for the diseases we analyzed. AD shows the largest magnitude of change, and is characterized by a steep exponential incidence rate increase continuing to very old age. T2D, CAD, and stroke show a comparatively smaller progression, while cancers show a relatively small magnitude increase, with lung cancer showing the least change. Fig. 1**b** shows the PRS distribution for diagnosed individuals by age when the simulation is re-run using the discovered aging coefficient without gene editing (a baseline validation). It matches the results of[38], where a simulation using a simpler, direct algorithm, which was more appropriate for that research, produced a similar outcome. A more precise indication of the accuracy of the discovered aging coefficient is shown in Fig. 4, where the simulation

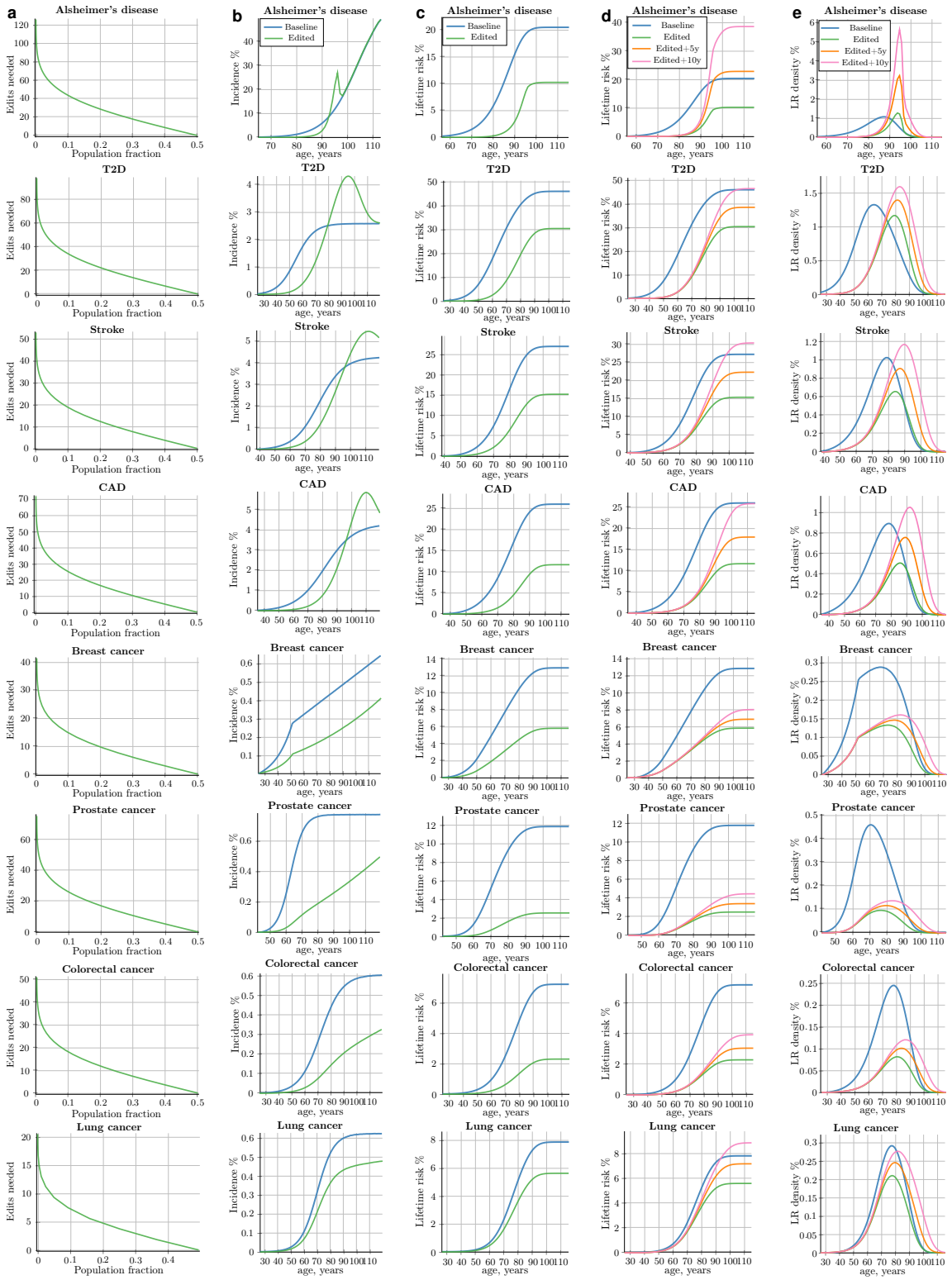


Fig. 2. Gene therapy Scenario 1: Treating all above-average-risk individuals.
a Per LOD number of edits for risk population. **b** Incidence rate pattern, baseline and after gene therapy. **c** Lifetime risk decrease, baseline and after gene therapy. **d** Lifetime risk, baseline and after gene therapy with lifespan increase of 0, 5, and 10 years. **e** Lifetime risk density Eq. 13, baseline and after gene therapy with lifespan increase of 0, 5, and 10 years.
See Supplementary Fig. 5, Supplementary Fig. 6, and Supplementary Fig. 7 for enlarged views.

aging reproduces the population incidence rate Eq. 1.

Knowing the hazard ratio for any given PRS value at any given age is exactly what the simulation needs in order to model gene therapy and the effects of lifespan increase. We applied the aging coefficient in two gene therapy scenarios, described in the Methods section:

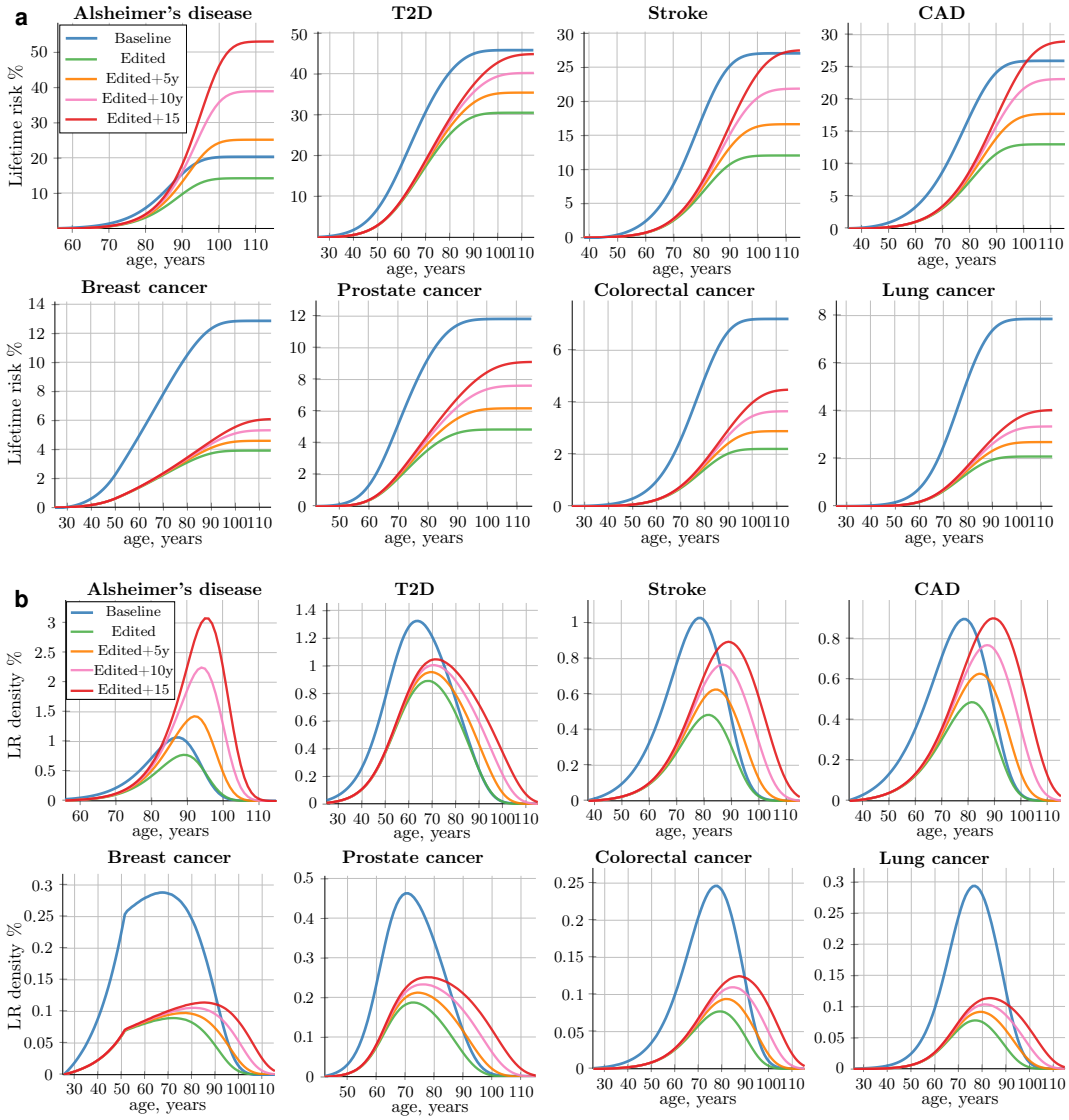


Fig. 3. Scenario 2: All individuals in a population have corrective gene therapy for, on average, 15 SNPs (corresponding to an OR of 0.25).
a Lifetime risk. **b** Lifetime risk density Eq. 13. Result of gene therapy and changes with lifespan increases of 5, 10, and 15 years. Supplementary Fig. 4 plots incident rates for baseline and treated populations.

Scenario 1: Prophylactically treating only above-average-risk individuals

The simulated prophylactic gene therapy corrected the PRS value of all individuals with above-average risk to the average PRS. Fig. 2**a** shows the number of gene variants that would need correction under the common low-effect genotype model. AD, the analyzed LOD with the highest heritability, would require over 120 edits for the highest risk individuals, while lung cancer would require over 20. The number is lower for the majority of the treated population: below 22 edits for half of the treated population for AD and fewer than three edits for lung cancer. As expected, the incidence rate after gene therapy drops significantly at younger ages for all LODs, which results in a greatly decreased lifetime risk for all analyzed LODs; see summary Table 1.

For the four non-cancer LODs, there is a spike in the incidence rate at a very old age, which exceeds the baseline incidence rate; see Fig. 2**b**. The AD incidence rate displays a steep spike around the age of 95 and merges with the baseline rate pattern

soon after. This can be explained by the continuing exponential incidence rate pattern[43]. T2D, CAD, and stroke also display significant later spikes above the baseline, which asymptotically lead toward the baseline incidence rate line from above with age progression. This pattern can be explained by the hazard ratio, represented by the aging coefficient $A(t)$, increasing to a point when a large fraction of the population with similar PRSs resulting from the therapy are diagnosed with these LODs. The maximum incidence spike occurs at the age of 95 for T2D and at about 110 for stroke and CAD. These are the ages at which, for AD and T2D, only a small fraction of the population is still alive, and, in the case of stroke and CAD, practically the whole population is deceased. As a result, with mortality remaining constant, the lifetime risk—the chance of a person being diagnosed but not necessarily dying from an LOD—drops significantly; see Table 1.

All cancer incidence rates remained consistently below the baseline through the maximum analysis age of 120 years. It is notable that the three most prevalent cancers show the highest proportionate improvement, which can be explained by a combination of their relatively high heritability and relatively low incidence rates.

Lung cancer showed the least improvement. This is due to the disease having the lowest reported heritability, which necessitates the fewest required edits for the treated population under this scenario, as seen in Fig. 2**a**.

In light of the hypothetical future advances in medicine, including prophylactic gene therapy procedures such as these, the average life expectancy is expected to continue to increase. We modeled each disease independently without accounting for co-morbidities. This simplification works because it allows us to track the diagnosis of the disease and consider mortality as being *from all causes*. In Fig. 2**d**, we consider an increase in the average lifespan of 5 and 10 years, which is sufficient for the purposes of this scenario. We find that AD lifetime risk regains the baseline level before 5 years and almost doubles in relation to the baseline with a 10-year life extension; see Table 1. T2D, stroke, CAD, and lung cancer all either match or slightly exceed the baseline with a lifespan increase of 10 years, with the incidence onset curve noticeably delayed. The most prevalent cancers—breast, prostate and colorectal—also show a significant improvement in population lifetime risk, which remains significantly lower than the baseline with a 10-year increase in lifespan.

Fig. 2**e** provides an additional view from the perspective of lifetime risk density, described as the incidence rate relative to the initial population count adjusted for mortality; see Eq. 13. The plot shows how the incidence peak will shift toward older ages with increased longevity.

Scenario 2: Effecting a fourfold decrease in all individuals' polygenic risk

A 0.25 OR multiplier was applied to all individual PRSs, and the simulation run from age zero to 120 years of age, Fig. 3**a**. This corresponds to between 10 and 15 edits, as explained in the Methods section, for all modeled LODs. In this scenario, the incidence rate remains below the baseline throughout the lifespan for all LODs analyzed. Lifetime risk is also lower for all LODs and shows a delay of approximately a decade in the incidence rate curve for T2D, stroke, and CAD, and an even more significant improvement for cancers. AD again benefits the least.

As in the previous scenario, the AD lifetime risk exceeds the baseline within five years of the increased lifespan. It took 15 years of longer life for T2D, stroke, and CAD to get close to or slightly exceed their respective baseline lifetime risks. All cancers stayed far below their baseline lifetime risks, even with a 15-year extension of lifespan. The increase in the fraction of older individual diagnoses leading to this lifetime risk increase alongside longer life in Scenario 2 is presented in Fig. 3**b** and Table 1.

Baseline scenario of increased lifespan (without gene therapy)

In this scenario, the lifespan increases by 5 and 10 years, with each LOD risk remaining unchanged. The summary is presented in Table 1 as a baseline for comparison with the two gene therapy scenarios. A graphical representation of the incidence rate and lifetime risk for this baseline case can be seen in Supplementary Fig. 2 and Supplementary Fig. 3.

Discussion

The hypothetical future gene therapy aimed at LOD prevention can be characterized by the resulting improvement in PRS of the individuals in the population, with the lowered polygenic risk of the prophylactically treated corresponding to a lower hazard of becoming ill at each particular age. To be able to quantify the population health effects, we needed to know the hazard ratio associated with a particular PRS value for each year of age in LOD progression. This is not a trivial calculation, due to the complex pattern of the diagnosed individuals' PRSs decreasing at a different rate for each LOD as age advances. We designed a population simulation that naturally accounts for the PRS change of individuals as they age, as well as other environmental effects that accumulate with age. In addition, we quantified how these diseases' lifetime risks will behave with increasing longevity, assuming an unchanged aging process for each particular LOD.

The modeled increased lifespan in the baseline scenario, without prophylactic gene therapy, reinforces the old wisdom that aging itself is the predominant risk factor for most late-onset diseases and conditions; see Table 1. As is well known, with a longer lifespan, the risk of LODs increases. This is most prominent with AD. It is a known fact that, in Japan, the lifetime

Table 1. **Lifetime risk: Baseline, Scenario 1, and Scenario 2 of prophylactic gene therapy**

	Highly prevalent LODs				Cancers			
	AD	**T2D**	**Stroke**	**CAD**	**Breast**	**Prostate**	**Colorectal**	**Lung**
Literature and clinical data:								
Heritability	0.795	0.69	0.55	0.41	0.57	0.40	0.31	0.10
Literature max yearly incidence rate	$> 20\%$	2.5%	4.4%	3.6%	$<0.5\%$	$<0.8\%$	$<0.6\%$	$<0.6\%$
Lifetime risk, baseline + longer life:								
+5 years life expectancy	160%	112%	128%	127%	115%	123%	127%	128%
+10 years life expectancy	228%	123%	156%	155%	130%	147%	156%	156%
Lifetime risk, Scenario 1 vs baseline:								
Edited, unchanged life expectancy	**50%**	**66%**	**57%**	**45%**	**45%**	**21%**	**32%**	**72%**
Edited, +5 years life expectancy	112%	84%	82%	69%	53%	29%	43%	92%
Edited, +10 years life expectancy	189%	101%	112%	100%	62%	38%	55%	114%
Lifetime risk, Scenario 2 vs baseline:								
Edited, unchanged life expectancy	**70%**	**67%**	**44%**	**50%**	**30%**	**41%**	**31%**	**27%**
Edited, +5 years life expectancy	124%	77%	61%	69%	36%	52%	40%	34%
Edited, +10 years life expectancy	191%	88%	81%	89%	41%	65%	51%	43%
Edited, +15 years life expectancy	260%	98%	101%	112%	47%	77%	62%	52%

LODs' heritability and clinical incidence summarized from Methods literature. The baseline is considered 100% in lifetime risk comparisons; 160% after a 5-year lifespan expansion for AD means an increase in lifetime risk (LR) of 1.6 times, and 50% after gene therapy means half the LR, compared to the baseline value.

risk of AD is higher than in many other countries[44], and the average lifespan in Japan is also higher. Our results confirm once more that, if mortality from all causes is lower, leading to a longer lifespan, AD is the LOD that exhibits the most rapid rise in late-age prevalence.

In both gene therapy scenarios, the incidence rate at younger onset ages diminishes significantly. This also corresponds to a much-improved lifetime risk for each LOD; see Fig. 2 and Fig. 3.

Modeling an increased lifespan due to all causes in gene therapy Scenario 1, we find that AD lifetime risk regains the level of baseline incidence within five years and almost doubles above the baseline with a 10-year life extension; see Table 1. T2D, stroke, CAD, and lung cancer match or slightly exceed the baseline with a 10-year life extension. It is important to note that, while the incidence rate and lifetime risk may increase with longer lives for the abovementioned diseases, from the population perspective, the onset age is delayed by years and even decades for the formerly highest-risk individuals.

The most prevalent cancers—breast, prostate and colorectal—showed a very significant decrease in lifetime risk, even when lifespan increased. Lung cancer requires the fewest edits in order for the at-risk half of the population to achieve the baseline average PRS in Scenario 1, but this is because low heritability implies low genetic variance, which is reflected in the lower number of gene variants possessed by the higher-than-average risk population. This naturally results in a lower initial preventative effect compared to other reviewed cancers, and lung cancer regains the baseline lifetime risk with a 10-year increase in average lifespan in this scenario.

Many LODs follow the exponential incidence rate increase pattern in initial onset decades[38], yet AD incidence continues to rise exponentially past the age of 95[43,45]. Unless a dramatic discovery is made for AD prevention, statistically, this LOD is one of the most persistent modes of human organism failure at old age, to which the majority of us are at risk of succumbing if we live long enough. Similarly, the Framingham General Cardiovascular Risk Score includes age as one of the major risk factors for stroke and CAD[37], and our modeling concurs.

T2D is characterized by an early initial onset age, and while the prevalence of T2D is high, by the age of 80, the incidence rate is essentially constant at 2.5% in the clinical dataset we used[46]. As a result, the incremental lifetime risk is relatively lower in 5- and 10-year lifespan extensions, starting from an already high level. In the case of prophylactic gene therapy, the population health effect also will be lasting. Even though we analyzed each LOD independently, T2D is one of the diseases that causes the most co-morbidities, accelerating incidences of cardiovascular and other diseases, sometimes by decades[46]. For this reason, the preventative treatment of T2D could mean health improvement or presentation delay for a range of LODs, independently of or in addition to treating their specific gene variants.

By precipitating a fourfold decrease in the LOD odds ratio for each individual, gene therapy Scenario 2 could be viewed as a model of a more constant effect gene therapy. Unsurprisingly, lower heritability LODs, such as lung cancer, gain a more equitable benefit from this scenario when compared with the Scenario 1 treatment. The most noticeable result for LODs with

lower cumulative incidence or prevalence is the fact that gene therapy has the longest-lasting effect on these diseases. We analyzed some of the most prevalent LODs, and there are a large number of LODs with a much lower prevalence that would similarly benefit. Based on known heritability values for many diseases, it may be within the realm of possibility to find a sufficient number of variants to achieve fourfold odds ratio reductions.

A question is often asked about the lifetime risk values for a particular LOD, and there may be a correct value at a particular moment in time for a specific locale. This number is subject to change with population lifestyle and environment changes, medical procedures, and particularly increases in the average lifespan. A decade later, the value may change for a hypothetical locale, and the value for one locale may differ from that of others around the world, both now and in the future. Our modeling shows how lower disease risk and increased lifespan act as opposing forces in this equilibrium. Treating the at-risk population would improve health, and the lifetime risk would regain ground within 10–15 years of increased lifespan for the most prevalent LODs, with the relative incidence rate becoming statistically higher for very old individuals.

Conclusions

Our computer simulation successfully correlated the aging processes with the individual's PRS at each age, finding what we call the aging coefficient for eight prevalent LODs. This aging coefficient was then applied to two model scenarios of gene therapy. Through the same simulation process, we quantified the outcomes for the population incidence rate and lifetime risk, also modeling the effect of increased longevity on LOD progression and lifetime risk.

We find that, were intensive gene therapy to become possible, it would dramatically delay the average onset of the LODs analyzed here and improve population lifetime risk.

The simulations highlighted the patterns that distinguish the analyzed LODs. It would be more difficult to limit the prevalence of AD if lifespan also increased. Perhaps a pharmaceutical intervention targeting a causal metabolic pathway or immune or inflammatory response may be more effective for AD; however, the past announcements giving false hope are too numerous to cite.

T2D, stroke, and CAD lifetime risk will regain their pre-treatment baselines within 10 to 15 years of longer life, which is equivalent to delaying the average onset age of these LODs by as many years.

Based on heritability and incidence rate combinations, our model predicts that prophylactic gene therapy has the potential to bring significant and longer-lasting benefits for cancer prevention, even with a similar or smaller number of edited gene variants than for the other diseases we modeled.

Methods

Simulation flow

The simulation is performed in two stages: discovery or mapping of the aging coefficient and analysis. Discovery iteratively maps an LOD aging coefficient based on the modeled genetic architecture and incidence rate statistics. The analysis simulation uses the discovered aging coefficient to quantify the health outcomes due to gene therapy and lifespan extension.

Aging coefficient discovery

1. The simulation initialization steps are performed, including allocating population objects, assigning individual PRSs based on the modeled genetic architecture, and building an incidence rate functional approximation.

2. The simulation works as an iterative procedure, where the values of $A(t)$ are matched for all ages, starting at the age at which the incidence rate first becomes noticeable for an LOD. We can call this age $t = T_0$. The initial value is immaterial, and we always start with the value $A(T_0) = 0.001$, because the iterative process, by following steps 3 and 4, rapidly finds a close match value. As an example, in the case of AD, age $T_0 = 39$ years, $I(39) = 2.06 \cdot 10^{-6}$, and resulting $A(39) = 5.85 \cdot 10^{-9}$. $A(t)$ discovery for the next year starts with the value for the previous year and as a result requires slightly fewer iteration steps to find a match.

3. This $A(t)$ value is applied to all individuals' PRSs, following Eq. 4, thus making a first estimate of each individual hazard ratio for that year. Based on the resulting probabilities, a number of individuals will be diagnosed at age t.

4. The simulation verifies how well the result matches the expected incidence rate for the LOD at this age, and a better matching approximation value is recalculated and reapplied to step 3. This process will iterate steps 3 and 4 until a predetermined level of accuracy is achieved, with the aim of 0.1% reproduction accuracy for ages with significant incidence rates.

5. This cycle of steps 3 and 4 is then re-run a predetermined number of times (we used 10 repetitions), with iterations alternatively starting with a value $A(t)$ 10% above and below the previously determined value to account for a potential determination bias. The results of these re-runs are averaged and the resulting variance is evaluated and recorded.

6. The validated aging coefficient $A(t)$ from step 5 is then applied to the step 3 operation one final time in order to sample the individuals diagnosed at this age out of the population. Statistically, the highest-PRS individuals are most likely to be diagnosed, and these individuals are excluded from the healthy population.

7. The age is advanced by one year, with a now-smaller population, as in Eq. 4, and the simulation repeats steps 3–6 until all ages are covered—in our case, until the age of 120.

At completion, we have discovered the aging coefficient that maps the modeled PRS value to the individual hazard ratio for each year of age.

Gene therapy and changing life-expectancy analysis

The prophylactic gene therapy experiment assumes that we can effectively edit a large number of SNPs. The therapy itself is a simple arithmetic operation of reducing the individual PRS according to the scenarios described below. We are interested in applying the derived PRS mapping as the population ages and finding the resulting population incidence rate and lifetime risk pattern. The therapy simulation was performed in two scenarios:

Scenario 1: Prophylactically treating above-average-risk individuals to match the polygenic risk average for the baseline population. Fig. 2**a** shows the average number of alleles that need correction for at-risk individuals in accordance with the population distribution. Their risk is adjusted to become equal to the baseline population mean at age zero. The simulation then proceeds through the complete age range with the treated population, while we analyze the resulting incidence rate produced in the population with the PRSs reduced by therapy.

Scenario 2: Simulating a fourfold reduction in all individuals' PRS (OR = 0.25). This scenario requires an edit of 15 SNPs in a common allele low-effect-size scenario with an average OR of 1.1, it is simply calculated as: $1.1^{15} \approx 4$. Choosing to edit only maximum-effect-size SNPs with OR = 1.15 would require approximately 10 edits: $1.15^{10} \approx 4$. While Scenario 1 treats current at-risk individuals, Scenario 2 may be viewed as an emulation of treating some fraction of at-risk individuals for a number of generations, leading to a decline in the number of risk alleles in the whole population as a result of previous generations undergoing inheritable prophylactic treatment.

To model the lifespan increase, we adjust the mortality from all causes by the number of years required. We shift all mortality rates from the US Social Security "Actuarial Life Table"[47] by 5, 10, and 15 years. This approximation is valid because we focus exclusively on LODs, with onsets commencing decades after birth.

This study does not evaluate potential interference due to pleiotropy[48,49], here defined as the possible negative influence on other phenotypic features while trying to prevent an LOD by modifying a subset of SNPs. The high-risk individual PRS is caused by a large number of variants, as can be seen in Supplementary Fig. 1. There is a relatively small difference in the absolute number of detrimental alleles between the population average and higher-risk individuals. Arguably, for personalized prophylactic treatment, it will be possible to choose a small fraction of variants from a large set of available choices, without antagonistic pleiotropy, or perhaps even choose SNPs that are agonistically pleiotropic to some other LOD.

Conceptual outline

We modified and built upon our earlier research simulation[38]. The implementation in that research focused on two main questions: (1) determining the allele distribution change as a function of PRS and the incidence rate pattern of the LODs using the genetic architecture models of[50,51] and (2) quantifying the resulting change in GWAS statistical discovery power and simulating a GWAS cohort study.

The elements related to (2) were not used in the current research. The first part of that work, dedicated to population allocation, population aging, PRS analysis, multithreaded processing architecture, and related statistical analysis, also applies here.

The aging coefficient concept

The earlier research[38] showed that the incidence rates for these LODs continuously increase or remain high as individuals age, and the PRS of individuals diagnosed at older ages, on average, declines. If we follow the additive model of PRS and environmental effects, from a statistical perspective, two processes are balancing disease incidence: the aging or environmental effect increases the average population susceptibility, and this makes it more likely for older individuals to become ill. Individuals with a higher PRS are more likely to become ill, and with higher incidence rates, a larger fraction of them become ill with every additional year of age and are therefore excluded from the high-risk pool for the following years of age. The aging process continues, as reflected in the incidence in subsequent years, which remains high or rises.

Statistically, individual PRS and environmental effects, including aging, affect the probability of a person becoming ill[25]. According to Cox's proportional hazards model[52], in every year of age, a hazard ratio can be assigned for an individual to be diagnosed with an LOD. Our goal is to find the mapping between PRSs and individual hazard ratios on a yearly basis:

$$R_u(t) \propto A(t) \cdot G_u, \tag{1}$$

where $R_u(age)$ is the hazard ratio of uth unaffected individual, and t is the individual's age in years. PRS G_u remains constant for each individual for life, and the multiplier $A(t)$ drives the age-related increase in LOD incidence. We call $A(t)$ the aging coefficient.

We should be able to map or "discover" $A(t)$ by applying the yearly incidence rate in a population to the individual's PRS distribution, based on an LOD genetic allele architecture, through the population simulation flow described above. The discovered $A(t)$, then applied to the population simulation on a yearly basis, should precisely reproduce the LOD incidence rate used to discover it. Later, if we build a population with a modified PRS, emulating the effect of gene therapy, and run the simulation applying $A(t)$ to this population, we will be able to find the corresponding incidence rate and other resulting statistics.

In addition to the age-related change in PRS distribution described above, the aging coefficient automatically incorporates other miscellaneous environmental effects that accumulate with age, because these are all reflected in the yearly incidence rate. We use the term "aging coefficient" throughout this publication rather than "environmental effect" to emphasize that it is an aggregate age-dependent parameter.

Discovery of the aging coefficient via probabilistic iterative simulation

Statistically, the average number of diagnosed N_d in a year t is a function of the incidence rate $I(t)$ and the remaining population unaffected by an LOD $N_u(t)$:

$$N_d(t) = I(t) \cdot N_u(t), \tag{2}$$

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

$$N_u(t+1) = N_u(t) - N_d(t) = N_u(t)(1 - I(t)). \tag{3}$$

By applying the incidence rate to the population number, we get a fractional average number of individuals projected to become ill per year. Because we run a population simulation that models aging as a process, we now need to map each unaffected individual's PRS-to-hazard ratio, symbolically described by Eq. 1. Assuming we know $A(t)$, the number of *simulation* individuals diagnosed in a year, $N_s(t)$ will be:

$$N_s(t) = \sum_{u=1}^{N_u(t)} \begin{cases} 1, & if \quad (A(t) \cdot G_u > UniformRealDistribution(0,1)) \\ 0, & else \end{cases} \tag{4}$$

Because we do not initially know the value of $A(t)$, we start with a relatively close by magnitude value. The iterative process will need to match $N_s(t)$ to $N_u(t)$ with the predetermined degree of accuracy. This is done by iterating through steps 3–6 of the previous subsection, *Simulation flow / Aging coefficient discovery*. That is, we iteratively adjust $A(t)$ sequentially for each year of age to achieve the closest match by minimizing ΔN:

$$\Delta N = N_s(t) - N_d(t) \tag{5}$$

As apparent from Eq. 4, because the equation is probabilistic, if we run the simulation multiple times, at each run we will have a different set of individuals diagnosed. While the higher-PRS individuals are more likely to be diagnosed, each run—particularly at lower incidence or population numbers—will produce a different distribution at each year of age, which then affects, to some extent, every subsequent year. To achieve results with high reproducibility, we need to use large populations and perform many iterations, as summarized in the *Accuracy and statistical analysis* section below.

Model genetic architecture

We use the additive logistic model of genetic architecture, where the polygenic risk is a combination of the sum of individual effects odds ratio logarithms and environmental effects. To avoid confusion, we must clarify that Chatterjee et al.[25], for instance, call this model *multiplicative* when the odds ratios (ORs) of effect sizes are multiplied and *additive* when logarithms are added. This is different from a model in which the probabilities of independent effects are summed in the combined risk assessment, as, for example, seen in sufficient component cause models in epidemiology[53]. A quote from Chatterjee et al.[25] is relevant to this research: "For case-control studies, if it can be assumed that environmental risk factors are independent of the SNPs in the underlying population, then case-only and related methods can be used to increase the power of tests for gene-environment interactions. To date, post-GWAS epidemiological studies of gene-environment interactions have generally reported multiplicative joint associations between low-penetrant SNPs and environmental risk factors, with only a few exceptions."

Out of six genetic architectures validated in earlier research[38], we used two, listed in Table 2. It has been determined that common genetic variants explain the majority of heritability for most complex traits[54,55]. Therefore, the *common allele low effect size* genetic architecture model (Scenario A, Table 2) is best suited for the LODs we analyze, and all results are reported

using this scenario. For verification, all simulations were also performed using a more extreme *rare allele medium effect size* model (Scenario B in Table 2), and the results for lifetime risk were essentially identical to Scenario A. Any material difference in the results would warrant additional investigation. Because the results do not differ, we do not present separate figures for Scenario B; the corresponding simulation data are available in the Supporting Information files.

Table 2. Genetic architecture scenarios with modeled allele frequencies and effect sizes.

Scenario	MAF range	OR range	MAF values	Allele OR values
A. Common low	0.073 - 0.499	1.05 - 1.15	0.073 0.18 0.286 0.393 0.5	1.05 1.075 1.1 1.125 1.15
B. Rare medium	0.0146 - 0.0998	1.28 - 2.01	0.0146 0.036 0.0572 0.0785 0.0998	1.28 1.463 1.645 1.828 2.01

To build the genetic architecture, minor allele frequencies (MAFs) and ORs are distributed in equal proportion using the values in the table. The blocks are repeated to achieve the required heritability. "Common low" stands for *common allele low effect size architecture* and "rare medium" for *rare allele medium effect size architecture*.

We calculated an individual PRS as the sum of all alleles' effect sizes, which is by definition a log(OR) (odds ratio logarithm) for each allele[51]:

$$\beta = log(OR) = \sum_k a_k log(OR_k), \tag{6}$$

where a_k is the number of risk alleles (0, 1, or 2), and OR_k is the OR of additional liability presented by the k-th allele.

Variance of the allele distribution is determined by:

$$var = 2\sum p_k(1-p_k)(log(OR_k))^2, \tag{7}$$

where p_k is the frequency of the k-th genotype[51]. The contribution of genetic variance to the risk of the disease is heritability:

$$h^2 = \frac{var(g)}{var(g)+\pi^2/3}, \tag{8}$$

where $\pi^2/3$ is the variance of the standard logistic distribution[50]. Most of LODs we analyze are characterized by high to medium familial heritability[56–64], with lung cancer known for its lower heritability[65]. We use the heritabilities of these LODs, as summarized in Table 1, with the resulting distribution depicted in Supplementary Fig. 1.

LOD incidence rate functional approximation

To evaluate each LOD's allele redistribution with age, it was necessary to approximate the yearly incidence from much rougher-grained statistics. The incidence rate functional approximations follow the earlier research[38].

The incidence approximation $I(t)$ is represented by Eq. 9. 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.

$$I(t) = \begin{cases} ae^{bt}+c, & \textit{until intersection with the line, below} \\ i+st, & \textit{thereafter} \end{cases} \tag{9}$$

The logistic approximation (with parameters a, b, c, and d) is characterized by the following equation:

$$I(t) = \frac{a}{1+e^{(c-t)/b}} + d, \tag{10}$$

We use the yearly incidence rate logistic and exponential regression for the functional approximation of the LODs we analyze based on the available clinical incidence statistics, seen in Supplementary Fig. 8, and as also appropriate here, we use the logistic approximation for all LODs except breast cancer, for which the exponential followed by linear regression approximates the incidence rate more accurately.

Additional statistical concepts

The population mortality from the US Social Security "Actuarial Life Table"[47], in conjunction with the above two determinants, produces the lifetime risk pattern characteristic for each of the LODs in Fig. 2**c**. The PRS distribution is built with the number of variants required to achieve the analyzed LODs' heritability summarized from the familial studies.

Implicit in our simulation is the Cox proportional hazards model[52], where the hazard ratio for an individual, in one-year increments t in our case, is represented as:

$$\lambda_u(t) = \lambda_0(t)e^{\sum_i^n b_i x_i}, \tag{11}$$

where b_i are the coefficients of risk or impact of the ith hazard cause, and the x_i are the covariates of presence or absence of the ith hazard for individual u, measured as 0 or 1. The occurrence or absence of a polygenic or GWAS case is the sum of risk alleles β, $\lambda_u(t)$ is the hazard ratio of an individual in year t, and $\lambda_0(t)$ is a baseline hazard multiplier common to all individuals. The chances for an individual not to be diagnosed with an LOD by age T is:

$$\prod_{t<T}(1-\lambda(t)), \tag{12}$$

which is an important statement of probability; to remain healthy to age T is the product of the probabilities of the same for each preceding year. Zhang's[66] adjustment can similarly be used to estimate the hazard ratio.

We introduce a parameter that is useful in depicting LOD lifetime risk contributions at yearly age increments, the *lifetime risk density*, $D(t)$. It can be calculated as the incidence rate at age t relative to the initial population count, adjusted for mortality:

$$D(t) = \frac{I(t)S(t)}{N(0)}, \tag{13}$$

where $I(t)$ is the yearly incidence, $S(t)$ is the survivor rate from actuarial tables[47], and $N(0)$ is the initial population count. It is particularly handy for comparing the age distribution changes of LODs when modeling increased lifespan. Integrating the area under the curve, or summing up discrete yearly values, exactly corresponds to an LOD lifetime risk $L(t)$:

$$L(t) = \sum_t D(t), \tag{14}$$

see Fig. 2**e**, Fig. 3**b**, Supplementary Fig. 3, and Supplementary Fig. 7.

Accuracy and statistical analysis

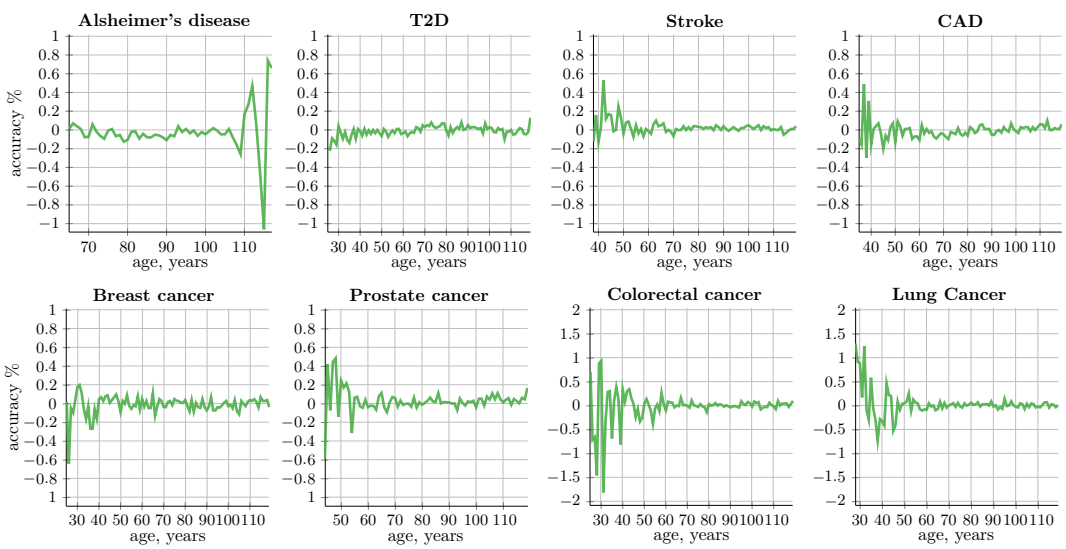


Fig. 4. Accuracy of simulation aging coefficient discovery in reproducing the input incidence rate approximation.

Fig. 4 shows the resulting accuracy produced by applying the aging coefficient to eight LODs, compared to the input incidence rate approximation. The error rate begins near 2% for colorectal and lung cancers, is noticeably better than 1% for the remaining LODs, and stays below 0.2% for most of the LODs' onset range. This validation shows that the combined discovery and analysis reproduce the input LOD incidence rate with a high degree of precision.

The values with lower accuracy occur at very early and very late onset ages, affecting a minute fraction of the total population, this itself being the cause of the variability. As seen in Fig. 2**e**, Fig. 3**b**, the incidence density is very low at younger ages, where the incidence rate is below $1 \cdot 10^{-5}$. Due to the population mortality, even though the incidence rate is high at old ages, the number of individuals alive is small after 105 years of age, even with an increase in the modeled lifespan of 15 years. Only in the case of AD, the incidence of which grows exponentially to a very late age, does the remaining unaffected population become small near 120 years of age, causing error rate to increase to close to 1%, a minuscule deviation with the particularly low lifetime risk density at this age.

The above precision is due to the use of large population sets: 25 million simulated individuals, with aggregation of 250 simulation loops (configurable) for each simulation iteration in the discovery stage, and a population of one billion for the gene therapy simulation. While the analysis could certainly have been performed with smaller simulation sets, leveraging the available computing equipment allowed us to achieve low statistical variance while simplifying analysis. This precision rendered the use of error bars in the graphical displays impractical.

Data sources, programming and equipment

We used population mortality numbers from the 2014 US Social Security "Actuarial Life Table"[47], 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: Alzheimer's disease[43,45,67,68], type 2 diabetes[46], coronary artery disease and cerebral stroke[69], and cancers[70,71].

To run simulations, we used an Intel Xeon Gold 6154 CPU-based 36-core computer system with 288GB of RAM. The simulation is written in C++, and the source code can be found in the Supplementary Information.

The final simulation data, additional plots, R scripts, and the executables are also 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 a corresponding execution slowdown.

Acknowledgments

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

Author contributions

Roman T. Oliynyk solely contributed to all aspects of this work.

Competing interests

The author declares that there are no competing interests.

Data availability

The author confirms that all relevant data are included in the paper and its supplementary information files.

References

1. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. & Yang, J. 10 years of gwas discovery: biology, function, and translation. *The Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. *Available at http://omim.org/statistics/geneMap (accessed May 31, 2018)* (2018).
3. Muller, H. J. Our load of mutations. *Am. journal human genetics* **2**, 111 (1950).
4. Morton, N. E., Crow, J. F. & Muller, H. J. An estimate of the mutational damage in man from data on consanguineous marriages. *Proc. Natl. Acad. Sci.* **42**, 855–863 (1956).
5. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome biology* **17**, 241 (2016).
6. Lynch, M. Mutation and human exceptionalism: our future genetic load. *Genetics* **202**, 869–875 (2016).
7. Gao, F. & Keinan, A. High burden of private mutations due to explosive human population growth and purifying selection. *BMC genomics* **15**, S3 (2014).
8. Ginn, S. L., Amaya, A. K., Alexander, I. E., Edelstein, M. & Abedi, M. R. Gene therapy clinical trials worldwide to 2017: An update. *The journal gene medicine* **20**, e3015 (2018).
9. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl. Acad. Sci.* **107**, 1752–1756 (2010).
10. Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J. *et al.* Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nat. genetics* **44**, 369–375 (2012).

11. Ribezzo, F., Shiloh, Y. & Schumacher, B. Systemic dna damage responses in aging and diseases. In *Seminars in cancer biology*, vol. 37, 26–35 (Elsevier, 2016).

12. Nelson, P. & Masel, J. Intercellular competition and the inevitability of multicellular aging. *Proc. Natl. Acad. Sci.* 201618854 (2017).

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

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

15. Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human genome sequencing in health and disease. *Annu. review medicine* **63**, 35–61 (2012).

16. Yashin, A. I., Arbeev, K. G., Wu, D., Arbeeva, L., Kulminski, A., Kulminskaya, I., Akushevich, I. & Ukraintseva, S. V. How genes modulate patterns of aging-related changes on the way to 100: Biodemographic models and methods in genetic analyses of longitudinal data. *North Am. Actuar. J.* **20**, 201–232 (2016).

17. Rzhetsky, A., Wajngurt, D., Park, N. & Zheng, T. Probing genetic overlap among complex human phenotypes. *Proc. Natl. Acad. Sci.* **104**, 11694–11699 (2007).

18. Steck, A. K. & Rewers, M. J. Genetics of type 1 diabetes. *Clin. chemistry* **57**, 176–185 (2011).

19. Ghani, M., Reitz, C., St George-Hyslop, P. & Rogaeva, E. Genetic complexity of early-onset alzheimer's disease. In *Neurodegenerative Diseases*, 29–50 (Springer, 2018).

20. Jager, R. D., Mieler, W. F. & Miller, J. W. Age-related macular degeneration. *New Engl. J. Medicine* **358**, 2606–2617 (2008).

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

22. Sobrin, L., Ripke, S., Yu, Y., Fagerness, J., Bhangale, T. R., Tan, P. L., Souied, E. H., Buitendijk, G. H., Merriam, J. E., Richardson, A. J. *et al.* Heritability and genome-wide association study to assess genetic differences between advanced age-related macular degeneration subtypes. *Ophthalmology* **119**, 1874–1885 (2012).

23. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbnsfp v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site snvs. *Hum. mutation* **37**, 235–241 (2016).

24. Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C. & Brookes, A. J. Gwas central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. journal human genetics* **22**, 949 (2014).

25. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392 (2016).

26. Wright, A. V., Nuñez, J. K. & Doudna, J. A. Biology and applications of crispr systems: harnessing nature's toolbox for genome engineering. *Cell* **164**, 29–44 (2016).

27. Carroll, D. Genome engineering with zinc-finger nucleases. *Genetics* **188**, 773–782 (2011).

28. Joung, J. K. & Sander, J. D. Talens: a widely applicable technology for targeted genome editing. *Nat. reviews Mol. cell biology* **14**, 49 (2013).

29. Chen, J. S., Dagdas, Y. S., Kleinstiver, B. P., Welch, M. M., Harrington, L. B., Sternberg, S. H., Joung, J. K., Yildiz, A. & Doudna, J. A. Enhanced proofreading governs crispr-cas9 targeting accuracy. *bioRxiv* 160036 (2017).

30. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by crispr–cas9 leads to large deletions and complex rearrangements. *Nat. biotechnology* (2018).

31. Thompson, D., Aboulhouda, S., Hysolli, E., Smith, C., Wang, S., Castanon, O. & Church, G. The future of multiplexed eukaryotic genome engineering. *ACS chemical biology* **13**, 313–325 (2017).

32. Kohman, R. E., Kunjapur, A. M., Hysolli, E., Wang, Y. & Church, G. M. From designing the molecules of life to designing life: Future applications derived from advances in dna technologies. *Angewandte Chemie* **57(16)**, 4313–4328 (2018).

33. Fisher, R. A. The genetical theory of natural selection: a complete variorum edition. *Dover Publ.* (1930).

34. Haldane, J. The part played by recurrent mutation in evolution. *The Am. Nat.* **67**, 5–19 (1933).

35. Kennedy, B. K., Berger, S. L., Brunet, A., Campisi, J., Cuervo, A. M., Epel, E. S., Franceschi, C., Lithgow, G. J., Morimoto, R. I., Pessin, J. E. *et al.* Geroscience: linking aging to chronic disease. *Cell* **159**, 709–713 (2014).

36. Vineis, P., Schulte, P. & McMichael, A. J. Misconceptions about the use of genetic tests in populations. *The Lancet* **357**, 709–712 (2001).

37. D'Agostino Sr, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Massaro, J. M., Kannel, W. B. *et al.* General cardiovascular risk profile for use in primary care. *Circulation* (2008).

38. Oliynyk, R. T. Age-related late-onset disease heritability patterns and implications for genome-wide association studies. *bioRxiv* (2018). DOI 10.1101/349019. https://www.biorxiv.org/content/early/2018/07/11/349019.1.full.pdf.

39. Association, A. *et al.* 2017 alzheimer's disease facts and figures. *Alzheimer's & Dementia* **13**, 325–373 (2017).

40. Lakatta, E. G. & Levy, D. Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises: Part i: aging arteries: a "set up" for vascular disease. *Circulation* **107**, 139–46 (2003).

41. Lakatta, E. G. So! what's aging? is cardiovascular aging a disease? *J. molecular cellular cardiology* **83**, 1–13 (2015).

42. Mitchell, T. J., Turajlic, S., Rowan, A., Nicol, D., Farmery, J. H., O'Brien, T., Martincorena, I., Tarpey, P., Angelopoulos, N., Yates, L. R. *et al.* Timing the landmark events in the evolution of clear cell renal cell cancer: Tracerx renal. *Cell* **173**, 611–623 (2018).

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

44. Wu, Y.-T., Beiser, A. S., Breteler, M. M., Fratiglioni, L., Helmer, C., Hendrie, H. C., Honda, H., Ikram, M. A., Langa, K. M., Lobo, A. *et al.* The changing prevalence and incidence of dementia over time - current evidence. *Nat. Rev. Neurol.* (2017).

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

46. Boehme, M. W., Buechele, G., Frankenhauser-Mannuss, J., Mueller, J., Lump, D., Boehm, B. O. & Rothenbacher, D. 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 Heal.* **15**, 855 (2015).

47. *Social Security Administration (US). Available at https://www.ssa.gov/oact/STATS/table4c6.html (accessed May 31, 2018)* (2018).

48. Stearns, F. W. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767–773 (2010).

49. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet.* **29**, 66–73 (2013).

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

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

52. Cox, D. Regression models and life-tables. *J. Royal Stat. Soc. Ser. B (Methodological)* **34**, 187–220 (1972).

53. Park, L. & Kim, J. H. A novel approach for identifying causal models of complex diseases from family data. *Genetics* genetics–114 (2015).

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

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

56. Wienke, A., Holm, N. V., Skytthe, A. & Yashin, A. I. The heritability of mortality due to heart diseases: a correlated frailty model applied to danish twins. *Twin Res. Hum. Genet.* **4**, 266–274 (2001).

57. 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. *J. internal medicine* **252**, 247–254 (2002).

58. Devan, W. J., Falcone, G. J., Anderson, C. D., Jagiella, J. M., Schmidt, H., Hansen, B. M., Jimenez-Conde, J., Giralt-Steinhauer, E., Cuadrado-Godia, E., Soriano, C. *et al.* Heritability estimates identify a substantial genetic contribution to risk and outcome of intracerebral hemorrhage. *Stroke* **44**, 1578–1583 (2013).

59. Aparicio, H. J. & Seshadri, S. Familial occurrence and heritability of stroke. In *Stroke Genetics*, 9–20 (Springer, 2017).

60. Naj, A. C. & Schellenberg, G. D. Genomic variants, genes, and pathways of alzheimer's disease: An overview. *Am. J. Med. Genet. Part B: Neuropsychiatr. Genet.* **174**, 5–26 (2017).

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

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

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

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

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

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

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

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

69. Rothwell, P., Coull, A., Silver, L., Fairhead, J., Giles, M., Lovelock, C., Redgrave, J., Bull, L., Welch, S., Cuthbertson, F. *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* **366**, 1773–1783 (2005).

70. *Available at http://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk (accessed May 31, 2018)* (2018).

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

SUPPLEMENTARY FIGURES

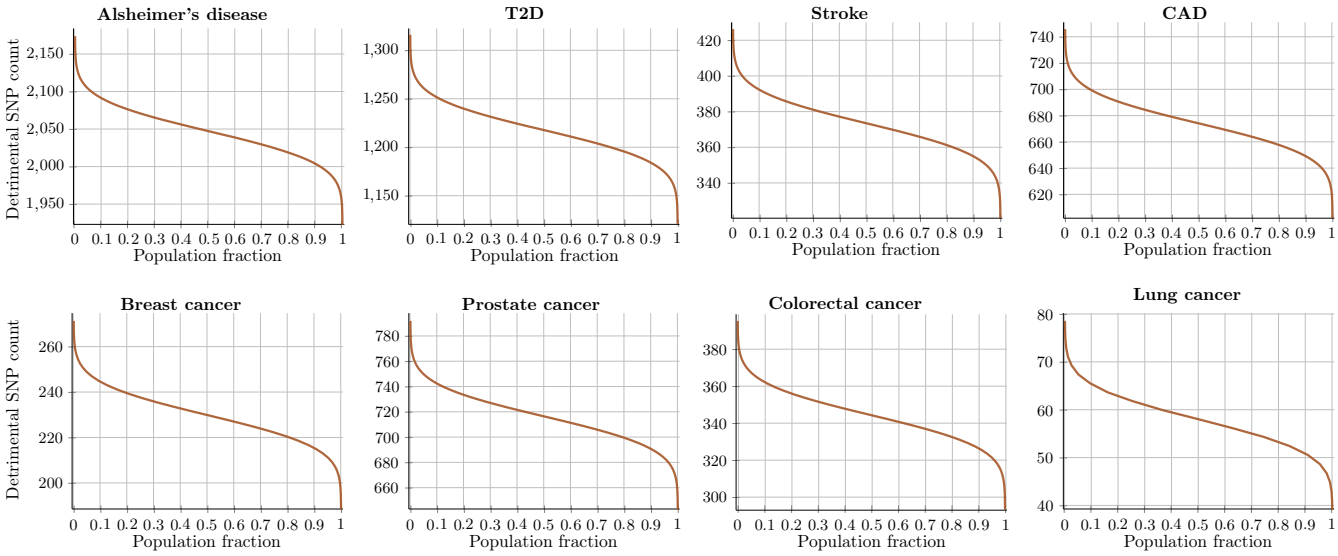


Fig. 5. Supplementary Fig. 1

Population distribution of detrimental variants for common low-effect-size genetic architecture. Based on initial heritability, the individuals in a population carry relatively high counts of detrimental low-effect alleles, resulting in the combined LOD PRSs. The higher the heritability, the higher the total number of detrimental LODs and the higher the PRS produced by a relatively smaller difference between the population mean and high- or low-risk individuals.

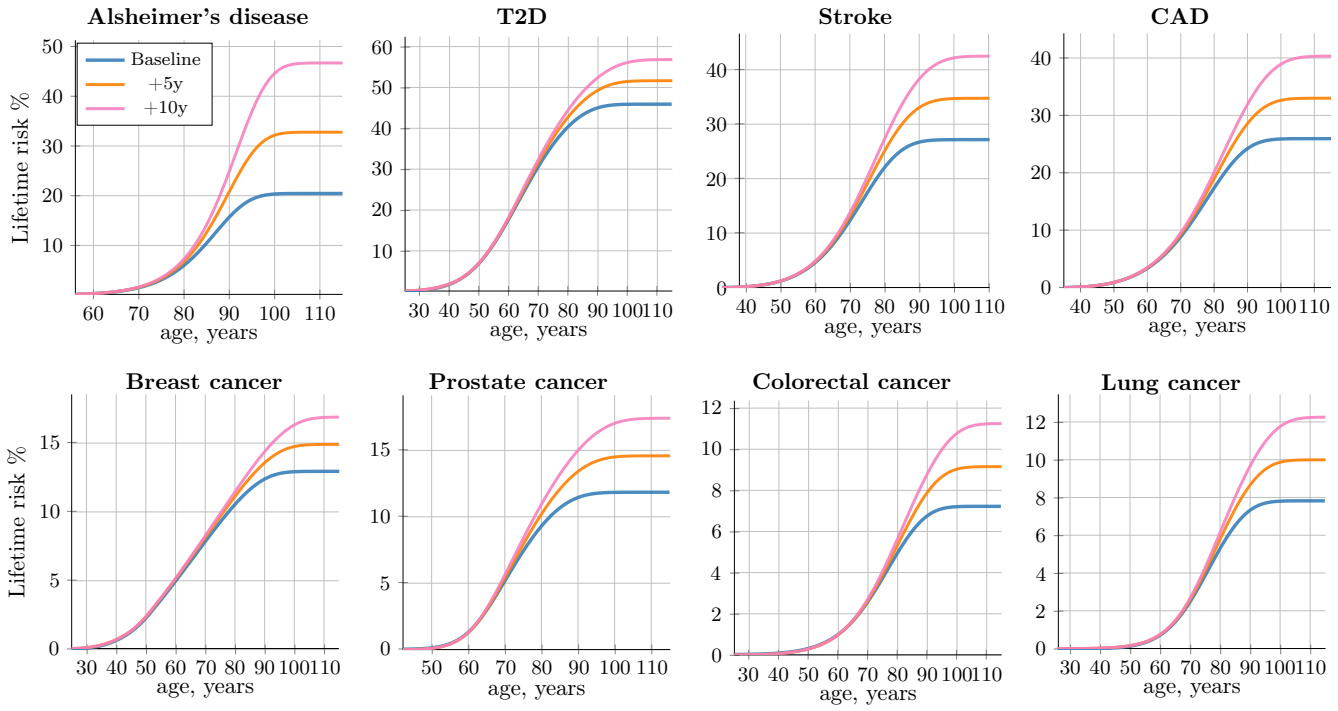


Fig. 6. Supplementary Fig. 2

Projected LOD lifetime risk increase if lifespan is increased by 5 and 10 years. Baseline without gene therapy. This is the baseline scenario without gene therapy or other health improvements for the plotted LOD. This represents the case in which the lifespan increases due to causes other than the plotted LOD. Mortality patterned from US Social Security "Actuarial Life Table"[47]

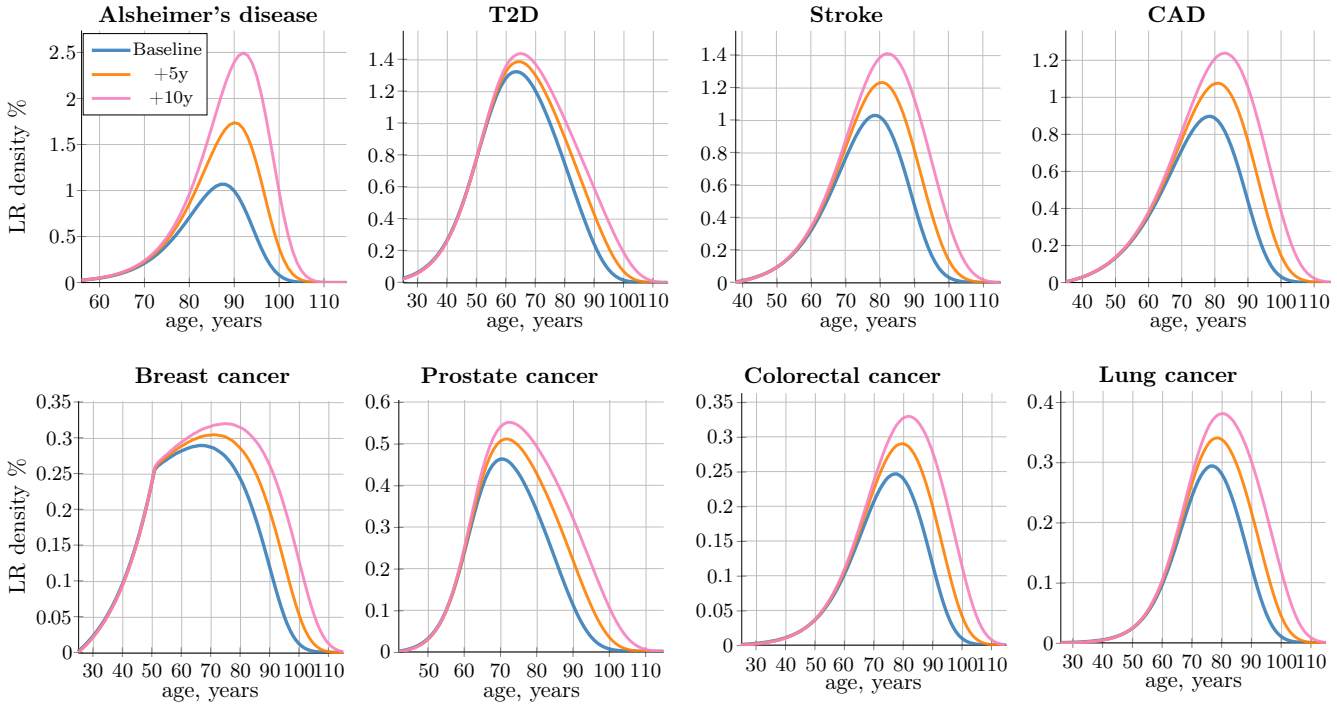


Fig. 7. Supplementary Fig. 3

Projected LOD incidence rate relative to the number of individuals at birth if the lifespan is increased by 5 and 10 years. Baseline without gene therapy. This is also the incidence rate density: The area under the curve is equal to the lifetime risk for each scenario, accounting for mortality[47]. This is the baseline scenario without gene therapy or other health improvements for the plotted LOD. This represents the case where lifespan increases due to causes other than the plotted LOD.

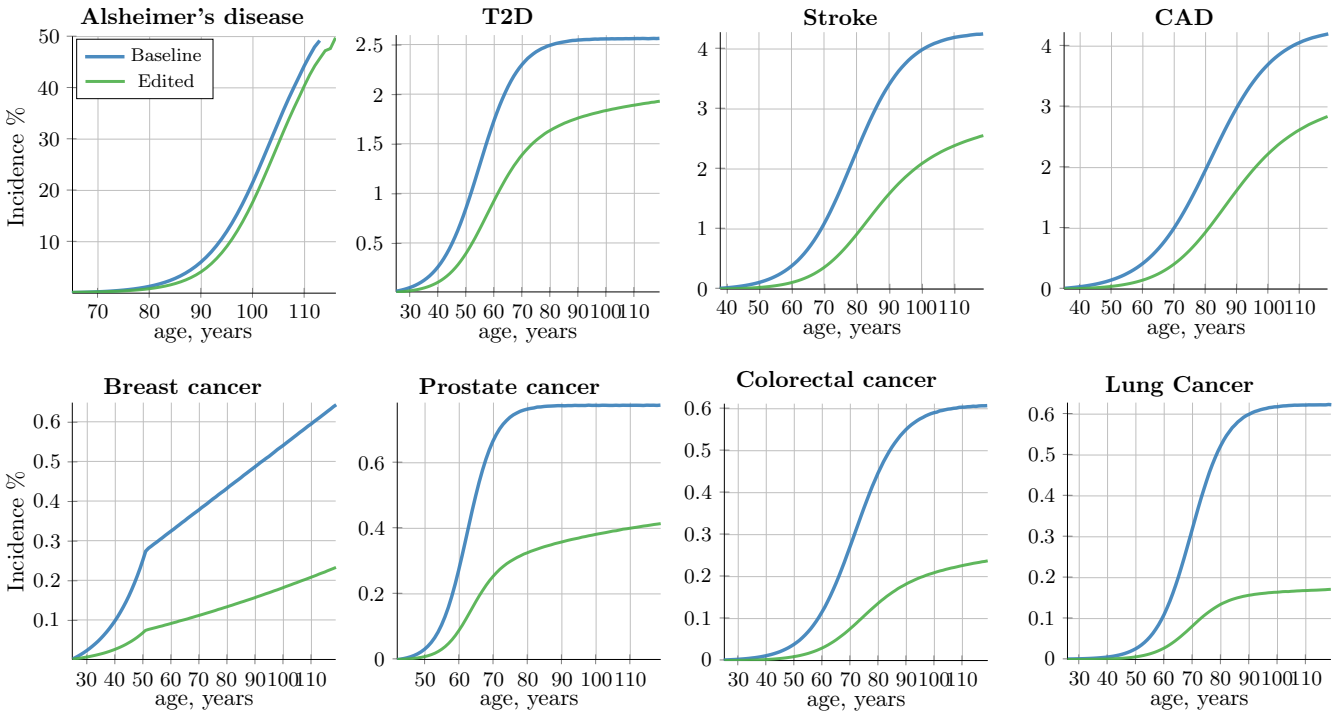
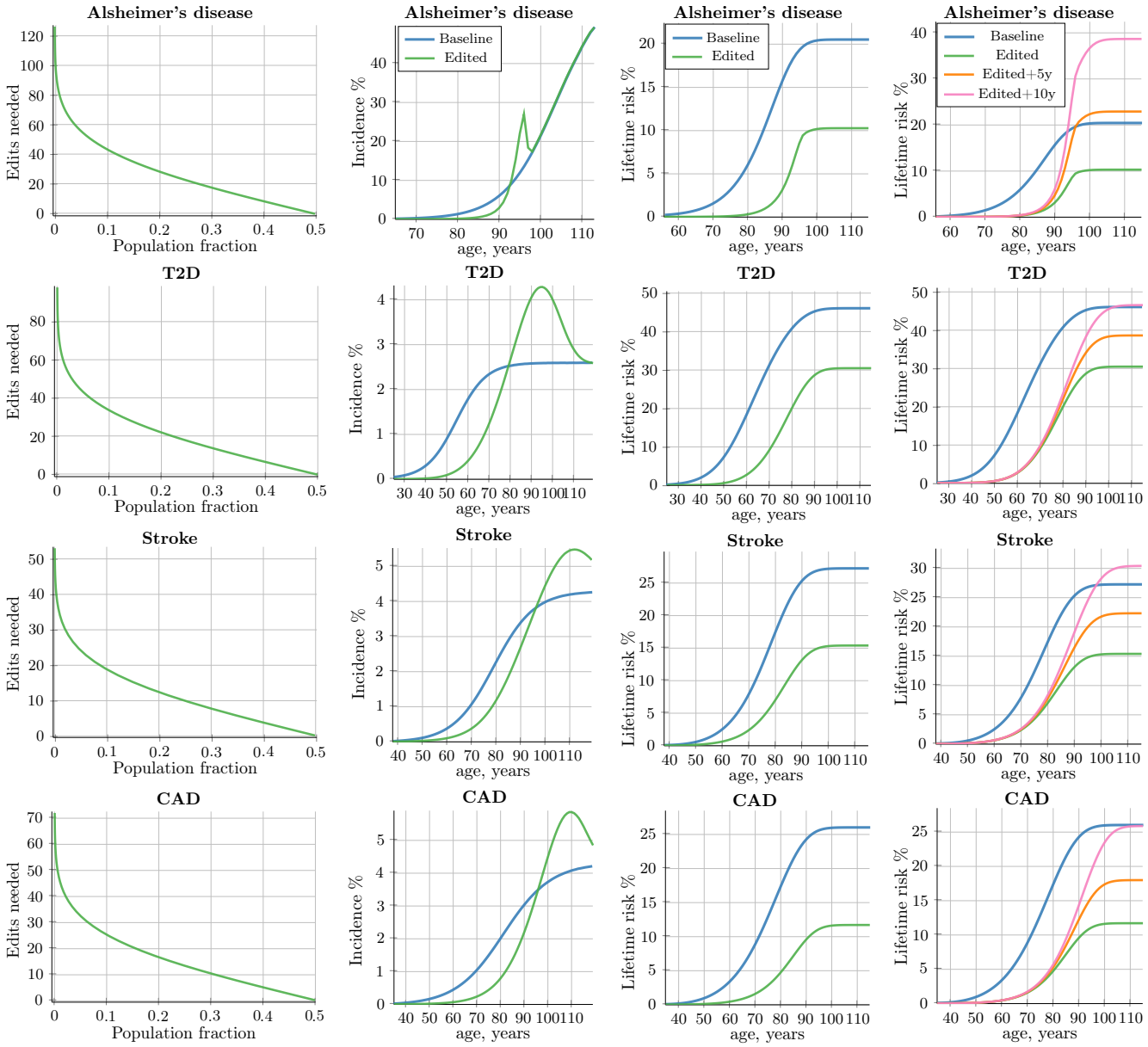


Fig. 8. Supplementary Fig. 4

Scenario 2: Gene therapy corrected on average 15 SNPs, corresponding to an odds ratio change of 0.25 for all individuals in a population.



(a) Pe LOD, number of edits for at-risk population

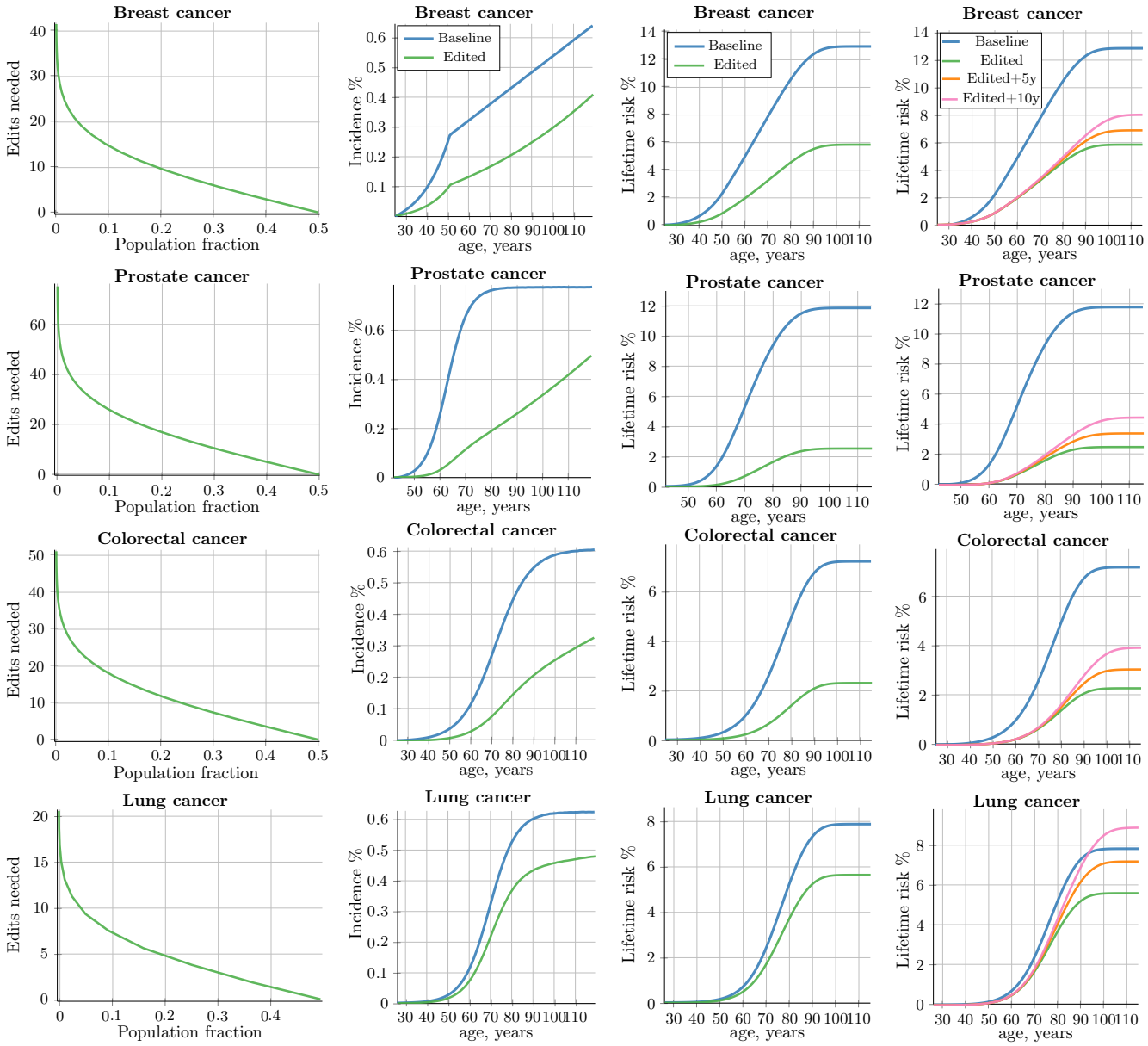
(b) Incidence rate decline from baseline after gene therapy

(c) Lifetime risk decrease from baseline after gene therapy

(d) Lifetime risk improvement from baseline and lifespan increase: 5- and 10- year scenarios

Fig. 9. Supplementary Fig. 5

Scenario 1: Enlarged view of gene therapy results with the incidence rate, lifetime risk, and lifetime risk with lifespan extension for prevalent LODs.



(a) Per LOD number of edits for risk population

(b) Incidence rate decline from baseline after gene therapy

(c) Lifetime risk decrease from baseline after gene therapy

(d) Lifetime risk improvement from baseline and lifespan increase: 5- and 10- year scenarios

Fig. 10. Supplementary Fig. 6

Scenario 1: Enlarged view of gene therapy results with the incidence rate, lifetime risk, and lifetime risk with lifespan extension for cancers.

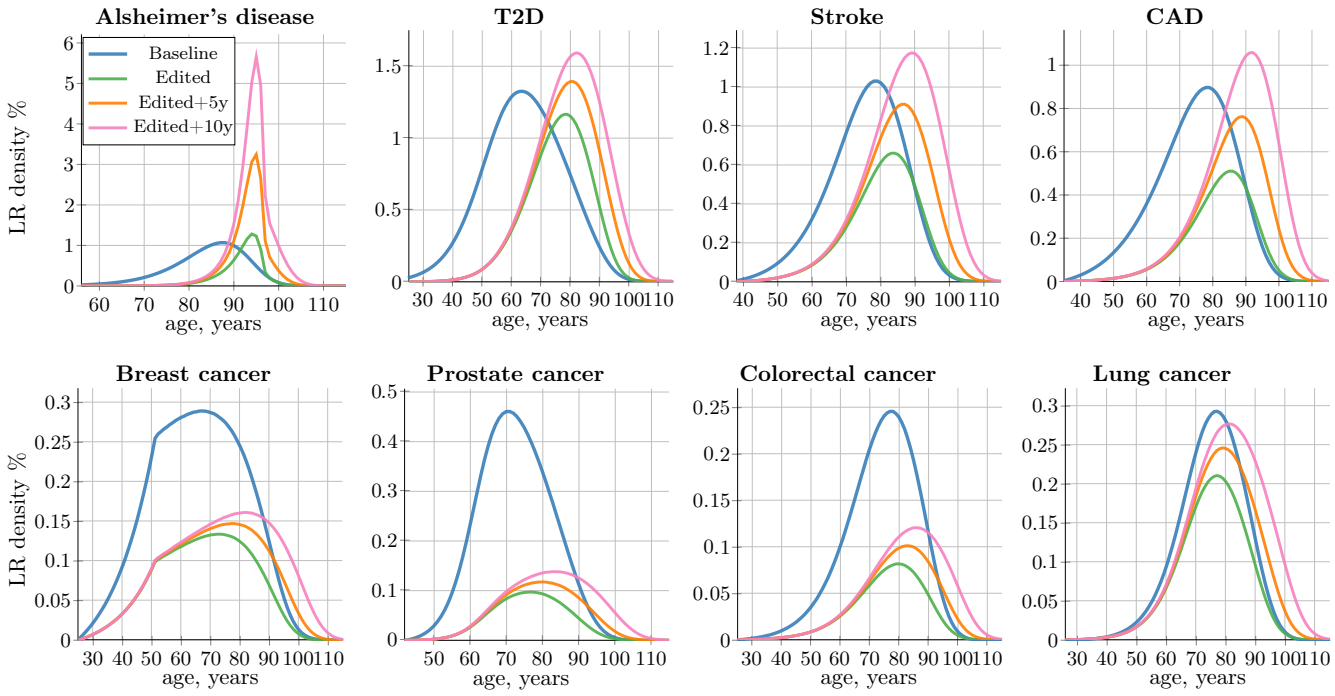


Fig. 11. Supplementary Fig. 7

Scenario 1: Enlarged view of incidence rate relative to the number of individuals at birth. Baseline, gene edited with 0-, 5- and 10-year lifespan extensions. This plot reflects incidence rate density, where the area under the curve is equal to the lifetime risk for each scenario, accounting for mortality. For comparison, see these LODs' projections of lifespan increase as a baseline without gene therapy in Supplementary Fig. 2 and Supplementary Fig. 3.

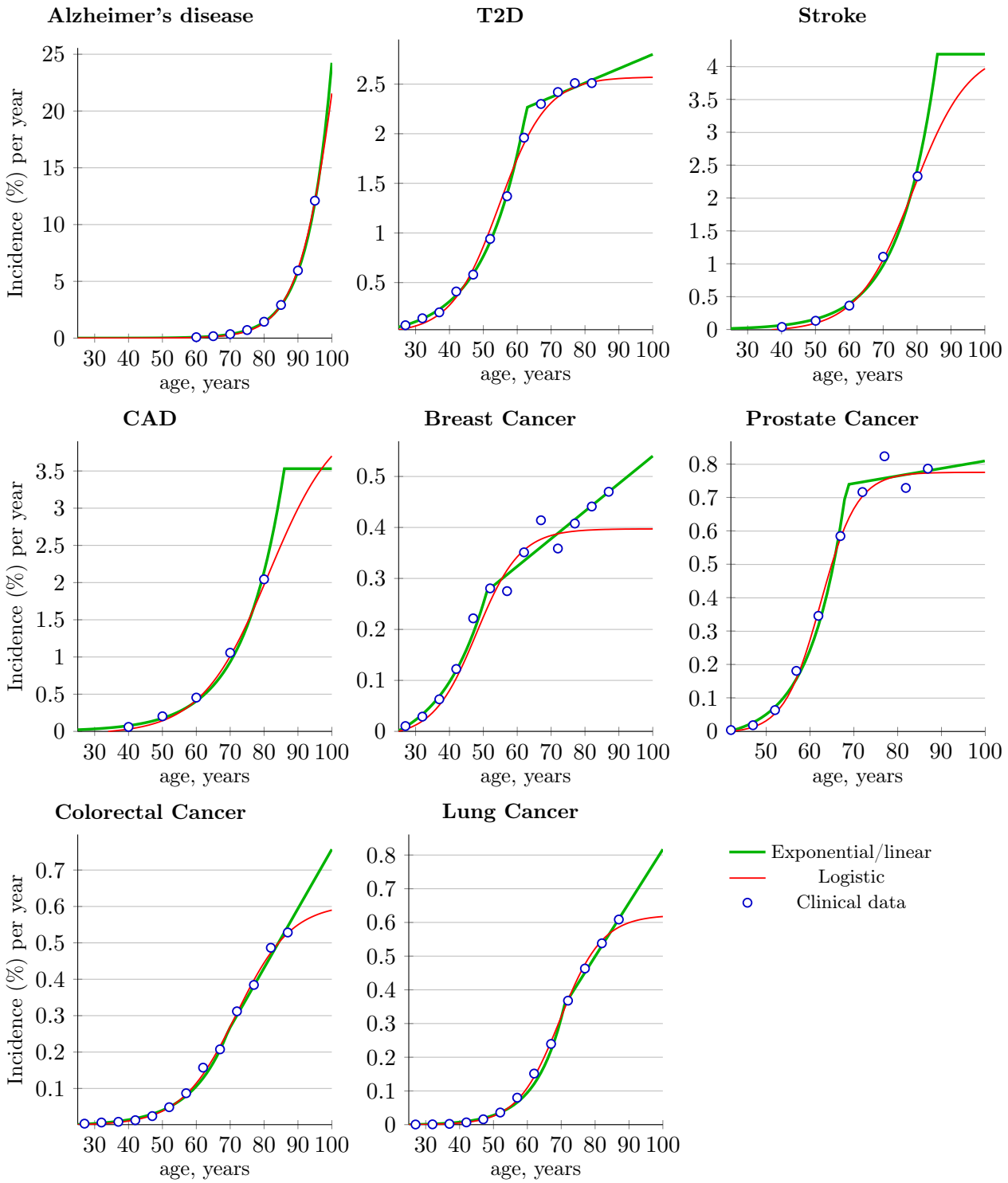


Fig. 12. Supplementary Fig. 8

LOD clinical incidence rates and functional approximations. Two functional approximations of clinical data: (green) exponential followed by linear and (red) logistic.