Could future gene therapy prevent aging 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

In a few decades, gene therapy techniques and genetic knowledge may sufficiently advance to support prophylactic gene therapy to prevent late-onset diseases (LODs).
Polygenic risk scores and risk allele distribution of diagnosed individuals change with LOD diagnosis age, which may complicate statistical risk estimates. Population simulation naturally accounted for this effect. It quantified the correlation between aging process, polygenic score and hazard ratio of LODs at all ages based on clinical incidence rate and familial heritability for eight highly prevalent diseases.
Simulations confirmed that the hypothetical gene therapy would be very beneficial in delaying the onset age and lowering lifetime risk of LODs. Longevity counterbalances the gains, with type 2 diabetes, stroke, and coronary artery disease regaining the pre-treatment baseline with 10-15 years of longer life. Alzheimer's disease proves most resistant, surpassing the baseline lifetime risk within 5 years longevity improvement. 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 (GWAS) of human and other organisms' genomes have become an everyday occurrence[1]. The knowledge of genetic variants, particularly single nucleotide polymorphisms (SNPs) that are 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. The Mendelian conditions cause high mortality and morbidity, but each of these conditions affects a tiny fraction of the population, however a large number of conditions is 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 that 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[3,4]. If these were to combine and become homozygous, they could be harmful or lethal[5–7]. Many of these conditions are well-characterized. In some cases, individualized genetic diagnoses are possible, where an SNP that needs to be edited can be specified precisely. 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 were conducted worldwide[8]. When the medical technology becomes available, individuals treated will be effectively cured and will not have to worry about the specific and single cause of their disease. For this reason, we do not analyze Mendelian conditions in this study.

We will focus on polygenic or complex late-onset diseases (LODs), which pose a more nuanced problem They 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], see the graphical demonstration in this study on model genetic architectures Supplementary Fig. 1. These diseases include 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 rare individuals makes it unlikely they will become ill from some but not all LODs[18]. 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 SNPs.

It is important to note that many diseases, including Alzheimer's disease (AD) and diabetes mellitus, also possess highly detrimental mutations that singularly can cause high disease liability. Usually, they have a special medical diagnosis associated with them. Diabetes mellitus caused by single defects in HLA-DQA1, HLA-DQB1, or HLA-DRB1 genes[19] develops in early childhood, affects a small percentage of overall diabetes mellitus sufferers, called type 1 diabetes to distinguish it from the common type 2 diabetes. Similarly, with Alzheimer's, the so-called early-onset AD (EOAD) is cased primarily by APP, PSEN1, or PSEN2 gene mutations and affects a relatively small portion of the population, starting in their thirties and affecting a majority of mutation carriers by age 65[20]. Macular degeneration[21–23] is primarily caused by a small number of high-effect variants and manifests at 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 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.

The computational techniques attempting to evaluate the effects of mutations or gene variants are being developed, though their accuracy needs to improve dramatically to be applicable to personalized human genetic evaluation or treatment[24]. Similarly, while there are extensive libraries of human SNPs, like dbSNP, HapMap, SNPedia and aggregating sites[25], the information is far from actionable as far as modifying multiple personal SNPs would be considered. Finding or being able to computationally estimate a complete set of the low effect causal SNPs is the knowledge that may take decades to gain. For purposes of this research, we assume that we know the "true" causal gene variants (see Chatterjee at al.[26]) distinct from GWAS SNPs that are in a 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 same low risk as applying an influenza vaccination, for modification of a large number of gene variants distributed over the human genome. The newest gene editing technique is called CRISPR-Cas9[27]. It supplemented and mostly replaced older technologies, zinc-finger nucleases[28] and TALEN[29], though, for some applications, these older techniques continue to be more appropriate. While the 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 succeed, using homologous repair with a template or a sister chromatid sequence. Successes announced by studies "Enhanced proofreading governs CRISPR-Cas9 targeting accuracy"[30], may be followed later by a discovery such as, "Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements"[31]. CRISPR is only 5 years old and is a rapidly developing technology with a great promise.

A potential future technology, currently in early exploratory stages, could be synthetic genomics studied in Harvard Church lab[32,33]. These techniques could also help in the precise mapping of gene variant effects on disease phenotypes. If all the approaches mentioned above ultimately fail to become reliable enough for all aspects of gene therapy, it is almost certain that a new suitable technique will be invented.

Notwithstanding the systematic media splashes about designer babies, many questions on 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 will assume 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 apply this hypothetical gene therapy model to see what would happen to LOD progression as the population ages. We are not considering additions of artificially designed genomic sequences but rather, only corrections done to a typically low effect heterozygous in population gene variants, a detrimental variant correction to a naturally occurring neutral state. For simplicity, our model operates on SNPs though the same would apply with higher complexity of gene therapy to other gene variant types.

Many changes in lifestyle and medical care led to increased longevity over the last century, and this trend is projected to continue; preventive gene therapies may also be one of the future factors. Therefore, we will model the age extension by a duration long enough to make projections but not so long as to render the model results questionable.

To be concise, further in 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 to have a person born with all cells treated, because all developmental stages are affected by the genome, which implies the germline and likely heritable editing. Yet, we do not know what technological possibilities and ethical considerations the future will bring; therefore, we assume 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 often a complex process can be probabilistically analyzed with a judiciously designed simulation, which is particularly applicable for population genetics analysis. The first models started a hundred years ago with analytical estimates[34,35] and later with the advent of computers, through simulations. The phenomena behind aging and LOD progression are complex. For example, individuals with a high polygenic risk score 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. 1b. Statistically, either scenario will happen with low likelihood, 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, pleiotropic[36], stochastic and methodologically explained by a concept of penetrance[37], et cetera. 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 allow to eliminate a large degree of uncertainty inherent in purely genomic predictions. A good example is the Framingham General Cardiovascular Risk Score[38], 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 a 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 use a simulation approach to a complex problem, and when the model is found that describes the baseline conditions well, we change the simulation input parameters moderately and evaluate the resulting projections.

We review and analyze eight of the most common diseases: Alzheimer's disease, type 2 diabetes, cerebral stroke, coronary artery disease (CAD), and four of the most prevalent cancers: breast, prostate, colorectal, and lung. From the methodological perspective, this current research is self-contained, yet, to avoid repeating a lengthy presentation, readers desiring to learn more about analyzed LODs from heritability, relevant clinical details, and an incidence rate progression perspective are referred to[18], together with that study supplementary documents dedicated to these subjects. Here we are going to use the same genetic architecture models and take advantage of the incidence rate functional approximations that we describe in Methods.

Similar to[18], 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 have to clarify that Chatterjee et al.[26], for example, call this model multiplicative when odds ratios (ORs) of effect sizes are multiplied and additive when logarithms are added. This is different from a model where the probabilities of independent effects are summed in the combined risk assessment, as, for example, viewed in *sufficient component cause* models in epidemiology[39]. A quote from Chatterjee et al. is relevant for 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."[26]

We set the goal to quantify the effects of hypothetical future gene therapy specifically in the case of LODs in an aging population. It was determined[18] that the average polygenic score of unaffected individuals diminishes with age while the risk of the disease increases with age, as reflected in rising incidence rates. This indicates that the environmental effects and aging make an increasing contribution to LOD incidence with age progression, which is particularly prominent in 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. They are failures of operation that may start at molecular, cellular, or organ levels and affect vital subsystems such as metabolism, circulation, and cognition. The failure does not happen immediately prior to the diagnosis age. For example, AD deterioration begins decades before symptoms become noticeable[40], which is similar for cardiovascular disease[41,42] and cancers[43].

While all the factors involved in an LOD causality and age of onset are complex, and vary between 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 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, we determine what we call the *aging coefficient*, 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 can be used to simulate the prophylactic gene therapy aimed at LOD prevention, and to quantify the populational change in the LOD's incidence and lifetime risk.

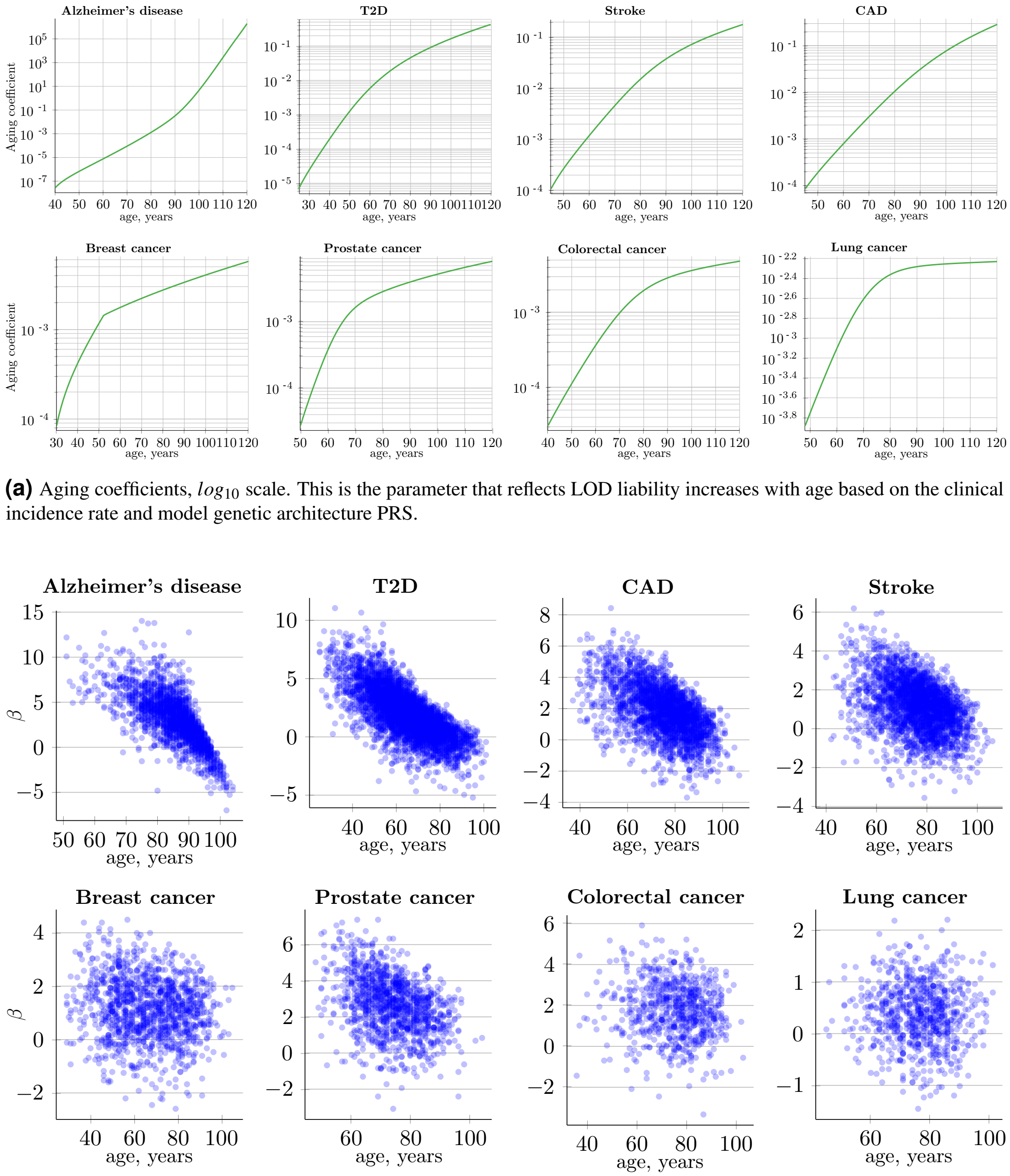
We considered a number of scenarios of prophylactic gene therapy. In the most extreme scenario, we could make thousands of edits for AD, and hundreds for most of the rest of the LODs, reversing all variants constituting individual PRSs to a neutral state; this would show a resulting zero disease risk 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 Methods:

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

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

To conclude this Introduction, it may appear that the scenarios we review could be simply calculated numerically. Among the reasons for this rigorous simulational approach is the finding[18] that the risk allele distribution and the polygenic risk of individuals diagnosed with an LOD diminishes with age. This PRS change is most prominent for highest heritability and prevalence LODs. Genetic statistical calculations do not account for this effect, particularly when quantifying health outcomes of a hypothetical genetic therapy, which changes the population PRSs, is concerned. Applying the novel approach, we quantify the correlation between environment/aging process, polygenic score and hazard ratio of selected LODs at all ages based on clinical incidence rate and familial heritability, and apply the resulting *aging coefficient* to simulated populations in two scenarios of hypothetical future gene therapy. Hence these results.



(a) Aging coefficients, log_{10} scale. This is the parameter that reflects LOD liability increases 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 $\beta = log(OddsRatio)$.

Fig. 1. Aging coefficients applied to reproduce the LOD incidence distribution by age based on individual PRS

Results

We use 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: Alzheimer's disease, type 2 diabetes, cerebral stroke, coronary artery disease, 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 Eq. 1 depicted in Fig. 1a, mapping individual PRSs to the hazard ratio specific to each LOD on a yearly basis. 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 aging coefficient changes with age for the diseases we analyze. AD shows the largest magnitude of change, and is characterized by steep exponential incidence rate increase continuing to very old age. T2D, CAD, and, stroke show a comparatively smaller progression, and cancers show a relatively small magnitude increase, with lung cancer showing the least change. Applying the aging coefficient back to the simulated population, Fig. 1b shows the simulation result of the PRS distribution for diagnosed individuals by age. It matches the results of[18], where a simulation using a simpler direct algorithm that was more appropriate for that research was performed. A more precise indication of the model accuracy is shown in Fig. 4.

Knowing the hazard ratio for any given PRS at any given age is exactly what the simulation needs for modeling gene therapy and lifespan increase. We applied the aging coefficient in two simulation scenarios described in Methods:

Scenario 1: Prophylactically treating above-average risk individuals to match polygenic risk averages for the baseline population

The simulated prophylactic gene therapy corrected the population with above-average risk to the average PRS. Fig. 2a shows the number of gene variants that would need correction under the common low effect genotype model. Alzheimer's disease, the analyzed LOD with the highest heritability, would require over 120 edits for the highest risk individuals and lung cancer 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 less than three edits for lung cancer, see summary Table 1. 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.

For the four LODs other than cancers, there is a spike in the incidence rate at a very old age, exceeding the baseline incidence rate, Fig. 2b. The AD incidence rate displayed a steep spike around age 95 and merged with the baseline rate pattern soon after. This can be explained by the continuing exponential incidence rate pattern[44].

T2D, CAD, and stroke also displayed significant, respectively later spikes above the baseline, which asymptotically lead toward the baseline incidence rate line from above with age progression.

This 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 are likely to be diagnosed with these LODs. 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 relatively high heritability and relatively lower incidence rate.

Lung cancer showed the least improvement. This is due to the lowest reported heritability, which leads to the fewest required edits for the treated population under our scenario, see Table 1.

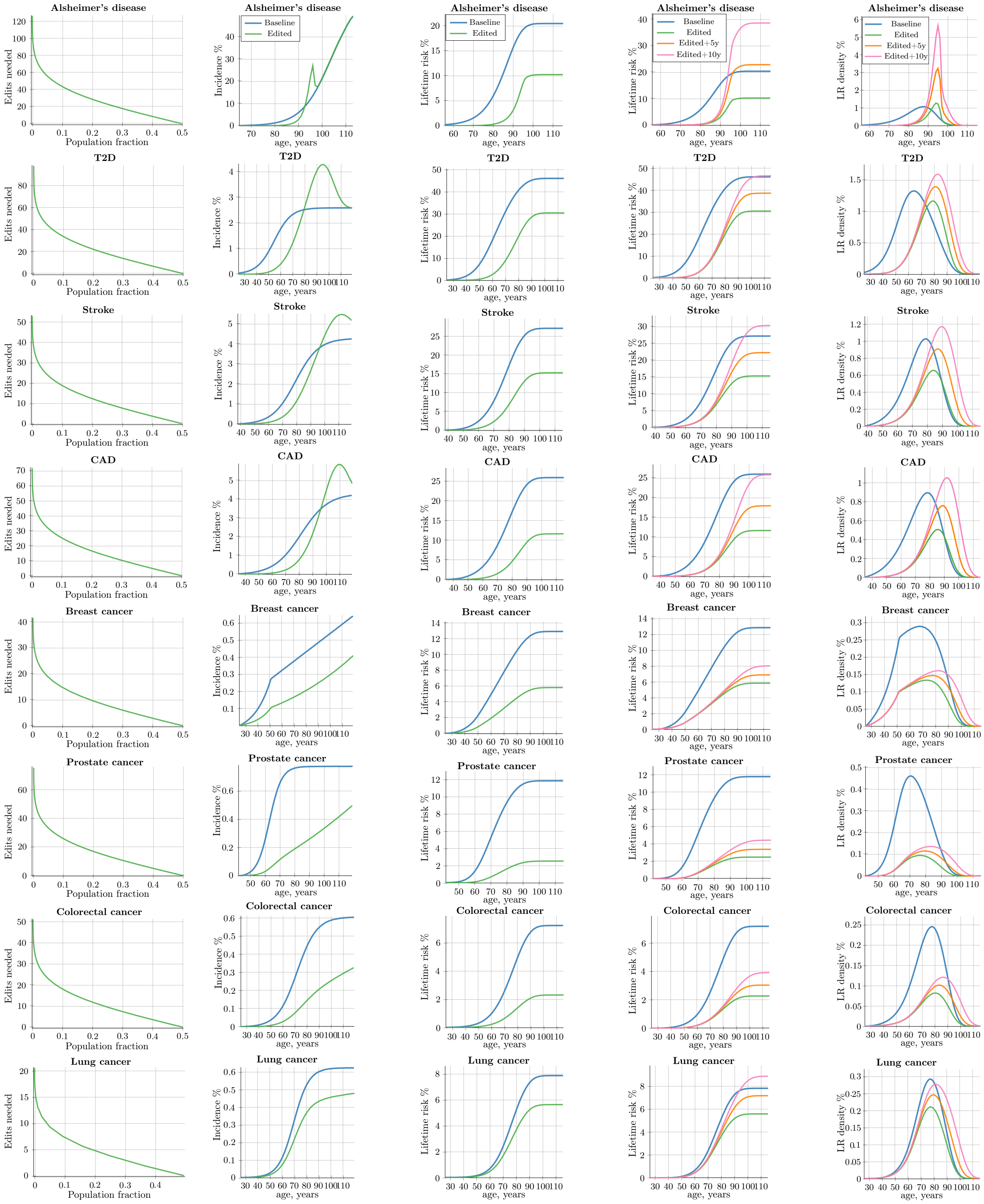
With the hypothetical future advances in the medicine, including these prophylactic gene therapy procedures, the average life expectancy is expected to increase. We model each disease without accounting for increased mortality caused by a LOD, which may be limiting the diagnoses of other diseases. This simplification works because we track the diagnosis of the disease, and taken independently, the count assumes there is a decrease in mortality from all causes. In Fig. 2d we consider an average lifespan increase by 5 and 10 years. We find that AD lifetime risk regains the baseline level within 5 years and almost doubles above the baseline with a 10-year life extension; see Table 1. T2D, stroke, CAD, and lung cancer all slightly exceed or match the baseline with 10 years longer lifespan, with incidence onset curve noticeably delayed. The most prevalent cancers—breast, prostate and colorectal—also show a significant improvement of population lifetime risk, which remains significantly lower than the baseline with 10 years longer lifespan.

Fig. 2e 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. 8. The plot shows how the incidence peak shifts toward older ages with increased longevity.

Supplementary information Supplementary Fig. 2 and Supplementary Fig. 3 show the modeled progression of the incidence rate and lifetime risk with only the increased lifespan, without gene therapy, the summary of which is presented in Table 1.

Scenario 2: Simulating all individual PRSs, lowered by an OR of 0.25 (fourfold OR decrease)

A 0.25 OR multiplier was applied to all individual PRSs, and the simulation ran from age zero to 120 years of age, Fig. 3a. In this scenario, the incidence rate stays below the baseline for all LODs analyzed throughout the lifespan. Lifetime risk is



(a) Per LOD number of edits for risk population

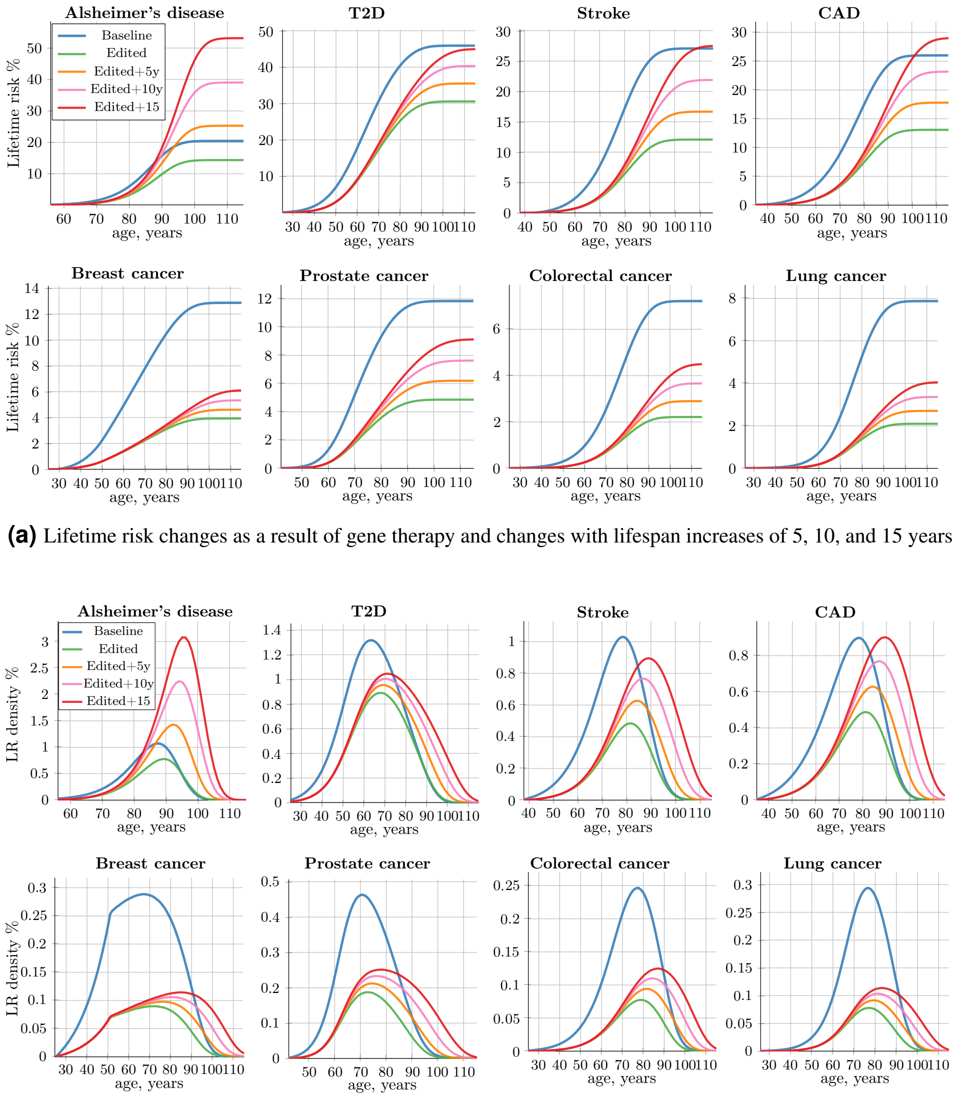
(b) Incidence rate pattern after gene therapy

(c) Lifetime risk decrease after gene therapy

(d) Lifetime risk, and lifespan increase 5,10 y

(e) Lifetime risk density Eq. 8 and lifespan increase

Fig. 2. Gene therapy Scenario 1: Treating all above average risk individuals
Population values and results with the incidence rate, lifetime risk, and lifetime risk with lifespan extension. See Supplementary Fig. 5, Supplementary Fig. 6, and Supplementary Fig. 7 for enlarged views.



(a) Lifetime risk changes as a result of gene therapy and changes with lifespan increases of 5, 10, and 15 years

(b) Lifetime risk density Eq. 8 with changes in lifespan increases of 5, 10, and 15 years

Fig. 3. Scenario 2: All individuals in a population have gene therapy corrected, on average, for 15 SNPs (corresponding to an OR of 0.25). Supplementary Fig. 4 shows the standard incidence rate plot.

also lower for all LODs and shows about a decade delay in the incidence rate curve for T2D, stroke, and CAD and even more significant improvement for cancers. AD again benefits the least.

As in the previous scenario, the AD lifetime risk exceeds the baseline within 5 years of the increased lifespan. It took 15 years of longer life for T2D, stroke, and CAD to get close to or slightly exceed the baseline lifetime risk. All cancers stayed far below the baseline lifetime risk even with 15 years longer life. The increase in the fraction of older individual diagnoses, leading to this lifetime risk increase with increased lifespan in Scenario 2, is presented in Fig. 3b.

Discussion

The aging coefficient was used to facilitate a hypothetical *Scenario 1*, where all individuals with an above average PRS were subjected to prophylactic gene therapy by setting their PRS to the average of the baseline population. The simulated aging of the population showed that, as expected, for all modeled LODs, the incidence rate at younger onset ages diminished significantly; see Table 1.

The incidence rate curve remains below the base incidence for all ages in the case of cancers. The four highly prevalent LODs display a spike in the incidence rate significantly above the baseline with maximums at about 95 years of age for AD and T2D and about 110 years of age for stroke and CAD. These are the ages when a small fraction of the population is still

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%

For our in-depth literature review and analysis of these LODs' incidence and other disease characteristics, see Supplementary Information in[18]. The baseline is considered 100% in lifetime risk comparisons; 160% after 5-year lifespan expansion for AD means lifetime risk (LR) increased 1.6 times, and 50% after gene therapy means 0.5 times the LR, comparing with the baseline value.

alive for the first two LODs, and practically the whole population is dead for the second two LODs. As a result, with mortality remaining constant, the lifetime risk—the chance of a person being diagnosed but not necessarily dying from an LOS—drops significantly, Table 1.

Modeling an increased lifespan due to all causes, we find that AD lifetime risk regains the level of baseline incidence within 5 years and almost doubles above the baseline with a 10-year life extension; see Table 1. T2D, stroke, CAD, and lung cancer also slightly exceed the baseline.

It is important to note that while the incidence rate and lifetime risk may increase with longer lives for the above-mentioned diseases, from the population perspective, the onset age is delayed by years and even decades.

The most prevalent cancers—breast, prostate and colorectal—showed a very significant decrease in lifetime risk, even when lifespan increased. Lung cancer requires the least edits to make the at-risk half of the population achieve the baseline average PRS in Scenario 1, but this is because low heritability implies a low genetic variance, which is reflected in the lower number of gene variants in the higher than average risk population. This naturally results in a lower initial 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.

We modeled life extension in the baseline scenario, without the prophylactic gene therapy, and it highlights the old wisdom that aging itself is the predominant risk factor for most late-onset diseases and conditions. It is a known fact that in Japan, the lifetime risk of AD is higher than in many other countries[45], and the average lifespan in Japan is also higher. Table 1 results confirm once more that, if mortality from all causes is lower, AD is the fastest LOD to increase in late age prevalence. Most LODs follow the exponential incidence rate increase pattern in initial onset decades[18], yet AD continues exponentially past age 95[44,46]. 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 that most of us are at risk of succumbing to if we live long enough. Similarly, Framingham General Cardiovascular Risk Score includes age as one of the major risk factors for stroke and CAD[38], 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[47]. As a result, the incremental lifetime risk is relatively lower in 5 and 10-year lifespan extensions, starting from an already high level. On the other hand, in the case of prophylactic gene therapy, the population health effect would also be lasting. Even though we analyze each LOD independently, T2D is one of the diseases that causes the most co-morbidities and accelerating incidences of cardiovascular and other diseases sometimes by decades[47]. It is worth noting that the preventative treatment of T2D could mean health improvement or presentation delay for a range of LODs, independently or in addition to treating their specific gene variants.

Scenario 2, decreasing LOD odds ratio for each individual fourfold, could be viewed as a model with more intensive gene editing. It corresponds to between 10 and 15 edits, explained in Methods, equal for all modeled LODs. This scenario can also be viewed as cumulative over a number of generations' ongoing lower scale gene therapy, where the average PRS drops due to the risk allele frequency gradually falling. Apparently, the lower heritability LODs like lung cancer benefit from this scenario more equitably, comparing with the Scenario 1 treatment. The noticeable result for the LODs with the lower cumulative incidence or prevalence is the longest lasting effect. We analyzed some of the most prevalent LODs, and there is a large number of LODs with a much lower prevalence that would similarly benefit. Based on familial studies, it may be not outside the range of possibility to find a sufficient number of genomic variants to achieve fourfold odds ratio reductions.

A question is often asked about the correct lifetime risk values for a particular LOD, and there may be a correct value in 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 this value may be different compared to other locales in 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 population at risk 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 the very old individuals. The outcome of this magnitude of gene therapy is comparable with improvements in health span that can be achieved by lifestyle improvements for T2D, stroke, and CAD - the highly environmentally influenced LODs[18].

In conclusion, if intensive gene therapy becomes possible, it could dramatically delay the average onset of the LODs analyzed here and improve individuals' lifetime risk. It would be difficult to prevent Alzheimer's disease if lifespan also increased. We can only hope that a pharmaceutical intervention targeting a causal metabolic pathway, immune or inflammatory response may be more effective for AD; however, the number of the past false hope announcements is too numerous to cite.

Based on heritability and incidence rate combinations, this kind of gene therapy may bring significant and longer lasting benefits for cancer prevention, even with a similar or smaller number of edited gene variants than for other diseases we modeled.

Methods

Foundations

The research goal for this paper is to build a genetic model based on clinical data of eight highly prevalent LODs. To do this, we need to map the PRSs to individual hazard ratios on a yearly basis:

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

where $R_u(age)$ is the hazard ratio of uth unaffected individual, and t is the age in years. PRS G_u remains constant for each individual, and the multiplier $A(t)$ drives the age-related increase in an LOD risk. $A(t)$ we call the *aging coefficient* and it will primarily depend on the incidence rate and the remaining unaffected individual's PRS. This means that, as discussed in the Introduction, we are following a proportionate—also known as a multiplicative—model of genetic architecture, just like in[18]; see also[26]. The simulation needs to find the $A(t)$ for each year of LOD progression resulting in decrease of unaffected individual count N due to being diagnosed as a function of their PRS: $N(t+1) = N(t)(1-I(t))$, where the yearly incidence rate $I(t)$ is the fraction of individuals diagnosed with a disease out of all the individuals as yet unaffected at the start of the year. Here we also will need to probabilistically simulate the individuals diagnosed each year of age based on their hazard ratio, in this approach as a product of changing $A(t)$ and changing distribution of G_u, as will be discussed in simulation design.

We reuse the yearly incidence rate logistic and exponential regression for the functional approximation of the LODs we analyze based on the available clinical incidence statistics from[18], and as also appropriate here, we use the logistic approximation for all LODs except breast cancer for which the exponential followed by linear regression is more accurate. The most important consideration, is that the age-related change in the allele frequency distribution for the unaffected population is a function of initial heritability and the cumulative incidence. The third variable, the population mortality from the US Social Security "Actuarial Life Table"[48], in conjunction with the above two determinants, produces the lifetime risk pattern characteristic for each of the LODs in Fig. 2c.

We used the genetic architecture proposed by[49] in order to generate the number of alleles to achieve a required heritability. Out of six genetic architectures in[49], the common low effect size genetic architecture model is most suitable for the LODs we analyze. All simulation runs 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—the common allele low effect size genetic architecture model—presented in reports and analysis here. We don't present separate figures for scenario B, the corresponding data is available in Supporting Information.

We calculated an individual polygenic risk score as a sum of all alleles' effect sizes, which is by definition a log(OR) (odds

Table 2. **Genetic architecture scenarios.**

Scenario	MAF range	OR range	MAF values	Allele ORs 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*, rare medium for *rare allele medium effect size architecture*.

ratio logarithm) for each allele,[49]:

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

where a_k is the number of risk alleles (0, 1, or 2), and OR_k is the OR of additional liability presented by the k-th allele. In our publication figures, for brevity, we also use the notation β for PRS, $\beta = log(OR)$. Variance of the allele distribution is determined by:

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

where p_k is the frequency of the k-th genotype[49].

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

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

where $\pi^2/3$ is the variance of the standard logistic distribution[50]. We use heritabilities of the LODs we analyze, as summarized in Table 1.

Implicit in our simulation is a Cox proportionate risk model[51], where the hazard ratio for an individual, in one-year increments t in our case, is represented as:

$$\lambda_u(t) = \lambda_0(t)e^{\sum_i^n b_i x_i}, \tag{5}$$

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. Chances for an individual not to be diagnosed with an LOD by age T is:

$$\prod_{t<T}(1-\lambda(t)), \tag{6}$$

which is an important statement of probability, as to remain healthy to age T is the product of probabilities of the same for each preceding year.

The values of polygenic scores for each individual in the simulated population are based on ORs built using the logit model[49]. For low prevalence within an age cohort, which is equivalent to the incidence rate for that time interval, and when the OR is close to 1, OR values are practically identical to hazard ratios. For example[52] does so in the case of breast cancer, considering breast cancer as a low-incidence LOD. While it is very likely the case for most cancers, the high incidence and high heritability values, particularly for AD and T2D, justify a more proactive approach in view of high average polygenic risk of cases at earlier ages for these diseases, and even higher risk outliers due to significant variance as a source of the high initial heritability[18]. We similarly use an adjustment formula, converting the OR to relative risk, subject to the known incidence, as in[53]:

$$H_u = \frac{G_u}{1+I(t)\cdot(G_u-1)}, \tag{7}$$

where H_u is an estimated hazard ratio for a polygenic score G_u of the uth unaffected individual. This formula is widely used in clinical practice and research as a good estimate conversion, and we apply it to all G_us in the simulation run. For verification, we also collect all simulation results without this adjustment. We find that, indeed, the results vary slightly in declining order of significance for AD, T2D, CAD, and stroke, and are practically identical for cancers with and without the adjustment. Qualitatively, the results would lead to the same conclusions.

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:

$$D(t) = \frac{I(t)S(t)}{N(0)}, \tag{8}$$

where $I(t)$ is the yearly incidence, $S(t)$ is the survivor rate from actuarial tables[48], 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 in our implementation, exactly corresponds to an LOD lifetime risk $L(t)$:

$$L(t) = \sum_t D(t), \tag{9}$$

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

Simulation design

The initialization, generation of population individuals, and genetic architecture are done identically to those of[18] and are similar to many population genetics simulations. Additionally, the yearly incidence approximation and the individual PRSs are done similarly.

The discovery of the aging coefficient is done using an iterative simulation process. The simulation operates in two stages: aging coefficient discovery and analysis of gene therapy with optionally changing life expectancy.

Aging coefficient A(t) discovery

The aging coefficient discovery proceeds iteratively through the modeled population individuals and years of age, starting at age zero.

This occurs by taking a first estimate run, starting with an initial value, and then running a set of 250 simulation loops over the population base of 25 million and aggregating the results (these settings are configurable, and these values allow a good balance between parallelization, accuracy, and discovery time on the computer system used). A large number of runs is particularly important at younger ages with low incidence rates and high variability in a number of diagnosed individuals between runs. The simulation uses an adaptive algorithm to achieve convergence on the value $A(t)$ within a target precision of 0.2% or better for most of the lifespan. To avoid or uncover any discovery bias, we repeat the determination for the same age, starting alternatively 10% up or down from the previously determined value, running 10 cycles of discovery like above (configurable), averaging the final result, and recording the standard deviation. This process continues through the predetermined life expectancy range for which we chose 120 years of age, close to the current utmost limit of the human lifespan. As seen in Fig. 2e, the population mortality makes it very insignificant for the final result whether the approximation accuracy becomes lower after 100 years of age, even with the modeled lifespan increase by 15 years. The resulting "roundtrip" accuracy is within two percent of the $A(t)$ magnitude at the initial onset ages, where the incidence rate is below $1 \cdot 10^{-5}$. This accuracy would be sufficient for the entire simulation and has an infinitesimal effect on overall accuracy because of the low cumulative incidence at ages with a low lifetime risk density, Eq. 8 and Eq. 9. The accuracy rapidly improves to the target better than a 0.2% range for most ages at higher incidence rates. Only in the case of AD, where incidence grows exponentially to very late age, the remaining unaffected population becomes small near 120 years of age, and accuracy decreases to close to 1%, a minuscule deviation with particularly low lifetime risk density at this age.

Gene therapy with changing life expectancy analysis

The analysis stage is run with a large population of one billion to achieve high precision statistics. Here we apply the discovered values of $A(t)$ to the PRS of this large population. Fig. 4 shows the resulting accuracy for eight LODs, starting near 2% for colorectal and lung cancers, below 1% for the remaining LODs, and staying 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 the desired precision.

The prophylactic gene therapy experiment assumes that we can effectively edit a large number of SNPs. This study is not making any pleiotropic consideration[54,55]. Just as a short remark, the high-risk individual PRS is caused by a large number of variants; see Supplementary Fig. 1. There is a relatively small difference in the absolute number of detrimental alleles between the population average and the high-risk individuals. Arguably, for personalized prophylactic treatment, there will be an option to choose a small fraction of variants without pleiotropy or even agonistic to other LODs pleiotropy from a large set of SNPs.

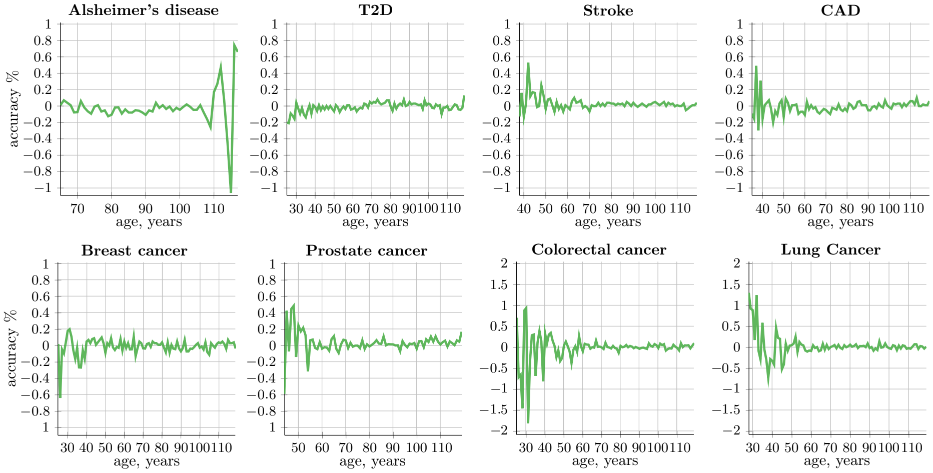


Fig. 4. Simulation aging coefficient discovery accuracy in reproducing the input incidence rate approximation.

We used the aging coefficient in two scenarios:

Scenario 1: Prophylactically treating above average risk individuals to match the polygenic risk average for the baseline population. Fig. 2a shows the average number of alleles that need editing for individuals at risk from population distribution, so their risk becomes equal to the baseline population mean at age zero. In these figures, we display the population frequency between $1\cdot10^{-4}$ on the low and high ends of the graph, without any rare outliers. The simulation then proceeds through the complete age range with this treated population, collecting and analyzing the resulting population statistics.

Scenario 2: Simulating all individuals' PRS lowered by an odds ratio 0.25 (fourfold). This scenario requires an edit of 15 SNPs in common allele low effect size scenario. With average OR=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 individuals at risk, scenario 2 may be viewed as a simulation of the effect of treating individuals at risk for a number of generations, where the number of risk alleles in the whole population becomes lower as a result of previous generations undergoing inheritable prophylactic treatment.

To model the lifespan increase, rather than making assumptions about the pattern of the lifespan increase age distribution, we assume that the mortality of all causes will be reduced proportionally. We shift all mortality rates from the US Social Security "Actuarial Life Table"[48] by 5, 10, and 15 years. This is valid because we are not concerned by early age disease analysis, focusing exclusively on LODs, with onset starting decades after birth.

Data sources, programming and equipment

We used population mortality numbers from the 2014 US Social Security "Actuarial Life Table"[48], 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 used extensively for our analysis and using the materials referenced in the Supplementary Information to[18]: Alzheimer's disease[44,46,56,57], type 2 diabetes[47], coronary artery disease and cerebral stroke[58], and cancers[59,60].

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 code can be found in Supplementary Files.

The final simulation data, additional plots and elucidation, source code, and the executables 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 a corresponding execution slowdown.

Statistical analysis

The core model parameter, the aging coefficient, is validated to fit within a two sigma band for 2% accuracy at low incidence values and within 0.2% for the majority of the LOD incidence range. The same precision values are achieved for the analysis stage of the simulation. This precision makes impractical the use of error bars in graphical displays.

Acknowledgments

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

Competing interests

The author declares that there are no competing interests.

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. 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`.
19. Steck, A. K. & Rewers, M. J. Genetics of type 1 diabetes. *Clin. chemistry* **57**, 176–185 (2011).
20. 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).
21. Jager, R. D., Mieler, W. F. & Miller, J. W. Age-related macular degeneration. *New Engl. J. Medicine* **358**, 2606–2617 (2008).
22. 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).

23. 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).

24. 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).

25. 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).

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

27. 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).

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

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

30. 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).

31. 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).

32. 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).

33. 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).

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

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

36. 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).

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

38. 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).

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

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

41. 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).

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

43. 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).

44. 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).

45. 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).

46. 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).

47. 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).

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

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

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

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

53. 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).

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

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

56. 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).

57. 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).

58. 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).

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

60. 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).

1 Supplementary Files

Source code, executable, scripts and configurations. This zip file contains the executable simulation, the source code and the project, R scripts and corresponding batch files for producing functional approximations of clinical incidence, and simulation results.

SUPPLEMENTARY FIGURES

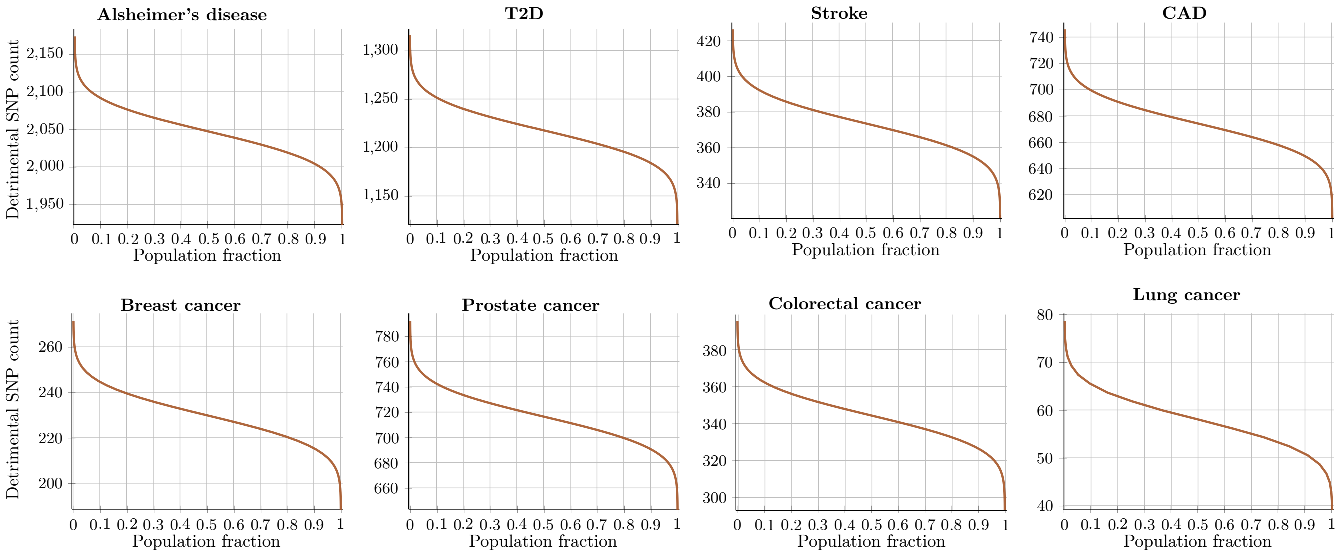


Fig. 5. Supplementary Fig. 1

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

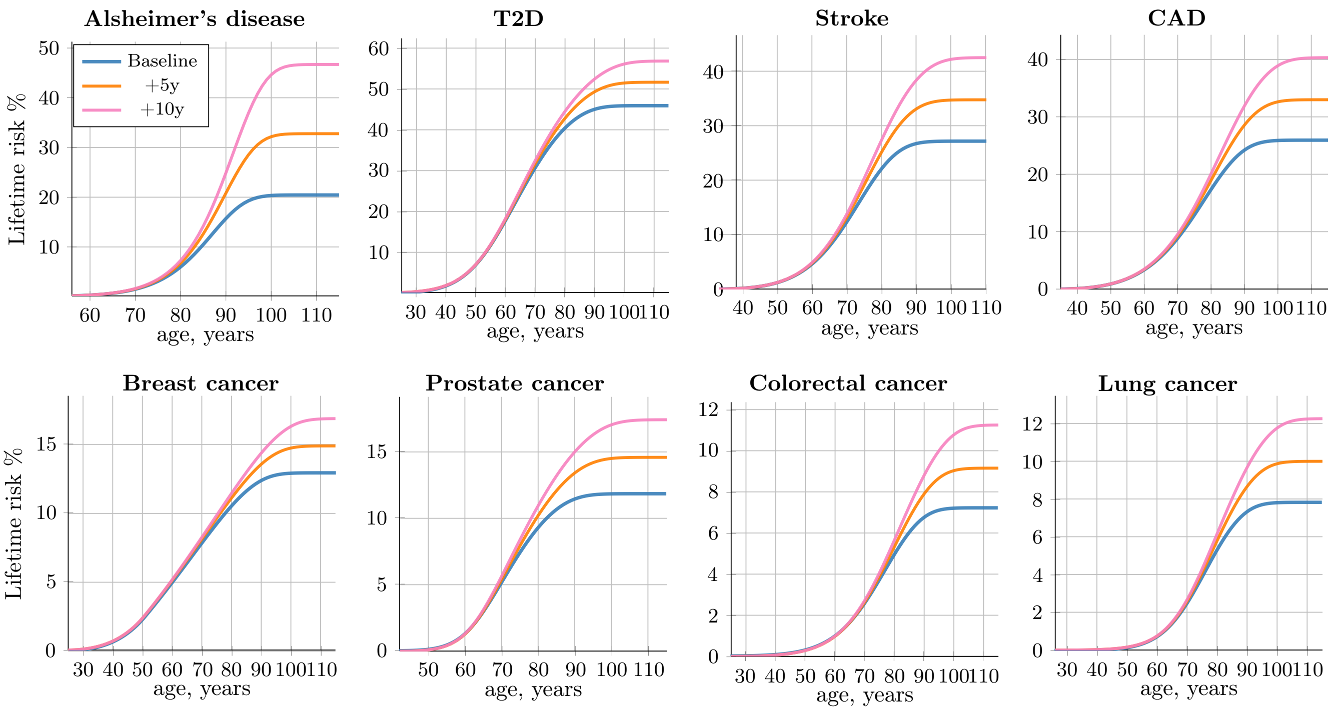


Fig. 6. Supplementary Fig. 2

Projected LOD lifetime risk increase if lifespan 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 where lifespan increases due to other causes than the plotted LOD. Mortality patterned from US Social Security "Actuarial Life Table"[48]

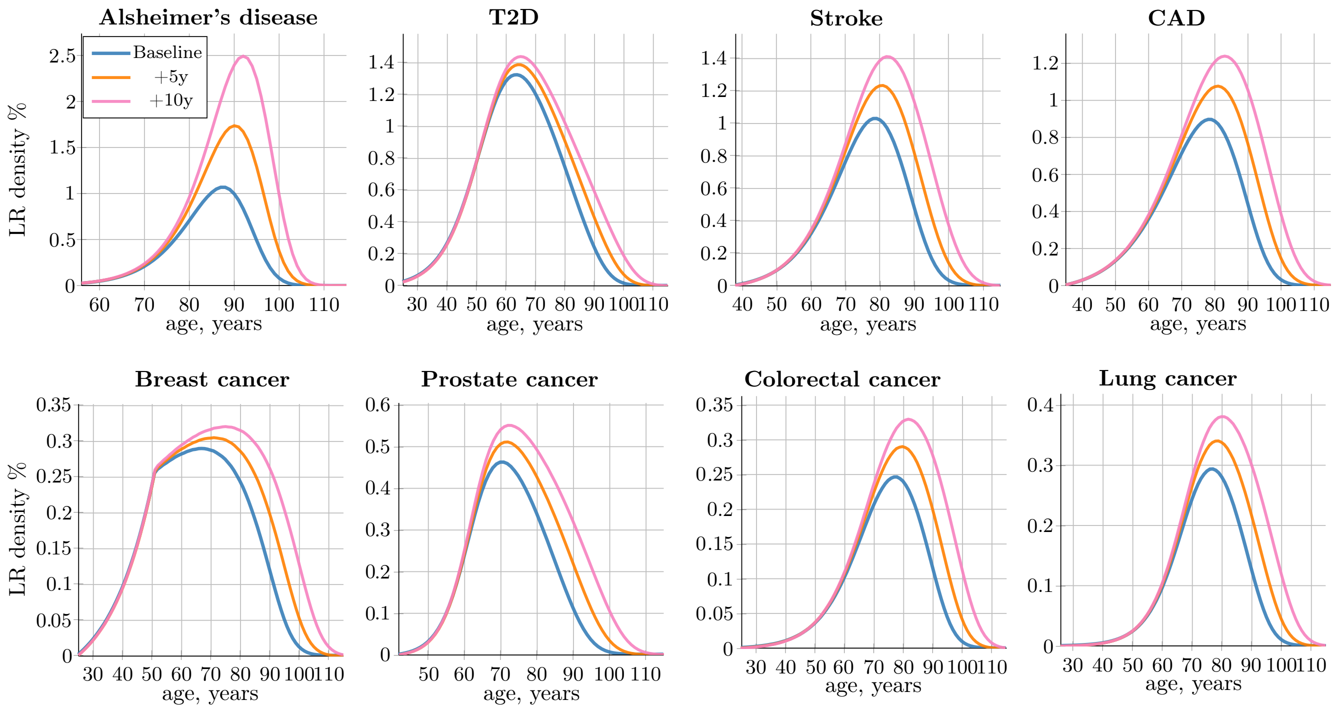


Fig. 7. Supplementary Fig. 3

Projected LOD incidence rate relative to the number of individuals at birth if the lifespan 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[48]. 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 other causes than the plotted LOD.

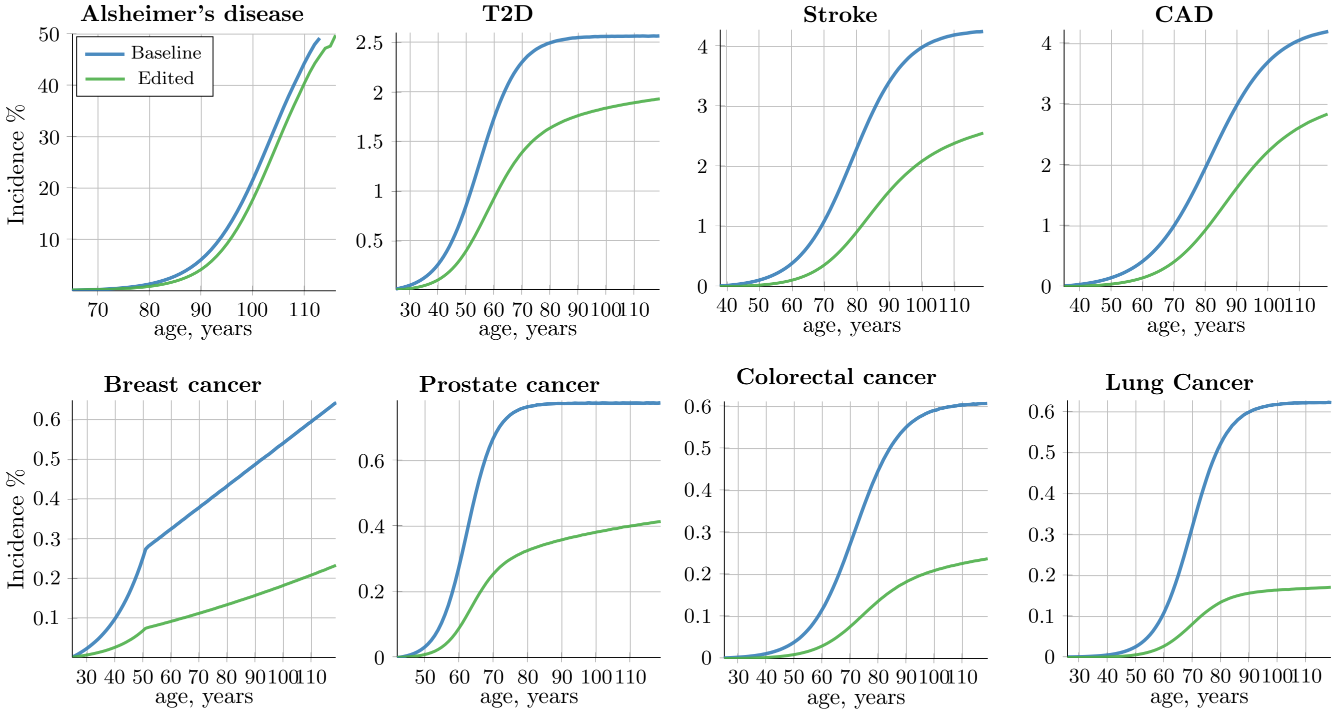
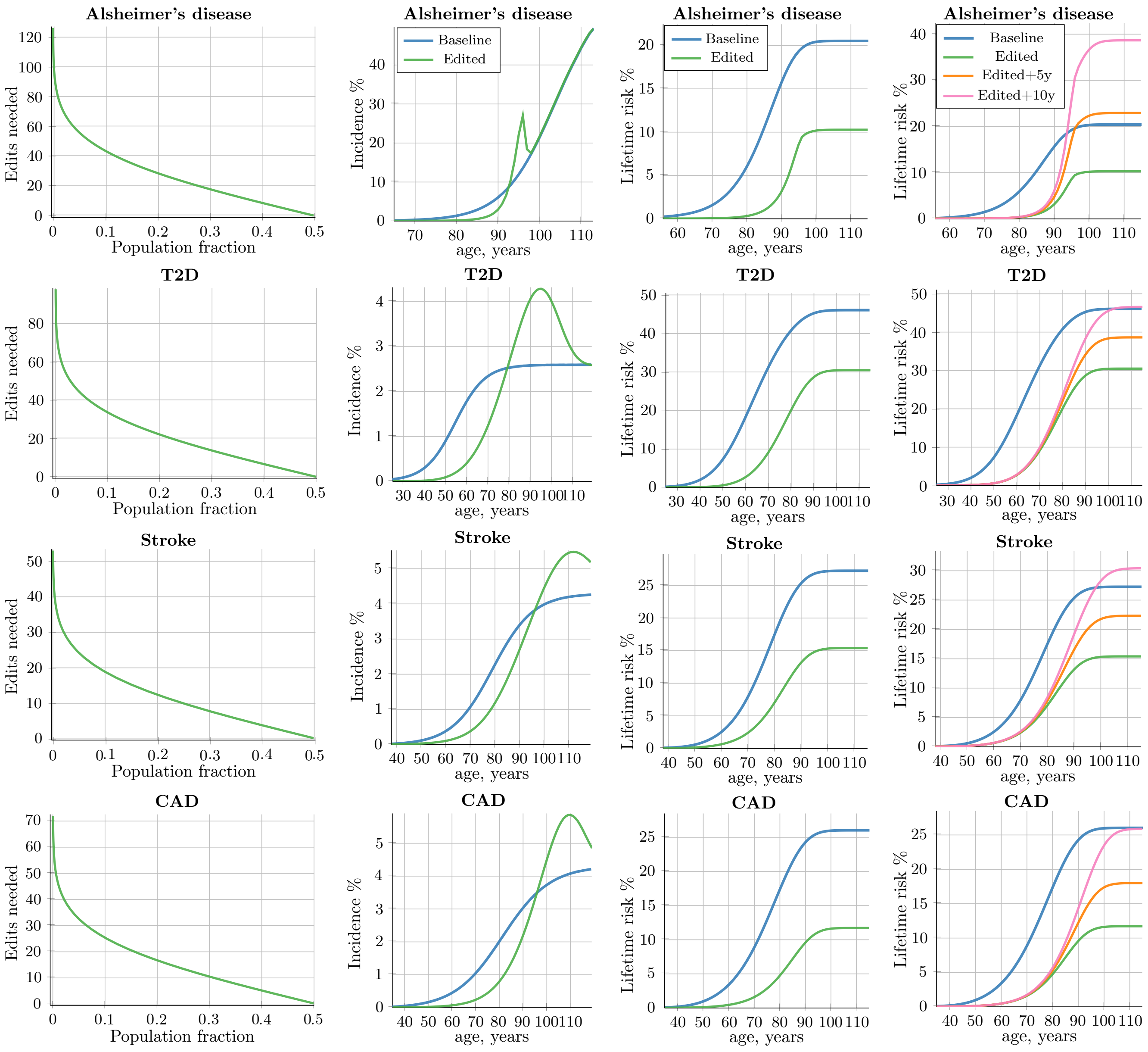


Fig. 8. Supplementary Fig. 4

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) Per LOD number of edits for risk population

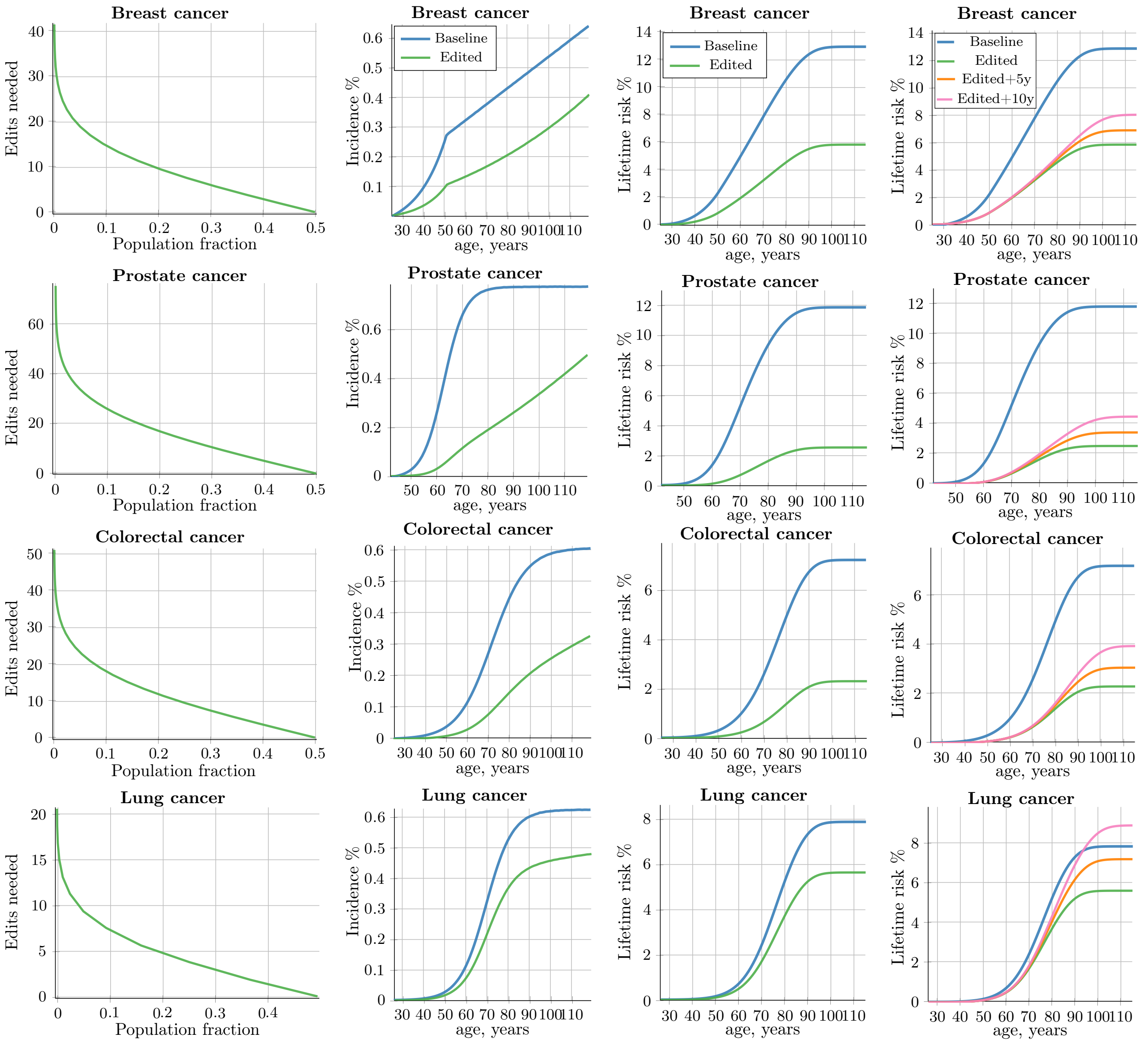
(b) Incidence rate decrease from baseline after gene therapy

(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 decrease 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.

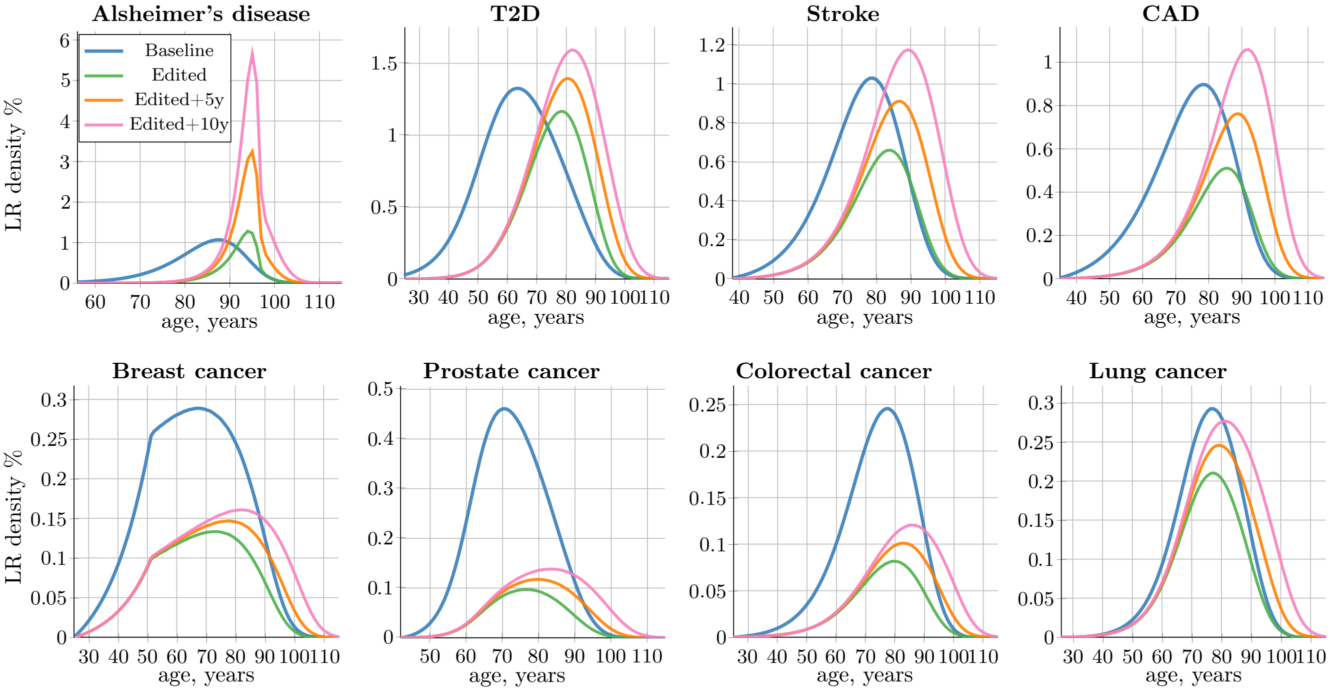


Fig. 11. Supplementary Fig. 7

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 extension. This plot has a meaning of 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.