



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Human Lifespan: Recent Trends and Genetic Determinants

Paul Reimund Hugo Jan Timmers, B.Sc., M.Res.


Doctor of Philosophy with Integrated Study

The University of Edinburgh

2020

Declaration

I declare that this thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work presented here is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below and in Chapters 2–4. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.



.....

Paul Timmers, 15 March 2019

Assistance

My supervisors Peter Joshi and Jim Wilson commented on drafts of the Chapters in this thesis, highlighting areas which needed rewording, needed to be expanded, or needed to be cut down, as well as suggesting specific references and recommending small tweaks to the tables and figures.

The work presented in Chapter 2 was approved for publication in *BMJ Open* as ‘*Trends in disease incidence and survival and their effect on mortality in Scotland: nationwide cohort study of linked hospital admission and death records 2001–2016*’, by **Paul RHJ Timmers**, Joannes J Kerssens, Jon W Minton, Ian Grant, James F Wilson (Supervisor), Harry Campbell, Colin M Fischbacher, and Peter K Joshi (Supervisor). Details of each author’s contributions are listed in Chapter 2.

The work presented in Chapter 3 was published in *eLife* as '*Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances*', by **Paul RHJ Timmers**, Ninon Mounier, Kristi Lall, Krista Fischer, Zheng Ning, Xiao Feng, Andrew D Bretherick, David W Clark, eQTLGen Consortium, Xia Shen, Tõnu Esko, Zoltán Kutalik, James F Wilson (Supervisor), and Peter K Joshi (Supervisor). Details of each author's contributions are listed in Chapter 3.

The work presented in Chapter 4 was submitted for publication as '*Multivariate genomic scan of human ageing traits reveals novel loci and identifies haem metabolism as a human ageing pathway*' by **Paul RHJ Timmers**, James F Wilson (Supervisor), Peter K Joshi (Supervisor), and Joris Deelen. Details of each author's contributions are listed in Chapter 4.

Abstract

Human lifespan is determined by a complex interplay of genetics, environment, lifestyle and chance. In the UK, life expectancy has increased by roughly three years every decade, but despite longer lives, individuals also spend more years living with chronic disease. With populations greying and periods of morbidity becoming more prolonged, the burden of ageing and age-related disease is set to become a major healthcare challenge. Understanding the factors underlying trends in human lifespan could guide policy interventions to mitigate the burden of disease, while an understanding of the genetics of lifespan could provide insight into the ageing process. The latter could in turn reveal potential therapeutic targets to delay age-related disease and inform which individuals to target based on their genetic risk.

In this thesis, I explore human lifespan from these two perspectives. First, I examined trends in mortality and morbidity in two million Scots using hospital admission and death records and found recent improvements in lifespan could be largely explained by improvements in the incidence and survival after hospitalisation of cancers and heart disease. However, I also found recent deteriorations in infectious disease, especially for individuals from lower socioeconomic classes, suggesting a need for a renewed public health focus in this area. Next, I performed a genome-wide association study (GWAS) to find genetic determinants of lifespan using DNA from 27 European cohorts and the lifespans of their parents (one million total). I identified 12 genomic regions affecting survival and found genetic variants across the genome, when aggregated into polygenic scores, could distinguish up to five years of survival between score deciles. Combining the lifespan GWAS with two other GWAS of lifespan-related traits, I identified 78 genes—some of which delay ageing in model organisms—which putatively influence both human lifespan and healthy years of life and which are enriched for haem metabolism. These findings present the most promising targets for therapeutic interventions to date, which may help delay the onset of age-related disease and extend the healthy years of life for all.

Lay Summary

People around the world are living longer, but for many older individuals the end of life is still marked by a long period of illness. With both the number of people over 65 and the number of years they spend in ill health on the rise, the healthcare system is put under increasing stress. As such, there is a pressing need to understand what is driving the changes in life expectancy and what can be done to extend the healthy years of life for everyone.

Lifespan is complex and can be influenced by many factors. Some of these factors change over time, like healthcare and public policy, while others don't change over time but do differ between people, like DNA. My work was focused on understanding lifespan from these two angles. I first studied the hospital and death records of all adults in Scotland and found that the greatest number of lives were saved because the number of new hospital admissions for heart conditions and cancers went down, as well as the likelihood of dying after being hospitalised (especially for cancers). In contrast, the largest increase in deaths was because of a rise in the number of new hospital admissions for infections and a rise in the likelihood of dying afterwards, especially for people from more deprived areas. These trends suggest public healthcare should renew their focus on combatting infectious disease.

I next studied the DNA of 500,000 Europeans and asked them about their parents: how old they were and whether they were still alive. Using this information, I found 12 regions in their DNA that often varied between individuals with shorter- and longer-lived parents. I then made a survival score for each individual, summing up the effects of all variations in the genome, and found that when we divided participants into ten groups based on this score, the top group lived an average of 5 years longer than the bottom. Lastly, combining this study with similar ones, I linked 78 genes to both lifespan and healthy years of life, with many of them being involved in processing iron. Some of the same genes have already been shown to extend lifespan in worms and mice, making them promising targets for development of drugs that could delay age-related disease and make people live healthier for longer.

Acknowledgements

I would like to thank

David Clark, Andrew Bretherick and Peter Joshi for building the GWAS pipeline, which greatly sped up the study of parent lifespans in UK Biobank, and will undoubtedly continue to contribute to discoveries in the future;

The authors who wrote the computing tools I used, including Tom Haller (RegScan), Futao Zhang (SMR-HEIDI), Alfred Aho, Peter Weinberger, and Brian Kernighan (AWK), the R project team (R language), Terry Therneau (R survival package), the R data.table team (R data.table package), and countless others;

The UK Biobank, LifeGen, and Longevity cohorts, including their investigators, analysts, and of course participants, whose deaths or parents' deaths made this work possible, and in doing so may improve the lives of the future generations to come;

The members of the Estonian Genome Centre and the Lausanne Statistical Genetics Group for contributing ideas and analyses, and being such a joy to collaborate with;

The members of the Wilson Group for making me feel a valued member of the team and making the office an enjoyable and exciting place to work;

The Medical Research Council and the University of Edinburgh for funding my scholarship;

My supervisor Jim Wilson for granting me access to their data, facilitating opportunities to collaborate, supporting my academic goals, and providing comments and feedback every step of the way; and

My supervisor Peter Joshi for his mentorship and enthusiastic collaboration, his persistence in driving my personal and professional skill development, and his unending effort to provide me with opportunities to achieve academic success.

You are without a doubt the best PhD supervisors and colleagues I could ask for.

Contents

Declaration	iii
Assistance	iii
Abstract	v
Lay Summary	vii
Acknowledgements	ix
Contents	x
Tables and Figures	xi
Chapter 1: Introduction	1
1.1 Ageing and age-related disease are an increasing burden to society	1
1.2 How to study ageing	4
1.2.1 Measuring a life	4
1.2.2 The genetic component of lifespan	7
1.2.3 Genome-wide association of lifespan.....	13
1.2.4 Datasets large enough to study lifespan.....	19
1.3 How precision medicine and therapeutic discovery will affect ageing	22
1.4 Conclusions	26
Chapter 2: Trends in disease incidence and survival and their effect on mortality in Scotland	27
2.1 Introduction	27
2.1.1 Context.....	27
2.1.2 Contributions	28
2.2 Manuscript accepted for publication	29
2.3 Conclusion	62
Chapter 3: Genome-wide association of lifespan in UK Biobank and LifeGen	67
3.1 Introduction	67
3.1.1 Context.....	67
3.1.2 Contributions	68
3.2 Published article	70
3.3 Conclusion	110
Chapter 4: Genome-wide multivariate association of healthspan, lifespan, and longevity	113
4.1 Introduction	113
4.1.1 Context.....	113
4.1.2 Contributions	114
4.2 Manuscript submitted to journal	115
4.3 Conclusion	136
Chapter 5: Discussion	139
5.1 Trends in disease and their effect on mortality	139
5.2 The role of genetic factors on human lifespan	142
5.3 Insights into the ageing process	147
5.4 Sex and socioeconomic determinants of lifespan	150
5.5 Future work	152
5.5.1 <i>Lifespan or longevity?</i>	152
5.5.2 Trans-ethnic studies of survival	155
5.5.3 Rare and recessive variants	156
5.5.4 Biomarkers of lifespan	158
5.7 Conclusion	160
Bibliography	163
Appendix	185

Tables and Figures

Table 1: Lifespan heritability estimates	9
Table 2: Characteristics of population cohorts used in this body of work.....	21
Table 3: Genomic regions robustly associated with human survival	143
Figure 1: The past, present, and projected demographics of the UK population shows the proportion of elderly individuals is growing rapidly.....	2
Figure 2: UK individuals aged 65 and above may spend more years living with severe disabilities in the future	3
Figure 3: Parent-offspring regression in Scotland estimates lifespan heritability in Scotland to be <17%.....	10
Figure 4: The growing list of studies and associations catalogued in the GWAS catalog	16
Figure 5: Breast cancer risk as a function of age and polygenic risk score percentage	25

Chapter 1: Introduction

1.1 Ageing and age-related disease are an increasing burden to society

The populations of the United Kingdom and many developed regions are greying. Fertility and mortality rates are declining while life expectancy is on the rise[1,2]. The result is an increasing number of elderly people working longer with pensions being supported by fewer working adults[3], mounting pressure on healthcare services that struggle to cope with the chronic diseases of old age[4], and a pressing need for ways to improve the health and well-being of the elderly[5].

These demographic shifts can be illustrated by nationwide statistics across the UK, a country which has seen the average births per woman decline from 2.2 only 50 years ago to 1.8 today, while the median population age increased from 35.1 years to 40.5 years. In the same period, life expectancy at birth rose by more than 10 years, with newborns in 1960 expected to survive on average until 70.6 years and newborns in 2020 expected to live on average until 81.8 [2]. In line with these trends, the Office for National Statistics projects that within the next 50 years, the number of people in the UK who are aged 65 years and above will increase by 8.6 million individuals, and make up 26.5% of the total population (up from 18.0% currently)[6] ([Figure 1](#)).

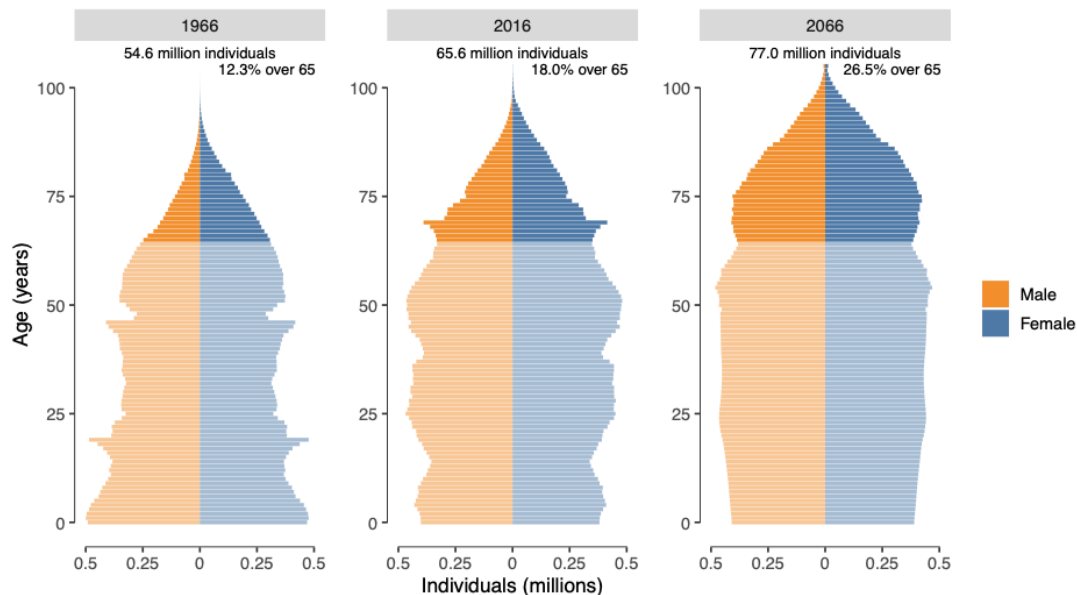


Figure 1: The past, present, and projected demographics of the UK population shows the proportion of elderly individuals is growing rapidly. Darker bars show individuals who are 65 years and older. Demographics for the year 2066 are projected assuming current trends in migration and mortality continue. Data from the UK Office for National Statistics[6].

While these improvements in life expectancy are clearly a positive trend, living longer does not necessarily imply enjoying an active and healthy life, free of disease. Individuals over 65 years of age can expect to live another 10 to 11 years on average, but they will spend only just over half of these years in good health, both in terms of self-rated well-being and disability-free life[7]. The other half of their remaining years of life is marked by chronic disease and functional impairment, requiring long periods of healthcare and disease management. A Lancet study by Guzman-Castillo *et al.* (which uses a more severe definition of disability) finds increases in disability-free life may be lagging behind increases in life expectancy[8]. If current trends continue, elderly individuals could end up spending more years living with severe disease than the elderly of the previous decade ([Figure 2](#)).



Figure 2. UK individuals aged 65 and above may spend more years living with severe disabilities in the future. Disabilities here include cardiovascular disorders, cognitive impairment, and moderate-to-severe loss of function. Depicted is the remaining life expectancy for individuals aged 65 in 2015 vs. 2025 (projected), with the number of years spent living with severe disability on the left and without severe disability on the right. Error bars represent 95% confidence intervals. Data from Table 4 from Guzman-Castillo *et al.* [8].

Notwithstanding the personal cost of extended periods of chronic disease, it is likely the greying population will exacerbate the pressures on the UK medical system, which already struggles to meet waiting time quotas and provide access to treatment[9]. Overall healthcare requirements and costs increase gradually with age, but after age 65 these costs start to increase exponentially. For example, 2013 data from the National Health Service on the costs of general and acute care shows the service spends an average of £278 per person per year on adult patients aged 25 to 45, and £568 on those aged 45 to 65. These costs increase to £1,362 and £2,179 per patient per year for those aged 65 to 85, and 85 and above, respectively[6].

Thus, greying populations and their associated increase in the burden of age-related disease are set to become one of the largest health challenges in the UK

and beyond. However, healthcare systems around the world are currently set up to treat individual diseases and manage age-related morbidities for as long as possible[10], with no clear holistic way to handle an increasingly elderly population. Similarly, much research funding has been allocated to the treatment of one age-related disease at a time rather than to the study and prevention of many[11,12]. It is clear a new approach is needed to understand the complex landscape of age-related diseases on a population level and identify and (pre-emptively) treat the individuals at most risk of morbidity and mortality. At the same time, improvements in age-related disease will not accelerate without advances in our understanding of the ageing process itself.

1.2 How to study ageing

1.2.1 Measuring a life

Human lifespan is defined as the length of time between birth and death. On a population level, this can be measured in terms of the average life expectancy at birth, and is calculated in two main ways: cohort and period life expectancy.

Cohort life expectancy is simply the average lifespan of the cohort, which is known exactly when the last individual of the cohort has died. It can also be estimated when some individuals are still alive by predicting future mortality rates of these remaining individuals. More specifically, cohort life expectancy is calculated as the sum of the probability an individual will be alive every year, based on that year's age-specific mortality (known or predicted)[13]. For example, the cohort life expectancy at birth in 1950 is the sum of the survival probability of a newborn in 1950, a 1-year-old in 1951, a 2-year-old in 1952, etc. Cohort life expectancy is the most accurate estimate of life expectancy for an age group, but it can have wide confidence margins when calculating life expectancy for more recent cohorts, as there is a high degree of uncertainty in projected mortality rates.

Conversely, period life expectancy refers to remaining life expectancy based on observed mortality rates of the current period, which is usually calculated from one or a small number of consecutive years. In contrast to cohort life expectancy, period life expectancy assumes future mortality rates are the same as the mortality rates from the period under study[13]. Thus, the remaining life expectancy ‘LE’ for individuals currently aged ‘n’ can be written as:

$$LE_n = \sum_{i=n}^{105} (1 - \lambda(i)) \quad (1)$$

Where $\lambda(i)$ is the yearly mortality rate of individuals aged ‘i’. These age-specific mortality rates are calculated from the observed deaths in the population, usually captured through nationwide death records gathered by governmental organisations such as the National Records of Scotland[14] and the UK Office for National Statistics[15]. It is known period life expectancies tend to underestimate individuals’ lifespan—as mortality rates decrease over time[13]—but they can be calculated without relying on projections and the life expectancy estimates they provide (life tables) can be easily compared between countries and time periods.

While population life expectancy estimates are useful to draw comparisons between countries and socioeconomic groups, they provide limited information about the underlying factors that determine the quality and length of life. On an individual level, lifespan can be influenced by a complex interplay of genetics, lifestyle, environmental exposures and pure chance. Humans, like most organisms, experience a progressive decline in the ability to maintain and repair their tissues after reaching adulthood, leading to a loss of function. This decline eventually manifests itself as age-related disease and ultimately results in death[16].

The dynamics of individual mortality have been described mathematically as early as the 19th century[17] and can be thought of as a combination of two types of hazards: an age-independent component (e.g. mortality due to external causes) and a component which increases exponentially with age (e.g. mortality due to

accumulation of molecular damage). Nowadays, there are a variety of models in use when studying survival, differing by their assumptions regarding the shape of the hazard function. On the one hand is the Kaplan-Meier model, which does not make any assumptions about the hazard but is limited to categorical covariates. On the other hand are Gompertz and Weibull models, which easily incorporate quantitative covariates but completely specify the shape of the hazard and are therefore susceptible to model misspecification[18]. A middle ground is the Cox Proportional Hazards model, which allows for multiple predictors without making any assumptions about the baseline hazard:

$$\lambda(t) = \lambda_0(t)e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n} \quad (2)$$

Here, $\lambda(t)$ is the hazard function: the probability of an individual dying at time 't', given they have been alive until time 't'. This function is dependent on two components, the baseline hazard $\lambda_0(t)$ at time 't', and an exponential component, which is modified by a number of covariates X_1 - X_n varying by individual, with their associated effects β_1 - β_n . The Cox model only examines the ratio of hazards between individuals, allowing the baseline hazard $\lambda_0(t)$ to cancel out (and remain unspecified), under the assumption that the effects of covariates on this hazard are proportional (i.e. do not vary over time).

The Cox proportional hazards model can be applied to any time-to-event data, making it a useful statistical tool to study both the incidence of disease and the incidence of death. Moreover, the variety of covariates the model can accommodate allows us to estimate the effect of a disease on mortality while accounting for differences in sex and socioeconomic status, or analogously, the effect of genetic variants on lifespan while correcting for covariates and population structure.

1.2.2 The genetic component of lifespan

Identifying genes which extend lifespan has the potential to reveal biological processes underlying the ageing process. Research into the genetics of lifespan started in earnest in the late 20th century with studies on model organisms rather than humans. In 1983, a set of long-lived *Caenorhabditis elegans* nematode worms were discovered after exposing them to the DNA-mutating compound ethylmethanesulfonate, demonstrating for the first time that certain genetic mutations could extend survival[19]. Since then, numerous researchers have studied *C. elegans* in an attempt to find longevity genes. Five years after the discovery of the first long-lived mutants, Friedman and Johnson[20] mapped the mutations in these worms to a region on chromosome 2 called *age-1*, which was later discovered to be homologous to the phosphoinositide 3-kinase catalytic subunits in mammals[21]. Another five years after the discovery of *age-1*, researchers found mutations in the *daf-2* region which could double nematode lifespan[22], as long as the activity of another region called *daf-16* was not disrupted. Homologues of these regions were later identified in humans as the *IGFR1* (insulin-like growth factor 1 receptor) gene and *FOXO* (Forkhead box O) transcription factor family[23,24], and highlighted the importance of the insulin signalling pathway in nematode longevity.

Similar mutation experiments were performed in a variety of organisms, including yeast[25], flies[26], and mice[27]. These revealed the longevity genes and pathways discovered in *C. elegans* were also important in regulating the survival of other eukaryotes. In 2013, a landmark paper put forward nine interconnected hallmarks of ageing, which summarised the common ageing pathways discovered in model organisms[16]. These hallmarks include three directly related to genetic information and gene expression: genomic instability, involving damage and repair of the nuclear and mitochondrial genomes; telomere attrition, the loss of the protective ends of chromosomes and subsequent cell death or cellular senescence; and epigenetic alterations, involving the redistribution of chromatin and methylation which in turn alters gene expression[16]. The remaining six hallmarks describe the age-related decline of

larger organelles—such as protein homeostasis machinery and mitochondria—and the decline of cells themselves. Thus far, advances in our understanding of the determinants of lifespan have been limited to model organisms, and whether these findings will translate well to humans remains unanswered.

The study of human genetics is different from model organisms in that experimental perturbations of genes are unethical and impractical, and researchers can therefore only study the effects of genetic variation occurring naturally within the population. It has long been observed that characteristics of long life and prolonged good health run in families[28,29]. However, the sharing of these traits within families could be due to similarities in lifestyle and environment, not just shared genetic factors.

More generally, heritability (h^2) is used as a measure to capture the relationship between genetic factors and variation in a trait. It estimates the amount of variation in a phenotype (V_P) that can be attributed to the additive effects of genetic variation (V_A), rather than the environment or chance:

$$h^2 = V_A/V_P \quad (3)$$

One approach to determine the heritability of a trait is to compare the correlation in the trait between sets of identical and non-identical twins[30]. When assuming no dominance or epistatic effects, this can be calculated as:

$$V_A = 2(\text{COV}_{MZ} - \text{COV}_{DZ}) \quad (4)$$

Where COV_{MZ} is the covariance between the trait for monozygotic twins, and COV_{DZ} the covariance between dizygotic (DZ) twins. This equation estimates the additive genetic variance under the assumptions that the shared environments between MZ and DZ twin pairs are the same, that the genetic similarity between DZ twins is on average 50% (which is not the case under assortative mating or inbreeding), and that there are no gene-environment interactions. In practice, the

heritability is often modelled using structural equations allowing for dominance effects and covariates to be taken into account.

For lifespan, studies on Danish and Swedish twins estimated the heritability of age at death to be upwards of 25% ([Table 1](#))[31,32], and these estimates have since been cited in hundreds of studies on human survival. However, there is debate whether twin studies systematically overestimate the amount of heritability, especially for traits relating to behaviour, as the environment shared between MZ twins may be more similar than that of DZ twins[33–35].

Study	Year	h^2	Data	N
Herskind <i>et al.</i> [31]	1996	26%	Danish twins	2,872 pairs
Skytthe <i>et al.</i> [32]	2003	27%	European twins	4,667 pairs
Mitchell <i>et al.</i> [36]	2001	25%	Amish community	1,655
Gögele <i>et al.</i> [37]	2011	15%	Alpine community	8,277
Gavrilova <i>et al.</i> [38]	1998	18%	European royal families	12,150 offspring
Joshi[39]	2015	<17%	Scottish families	4,642 offspring
Kaplanis <i>et al.</i> [40]	2018	15-16%	<i>Geni</i> online pedigree	~3,000,000
Ruby <i>et al.</i> [41]	2018	<10%	<i>Ancestry</i> online pedigree	~439,000,000

Table 1. Lifespan heritability estimates from twin, cohort, and online genealogy studies

As an alternative to twin studies, parent-offspring sets within population cohorts have also been used to estimate heritability. One study on the relationship between the age at death of Scottish individuals and their parents found each 10-year increase in the lifespan of one parent associated with around a 10-month increase in offspring lifespan, indicating the heritability of lifespan in Scotland was around 17% (95% CI 13%–21%). While this estimate is lower than twin studies, the same study also found correlations in age at death between spouses, suggesting the heritability estimate from parent-offspring duos could still be inflated due to shared environment factors and/or assortative mating and should be regarded as an upper limit[39] ([Figure 3](#)).

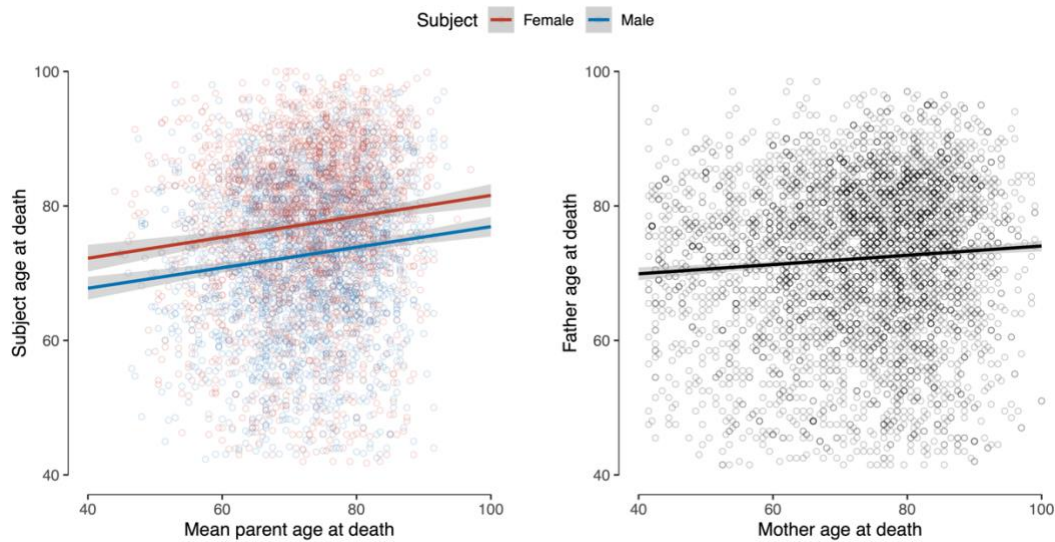


Figure 3. Parent-offspring regression in Scotland estimates lifespan heritability in Scotland to be <17%. The left graph shows the age at death of the subject (stratified by sex) as a function of the mean age at death of their parents. The right graph shows the regression of father age at death against mother age at death. Data from ORCADES and EASTER cohorts ($N = 4,642$) from Joshi[39].

Similar to parent-offspring research, there have been population studies which have attempted to estimate the heritability of lifespan using more distantly related individuals using the correlations between individuals as a function of their relatedness. Here, a genetic relationship matrix (calculated based on pedigrees or SNPs across the genome) is fitted as a random effect within a linear mixed model:

$$\mathbf{y} = \mu + \beta\mathbf{X} + \mathbf{g} + \epsilon \quad (5)$$

$$\mathbf{g} \sim \mathcal{MVN}(0, \mathbf{KV}_A)$$

Where \mathbf{y} is an $(n \times 1)$ vector of trait values for n individuals; μ is the intercept; \mathbf{X} is an $(n \times p)$ matrix of p covariates with the associated vector β of $(p \times 1)$ fixed effects; \mathbf{g} is a vector of Gaussian random effects, drawn from a multivariate normal distribution with mean zero and variance \mathbf{KV}_A ; ϵ is an $(n \times 1)$ vector of residuals; \mathbf{K} is the $(n \times n)$ genomic relationship matrix; and V_A is the additive genetic

variance. KV_A is thus the covariance matrix between individuals' genetic effects. Maximising the likelihood of this function provides an estimate of the additive genetic variance, which can be used together with the phenotypic variance to calculate the heritability ([Equation 3](#)). Population cohort estimates of the heritability of lifespan are in the range of 15–25%[36,38,42].

During the course of this PhD, the question of lifespan heritability was revisited by other authors. The popularisation of online genealogy websites led to the construction of family trees of millions of individuals, providing huge resources which allowed for more accurate inference of the genetic component to lifespan. First, a 2018 study of 3 million genealogy profiles estimated the additive heritability of lifespan to be 15-16%[40], although they assumed limited effects of shared family environment, despite substantial correlations in lifespan between spouses. More recently still, Ruby *et al.*[41] analysed their own genealogy pedigree of 439 million profiles using structural equation models which took into account non-genetic inherited factors, such as assortative mating, cultural influences, and socioeconomic status. Their model suggests less than 10% of lifespan is determined by genetics and this more modest estimate appears to be the most robust available to date.

The discrepancy in heritability estimates between Ruby *et al.*[41] (less than 10%) and the next largest study—Kaplanis *et al.*[40]—(around 15%) can be explained by the latter taking into account only inbreeding and shared family environment, which do not fully account for the effects of assortative mating. Indeed, Ruby *et al.*[41] show correlations in lifespan can be seen among siblings-in-law and even cousins-in-law, suggesting there are traits which influence lifespan which are commonly selected for in partners. For example, socioeconomic factors such as wealth and education, both subject to strong assortative mating[43] and linked to longer life[44], could inflate lifespan heritability estimates if their inheritance is considered to be completely genetic. Given assortative mating on such sociocultural factors is common across populations[45], heritability estimates

derived from the smaller population-based studies—none of which fully account for this source of phenotypic correlation—are likely inflated as well.

In contrast, twin studies rely on the comparison of monozygotic and dizygotic twins, which should have inherited largely identical sociocultural factors and are therefore expected to be less susceptible to inflation due to assortative mating. However, Ruby *et al.*[41] argue these studies generally recruit small numbers of individuals from limited geographical areas, and therefore could lack much of the variation found in large-scale population studies. Lower environmental variation reduces total variation in lifespan, and as a result increases the proportion explained by genetics (i.e. increasing the heritability estimate). At the same time, some argue the environment shared by monozygotic twins during embryonic development could lead to higher phenotypic similarities, which will also inflate heritability estimates if wrongly attributed to genetics[46].

However, regardless of the true heritability of lifespan, a number of geneticists argue estimation of the heritable component of a trait in human populations has limited value. While heritability describes the statistical properties of trait variation in a population, Rose argues it cannot describe the degree to which genetics act to determine any individual phenotype, and heritability estimates do not reveal which genes or how many are involved[47]. Similarly, Gamma and Liebreinz provide the example of phenylketonuria—a disease caused by a monogenic mutation which can be completely mitigated with a diet lacking phenylalanine—to argue that heritability also does not describe the degree to which a trait can be altered by the environment[48]. Lastly, Burt and Simons go one step further and argue that for complex traits, the genetic and environmental components are completely interdependent and trying to separate them is illogical[49]. For the purposes of this thesis, these limitations mean lifespan heritability estimates may not be able to describe the degree to which genes determine individual survival very well. However, as a statistical property, a low heritability estimate suggests large sample sizes may be needed to successfully study the genetic determinants of human lifespan.

1.2.3 Genome-wide association of lifespan

Although human lifespan only has a modest genetic component, it might still be possible to identify this component and study its function. Linking variation in a trait to specific regions in the genome is more generally referred to as quantitative trait locus mapping, and has historically been done using linkage analysis[50], which uses the segregation of genetic markers (variable DNA segments of which the physical location is known) with traits of interest to narrow down the region of the genome influencing the trait. Linkage analysis is performed using family pedigrees, with mapping resolution depending on the number of meiosis and recombination events. Herein lies the limitation of these studies as well: large family-based cohorts spanning multiple generations are needed to detect linkage of genetic markers and map them to a small region of the genome.

Due to this disadvantage, linkage analysis has provided limited insights into the genetics of quantitative traits[51], especially those determined by many variants of small effects. For lifespan, only a single genomic locus has been discovered by linkage analysis: a region within the 4q25 cytogenetic band, associated to exceptional survival[52] and survival free of major disease[53]. However, even when regions are mapped by linkage studies, identification of a causal variant is not straightforward. The human genome contains close to 650 million documented variations, with roughly 11 million of these occurring commonly (MAF > 5%) (dbSNP[54] build 153). Observing an association between a quantitative trait and a genomic locus suggests one or more genetic variants inherited together with the marker may influence the trait, but depending on the size of the region, hundreds to thousands of variants could be tagged with no clear indication of which one is causal (that is, if the signal is not chance or confounding to begin with).

Two technical innovations caused a breakthrough in human genetic mapping. The first was the development of DNA microarray technology, which were chips dotted with thousands of oligonucleotide sequences that could hybridise with specific DNA segments to create a signal due to their fluorescent tags[55]. The

second was the completion of the Human Genome Project[56], which gave researchers a comprehensive reference of the human genome for the first time and in turn accelerated the discovery of first hundreds of thousands and then millions of single nucleotide polymorphisms (SNPs). Together, these advancements led to the creation of genotyping arrays: DNA microarrays designed to capture hundreds of thousands of SNPs across the genome[55]. While these still did not capture all variation, many genetic variations are in physical proximity to each other and inherited together in large blocks, resulting in regions of linkage disequilibrium (LD). Therefore, SNPs captured by genotyping arrays could be used to inform allelic status of known but untyped variants in LD as well—a process called imputation—resulting in a genome-wide coverage of most of the commonly carried genetic variants[57].

More recently, SNP genotyping arrays have evolved to capture over 1 million SNPs as well as copy number variations, at an accuracy of over 99.5%[58]. Such arrays are also now used to calculate genetic relationships between individuals, reducing the need for population-wide pedigrees to account for relatedness[59]. Meanwhile, advances in DNA sequencing technology led to the generation of a larger variety of reference genomes, which facilitated the creation of larger reference panels (e.g. 1000 Genomes, Haplotype Reference Consortium)[60,61]. In turn, these reference panels allowed for more accurate imputation and greater coverage of the genome, with studies using state-of-the-art imputation panels now able to examine hundreds of millions of variants[62]. These cost-efficient and high-resolution genotyping methods provided a step change in our ability to study the genetics of quantitative traits compared to linkage analysis, with new, genome-wide association studies (GWAS) becoming the preferred method to map quantitative trait loci.

At its core, a GWAS of a quantitative trait employs a linear regression methodology, regressing trait values against allelic dosage for each individual while taking into account covariates like relatedness and population genetic

structure. The model is identical to the one described in **Equation 5**, with the addition of the SNP of interest (SNP):

$$\mathbf{y} = \mu + \beta\mathbf{X} + \mathbf{g} + \text{SNP} + \epsilon \quad (6)$$

Where SNP is coded as 0, 1, or 2 to reflect the number of effect alleles compared to the reference (e.g. TT, TC, CC, with reference to T), and as before, \mathbf{X} is a matrix of fixed covariates and \mathbf{g} is a vector of random effects based on genomic relationships. When using imputed data, the allelic dosage can take an intermediate value to take into account the uncertainty in the imputation. In practice, fitting all SNPs at once within this model is computationally intractable, so instead the regression is repeated separately for every single SNP to be tested[63].

While many of the tools developed to perform this association are designed to fit the fixed covariates and random components for every SNP[63–65], alternative methods have been developed to split the analysis into two steps to reduce the computational burden[66,67]. These methods first run the model without any SNPs, and then regress the residuals of the model—which are thus corrected for covariates and population structure—against SNP dosages, one at a time. Residualising the trait is not a trivial step: random-effects residuals must be scaled appropriately to maintain power and prevent bias[66,68]. In the context of lifespan, the Cox model (Equation 2) can be residualised using martingale-based residuals[69], which, assuming no time-dependent covariates, can be written as:

$$\widehat{M}_i = \delta_i - \Lambda_0(\tau_i)e^{\widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \dots + \widehat{\beta}_n X_n} \quad (7)$$

Where the martingale residual \widehat{M} for individual ‘i’ is calculated as the difference between the observed status δ of the individual (0 = alive; 1 = dead) and the cumulative baseline hazard Λ_0 up until time τ , as modified by the vectors of covariates \mathbf{X}_1 – \mathbf{X}_n and their estimated effects $\widehat{\beta}_1$ – $\widehat{\beta}_n$. Dividing martingale residuals

by the proportion of events in the population approximates them to log hazard ratio units[70], and allows them to be regressed against SNP dosages in a linear regression framework.

GWAS have been very successful in their discovery of genetic associations with traits of interest. As of 2020, around 160,000 trait-variant associations from more than 3,000 studies have been documented in the GWAS catalog[71], an online repository of GWAS results (Figure 4). However, prior to 2016, there had been exceedingly few discoveries for lifespan traits, with only two genomic regions near *APOE* and *CHRNA3/5* being discovered and replicated. This is likely because in comparison to other traits, lifespan is challenging to study genetically because the collection of genetic information is generally done on living individuals, who take a long time to die.

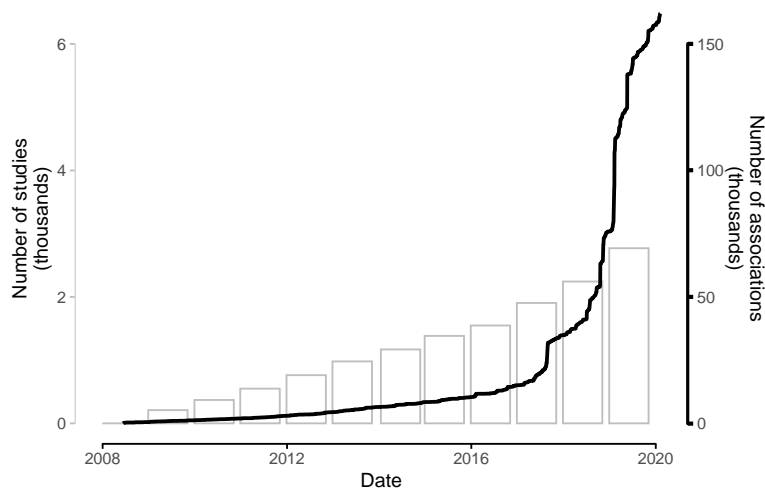


Figure 4. The growing list of studies and associations catalogued in the GWAS catalog. The bars show the cumulative number of genome-wide association studies (left axis) with at least one genome-wide significant association ($P < 5 \times 10^{-8}$) recorded within the GWAS catalog, while the line shows the total number of genome-wide significant associations (right axis). Data from GWAS catalog[71], accessed February 2020.

One line of lifespan research has attempted to address this problem by gathering cohorts of extremely long-lived individuals (commonly referred to as individuals reaching 'longevity'), under the assumption that the genetics of longevity cases

can be compared to a control cohort who will likely die before reaching the same age (or have already died at a younger age)[72]. Longevity research is supported by observations that the trait of reaching exceptionally old age runs in families and these long-lived families are generally healthy for longer[73–75]. In addition, there is some evidence suggesting the inherited component to lifespan may be larger at advanced age[38,76]. However, the study of exceptional individuals is, by definition, limited to only a small percentage of a population and therefore suffers from recruitment challenges[77,78]. In addition, there is some question regarding the trustworthiness of longevity outliers (particularly supercentenarians)[79] with evidence of fraud despite valid documents[80], although less extreme longevity cases are more likely to be valid. Even if most records are accurate, differences in environmental exposures experienced by the extremely old generation compared to control individuals can complicate results[81].

Nonetheless, longevity has been studied using a case-control framework, initially in linkage studies[52], and later using GWAS[82–84]. These first GWAS were performed on samples from the US and Europe containing up to 4,149 longevity cases, with each study finding or confirming only a single locus on chromosome 19 associating with longevity, near *APOE*. Despite the considerable resources invested in collecting and characterising these exceptional individuals, GWAS of longevity have yielded very few results compared to GWAS performed at the same time on less extreme phenotypes, such as a 2008 study on the height of ~25,000 individuals which discovered 10 loci[85]. The paucity of longevity results seems likely to stem from the modest heritability and high polygenicity of the trait: there may be many genetic variants with small effect sizes which are simply not detectable with current sample sizes. In addition, there has been considerable debate regarding the appropriate cut-off to use for defining longevity[81], calling for population- and sex-specific longevity thresholds to homogenise phenotypes. Recently, this was done for the largest longevity GWAS to date; however, despite a gain in power, this study only provided putative evidence for one new locus[72].

The alternative to recruitment of longevity cohorts is to study lifespan as a quantitative trait in more general population cohorts. While these cohorts tend to be much larger, almost all individuals will be alive at time of recruitment. Depending on the average age of the cohort, it may take 30 to 50 years before enough individuals have died to perform a lifespan study. For example, the UK and Estonian Biobank population cohorts were established more than 10 years ago but fewer than 5% of the genotyped individuals in each cohort have died thus far[86,87] ([Table 2](#)). Recently, a new approach has been put forward to deal with the low number of deaths in genotyped population cohorts: study of the lives of parents instead[39]. Parents and their children share half their DNA, meaning lifespans of parents can be tested against the alleles of genotyped cohorts, although the effective sample size will only be 1/4th of the original sample for each parent[88]. While this means larger sample sizes are necessary to identify variants affecting lifespan, there is no requirement for exceptional cases of survival, and almost all individuals within population cohorts can contribute information to the study, as long as they report the age and survival status of their biological father and/or mother. The power of this method was demonstrated in two studies published in 2016, showing parental age at death in a linear regression model[89] and parental survival in a Cox proportional hazards model[90] could be studied effectively in large population cohorts.

Despite the advantages of ease of recruitment and increases in death counts, there are also multiple drawbacks to using parental information. Without full knowledge of parental genotypes, it is impossible to construct an accurate kinship matrix for the parental generation, and as a result, genetic studies into parental lifespan cannot properly account for relatedness[88]. The easiest solution to this problem is to exclude related individuals, assuring their parents are unrelated as well, but this does result in a lower sample size. On a phenotypic level, the drawback of studying parents is that their survival reflects the causes of death in the parental generation and, given that these causes have changed over time[91], parental risk factors may not accurately translate to the current generation. Moreover, important lifespan-related covariates, such as smoking status and use

of medications, may not be readily available for the parental generation and can therefore not always be adjusted for.

Lastly, by definition, a parental analysis is limited to individuals who have survived until adulthood and who have been able or willing to conceive at least once. Developmental and fertility traits are heritable to some extent[92] and conditioning on them through sample selection can introduce collider bias[93,94]. That is, genotypes causing developmental and fertility traits may falsely appear to be associated with parental survival or have upwardly biased effect size estimates. However, unless genotype effects on colliders are very large or the correlation between fertility and parental lifespan is large, this bias may not have a material effect on genotype-phenotype associations[93].

1.2.4 Datasets large enough to study lifespan

Large population cohorts have been instrumental in driving modern discoveries, both in terms of traditional and genetic epidemiology. Recently, the linkage of nation-wide databases such as electronic health records and death records have created rich datasets of hundreds of thousands to millions of individuals. Scotland is unique in this regard as its government specifically set forth strategies and action plans in 2012 to link up nationwide datasets and make them accessible to researchers to use in the public's interest[39]. According to the National Data Catalogue, around 100 datasets have been catalogued, spanning everything from A&E visits to cancer treatment waiting lists to GP prescriptions to postcodes[95]. Where available, these datasets have been linked using Community Health Index (CHI) identifiers, although many of those without explicit CHI numbers can be linked by the electronic Data Research and Innovation Service upon request[95]. This means medical and demographic data spanning multiple decades is available to eligible researchers on the whole population of Scotland—over 5 million individuals[96].

In the context of lifespan, linkage of data such as dates of birth and death (if deceased), measures of deprivation, and dates and diagnoses of hospital visits (coded through international classification of disease formats ICD9 and ICD10) is especially relevant. For instance, nationwide death records are available for almost 1.5 million individuals since 1990 ([Table 2](#)) and can be used to infer long-term trends in mortality. Linkage of those records with hospital admissions data then allows mortality trends to be linked to trends in disease. While the sensitivity of the data requires additional security measures to maintain the privacy of the individuals involved, the scale and scope of data permit researchers to investigate the diseases determining lifespan and their trends with unprecedented precision.

In terms of genetic epidemiology, a number of cohorts with genetic data have recently gained global prominence and have allowed for detailed study of the genetics of lifespan. One such cohort is UK Biobank, a nationwide, prospective cohort study of around half a million individuals. These individuals were aged 40 to 69 and were recruited between 2006 and 2010 across 22 assessment centres in Scotland, Wales, and England[87]. Individuals were characterised using a combination of verbal interviews, questionnaires, and medical assessments. They also donated biological samples and a subset of individuals has since been imaged using magnetic resonance imaging or dual-energy X-ray absorptiometry. While individuals themselves have been anonymised, all individuals consented to have their UK Biobank record linked to NHS medical records and death records. Death records show that from the time of recruitment until 2019, only 4.1% of UK Biobank individuals died ($N = 20,442$; [Table 2](#)). However, at recruitment, individuals also reported the survival of their parents via touchscreen questionnaire, which showed around 78.8% of parental generation was deceased ($N_{\text{father}} = 433,444$; $N_{\text{mother}} = 345,644$).

In terms of genetics, the UK Biobank individuals were genotyped on two custom-built, largely identical (95% similarity), genotyping arrays. The arrays contained around 800,000 markers designed to provide good coverage of the genome for

imputation purposes, as well as a number of specific markers previously linked to phenotypic variation or known to be missense or protein-truncating variants[97]. Imputation was performed using both low-coverage UK-specific haplotype references (UK10K) and haplotypes from the 1000 Genomes Project[98]. Genotypic information was released to the scientific community in two phases, with array and imputed genotypes for the first 150,000 subjects being released in May 2015 and the remainder following in July 2017[87].

Cohort	Dates (recruitment)	Age	N	Deaths	Genotyped	Reference
National Records of Scotland (Deaths)	1990–2016	35+	1,477,796	1,477,796	No	Timmers <i>et al.</i> [91]
UK Biobank	2006–2010	40–69	502,506	20,442	Yes	UKB Data Showcase[99]
Estonian Biobank	1999 (ongoing)	18+	50,916	2,333	Yes	Leitsalu <i>et al.</i> [86]

Table 2. Characteristics of population cohorts used in this body of work.

Another dataset of importance is the Estonian Biobank, a longitudinal, prospective study of over 50,000 individuals resident in Estonia, equating to 5% of the adult population[86]. Unlike individuals in UK Biobank, the minimum age for individuals to participate in the Estonian Biobank was 18, with no maximum age. Individuals were recruited in private practices, hospitals, and recruitment offices across all 15 Estonian counties[86], which administered baseline questionnaires and anthropometric/blood tests. Similar to UK Biobank, the Estonian cohort benefits from record linkage to national databases, including the Population Register, the Estonian Causes of Death Registry, and Estonian Tuberculosis Registry, and health insurance/medical records, which are all updated semi-annually.

In terms of genetics, Estonian individuals' DNA was read using four different methods. Whole genome sequencing was performed on a subsample of around 2,500 participants, which were subsequently used to create a population-specific haplotype reference[100]. Around 8,000 individuals were genotyped using the HumanOmniExpress beadchip and the remainder of the cohort was or is being genotyped using the Global Screening Array from Illumina[101]. The cohort is currently still expanding, set to reach a total sample of 150,000 genotyped individuals by 2020. As of 2015, around 4.6% of the cohort was deceased (N = 2,333; [Table 2](#)), while 38.9% of their parents were reported as dead at time of recruitment.

1.3 How precision medicine and therapeutic discovery will affect ageing

It is likely that large population datasets will transform our understanding of diseases and the ageing process, but questions remain on how this knowledge should be implemented in clinical settings and the effect it will have on healthcare.

Understanding which diseases are responsible for the most deaths and predicting which ones will be important in the future will have the most immediate benefit. Healthcare policy can be directed to improve treatment and prevention of the most common diseases with the highest mortality. Additionally, this research allows diseases with the largest socioeconomic disparity to be highlighted and addressed. At the same time, monitoring which disease categories show increasing mortality rates over time and taking preventative action to mitigate these trends can halt slowdowns or reversals in population life expectancy.

Investigating the genetics of lifespan has more long-term benefits on healthcare, as genetic screening of individuals for risk of disease prior to its occurrence has

thus far been limited to congenital disorders, such as phenylketonuria and congenital hypothyroidism[102], and to high-penetrance variants, such *BRCA1* and *BRCA2* for breast and ovarian cancer[103], and *HTT* polyglutamine expansion for Huntington’s disease[104]. However, with the advent of GWAS and large datasets, it has now become possible to calculate individuals’ genetic risk for more complex diseases, such as myocardial infarction and diabetes[105], using many variants spread across the genome. Summing the effects of tens to millions of variants creates an overall score for each individual, known as a polygenic risk score:

$$PRS_i = \sum g_{ij}\beta_j \quad (8)$$

Where PRS_i is the polygenic risk score for the i th individual, calculated as the sum of j effect alleles g (denoted 0, 1, or 2) times its associated effect β .

There is debate regarding which variants to include in the risk score, with one side advocating a sparse model of only the most informative SNPs, and the other side suggesting the inclusion of all SNPs while accounting for the correlation amongst them[106,107]. From a bioinformatics perspective, it makes sense to include as many SNPs as are available to improve the predictive performance of the score, as long as the information they contribute outweighs the noise. Effect size estimates become noisier with more weakly associated variants, but Bayesian priors can be applied to shrink the weights of uncertain variants, as is done in LDpred[107]. However, from a clinical point of view, the small improvements gained from including genome-wide information may not outweigh the downsides. For one, robustly replicated, genome-wide significant SNPs are more likely to have a biologically plausible pathway to the phenotype, which should increase their predictive power in populations other than the training set. This is especially important when predicting phenotypes in a population with a different ancestry from the original study, where differences in LD and allele frequencies can lead to misestimation of scores[108]. A third method is to start from genome-wide significant SNPs and sequentially add more

putative SNPs into the model, up to a level of uncertainty that balances both prediction and transferability. PRSice[109] strikes this middle ground by testing a multitude of SNP cut-offs to identify the list of SNPs that is optimal at predicting the phenotype in a validation sample (rather than the training sample).

Once polygenic risk scores have been calculated for individuals in a population, individuals can be stratified into groups and tested for their susceptibility to disease. For example, a study creating polygenic risk scores for cardiovascular disease found that individuals with extreme scores were much more likely to be diagnosed with the disease[105]. As GWAS sample sizes have grown, these scores have become more informative, reaching a point where the combined risk from common variants can match or exceed the disease risk associated with routinely tested monogenic mutations[105].

One application of polygenic risk scores is their use in disease screening and preventative care for at-risk individuals. For example, one study found individuals with higher polygenic risk scores for atherosclerosis were more likely to get the disease but also benefitted the most from statin therapy in preventing an ischaemic heart attack[110]. Another study found a polygenic risk score using several hundred breast cancer SNPs could be used to calculate breast cancer risk by age and use this to make group-specific recommendations for age of first breast cancer screening[111] ([Figure 5](#)). Similarly, polygenic risk scores for lifespan may prove to be useful in stratifying individuals in terms of their likelihood of disease and death, and inform those who are at the highest risk to take preventative measures. On the other hand, individuals with positive scores for lifespan could be more resilient and may be able to withstand more aggressive treatment.

Secondly, knowledge of the genetics of lifespan may be used to inform therapeutic drug development to mitigate the burden of age-related disease. Discovery of new drugs and their development into effective and safe treatments has high failure rates, with only around 1 in 9 compounds entering clinical trials being ultimately

approved[112]. However, it is estimated that incorporating genetic information into the drug development process could double the success rate, as the genes targeted by approved drugs tend to be enriched for GWAS signals[113]. These enrichments appear to be most significant for genes related to musculoskeletal, metabolic, and blood phenotypes.

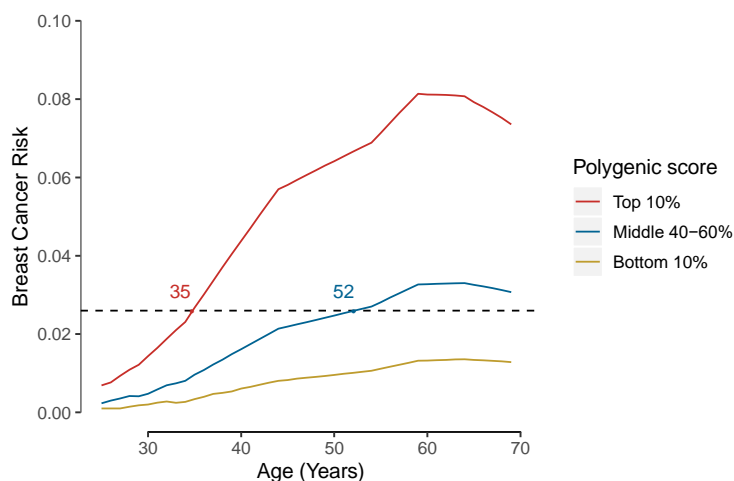


Figure 5. Breast cancer risk as a function of age and polygenic risk score percentage. Solid lines show the absolute 10-year risk of developing breast cancer for individuals with varying levels of polygenic risk based on 313 breast cancer SNPs. Dotted line represents the average 10-year risk of breast cancer for all 47-year-old women (i.e. when they first become eligible for screening in the UK) (2.6%). Individuals in the highest score decile will reach this level of risk by age 35, while individuals in the lowest decile will never reach this level of risk. Data from Mavaddat, N. *et al.* [111]

However, while this retrospective evidence for the utility of GWAS in drug discovery is promising, it is unclear whether the success rate of drug development would shift if more compounds were submitted to clinical trials based on GWAS signals. This is because interpretation of GWAS signals is very difficult: many GWAS hits are located in intergenic regions and could affect anything from the splicing or expression of a gene to the stability of messenger RNA or protein[114].

Given the modest heritability of human lifespan, it may be possible that no single longevity gene can be found and targeted to delay ageing and age-related disease

using human genetic studies alone. Nonetheless, the genetics of lifespan can highlight which biological pathways are important in the human ageing process and provide insight into the differences and similarities with model organisms. Several therapeutics are in development to target known hallmarks of ageing, including metformin[115], rapamycin[116], and nicotinamide[117], which have shown to extend healthy life in model organisms. Whether these drugs will be successful in humans will depend largely whether the pathways they target are relevant to human survival as well. Analogously, knowledge of the biology of human ageing can inform which targets to examine in model organisms.

1.4 Conclusions

Greying populations and their associated burden of chronic disease are set to become one of the greatest global challenges. There is a pressing need to understand how disease affects lifespan on a population level, and genes and biological pathways affect lifespan on an individual level. Applying statistical survival modelling to some of the largest datasets to date has made it possible to address these challenges, with an improved understanding of diseases limiting lifespan leading to new policy, and an improved understanding of the genes influencing lifespan leading to personalised, preventative treatment and novel therapeutics.

In order to address the challenges of an ageing population, in this body of work, I use hospital and death records across Scotland to investigate how morbidities influence lifespan, highlighting the diseases responsible for the greatest burden of mortality, and how they have changed over time. I then use genetic data on individuals from the UK and Europe to identify and quantify the genetic component to lifespan, examining the role of genes and pathways on disease and death. Finally, I meta-analyse large-scale genetic studies on ageing-related phenotypes to shed light on the processes underlying healthy ageing in humans.

Chapter 2: Trends in disease incidence and survival and their effect on mortality in Scotland

2.1 Introduction

2.1.1 Context

Risk of disease and death increase exponentially with age, making ageing and age-related morbidity the strongest determinants of human lifespan[118]. However, morbidities have varied over time; therefore, an analysis of morbidity trends may give insight into changes in the determinants of lifespan. The aim of the following study was to explain the observed trends in mortality through trends in morbidities.

It is well-known that life expectancy in high-income countries such as the UK has increased by roughly three years every decade[119]. So far, studies into the determinants of this trend have been limited to changes in cause of death[120], which are often poorly recorded and of which the quality can vary between sex and socioeconomic deprivation. In addition, the retrospective nature of a causes of death study is unable to assess changes in the incidence of disease nor the likelihood of death following disease.

In this Chapter, I sought to study how changes in disease incidence and survival have influenced human lifespan between 2001 and 2016. Examining the effect of diseases on mortality and tracking how the incidence and survival rates of these diseases have changed allows us to model future changes in the determinants of lifespan and predict mortality. Having examined the influence of recent disease trends on human lifespan, the following chapters go on to examine how—at a single point in time—genetic factors can also influence the incidence of disease and death, ultimately determining one's length and quality of life.

2.1.2 Contributions

The idea for the prediction of mortality through morbidities was first put forward by Craig Butler and Stuart McDonald (Lloyds Banking Group), who approached and later funded Peter Joshi and Colin Fischbacher to perform such a study in Scotland. After a basic project plan was formed, Jan Kerssens (electronic Data Research and Innovation Service) performed the linkage of raw healthcare and death records and made these available in the Scottish National Safe Haven. Peter then performed preliminary quality checks and data analysis in the Safe Haven, which included writing scripts to test if data could be accessed and cox survival models could be constructed.

I picked up the work from here, refining these basic scripts into a pipeline that could run the analysis for a subset of diseases and stratify analyses by sex and socioeconomic status. I was also responsible for strengthening the preliminary data quality control, which involved applying additional exclusions criteria, producing data descriptives, and creating figures to visualise the exclusions. I then performed dozens of runs to generate disease-specific results and wrote the scripts to combine these into overall measures of morbidity, format the results for publication, and draw all the figures.

The electronic Data Research and Innovation Service checked the results for any potentially identifiable information before extracting them from the Safe Haven. I wrote the draft manuscript with help from Peter and Jan. Specifically, Peter wrote the ethical approval section and contributed substantially to the discussion of the study limitations, while Jan contributed the section on data linkage. All co-authors provided useful comments and feedback on the initial draft, which included requests for rephrasing (especially to distinguish “disease survival” from “all-cause mortality after hospitalisation for a disease”) and providing relevant references.

Finally, I would also like to acknowledge comments from four named *BMJ Open* reviewers, David Roder (University of South Australia, Australia), Yuling Hong (Centers for Disease Control and Prevention, USA), Qingfeng Li (Johns Hopkins University, USA), and Rosie Cornish (University of Bristol, UK), which improved the manuscript prior to publication.

2.2 Manuscript accepted for publication

This manuscript was submitted to the journal *BMJ Open* and went through formal peer review. It was accepted for publication on 24 February 2020. A copy of the Author Accepted Manuscript prior to proofing is included below, provided under the terms of the Creative Commons Attribution License CC BY 4.0.

Trends in disease incidence and survival and their effect on mortality in Scotland: nationwide cohort study of linked hospital admission and death records 2001–2016

Paul RHJ Timmers¹, Joannes J Kerssens², Jon W Minton³, Ian Grant², James F Wilson^{1,4} (Supervisor), Harry Campbell¹, Colin M Fischbacher², Peter K Joshi¹ (Supervisor)

1) Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK

2) Information Services Division, NHS National Services Scotland, Edinburgh, UK

3) Public Health Observatory, NHS Health Scotland, Edinburgh, UK

4) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

The formatted article and supplementary information can be found at *BMJ Open*.

Available at: <http://dx.doi.org/10.1136/bmjopen-2019-034299>

Abstract

Objectives: Identify causes and future trends underpinning Scottish mortality improvements and quantify the relative contributions of disease incidence and survival.

Design: Population-based study.

Setting: Linked secondary care and mortality records across Scotland.

Participants: 1,967,130 individuals born 1905–1965 and resident in Scotland 2001–2016.

Main outcome measures: Hospital admission rates and survival within five years post-admission for 28 diseases, stratified by sex and socioeconomic status.

Results: “Influenza and pneumonia”, “Symptoms and signs involving circulatory and respiratory systems”, and “Malignant neoplasm of respiratory and intrathoracic organs” were the hospital diagnosis groupings associated with most excess deaths, being both common and linked to high post-admission mortality. Using disease trends, we modelled a mean mortality hazard ratio of 0.737 (95% CI 0.730–0.745) from one decade of birth to the next, equivalent to a life extension of ~3 years per decade. This improvement was 61% (30%–93%) accounted for by improved disease survival after hospitalisation (principally cancer) with the remainder accounted for by lowered hospitalisation incidence (principally heart disease and cancer). In contrast, deteriorations in infectious disease incidence and survival increased mortality by 9% (~3.3 months per decade). Disease-driven mortality improvements were slightly greater for men than women (due to greater falls in disease incidence), and generally similar across socioeconomic deciles. We project mortality improvements will continue over the next decade but slow by 21% because much progress in disease survival has already been achieved.

Conclusion: Morbidity improvements broadly explain observed mortality improvements, with progress on prevention and treatment of heart disease and cancer contributing most. The male-female health gaps are closing, but those between socioeconomic groups are not. Slowing improvements in morbidity may explain recent stalling in improvements of UK period life expectancies. However, these could be offset if we accelerate improvements in the diseases accounting for most deaths and counteract recent deteriorations in infectious disease.

Strengths and limitations of this study

- The individual-level linkage of hospital and death records in this population-wide dataset allows for direct modelling of improvements in 28 disease categories in terms of improvements in disease incidence and subsequent survival, stratified by sex and socioeconomic status.
- Exclusion of migrating individuals means changes in disease are unaffected by population shifts, and allow for diseases to be compared with each other and summarised into trends in mortality based on morbidity.
- Hospital admission diagnosis and subsequent survival avoid issues with cause of death recording; however, they do not provide evidence of the causal effect of disease on mortality and may in some cases track changes in underlying frailty.
- This study is limited to the assessment of diseases which result in a hospital admission prior to death.

Introduction

In recent decades, there has been a substantial improvement in life expectancies at birth in the UK[121]. More recently, several studies have suggested that there has been slowdown in improvements in the USA, UK, France, Germany, Sweden, the Netherlands and other Organisation for Economic Co-operation and Development countries; however, the causes are less clear, with speculation that they may arise from slowing improvements in cardiovascular disease, increased influenza mortality and/or pressure on health and social care services[121–128]. Understanding trends in disease incidence and subsequent survival could illuminate such trends in mortality, and disentangling how and how much different diseases contribute has the potential to reveal whether investment in healthcare and research is directed at the most urgent diseases and most affected individuals.

Through its electronic Data Research and Innovation Service (eDRIS), Scotland has linkage of historical individual death and electronic health records in a controlled environment, with specific study approvals by the Public Benefit and Privacy Panel. This allows direct modelling at an individual level of the incidence of disease and subsequent death or survival of subjects. Furthermore, because historic records are available and the whole population is covered, a retrospective cohort study can be constructed (with inherent representativeness of the initial sample, with very complete levels of follow-up, and without survivor bias).

Here, we use population-wide data between 2001 and 2016 on residents of Scotland born before 1966 to explore how trends in longevity were driven by different trends in broad classes of disease incidence or survival, and highlight diseases which have shown more or less improvement in their contribution to overall mortality. We partition overall mortality by sex and socioeconomic status and, assuming past disease improvements continue to the same extent in the future, use these results to project future improvements in mortality and their changing sources.

Methods

All methods and results are reported in line with RECORD guidelines[129].

Data sources

We received ethical approval to access administration and care records from NHS National Services Scotland (NSS) from 2001 up to 2016. The final study population included all 1,967,130 individuals born between 1905 and 1965 who registered with the NSS, were resident in Scotland during the study period, and had complete and reliable records on their date of birth, socioeconomic status, and death (if applicable). Linkage and quality control of the data are described below.

Community Health Index dataset

Records were extracted from the historical and current Community Health Index (CHI) dataset. This is a register of all patients in NHS Scotland and is fed by eight regional databases (e.g. GP database, cancer screening). The register is considered complete from 2001 onwards. The CHI number, contained in the dataset, is effectively a patient identifier and added to other health datasets to make linkage possible, for instance between hospital admissions, death records and the Scottish cancer registry[130]. Our extract consisted of 2,691,304 de-identified records, constituting the identified population of Scotland in 2001 who had been born between 1905 and 1965. The Scottish Index of Multiple Deprivation (SIMD)[131] was used to quantify socioeconomic status, determined by individuals' full postcode, and subsequently converted into deciles. The dataset we received contained only records with district-level postcodes and SIMD deciles, of which we excluded individuals with district codes with less than 5,000 individuals (thereby excluding anomalous postcodes, often with special meanings, such as "marketing campaign"; N = 11,564). We also excluded individuals missing from the CHI database in 2016, but not recorded as dead (and therefore likely transferred out of Scotland; N = 573,711), individuals with record discrepancies between the CHI and National Records of Scotland databases (N =

79,131), and individuals with records outside of the study dates or missing information on socioeconomic class (N = 59,767), giving 1,967,130 individuals for analysis after quality control ([Supplementary file 1](#)). Characteristics of the excluded individuals were similar to the rest of the population, except for postcode exclusions and database transfers, which were missing socioeconomic information and death records, respectively, as expected ([Supplementary file 2](#)).

National Registry of Scotland death records

We received 1,477,796 death records from the National Registry of Scotland (NRS) of all deaths occurring between 1990 and 2016, of which 699,093 could be matched to the CHI database before quality control. Unmatched records were usually for deaths occurring prior to the study start (2001). Of the matched records, 176,197 belonged to individuals that were excluded during CHI quality control, leaving 602,506 total deaths for analysis ([Table 1](#)).

Acute hospital admission

Health records were also linked to 30,054,191 acute hospital admissions, of which 17,264,379 were dated between 2001 and 2016 and could be matched ([Supplementary file 3 & 4](#)).

Disease classification

The main diagnosis of acute hospital admission records, excluding any secondary diagnoses, was used to classify records into disease categories, which corresponded to disease blocks as described in the chapters of the ICD10[132]. In order to model the effect of disease incidence and avoid double counting of chronic conditions, we used only the first admission of a disease category for each individual, excluding subsequent visits to the hospital for diseases within the same category. The term “incidence” is used throughout this study to refer to the first recorded hospital admission of any disease within the disease category during the study period.

Design

Mortality trends were modelled using morbidity trends: we first determined the major disease categories (ICD10 blocks) associated with the most lives lost by taking into account the frequency of the disease (as measured by hospitalisation) and its effect on survival (as measured by the subsequent all-cause mortality of patients admitted for the disease compared to the mortality of everyone else). The effect of disease incidence has previously been modelled based on one-, five-, or ten-year mortality[133]; we chose five years as this captured the great part of excess mortality attributable to the incidence, rather than common underlying factors, although this does vary by disease (area under graphs in [Supplementary file 5](#), in excess of asymptotic rates) while leaving a range of 10 years in our study to examine trends over time. We combined disease frequency and 5-year excess age-adjusted death rates to calculate a burden of death weighting for each disease block. We then looked at how the age-adjusted trends in hospitalisation rates (as a proxy for incidence) changed for each disease, by decade of birth, projecting that if incidence of a disease fell by a given percentage, its contribution to mortality would fall similarly. The use of a cohort model for the incidence of disease was driven by empirical investigation. Specifically, the distinctions we found by decade of birth in cancers, especially “Malignant neoplasm of respiratory and intrathoracic organs” (C30-C39) in [Supplementary file 6 & 7](#), show a clear cohort effect. However, it should be recognised that the cohorts have only been observed over the study period (2001-2016). After calculating hospitalisation rates between decades of birth, we calculated their weighted average, reflecting the expected effect of all measured disease incidence changes on mortality rates, driven by decade of birth. Similarly, we looked at how the (age-adjusted) 5-year survival rates following first hospitalisation changed by year of hospitalisation. For each block this again gives a contribution towards reduced mortality, and the weighted average, the expected effect of changes in survival of the combined diseases on overall mortality. Adding these effects (and noting we assessed changes in survival from incidences over one decade), gives the expected effect on overall mortality from decade of birth to subsequent decade of birth from the effect of changes in disease incidence and survival, under the

(necessarily simplified) model that incidence is a function of birth cohort and survival post incidence is a function of year of incidence.

Statistical analysis

Mortality

A Cox proportional hazards model using NRS mortality data – fitting sex, decade of birth, and deprivation – was used to quantify mortality in the Scottish population during the study period. The same analysis was run stratified by sex and deprivation. Unless otherwise stated, (for example median age differences in Kaplan-Meier curves), years of life of a hazard effect have been calculated by multiplying the \log_e hazard ratio (lnHR) by 10 [70]. Only individuals with complete records were included in the analysis.

Morbidity

We grouped the main diagnoses of each NSS hospital admission into categories, as laid out by the ICD10 Chapters, and included only the first instance of admission for a category per individual (discarding subsequent repeat visits to hospital for a disease within the same disease category). Analysis was restricted to more common disease blocks. Visual inspection suggested a pragmatic threshold of at least 15,000 first-time admissions (see [Supplementary file 8](#) for all disease categories meeting this threshold).

Effects on the incidence of hospitalisation for the more common disease blocks was quantified using Cox proportional hazard models based on age, with events defined as the first incidence of hospitalisation. We fitted sex, deprivation, and decade of birth as covariates. Again, the same analyses were performed stratifying by sex and deprivation.

In order to quantify all-cause mortality in the five years following hospitalisation, person-time of individuals was divided into phases, corresponding to the study start until hospitalisation, the first five years after hospitalisation, and the

remaining time in the study. For example, an individual admitted to hospital in 2004 for ischaemic heart disease I20-25 (IHD) and surviving until 2010 would contribute three phases to the model: one for the period until hospitalisation (no event), one for the first 5 years after hospitalisation (no event), and one >5 years after hospitalisation (event after 1 year). The status of each phase was fitted as a covariate in a Cox proportional hazards model[134] with death as the event, adjusting for sex and deprivation:

$$h(x) = h_0(x) e^{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4} \quad (1)$$

Where h_0 is the baseline hazard, x the patient age, and X_1 - X_4 the covariates sex, deprivation, and logically coded phase status (0-5 years True/False and >5 years True/False), with corresponding effect sizes β_1 - β_4 . This yielded estimates of the proportional hazard of status (0-5 and >5 years) after hospitalisation compared to pre-hospitalisation mortality. Thus, the baseline hazard is a function of age and the hazard ratios reflected the effects of the other covariates. The same model was run, stratified by sex and deprivation.

Burden

For disease blocks with at least 15,000 first admissions during the study period, the relative mortality burden of each disease block was calculated as the excess mortality in the 5 years after hospital admission (Equation 1) multiplied by the number of first-time admissions for the disease block, as follows:

$$N_{firstadmission}/N_{total} * h_{(0,5)} \quad (2)$$

Where $N_{firstadmission}$ is the total number of first hospital admissions of the disease category during the study period, N_{total} is the total number of individuals in the study, and $h_{(0,5)}$ is the mortality of individuals in the first five years following hospitalisation compared to individuals who were never hospitalised for the disease category, measured in \log_e hazard ratios. The resulting value was then scaled to total 1 and provides a relative measure of the number of lives lost due

to the diseases within the category, with higher values indicating a disease category with more common diseases or diseases associated with higher subsequent mortality, and lower values indicating a disease category with rare diseases or diseases associated with lower subsequent mortality. Whilst this measure may in principle be affected by differing age patterns on incidence, it was judged sufficient for our purpose – to establish broad relative weightings of the importance of each disease category.

To maintain a feasible computational burden within the national safe haven, subsequent analysis was restricted to the 25 blocks with the highest burden of death on the population ([Table 2](#)). We added C50-C50 malignant neoplasm of the breast, C60-C63, malignant neoplasms of male genital organs, and G30-G32 other diseases of the nervous system to this list, out of specific interest: in the sex-specific effects and awareness of the limitations of our method for Alzheimer's disease (see discussion). All further analyses were performed on these top 28 blocks (T-28). The use of (first) hospitalisation for a disease as our definition of incidence is imperfect (e.g. for Alzheimer's disease where hospitalisation following incidence is rare or delayed, and even first diagnosis in the community will often be preceded by a long latent period)[135].

Disease survival

Improvements in disease outcomes by ICD10 block were calculated by comparing 5-year all-cause mortality ([Equation 1](#)) following hospitalisation in 2001 with 5-year all-cause mortality following hospitalisation in 2011. As 5-year mortality estimates in 2011 had more uncertainty (due to fewer deaths in 2011–2016), we also calculated 5-year mortality following first-time hospitalisation for every year between 2001 and 2011 (i.e. mortality of patients admitted in those years), and used the trend in mortality over time to inform the 2011 estimate. To do so, we regressed the yearly mortality estimates against year of hospital admission, fitting a 3rd order polynomial to allow for non-linear relationships, and weighted the estimates by the inverse of their variances to account for uncertainty:

$$y = \beta_1x + \beta_2x^2 + \beta_3x^3 + \epsilon \quad (3)$$

Where y is the 5-year mortality hazard after hospital admission in year x , and $\beta_1, \beta_2, \beta_3$ are the coefficients describing the relationship between y and x . We then used the value and standard error predicted for 2011 by the model as our estimate for 5-year all-cause mortality for hospitalisation in 2011.

Mortality estimates from morbidity

Estimates of the improvement in incidence of hospitalisation between decades of birth was combined into an overall improvement by performing a weighted sum of all diseases, with weights derived from the relative burden of death of each disease (see above).

These ten year improvements (I) due to incidence were added to the ten year improvements due to post incidence survival (S) to give a total improvement due to all morbidities, and proportions due to incidence/survival were calculated as the S or I / (I+S).

Ethical approval

This study was approved by the Public Benefit and Privacy Panel for Health and Social Care under application number 1617-0255/Joshi. As clinical records are provided without explicit patient consent, the panel requires the public benefit of the research to clearly outweigh any impact on individual patient privacy, and appropriate safeguards and security to be in place to protect patients. The panel granted access to de-identified patient data, accessible only through the National Safe Haven and only by University of Edinburgh researchers with data safeguarding qualifications. In addition, all results were reviewed by eDRIS before extraction from the safe haven to ensure no potentially identifiable information was made public.

Patient and public involvement

Patients and the public were not involved in the study or its design, beyond their contribution of health records. Due to the retrospective study design and anonymised nature of the records, it was not feasible to contact individual patients nor involve them in the dissemination of results.

Summary of outcomes

Mortality improvements

Age-adjusted falling mortality rates observed directly from NRS death records.

Disease burden of death

Prevalence of a disease category (total number of individuals admitted at least once 2001–2016), multiplied by the age-adjusted all-cause mortality within five years (in lnHR) after the first diagnosis of the disease category.

Disease weight

Disease burden of death, scaled 0-1, denoting the relative importance of a disease category.

Disease incidence improvement

Age-adjusted hazard of being admitted to hospital for a disease category (excluding subsequent hospital visits for the same disease category) from one decade of birth to the next.

Disease survival improvement

Age-adjusted hazard of dying within five years after the first hospital admission for a disease category in 2011 compared to having the first hospital admission in 2001.

Disease improvements

Linear combination of the disease survival and disease incidence (averaged across decades of birth) in units of lnHR.

Morbidity-driven mortality

The change in mortality rates expected from the improvements in morbidity (i.e. weighted sum of disease survival and incidence for all 28 diseases).

All model coefficients used in the results can be found in [Supplementary file 9](#).

Results

Mortality

The population consisted of 1,967,130 Scottish individuals aged 35 years or older at the start of the study period (1 December 2000). 53.3% were female, 78.5% had been admitted to hospital at least once within the study period, and 30.6% died over the course of the study (31 January 2016). See [Table 1](#) for detailed population characteristics.

Sex	Dead	N			Age Entry			Age Exit			Hospital Visits		
		Individuals	Admitted	Hos. Visits	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Male	FALSE	633,953	429,659	2,315,915	50.9	49.4	10.3	65.6	64.3	9.8	3.7	2.0	6.8
Male	TRUE	283,835	283,835	2,742,686	68.0	69.3	11.7	75.6	77.1	11.4	9.7	7.0	11.4
Female	FALSE	730,671	511,632	2,764,040	52.7	51.1	11.6	67.4	66.1	10.8	3.8	2.0	7.0
Female	TRUE	318,671	318,671	2,895,443	72.1	73.8	12.0	79.7	81.7	11.5	9.1	6.0	11.0
Both	FALSE	1,364,624	941,291	5,079,955	51.9	50.3	11.1	66.6	65.3	10.4	3.7	2.0	6.9
Both	TRUE	602,506	602,506	5,638,129	70.1	71.7	12.0	77.8	79.5	11.6	9.4	6.0	11.2
Both	ALL	1,967,130	1,543,797	10,718,084	57.5	55.4	14.1	70.0	69.2	11.9	5.4	3.0	8.8

Table 1: Description of the data. The population included almost 2 million individuals (one-third of whom died during the study). See [Supplementary file 3](#) for descriptives by deprivation including ICD10 codes. N – Number of records; Admitted – Individuals admitted to hospital at least once; Hos. Visits – All records of visiting the hospital; Age Entry – Age at the start of the study period (1 December 2000); Age Exit – Age at the end of the study period (31 January 2016) or at the end of life.

Quantifying mortality effects using Cox proportional hazard models, we observed statistically significant associations ($P < 1 \times 10^{-26}$) between mortality and sex, deprivation and decade of birth ([Supplementary file 10](#)). Women showed lower overall age-adjusted mortality rates compared to men (hazard ratio 0.71; 95% CI 0.70–0.71), corresponding to an expectation of life of 3.5 years longer than their male counterparts, whilst individuals from the most deprived areas (top decile) suffered mortality rates more than twice as severe (2.07; 95% CI 2.04–2.09) as those from the least deprived areas (bottom decile), corresponding to a difference in around 7 years of life. Median survival of men and women in the most deprived areas was 71.1 and 76.6 years, respectively, compared to 82.2 and 85.2 in the least

deprived areas ([Supplementary file 11](#)). A wide gap between the most deprived decile and the adjacent one for men is apparent visually: the difference in median survival between deprivation deciles 1 to 9 is roughly constant (0.82/1.05 years per decile for women/men), but moving from the 9th to 10th deprivation decile has a greater effect, especially for men (1.99/2.67 years for women/men). Lastly, individuals born in the decade commencing 1935 had age-adjusted mortality rates 2.45 (95% CI 2.39–2.51) times those born three decades later, corresponding to a difference in life expectancy of around 9 years of life.

Morbidities and consequent mortality

Multiplying total number of hospitalisations during the study period (as a proxy for disease prevalence) by 5-year mortality after hospital admission (as a proxy for disease severity) provided a weight for the death burden of hospitalisation of each ICD10 block. We restricted our analyses to 28 of the top disease blocks for burden of death (T-28, see methods). Among the T-28, total cases of disease incidence (i.e. first-time admissions) during the study period ranged from 33,613 (A30-A49, “Other bacterial disease”) to 225,504 (R00-R09, “Symptoms and signs involving the circulatory and respiratory systems”) ([Table 2](#)). Per-person total cases of disease incidence (not age-adjusted) were 68.0% higher for the most deprived decile (188,905 individuals with 331,701 first-time admissions) compared to the least deprived decile (187,193 individuals with 195,617 first-time admissions). Between sexes, per-person incidence was 2.2% higher for men (917,788 individuals with 1,257,417 first-time admissions) compared to women (1,049,342 individuals with 1,407,223 first-time admissions) ([Supplementary file 12](#)).

In the first five years, the highest all-cause mortality rate was for patients admitted for C76-C80 (“Malignant neoplasms of ill defined, secondary and unspecified sites”; hazard ratio 26.1) compared to all-cause mortality rates for those not admitted for C76-C80. The lowest 5-year all-cause mortality rate was for those admitted for K20-K31 (“Diseases of oesophagus, stomach and duodenum”; hazard ratio 1.8) compared to mortality rates for those not admitted

for K20-K31. Ordering diseases by their burden of death weights, we found “Influenza and pneumonia” (J09-J18), “Symptoms and signs involving circulatory and respiratory systems” (R00-R09), and “Malignant neoplasm of respiratory and intrathoracic organs” (C30-C39) were the disease categories responsible for the most death ([Table 2](#)), together accounting for 19% of the total death burden of the T-28 diseases.

Apart from sex-specific cancers, we observe significant differences in burden of death between men and women for injuries to the hip and thigh (S70-S79) and head (S00-S09), with the former having a higher burden in women due to more female cases and the latter having a higher burden in men due to more male cases. For both disease blocks, the effect of hospitalisation on subsequent mortality is greater in men than women (S70-S79 hazard ratio men: 3.19, women: 2.44; S00-S09 hazard ratio men: 2.32, women: 1.88). Strikingly, 5-year mortality after hospital admission for IHD is higher for women (hazard ratio 2.01/1.70 women/men), but this is offset by the lower prevalence of hospitalisation in women ([Supplementary file 12](#)).

ICD10	Disease grouping	Disease Importance			Average 10-year improvements HR (95%CI)			Survival to Incidence Ratio (SE)
		Total hospital visits	5-year mortality (HR)	Relative weight	Incidence	Survival	Combined	
J09-J18	Influenza and pneumonia	110,985	5.28	0.068	1.19 (1.11-1.27)	0.86 (0.80-0.92)	1.02 (0.92-1.12)	0.47 (0.13)
R00-R09	Symptoms and signs involving the circulatory and respiratory systems	225,504	2.08	0.061	0.88 (0.85-0.92)	0.79 (0.73-0.86)	0.70 (0.64-0.76)	0.65 (0.14)
C30-C39	Malignant neoplasm of respiratory and intrathoracic organs	54,178	21.12	0.061	0.83 (0.75-0.91)	0.81 (0.74-0.89)	0.67 (0.59-0.77)	0.52 (0.15)
R10-R19	Symptoms and signs involving the digestive system and abdomen	174,055	2.56	0.060	0.87 (0.83-0.91)	0.89 (0.82-0.97)	0.77 (0.70-0.85)	0.45 (0.19)
R50-R69	General symptoms and signs	157,357	2.67	0.057	0.90 (0.85-0.94)	0.97 (0.93-1.02)	0.87 (0.81-0.94)	0.19 (0.19)
C15-C26	Malignant neoplasms of digestive organs	71,981	8.13	0.056	0.79 (0.73-0.85)	0.67 (0.63-0.72)	0.53 (0.48-0.59)	0.63 (0.08)
I30-I52	Other forms of heart disease	142,898	2.70	0.052	0.78 (0.74-0.83)	0.84 (0.80-0.87)	0.66 (0.61-0.70)	0.43 (0.06)
C76-C80	Malignant neoplasms of ill defined, secondary and unspecified sites	39,339	26.13	0.047	0.84 (0.76-0.93)	0.69 (0.61-0.79)	0.58 (0.49-0.69)	0.68 (0.16)
Z40-Z54	Persons encountering health services for specific procedures & care	157,841	2.15	0.044	1.00 (0.95-1.05)	0.74 (0.69-0.80)	0.74 (0.68-0.81)	1.00 (0.19)
I60-I69	Cerebrovascular diseases	100,907	3.06	0.042	0.79 (0.74-0.85)	0.82 (0.76-0.87)	0.65 (0.59-0.71)	0.46 (0.09)
K55-K63	Other diseases of intestines	206,178	1.72	0.041	0.94 (0.91-0.98)	0.80 (0.77-0.84)	0.76 (0.71-0.80)	0.80 (0.12)
J40-J47	Chronic lower respiratory diseases	78,467	3.99	0.040	0.78 (0.73-0.84)	0.79 (0.72-0.86)	0.62 (0.55-0.69)	0.49 (0.11)
I20-I25	Ischaemic heart diseases	175,605	1.83	0.039	0.65 (0.63-0.68)	0.77 (0.72-0.82)	0.50 (0.46-0.55)	0.38 (0.05)
N30-N39	Other diseases of the urinary system	126,329	2.31	0.039	1.04 (0.99-1.10)	1.24 (1.15-1.34)	1.29 (1.17-1.42)	0.84 (0.22)
K20-K31	Diseases of oesophagus, stomach and duodenum	172,206	1.83	0.038	0.72 (0.69-0.75)	0.88 (0.79-0.96)	0.63 (0.56-0.70)	0.29 (0.11)
J20-J22	Other acute lower respiratory infections	77,520	3.57	0.036	1.11 (1.03-1.20)	1.11 (1.06-1.17)	1.24 (1.13-1.36)	0.50 (0.15)
S70-S79	Injuries to the hip and thigh	78,231	2.64	0.028	0.87 (0.78-0.96)	1.02 (0.98-1.06)	0.89 (0.79-0.99)	0.12 (0.14)
A30-A49	Other bacterial diseases	33,613	6.60	0.023	1.56 (1.39-1.74)	0.94 (0.77-1.15)	1.46 (1.16-1.84)	0.13 (0.21)
T80-T88	Complications of surgical and medical care, not elsewhere classified	75,217	2.32	0.023	1.04 (0.97-1.11)	0.83 (0.74-0.94)	0.86 (0.75-0.99)	0.84 (0.40)
N17-N19	Renal failure	37,213	5.14	0.022	1.02 (0.91-1.15)	0.82 (0.73-0.91)	0.83 (0.71-0.98)	0.90 (0.42)
I80-I89	Diseases of veins, lymphatic vessels and nodes, not elsewhere classified	84,073	1.94	0.021	0.80 (0.75-0.85)	0.79 (0.72-0.86)	0.63 (0.56-0.71)	0.52 (0.12)
K90-K93	Other diseases of the digestive system	47,091	2.98	0.019	0.98 (0.90-1.07)	0.83 (0.72-0.95)	0.82 (0.69-0.96)	0.91 (0.50)
I70-I79	Diseases of arteries, arterioles and capillaries	47,410	2.95	0.019	0.67 (0.61-0.74)	0.91 (0.85-0.99)	0.61 (0.54-0.69)	0.19 (0.08)
K50-K52	Non infective enteritis and colitis	59,183	2.27	0.018	0.73 (0.68-0.79)	0.89 (0.84-0.95)	0.65 (0.59-0.72)	0.27 (0.08)
S00-S09	Injuries to the head	64,925	2.09	0.018	0.95 (0.88-1.02)	0.98 (0.90-1.08)	0.93 (0.82-1.05)	0.22 (0.72)
C50-C50	Malignant neoplasm of breast	39,358	3.21	0.017	0.81 (0.74-0.89)	0.31 (0.27-0.36)	0.25 (0.21-0.30)	0.85 (0.07)
C60-C63	Malignant neoplasms of male genital organs	22,312	3.23	0.010	0.91 (0.79-1.05)	0.50 (0.44-0.57)	0.45 (0.37-0.55)	0.88 (0.14)
G30-G32	Other degenerative diseases of the central nervous system	4,655	3.29	0.002	0.75 (0.51-1.10)	0.96 (0.73-1.28)	0.78 (0.49-1.24)	0.11 (0.44)
TOTAL		2,664,631	2.77	1.000	0.89 (0.88-0.90)	0.83 (0.76-0.90)	0.74 (0.72-0.75)	0.61 (0.16)

Table 2: Relative mortality burden of hospital admission by disease grouping and improvements in hospitalisation incidence and survival. ICD10 – Diseases contained within the disease grouping, coded by International Classification of Disease Codes, Tenth Revision. See [Supplementary file 4](#) for counts of 3-letter ICD10 records within each ICD10 block. Total hospital visits – Number of first-time admissions with main diagnosis falling within the disease block. 5-year mortality – Mortality within the first five years after admission compared to individuals who had not yet or ever been admitted for the disease group. Relative weight – Relative burden of death as a function of hospital admissions and 5-year mortality, scaled to [0-1]. Incidence – Average hazard ratio of being admitted to hospital for each subsequent decade of birth. Survival – All-cause mortality hazard ratio after being admitted for the disease in 2011 compared to 2001. Combined – linear combination of changes in disease incidence and survival. 95% confidence intervals are listed in parentheses. Ratio – the ratio of changes in disease survival to incidence of hospital admission. Standard error is listed in parentheses. See [Supplementary file 12](#) for these data by sex and deprivation. See [Supplementary file 8](#) for the relative burdens of all disease groupings with more than 15,000 first-time hospital admissions.

Trends in disease

To understand changes in disease survival rates, we next modelled the effects of a disease on all-cause mortality by year of hospital admission for admissions between 2001 and 2011 and 5-year survival subsequent to admission. We find an overall improvement over time in patient survival following hospitalisation, with a median decline between 2001 and 2011 in the 5-year hazard ratio of 16.8% for admitted cases across the T-28 diseases. The biggest improvements were for malignant neoplasms of the breast (C50) and male genital organs (C60-C63), which have seen 68.7% (95% CI 64.1%–72.7%) and 50.2% (95% CI 42.9%–56.6%) declines in the 5-year hazard ratio between 2001 and 2011 mortality, respectively. On the other hand, “Other acute lower respiratory infections” (J20-J22) and “Other diseases of the urinary system” (N30-N39) have seen increases in mortality hazard of 11.3% (95% CI 7.0%–15.7%) and 24.0% (95% CI 14.6%–34.1%), respectively ([Table 2](#); [Supplementary file 13 & 14](#)).

We next modelled age-adjusted incidence of hospitalisation for a disease by birth decade, under the simplified model that incidence is a cohort rather than period effect – essentially modelling that current incidence is the effect of (previous) lifetime exposures, rather than current exposures. We find disease incidence has fallen decade on decade of birth for cancers, cardiovascular, and intestinal diseases, but this improvement appears to have slowed down in the last decade of birth (1955-1965) considered. Age-adjusted incidence of “Influenza and pneumonia” (J09-J18) and “Other bacterial diseases” (A30-A49) has worsened by decade on decade of birth, over the whole range of births considered ([Supplementary file 6 & 7](#)).

When taking both trends in incidence and survival into account – adding 1) the average age-adjusted incidence rate reductions between decade on decade of birth to 2) the 2001-2011 reductions in 5-year disease mortality ([Supplementary file 15 & 16](#)) – we observe the death burden of cancers is declining most ([Figure 1](#)). Notably, breast and prostate cancers have seen the largest improvement of all

disease categories in the last decade. “Other diseases of the urinary system” (N30-N39), “Other bacterial diseases” (A30-A49), “Other acute lower respiratory infections” (J20-J22), and “Influenza and pneumonia” (J09-J18) have all seen increases in their effect on age-adjusted all-cause mortality.

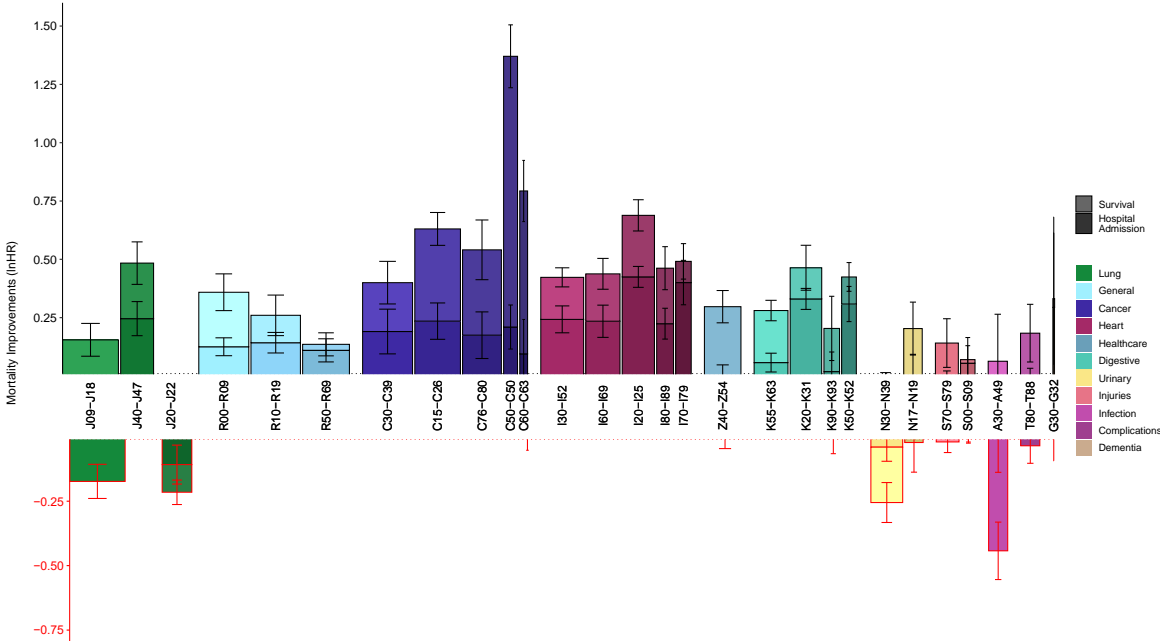


Figure 1: Modelled decade-of-birth upon previous decade-of-birth hospitalisations and survival show large improvements in cancer survival and heart disease incidence but deteriorations in infectious disease. Bars represent the mean improvements in hospital admission rate across decades of birth (darker bars), added to changes from 2001 to 2011 in 5-year survival rates following hospital admission (lighter bars). Both measures are expressed in age-adjusted. For definitions of each ICD10 block, see [Table 2](#). Width of the bars represents the relative burden of death of each disease based on total first-time hospital admissions and 5-year mortality; as such, the total area of each bar represents the relative contribution to improvements – or deteriorations – in population mortality. Error bars are standard errors of the Cox model coefficient. G30-G32 had too few hospital admissions to accurately model improvements (Survival: lnHR 0.04, SE 0.14; Hospital admission lnHR 0.29, SE 0.20). Z40-Z54 only showed improvements in survival.

Overall, we see broad consistency in the scale of improvements across decades of birth, except for “Malignant neoplasms of respiratory and thoracic organs” (C30-C39), where we see greater decade-on-decade improvements amongst later decades ([Figure 2](#)). Averaging these individual disease effects on death, using burden of death weightings, we can then compare the modelled death rates with those observed, and see broad correspondence, with the 1935 and 1945 decades, showing the greatest improvements. Overall, our morbidity model suggests individuals from each successive decade of birth experience an average mortality rate of 0.74 (gaining ~3 years of life) compared to the previous decade of birth ([Table 2](#)).

The shape of these disease-modelled mortality improvements by decade of birth broadly track the observed changes ([Figure 2](#)). This is especially apparent when stratifying the improvements by sex: [Supplementary file 17](#) shows a reasonable relationship between the projected morbidity driven mortality and observed mortality (i.e. mortality trends in the study can largely be explained by trends in disease incidence and survival). Across sex and deprivation strata, taking into account disease survival improvements between 2001 and 2011 and all improvements in disease incidence between decades of birth, we find the largest reductions in death are due to improvements in “Ischaemic heart diseases” (I20-I25), “Malignant neoplasms of digestive organs” (C15-C26), and “Malignant neoplasm of respiratory and intrathoracic organs” (C30-C39), while the largest increases in death are due to “Other bacterial diseases” (A30-A49) and “Influenza and pneumonia” (J09-J18) ([Supplementary file 18](#)). In addition, the deterioration in “Other diseases of the urinary system” (N30-N39) morbidity shows a consistent increase with deprivation, while “Other diseases of the digestive system” (K90-K93) shows consistently larger improvements in more deprived classes ([Supplementary file 12 & 19](#)).

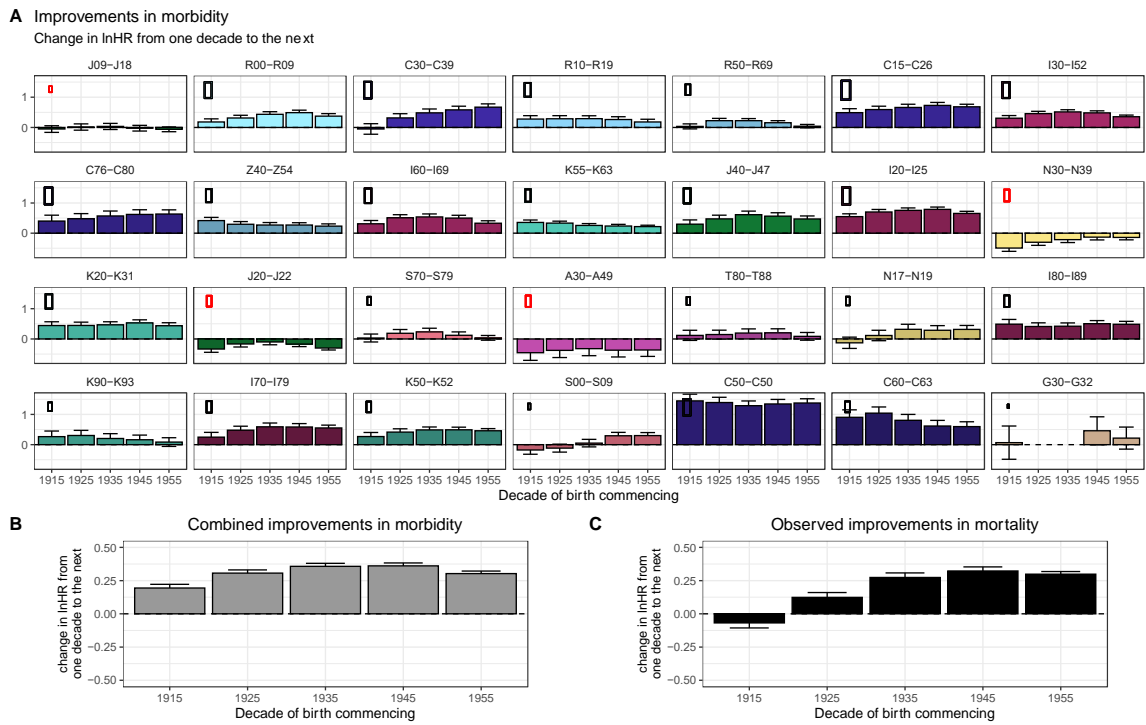


Figure 2: Modelled decade-of-birth upon previous decade-of-birth mortality reductions due to morbidity changes broadly track observed trends in mortality. Panels represent the combined improvements in hospital admission rate and 5-year mortality rates following hospital admission, expressed in age-adjusted lnHR and split by decade of birth - under the model where change in incidence of disease is modelled by decade of birth and added to the survival effect is the change in subsequent 5 year survival rates from incidences in 2001 and 2011. A) Improvements for each ICD10 disease block (for definitions see [Table 2](#)). Dots here represent the relative contribution of the disease to the overall improvements in morbidity-driven mortality, with larger dots indicating a greater contribution to morbidity improvements. A red circle around the dot indicates a negative contribution (i.e. deterioration). B) Modelled trend in deaths based on the weighted morbidities from the panels above. Diseases have been ordered by their burden of death ([Table 2](#)), so smaller bars in early panels may have similar effect on the grey bar average (indicated by the dot size) as larger bars in later panels. C) Observed trend in actual deaths from death records, by decade of birth, for comparison. See [Supplementary file 18](#) for this graph stratified by sex and deprivation.

Overall, we estimate 61.2% (95% CI 29.9%–92.6%) of the improvement in mortality rates was due to improvements in survival following hospital admission, with the balance arising from reduced (age-adjusted) admission rates ([Table 2](#)). Improved outcomes for cancers (C) were particularly driven by post-admission survival, especially C60-C63 (88% of mortality improvement attributable to survival rather than incidence), C50 (85%) and C76-80 (68%), whereas for cardiovascular diseases (I) the balance was more even, as seen in I80-89 (52%), I60-I69 (46%), I30-I32 (43%), I20-I25 (38%).

As previously noted, disease severity was defined as the log hazard ratio for subsequent all-cause mortality among those with a previous admission for an index group of conditions compared with those with no such admission. We regarded the rate of improvement in disease severity over time as being constant if there was the same relative fall in log hazard rate over successive time periods (so for example we regarded a fall in lnHR from 0.6 to 0.3 as equivalent to a fall from 0.3 to 0.15). Assuming the improvements in survival following hospitalisation continue for the coming decade, and differences between incidence in birth cohorts remains the same, we project a 21% slowing of improvements in mortality (-0.242 lnHR c.f. -0.305 lnHR; [Table 3](#)). Essentially, at least arithmetically, the population mortality benefits from improved cancer treatments in 2001-2011 will be hard to repeat as so much benefit has already accrued. Admittedly, this is a consequence of our model: essentially judging it equally difficult to reduce 50 excess deaths following cancer hospital admission associated to 25, as it was to reduce from 100 to 50, and as such should be considered speculative. On the other hand – our model is clearly valid *in extremis*: if all excess cancer deaths were eliminated, no further cancer driven improvement in mortality would be possible.

Stratified	Group	Current Improvements			Projected Improvements		
		Hospital	Five Year	Combined	Hospital	Five Year	Combined
		Admission	Mortality After		Admission	Mortality After	
Rate	Admission	Rate	Admission	Rate	Admission	Combined	
None		-0.1182	-0.1866	-0.3047	-0.1182	-0.1235	-0.2418
Sex	M	-0.1428	-0.1913	-0.3340	-0.1428	-0.1273	-0.2701
Sex	F	-0.0971	-0.1823	-0.2794	-0.0971	-0.1130	-0.2101
SIMD	1	-0.1154	-0.2204	-0.3360	-0.1154	-0.1280	-0.2433
SIMD	2	-0.1157	-0.1592	-0.2743	-0.1157	-0.0456	-0.1610
SIMD	3	-0.1202	-0.1186	-0.2392	-0.1202	-0.0370	-0.1572
SIMD	4	-0.1418	-0.1759	-0.3201	-0.1418	-0.0541	-0.1982
SIMD	5	-0.1231	-0.1731	-0.2958	-0.1231	-0.0914	-0.2145
SIMD	6	-0.1305	-0.1928	-0.3227	-0.1305	-0.0998	-0.2303
SIMD	7	-0.1233	-0.1787	-0.3044	-0.1233	-0.1109	-0.2343
SIMD	8	-0.1305	-0.2097	-0.3402	-0.1305	-0.1422	-0.2726
SIMD	9	-0.0926	-0.1619	-0.2546	-0.0926	-0.1033	-0.1959
SIMD	10	-0.0950	-0.1970	-0.2920	-0.0950	-0.1092	-0.2043

Table 3: Mean (over birth decades) decade of birth on decade of birth improvements in morbidity for the study period, and projections into the subsequent decade by sex and deprivation. Mortality improvements were estimated from morbidity records by combining the mean improvement in hospitalisation rate across birth cohorts and the improvement in disease severity between 2001 and 2011. This was then projected forward assuming improvements in age-adjusted hospitalisation rate between birth cohorts remained constant and improvements in severity remained proportional to the (now reduced) overall mortality of the disease group. Units are in lnHR.

Discussion

In a study of 1,967,120 lives and 10,718,084 hospital admissions, we observed a median age at death of 82.2/85.2 for men/women in the highest socioeconomic decile, and 11.1/8.6 years less for the lowest decile. Cancers (C), cardiovascular disease (I), respiratory diseases (J), and unclassified symptoms and signs (R) were the principal ICD10 chapters recurring in the top 28 disease blocks where hospital admission was associated with the greatest subsequent all-cause mortality, which was a product of the rate of first hospital admission with group of conditions and of all-cause mortality in the five years following admission with that condition. Specifically, our top five causes of hospitalisations associated with subsequent burden of all-cause deaths were, in descending order, 'Influenza and pneumonia' (more common and with higher subsequent mortality than the average T-28 disease), 'Symptoms and signs involving the circulatory and respiratory systems' (common), 'Malignant neoplasm of respiratory and intrathoracic organs' (higher mortality), 'Symptoms and signs involving the digestive system and abdomen' (common), and 'General symptoms and signs' (common and higher mortality). Whilst the latter might appear a benign diagnosis, our results suggest it is a fairly strong and frequent marker of subsequent all-cause mortality.

Across decades of birth, we modelled a reduction in mortality hazard of 0.737 (95% CI 0.730–0.745) due to improvements in morbidity, which broadly tracked improvements in observed mortality. The modelled improvement was 61% accounted for by reduction in excess mortality subsequent to admission and 39% accounted for by a fall in incidences of disease (as measured by hospital admission rates). The important (i.e. burden-of-death weighted) improvements in incidence were driven by cancers and heart disease, whilst improvement in outcomes following admission were mostly driven by cancer, particularly breast and prostate cancer. In contrast, we found deteriorations in the incidence of bacterial disease and in mortality following admission for respiratory and urinary infections. Levels of morbidity and mortality varied strongly across socioeconomic groups, but patterns in **changes** of such were generally less

apparent. Men showed greater rates of improvement in mortality and morbidity than women, with lung and throat cancers contributing most to male improvements and IHD contributing most to female improvements.

In conclusion, we find trends in morbidity appear to partly explain trends in mortality. The progress in prevention and cure within oncology and prevention of heart disease account for the greatest parts of mortality improvement in 2001-2016, and our model suggests mortality improvements may slow, simply because the absolute effect of progress in treatment of these diseases will be difficult to repeat. However, there is scope for further improvements in life expectancy, especially if new progress is made in the treatment of other diseases associated with death, or if prevention initiatives accelerate.

Strengths and weaknesses of the study

This study has avoided some of the known issues with cause of death recording[136] since it does not use cause-specific mortality and tracks wider disease effects and subsequent mortality (such as frailty) beyond direct causes of death, by combining hospitalisation and death records. Implicit tracking of underlying causes through an associated effect (admission to hospital for a disease) may improve estimates of trends in mortality, even if the underlying cause is obscure. We are also able to partition trends in deaths due to a disease based on trends in prevalence and incidence, which has been done for IHD[137], but not simultaneously across diseases in the same dataset. Also, our results are unaffected by population shifts as we excluded immigrants into Scotland after 2001, and instead reflect trends within the defined groups. Combined with the scale of our data, this consistent tracking has enabled us to make like-for-like comparisons of the mortality outcomes of different disease classes across socioeconomic groups and their trends over time.

However, this study also has a number of limitations, relating to the population under study, the definition of diseases, considerations with hospital admission

data, and modelling assumptions. Further discussion of these limitations can be found in [Supplementary file 20](#).

In brief, we excluded migrants out of Scotland because their subsequent trajectory (especially death) could not be tracked. Migrants may be healthier than the average individual and excluding them could therefore overestimate the incidence of disease and death in the population we studied. However, our observed trends should remain unaffected if migration patterns did not change significantly during the study.

Secondly, for practical reasons we grouped the main diagnoses of hospital admissions by ICD-10 chapters and excluded any secondary diagnoses. As a result, we are not able to comment on the trends of individual diseases within chapters (which could offset each other) nor the trends or effects of comorbidities. The latter may affect our results if comorbidities have changed over time or by socioeconomic status; for example, a decline in lung cancer over time as a competing risk for heart disease would inflate the observed improvements for heart disease. However, this is likely to be partially mitigated by reductions in mortality for individuals not admitted for heart disease. Future work may account for comorbidities more explicitly by using competing risk regression and site-specific survival.

Thirdly, the first hospital admission on record and its date is only a proxy for incidence of severe disease. Excluding subsequent hospital admissions may understate the burden of diseases which have recurring episodes (such as influenza), although trends in these diseases are unlikely to be affected given our definitions remained constant. Conversely, diseases such as dementia and multiple sclerosis which are generally managed in the community are unlikely to result in an (immediate) admission to hospital and are therefore not captured accurately in our study. Examining trends in these chronic diseases through GP records was outside the scope of this study, but integration of our results with future work on GP records is likely to refine overall morbidity estimates.

Another consideration with hospital records is their indirect link to death. This relationship can be confounded on the one hand by other health risks and lifestyle factors, and on the other hand by coding inaccuracies and changes in admission policies and screening. Our stratified analysis by sex and socioeconomic status partially mitigates the former, and coding inaccuracies are unlikely to affect disease trends if these inaccuracies are stable over time. There is evidence that screening policies and hospital usage has changed during the study period, but their influence is limited and the opposite effects on disease incidence and survival will mostly offset each other when looking at the effect on morbidity (e.g. influenza) ([Supplementary file 20](#)). However, some caution is needed when interpreting the exact split between improvements in disease incidence and survival.

Lastly, our model assumed 1) disease incidence is a function of year of birth, 2) survival after hospital admission is a function of year of incidence, and 3) these hazards are proportionate. The first two assumptions are a simplification, but necessary given year of birth and year of incidence are completely confounded for a given age at incidence. As to the third point, while disease status itself is not always strictly a proportional hazard, trends in incidence hazard ratios between birth decades and survival hazard ratios between years of hospital admission should still be captured appropriately ([Supplementary file 20](#)).

Strengths and weaknesses in relation to other studies

There was a degree of correspondence in the principal burdens assessed here and a recent study by the Scottish Burden of Disease study (SBD)[120]. This study used the same population and the same study period but assessed YLL (weighting young deaths more as opposed to our method which counted all deaths equally), included individuals younger than 35 years old, and used different disease groupings. Their principal burdens were IHD (ranked 13th in our list of burdens), tracheal, bronchus and lung cancers (3rd), chronic obstructive pulmonary disease (12th), stroke (10th), and Alzheimer's disease (-). Aside from Alzheimer's disease,

discussed below, much of the distinction appears to arise from our observation of an association between death and admissions with indistinct diagnosis (not considered a valid specific cause of death by SBD). In the case of influenza and pneumonia, differences arise due to our study identifying a marker of frailty as well as a direct cause of death, combined with SBD grouping influenza and pneumonia under lower respiratory infections. A relative strength of our study stems from usage of incident morbidity (as marked by hospitalisation) in advance of death, based on recorded diagnosis at the time of hospital visit, thus tracking remote effects such as long term frailty rather than cause of death (which has known limited accuracy, particularly at older ages[136]). However, the principal strength arises from the ability to distinguish trends in incidence of morbidity from trends in subsequent survival. On the other hand, a relative weakness is that we are reliant on hospital admission as a marker of incidence; therefore, diagnosed or latent (presumably milder) cases in the absence of admission are not visible to us, leading for example to significant discrepancy with SBD in the apparent relative burden of Alzheimer's disease, likely due to an understatement of its importance in our results.

The closing gap in mortality between the sexes and its widening across social classes observed in our study is consistent with recent findings from the Office of National Statistics, summarised by Torjesen[138], which looked at socioeconomic deprivation in England and Wales. Similarly, a recent study of health inequality in England found rising levels of lifespan inequality across socioeconomic groupings arising from increasing inequalities across a broad span of causes of death[139]. These studies had the advantage of a larger sample size (~7.5 million deaths cf. 600,000 in our study) and could therefore track trends in mortality and cause of death between stratified groups more accurately. However, Scotland's unique linkage of death records and electronic health records through eDRIS allowed us to directly model changes disease mortality at an individual level (avoiding issues with cause of death recordings and shifts in population demographics). Our study has the advantage of partly explaining these trends in mortality inequality through changes in disease incidence and survival: men experienced greater

improvements in incidence of lung cancer and survival following heart disease hospitalisation compared to women, while more socially deprived individuals (men and women combined) suffered worse deteriorations in infectious disease, especially for the incidence and survival of hospitalisation for urinary tract infections. However, in contrast to Bennet *et al.*[139], we do not find a clear pattern in overall morbidity improvements across socioeconomic deciles in Scotland, and we do not observe a widening inequality in cancer, respiratory and Alzheimer's disease morbidity within our study population, although we are underpowered to detect the latter and our disease groupings were not identical.

Lastly, a recent study of coronary heart disease mortality in Scotland, using a sophisticated model to apportion improvement between prevention and treatment, found improvements for coronary heart disease between 2001-2010 were similar across social classes, and reported 33%–61% of these improvements could be attributed to advances in treatment[137]. Given the very different methods, albeit studying the same population, there is reasonable concordance with our own study: we find roughly equal improvements in heart disease across social classes and estimate 38% (95% CI 28%–48%) of these improvements stem from increased survival after hospitalisation for ischaemic heart disease. Hotchkiss *et al.*[137] are able to further partition improvements by uptake of primary and secondary prevention drugs and treatments. Such detailed analysis of specific diseases has been beyond the scope of our study.

Implications for clinicians and policymakers

Much of the improvements in mortality observed in Scotland between 2001–2016 can be attributed to reductions in morbidity, as captured by hospital admissions. While this study examined mortality and morbidity in the Scottish population only, there is a substantial concordance in mortality trends across high-income countries[127], as well as similarities in disease-related mortality trends between Scotland and the rest of the UK[126], warranting similar studies to be performed in other high-income countries. It is a testament to healthcare services that the majority of mortality improvements appear to stem from

advances in disease survival post-admission. Observed improvements in cancer incidence and survival – especially breast and prostate – coincide with a continued effort within Scotland[140], the UK[141], and other high-income nations[142] to improve prevention and care of these diseases. However, the rapid advances in survival of both heart disease and cancer modelled by our study between 2001 and 2011 will be hard to continue to the same extent, as so much progress has already been made. At the same time, the observed deteriorations in infectious disease coincide with global increases in antimicrobial resistance[143] and emphasise the need to prioritise research in this area: infectious disease will become a larger contributor to mortality and may contribute to a widening of health inequalities between socioeconomic classes. If these current trends in morbidity continue, we expect morbidity-driven improvements in mortality to slow down. However, the life expectancy gap between Scotland and other high-income countries[144] suggests further mortality improvements are possible. The rate of this improvement will hinge upon whether advances in all major diseases categories – especially infectious disease – can catch up with the progress we have recently seen on heart disease and cancer, and whether preventable deaths from external causes (such as suicide and drug-related deaths), which cannot be accurately tracked using hospital admissions, decrease rather than rise.

Funding

The study was funded by the Lloyds Banking Group, for the creation, curation and dissemination of knowledge in the public interest, in particular to improve estimates of future population size and morbidity and mortality rates to facilitate healthcare and other government planning. All analyses stratified by socioeconomic deprivation (i.e. [Table 2](#), [Table 3](#), [Supplementary files 7, 9–12, 14, 16, 18, and 19](#)) were confidentially re-analysed using 10 socioeconomic groups specified by Lloyds, rather than the Scottish Index of Multiple Deprivation. No Lloyds employees were granted access to individual patient data. Data access was granted to University of Edinburgh researchers only and only through the

national safe haven by the Public Benefit and Privacy Panel for Health and Social Care under application number 1617-0255/Joshi.

The study and its design were conceived by the last author. The funder reviewed the design and said that adding stratification by socioeconomic status was a key requirement for meaningful analysis and their funding was conditional upon this, a request to which the authors readily agreed. The funder was kept informed of interim analyses and reviewed the draft manuscript, occasionally suggesting additional analyses or requesting clarifications. The funders were not permitted to and did not request the removal of any results.

The last author and authors affiliated with eDRIS and NHS Scotland were compensated by the University of Edinburgh using the grant from Lloyds Banking Group, in line with normal pricing for the work undertaken. PRHJT was funded by the MRC Doctoral Training Programme (MR/N013166/1) and the University of Edinburgh College of Medicine and Veterinary Medicine. JFW was funded by the MRC QTL in health and disease. Apart from Lloyds Banking Group, the funders had no role in the study design, data collection, analysis, and interpretation, or the decision to submit the work for publication.

Acknowledgements

We would like to thank Lloyds for the funding and in particular Craig Butler and Stuart McDonald for their engagement with the project. We would also like to thank Steve Pavis and Doug Kidd for their guidance with the Public Benefit and Privacy Panel for Health and Social Care.

Competing interests

PKJ reports grants from Lloyds Banking Group, PLC, during the conduct of the study. This grant was awarded to the University of Edinburgh and used in part to pay for research costs of JJK and CMF. PKJ also reports shares in Lloyds Banking

Group, PLC, as part of a diversified portfolio. The remaining authors declare no competing interests.

Author contributions

PRHJT—Formal Analysis, Software, Visualisation, Writing - Original Draft, Writing - Review & Editing. JJK—Data Curation, Writing - Original Draft. JWM - Writing - Review & Editing. IG—Writing - Review & Editing. JFW—Supervision, Writing - Review & Editing. HC—Writing - Review & Editing. CMF—Conceptualization, Writing - Review & Editing. PKJ—Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Supervision, Validation, Writing - Original Draft, Writing - Review & Editing.

Data sharing

Individual de-identified participant data cannot be shared. Statistical code and technical details are available upon request from the corresponding author.

2.3 Conclusion

Studying hospital records and death records from 2 million adults residing in Scotland, we found substantial differences in age-adjusted mortality between men and women, between individuals from different socioeconomic backgrounds, and between individuals born in different decades of birth. The latter shows population mortality has fallen over time, with individuals born in 1965 living on average 3 years longer than those born a decade earlier, who in turn live an average of 3 years longer than individuals born the decade before. Regardless of their year of birth, Scottish women lived on average 3.5 years longer than men, and individuals from the least deprived areas lived on average 7 years longer than those from the most deprived areas. As such, decade of birth, socioeconomic status, and sex can be considered major determinants of human lifespan.

We were able to explain the differences in survival in part by examining the diseases responsible for the most deaths: infections, CVD, cancers, hip/neck injuries, and disorders involving the lungs, digestive tract, and kidneys. For example, the higher mortality in men compared to women coincided with a higher incidence of most of these deadly diseases in men. Similarly, individuals from more socioeconomically deprived areas had a far greater incidence of disease, although their survival after hospital admission was not substantially different from individuals from less socioeconomically deprived areas. When looking at the effect of trends in disease on trends in mortality over time, we found the greatest gains in mortality improvements were achieved by lowering the incidence of CVD and cancers, as well as improving the survival of patients after a cancer diagnosis (especially for breast and prostate cancers). For men, improvements in CVD incidence were greater than women, coinciding with a narrowing of the life expectancy gap between men and women in the UK. We did not see a clear pattern in changing disease trends between socioeconomic strata from their different initial levels, even though data from the UK National Office

for Statistics suggested the discrepancies in life expectancy between socioeconomic strata are increasing.

Improvements in non-communicable diseases were contrasted by a deterioration in viral and bacterial infections, across sex and socioeconomic strata. For urinary infections, the deterioration was more pronounced for more socioeconomically deprived individuals. While we could not capture Alzheimer's disease through hospital admission records, related studies highlight it as a growing contributor to population mortality. Extrapolating current trends forwards, we estimated a 21% slowdown in morbidity improvements in the next decade, as we have already made significant advances towards preventing CVD and cancer-related deaths.

Related work on life expectancy in Scotland has highlighted the steady three year per decade growth in lifespan has slowed since 2011. We did not observe a slowing trend in 5-year mortality during our study period, but this could be because we lacked the data to reliably estimate mortality rates after 2011. However, our results do predict a slowdown in morbidity improvements should current trends continue. If they do, it is likely the rate of mortality improvements will decline in response. Mitigating this slowdown in mortality improvements and preventing deteriorations in life expectancy in the future can be achieved if the focus of public health policy is renewed towards preventing and treating infectious disease, taking note of the recent rise of antibacterial resistance and the greater burden of these infectious disease in the elderly. In addition, we recommend further efforts towards improving the diseases we have highlighted in this study: despite recent improvements, CVD and cancers remain some the largest contributors to mortality and addressing them has the greatest potential to extend healthy lives.

Finally, it is clear the importance of diseases in determining lifespan has shifted as preventative measures and treatment have reduced disease incidence and increased survival after diagnosis. However, it is less clear whether the

improvements in disease-driven mortality are due to a delay in the ageing process itself, or if patients have simply exchanged acute and deadly diseases for chronic and progressive diseases. This distinction is crucial, as the former indicates the period of morbidity is being delayed further towards the end of life, while the latter indicates a trend of extended age-related morbidity and disability. While our study can only comment on 28 disease categories, the advances in age-adjusted incidence we observed for most of these diseases suggest that, at a minimum, the (age-standardised) health of individuals has improved compared to previous decades. Future studies may stratify trends in disease by age to more explicitly test the hypothesis of a delay in the ageing process.

The nature of the ageing process and ways to slow it down is under active discussion in the field of biogerontology[145]. While researchers agree that ageing involves a progressive loss of functionality resulting in disease and eventually death, the origin of this decline is under debate. It is generally accepted that organisms are increasingly likely to die of predation or external causes with age (thus no longer contributing to the gene pool), which results in strong natural selection during development and a progressive decline in the strength of selection in later age[146].

Among the leading theories of ageing is the mutation accumulation theory[147], which proposes that the declining strength of natural selection implies mutations which are harmful later in life are not effectively eliminated. The antagonistic pleiotropy theory extends this idea to suggest mutations that lead to faster growth and higher fertility early in life will be selected for, even if these same mutations are detrimental to late life survival[146]. These theories suggest the ageing process is a stochastic accumulation of damage stemming from late-acting, harmful mutations and implies that improving cellular repair mechanisms to counteract the accumulation of damage could slow down ageing and prevent age-related disease.

In contrast to these so-called error theories is the hyperfunction theory of ageing[148], which proposes that ageing is programmed to some extent. Experiments on model organisms have found that developmental pathways, such as yolk production, continue to be active in late life[149,150]. The theory proposes that the declining strength of natural selection provides an evolutionary drive for effective development but no such drive to switch these pathways off after completion, when they actively start to damage the organism. By extension, this implies age-related damage and disease can be prevented by switching off or decreasing the activity of developmental pathways at the appropriate time. In the following Chapters, I explore the human ageing process further by examining the link between lifespan and genetics.

Chapter 3: Genome-wide association of lifespan in UK Biobank and LifeGen

3.1 Introduction

3.1.1 Context

The existence of a heritable component to human lifespan, albeit small, suggests there may be genetic variants which play a role in determining lifespan. Identifying these variants and characterising exactly how they influence survival has the potential to improve our understanding of the interplay between morbidities and mortality on a biological level, which the epidemiological work on Scottish patients in Chapter 2 did not reveal. Knowledge of the genetic determinants of lifespan could also uncover targets for clinical or pharmaceutical intervention, such as genes and pathways, capable of preventing the slowdown in mortality improvements and accelerating prevention of age-related disease.

After a decade of limited success trying to study cases of exceptional longevity, advances in genetic discovery for human survival were recently made by using a kin-cohort study design in which the genotypes of subjects were tested against a parental lifespan phenotype[89,90]. Less than a year later, a large consortium was established to study parental lifespan in 24 cohorts averaging 14,000 lifespans each (LifeGen). Together with the first release of UK Biobank, this increased the total sample to 600,000 and doubled the number of loci that could be shown to determine lifespan in multiple populations from two to four[70].

The full release of the genetic information of UK Biobank individuals in July 2017 brought with it the opportunity to expand previous analyses, both in terms of sample size and scope. In this Chapter, I use UK Biobank and LifeGen data to reveal new genetic variants, genes, and pathways important in determining lifespan, and compare the genetics of age-related disease to mortality.

The results from this Chapter are also used in the next Chapter, where I combine our findings with recent work on other ageing-related traits to more closely examine how lifespan genetics relates to human ageing.

3.1.2 Contributions

The idea of studying parental lifespan using cox survival models was originally conceived by Peter Joshi and made possible through the use of Martingale residuals (suggested by Krista Fischer)[90]. David Clark downloaded and decrypted the raw UK Biobank phenotypic and genotypic data for the study.

Prior to the public release of this data, David, Andrew Bretherick, Peter, and I developed a pipeline to perform the genome-wide association study. The initial framework of the pipeline was composed of scripts by David and Peter from previous GWAS they performed and included scripts to perform quality control on UK Biobank lifespan data. I optimised the pipeline to overcome the technical hurdles associated with the scale of the full UK Biobank phenotypic and genotypic datasets. Specifically, I wrote adaptive functions that would read in only the phenotypes necessary for the GWAS using superior data reading software and split the genotypic data into chunks that could be analysed in parallel.

This pipeline performed the quality control steps and regression analyses of the study, including the GWAS and age- and sex-specific analysis. The multivariate analysis between mother and father lifespans was performed by Xia Shen. The iGWAS analysis, DEPICT and PASCAL pathway analyses, and the age-related QTL enrichment analysis were performed by Ninon Mounier and Zoltán Kutalik, with the eQTL data originating from the eQTLgen consortium. The longevity replication and polygenic risk score associations with lifespan were done by Peter. The SMR-HEIDI eQTL/mQTL prioritisation and SOJO fine-mapping were done by Zheng Ning and Xiao Feng. The replication of polygenic risk scores in the Estonian Biobank were performed by Kristi Läll and Krista Fischer. Each co-

author also contributed by writing up the details of their analysis in the method section.

I performed all other analyses, which included writing scripts to perform quality control on the input data, reformatting data if using existing software (VEGAS2Pathway, Stratified LDSC, PRSice), checking results and compiling them for publication, and drawing figures.

For clarity, I was responsible for the following Results sections of the article:

- Genome-wide association analysis (excluding SSE)
- Sex and age-specific effects
- Disease and lifespan
- Cell type and pathway enrichment (excluding DEPICT, PASCAL, and eQTLs)
- Out-of-sample lifespan PRS associations (excluding results described in first three paragraphs)

I compiled the results provided by co-authors, creating all tables and figures (except Figure 8). I also drafted the majority of the manuscript, excluding sections by co-authors described above. Peter wrote the *eLife* digest and made significant contributions to the Discussion regarding assortative mating, differences in lifespan and longevity, transcriptomic age, actuarial use of polygenic risk scores, and the considerations of a linear mixed model in the context of a kin-cohort study. All co-authors provided feedback on the draft manuscript, especially Peter, Jim Wilson, and Zoltán Kutalik.

Lastly, I would also like to acknowledge the comments from unnamed reviewers from *Nature Communications* and *eLife*, whose feedback prompted a large restructuring of the article which ultimately made the results much clearer and easier to follow.

3.2 Published article

This work was published as an article in the journal *eLife* on 15 January 2019 after completing formal peer review. A copy of the Author Accepted Manuscript prior to proofing is included below, provided under the terms of the Creative Commons Attribution License CC BY 4.0.

Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances

Paul R.H.J. Timmers^a, Ninon Mounier^{b,c}, Kristi Läll^{d,e}, Krista Fischer^d, Zheng Ning^f, Xiao Feng^g, Andrew Bretherick^h, David W. Clark^a, eQTLGen Consortium¹, Xia Shen^{a,f,g}, Tõnu Esko^{d,i}, Zoltán Kutalik^{b,c}, James F. Wilson^{a,h} (Supervisor), Peter K. Joshi^{a,b} (Supervisor)

a) Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, EH8 9AG Edinburgh, United Kingdom

b) Institute of Social and Preventive Medicine, University Hospital of Lausanne, 1010 Lausanne, Switzerland

c) Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

d) Estonian Genome Center, University of Tartu, 51010 Tartu, Estonia

e) Institute of Mathematics and Statistics, University of Tartu, 50409 Tartu, Estonia

f) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden

g) State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

h) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, EH4 2XU Edinburgh, United Kingdom

i) Broad Inst of Harvard and MIT, Cambridge, 02142 MA, USA

1) See additional information

The formatted article, detailed methods, and supplementary information can be found at *eLife*. Available at: <https://doi.org/10.7554/eLife.39856>

Abstract

We use a genome-wide association of 1 million parental lifespans of genotyped subjects and data on mortality risk factors to validate previously unreplicated findings near *CDKN2B-AS1*, *ATXN2/BRAP*, *FURIN/FES*, *ZW10*, *PSORS1C3*, and 13q21.31, and identify and replicate novel findings near *ABO*, *ZC3HC1*, and *IGF2R*. We also validate previous findings near 5q33.3/*EBF1* and *FOXO3*, whilst finding contradictory evidence at other loci. Gene set and cell-specific analyses show that expression in foetal brain cells and adult dorsolateral prefrontal cortex is enriched for lifespan variation, as are gene pathways involving lipid proteins and homeostasis, vesicle-mediated transport, and synaptic function. Individual genetic variants that increase dementia, cardiovascular disease, and lung cancer – but not other cancers – explain the most variance. Resulting polygenic scores show a mean lifespan difference of around five years of life across the deciles.

Introduction

Human lifespan is a highly complex trait, the product of myriad factors involving health, lifestyle, genetics, environment, and chance. The extent of the role of genetic variation in human lifespan has been widely debated[151], with estimates of broad sense heritability ranging from around 25% based on twin studies[31,152,153] (perhaps over-estimated[154]) to around 16.1%, (narrow sense 12.2%) based on large-scale population data[40]. One very recent study suggests it is much lower still (<7%)[41], pointing to assortative mating as the source of resemblance amongst kin.

Despite this modest heritability, extensive research has gone into genome-wide association studies (GWAS) finding genetic variants influencing human survival, using a variety of trait definitions and study designs[83,89,90,155–159]. GWAS have primarily focused on extreme cases of long-livedness (longevity) – individuals surviving past a certain age threshold – and scanning for differences in genetic variation from controls. While this case-control design has the advantage of focusing on highly statistically-informative individuals, who also often exhibit extreme healthspan and have potentially unique genetic attributes[29,160], the exceptional nature of the phenotype precludes collection of large samples, and differences in definitions of longevity complicate meta-analysis. As a result, only two robustly replicated, genome-wide significant associations (near *APOE* and *FOXO3*) have been made to date[157,161].

An alternative approach is to study lifespan as a quantitative trait in the general population and use survival models (such as Cox proportional hazards[134]) to allow long-lived survivors to inform analysis. However, given the incidence of mortality in middle-aged subjects is low, studies have shifted to the use of parental lifespans with subject genotypes (an instance of Wacholder's kin-cohort method[162]), circumventing the long wait associated with studying age at death in a prospective study[89,90]. In addition, the recent increase in genotyped population cohorts around the world, and in particular the creation of UK Biobank[163], has raised GWAS sample sizes to hundreds of thousands of

individuals, providing the statistical power necessary to detect genetic effects on mortality.

A third approach is to gather previously published GWAS on risk factors thought to possibly affect lifespan, such as smoking behaviour and cardiovascular disease (CVD), and estimate their actual independent, causal effects on mortality using Mendelian Randomisation. These causal estimates can then be used in a Bayesian framework to inform previously observed SNP associations with lifespan[164].

Here, we blend these three approaches to studying lifespan and perform the largest GWAS on human lifespan to date. First, we leverage data from UK Biobank and 26 independent European-heritage population cohorts[70] to carry out a GWAS of parental survival, quantified using Cox models. We then supplement this with data from 58 GWAS on mortality risk factors to conduct a Bayesian prior-informed GWAS (iGWAS). Finally, we use publicly available case-control longevity GWAS statistics to compare the genetics of lifespan and longevity and provide collective replication of our lifespan GWAS results.

We also examine the diseases associated with lifespan-altering variants and the effect of known disease variants on lifespan, to provide insight into the interplay between lifespan and disease. Finally, we use our GWAS results to implicate specific genes, biological pathways, and cell types, and use our findings to create and test whole-genome polygenic scores for survival.

Results

Genome-wide association analysis

We carried out GWAS of survival in a sample of 1,012,240 parents (60% deceased) of European ancestry from UK Biobank and a previously published meta-analysis of 26 additional population cohorts (LifeGen[70]; [Table 1—source data 1](#)). We performed a sex-stratified analysis and then combined the allelic effects in fathers and mothers into a single parental survival association in two ways. First, we assumed genetic variants with common effect sizes (CES) for both parents, maximising power if the effect is indeed the same. Second, we allowed for sex-specific effect sizes (SSE), maximising power to detect sexually dimorphic variants, including those only affecting one sex. The latter encompasses a conventional sex-stratified analysis, but uses only one statistical test for the much more general alternative hypothesis that there is an effect in at least one sex.

We find 12 genomic regions with SNPs passing genome-wide significance for one or both analyses ($P < 2.5 \times 10^{-8}$, accounting for the two tests CES/SSE) ([Figure 1; Table 1](#)). Among these are 5 loci discovered here for the first time, at or near *MAGI3*, *KCNK3*, *HTT*, *HP*, and *LDLR*. Carrying one copy of a life-extending allele is associated with an increase in lifespan between 0.23 and 1.07 years (around 3 to 13 months). Despite our sample size exceeding 1 million phenotypes, a variant had to have a minor allele frequency exceeding 5% and an effect size of 0.35 years of life or more per allele for our study to detect it with 80% power.

We also attempted to validate novel lifespan SNPs discovered by Pilling *et al.* in UK Biobank at an individual level[159] by using the LifeGen meta-analysis as independent replication sample. Testing 20 candidate SNPs for which we had data available, we find directionally consistent, nominally significant associations for 6 loci ($P < 0.05$, one-sided test), of which 3 have sex-specific effects. We also provide evidence against 3 putative loci but lack statistical power to assess the remaining 11 ([Figure 2—source data 1](#)).

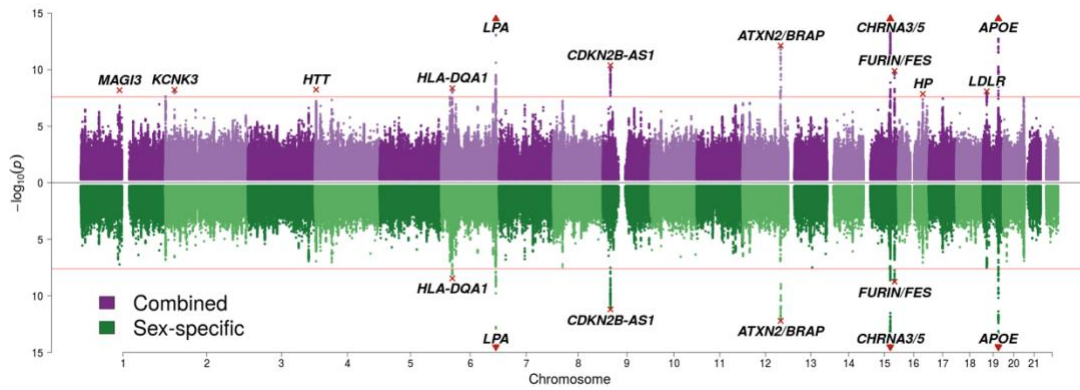


Figure 1: SNP associations with lifespan across both parents under the assumption of common and sex-specific effect sizes. Miami plot of genetic associations with joint parental survival. In purple are the associations under the assumption of common SNP effect sizes across sexes (CES); in green are the associations under the assumption of sex-specific effect sizes (SSE). P refers to the two-sided P values for association of allelic dosage on survival under the residualised Cox model. The red line represents our multiple testing-adjusted genome-wide significance threshold ($P = 2.5 \times 10^{-8}$). Annotated are the gene, set of genes, or cytogenetic band near the index SNP, marked in red. P values have been capped at $-\log_{10}(p) = 15$ to better visualise associations close to genome-wide significance. SNPs with P values beyond this cap (near *APOE*, *CHRNA3/5* and *LPA*) are represented by triangles.

At or near	rsID	Chr	Position	A1	Freq1	Years1	SE	CES P	PDES P	Disease
MAGI3	rs1230666	1	114173410	G	0.85	0.3224	0.0555	6.4E-09	6.1E-08	Autoimmune
KCNK3	rs1275922	2	26932887	G	0.74	0.2579	0.0443	6.0E-09	2.7E-07	Cardiometabolic
HTT	rs61348208	4	3089564	T	0.39	0.2299	0.0395	5.8E-09	1.2E-07	-
HLA-DQA1	rs34967069	6	32591248	T	0.07	0.5613	0.0956	4.3E-09	3.6E-09	Autoimmune
LPA	rs10455872	6	161010118	A	0.92	0.7639	0.0743	8.5E-25	3.1E-24	Cardiometabolic
CDKN2B-AS1	rs1556516	9	22100176	G	0.50	0.2510	0.0386	7.5E-11	6.4E-12	Cardiometabolic
ATXN2/BRAP	rs11065979	12	112059557	C	0.56	0.2798	0.0393	1.0E-12	6.2E-13	Autoimmune Cardiometabolic
CHRNA3/5	rs8042849	15	78817929	T	0.65	0.4368	0.0410	1.6E-26	1.9E-30	Smoking-related
FURIN/FES	rs6224	15	91423543	G	0.52	0.2507	0.0390	1.3E-10	1.8E-09	Cardiometabolic
HP	rs12924886	16	72075593	A	0.80	0.2798	0.0493	1.4E-08	9.1E-08	Cardiometabolic
LDLR	rs142158911	19	11190534	A	0.12	0.3550	0.0616	8.1E-09	3.3E-08	Cardiometabolic
APOE	rs429358	19	45411941	T	0.85	1.0561	0.0546	3.1E-83	1.8E-85	Cardiometabolic Neuropsychiatric

Table 1: Twelve genome-wide significant associations with lifespan using UK Biobank and LifeGen. Parental phenotypes from UK Biobank and LifeGen meta-analysis, described in [Table 1—source data 1](#), were tested for association with subject genotype. See [Table 1—source data 2](#) for LD Score regression intercept of each cohort separately and combined. Displayed here are loci associating with lifespan at genome-wide significance ($P < 2.5 \times 10^{-8}$). At or near – Gene, set of genes, or cytogenetic band nearest to the index SNP; rsID – The index SNP with the lowest P value in the standard or sex-specific effect (SSE) analysis. Chr – Chromosome;

Position – Base-pair position on chromosome (GRCh37); A1 – the effect allele, increasing lifespan; Freq1 – Frequency of the A1 allele; Years1 – Years of life gained for carrying one copy of the A1 allele; SE – Standard Error; P – the P value for the Wald test of association between imputed dosage and cox model residual; Disease – Category of disease for known associations with SNP or close proxies ($r^2 > 0.6$), see [Table 1—source data 3](#) for details and references. Despite the well-known function of the *HTT* gene in Huntington’s disease, SNPs within the identified locus near this gene have not been associated with the disease at genome-wide significance.

We then used our full sample to test 6 candidate SNPs previously associated with longevity[158,161,165,166] for association with lifespan, and find directionally consistent evidence for SNPs near *FOXO3* and *EBF1*. The remaining SNPs did not associate with lifespan despite apparently adequate power to detect any effect similar to that originally reported ([Figure 2—source data 1](#); [Figure 2](#)).

Finally, we tested a deletion, d3-GHR, reported to affect male lifespan by 10 years when homozygous[167] by converting its effect size to one we expect to observe when fitting an additive model. We used a SNP tagging the deletion and estimated the expected effect size in a linear regression for the (postulated) recessive effect across the three genotypes, given their frequency (see Methods). While this additive model reduces power relative to the correct model, our large sample size is more than able to offset the loss of power, and we find evidence d3-GHR does not associate with lifespan with any (recessive or additive) effect similar to that originally reported ([Figure 2—source data 1](#); [Figure 2](#)).

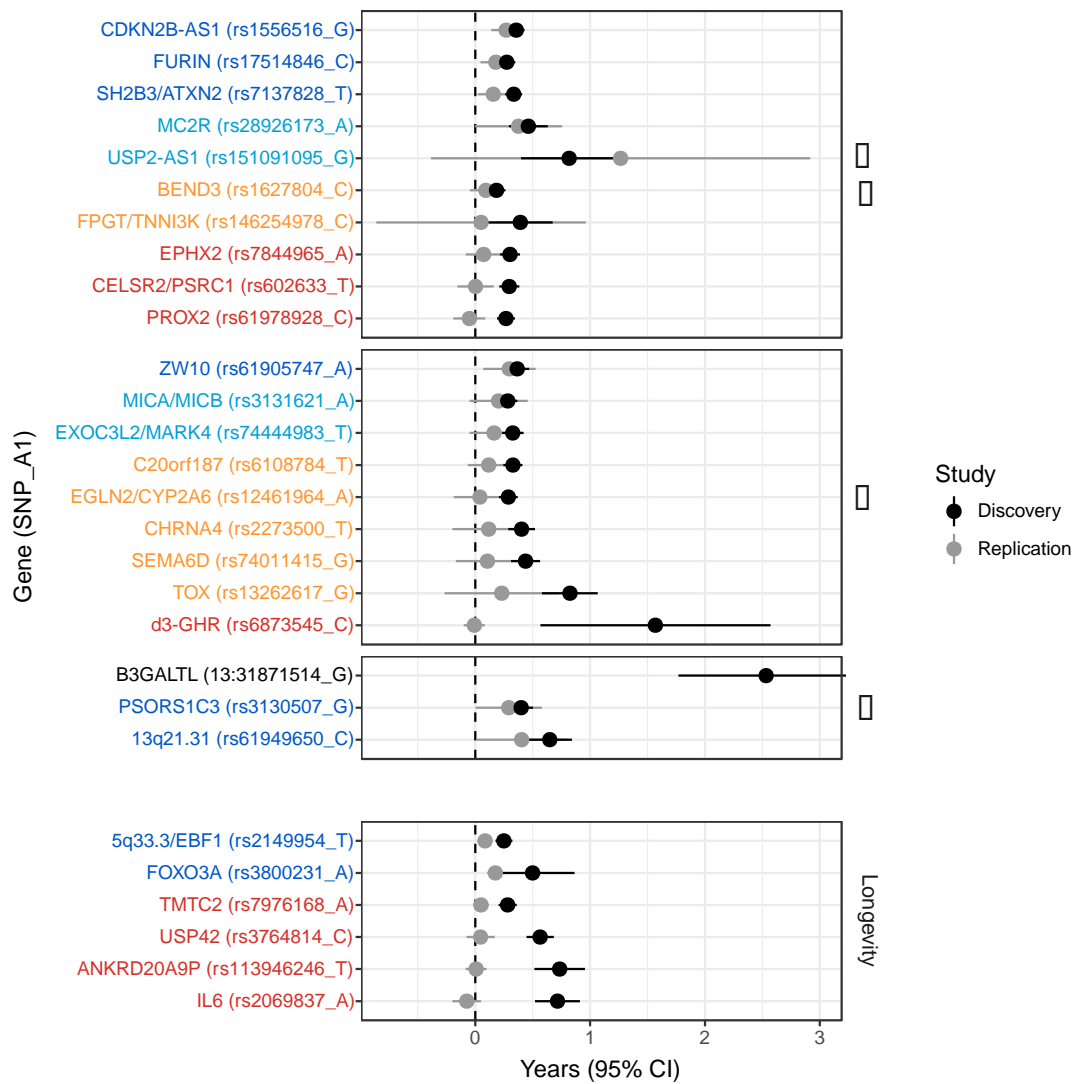


Figure 2: Validation of SNPs identified in other studies using independent samples of European descent. Discovery – Candidate SNPs or proxies ($r^2 > 0.95$) associated with lifespan (top panels, stratified by sex) and longevity (bottom panel) by previous studies^{14,15,18,24–26}. Effect sizes have been rescaled to years of life to make direct comparisons between studies (see Methods and [Figure 2—figure supplement 1](#)). Replication – Independent samples, either the LifeGen meta-analysis to replicate Pilling *et al.*¹⁵, or the full dataset including UK Biobank. Gene names are as reported by discovery and have been coloured based on overlap between confidence intervals (CIs) of effect estimates. Dark blue – Nominal replication ($P < 0.05$, one-sided test). Light blue – CIs overlap ($P_{\text{het}} > 0.05$) and cover zero, but replication estimate is closer to discovery than zero. Yellow – CIs overlap ($P_{\text{het}} > 0.05$) and cover zero, and replication estimate is closer to zero than discovery. Red – CIs do not overlap ($P_{\text{het}} < 0.05$) and replication estimate covers zero. Black – no replication data.

Mortality risk factor-informed GWAS (iGWAS)

We integrated 58 publicly available GWAS on mortality risk factors with our CES lifespan GWAS, creating Bayesian priors for each SNP effect based on causal effect estimates of 16 independent risk factors on lifespan. These included body mass index, blood biochemistry, CVD, type 2 diabetes, schizophrenia, multiple sclerosis, education levels, and smoking traits.

The integrated analysis reveals an additional seven genome-wide significant associations with lifespan (Bayes Factor permutation $P < 2.5 \times 10^{-8}$), of which SNPs near *TMEM18*, *GBX2/ASB18*, *IGF2R*, *POM12C*, *ZC3HC1*, and *ABO* are reported at genome-wide significance for the first time ([Figure 3](#); [Table 2](#)). A total of 82 independent SNPs associate with lifespan when allowing for a 1% false discovery rate (FDR) ([Table 2—source data 2](#)).

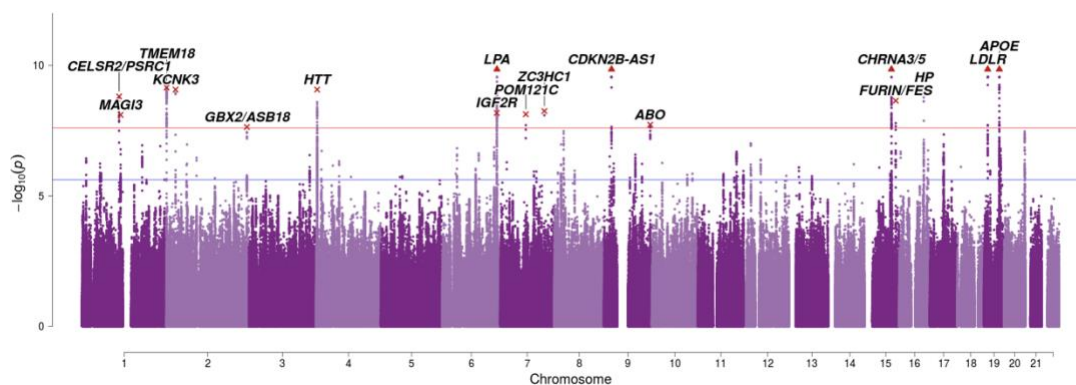


Figure 3: SNP associations with lifespan across both parents when taking into account prior information on mortality risk factors. Bayesian iGWAS was performed using observed associations from the lifespan GWAS and priors based on 16 traits selected by an AIC-based stepwise model. As the P values were assigned empirically using a permutation approach, the minimum P value is limited by the number of permutations; SNPs reaching this limit are represented by triangles. Annotated are the gene, cluster of genes, or cytogenetic band in close proximity to the top SNP. The red line represents the genome-wide significance threshold ($P = 2.5 \times 10^{-8}$). The blue line represents the 1% FDR threshold. [Figure 3—figure supplement 1](#) shows the associations of each genome-wide significant SNP with the 16 risk factors.

At or near	rsID	Chr	Position	A1	Freq1	Years1	SE	CES P	Risk	BF P
CELSR2/PSRC1	rs4970836	1	109821797	G	0.23	0.2234	0.0463	1.4E-06	LDL HDL CAD	1.6E-09
TMEM18	rs6744653	2	628524	A	0.17	0.2772	0.0511	5.8E-08	BMI	7.0E-10
GBX2/ASB18	rs10211471	2	237081854	C	0.80	0.2401	0.0493	1.1E-06	Education	2.3E-08
IGF2R	rs111333005	6	160487196	G	0.98	0.8665	0.1577	3.9E-08	LDL CAD	6.6E-09
POM121C	rs113160991	7	75094329	G	0.78	0.2541	0.0495	2.8E-07	BMI Insulin	7.5E-09
ZC3HC1	rs56179563	7	129685597	A	0.39	0.2107	0.0406	2.1E-07	CAD	5.6E-09
ABO	rs2519093	9	136141870	C	0.81	0.2244	0.0497	6.3E-06	LDL CAD	1.9E-08

Table 2: Bayesian GWAS using mortality risk factors reveals seven additional genome-wide significant variants. At or near – Gene or set of genes nearest to the index SNP; rsID – The index SNP with the lowest P value in the risk factor-informed analysis. Chr – Chromosome; Position – Base-pair position on chromosome (GRCh37); A1 – the effect allele, increasing lifespan; Freq1 – Frequency of the A1 allele; Years1 – Years of life gained for carrying one copy of the A1 allele; SE – Standard Error; CES P – the P value for the Wald test of association between imputed dosage and cox model residual, under the assumption of common effects between sexes. Risk – mortality risk factors associated with the variant ($P < 3.81 \times 10^{-5}$, accounting for 82 independent SNPs and 16 independent factors). BF P – Empirical P value derived from permutating Bayes Factors. See [Table 2—source data 1](#) for the causal estimate of each risk factor. See [Table 2—source data 2](#) for all SNPs significant at FDR < 1%.

As has become increasingly common[159], we attempted to replicate our genome-wide significant findings collectively, rather than individually. This is usually done by constructing polygenic risk scores from genotypic information in an independent cohort and testing for association with the trait of interest subject-by-subject. We used publicly available summary statistics on extreme longevity as an independent replication dataset[157,161], but lacking individual data from such studies, we calculated the collective effect of lifespan SNPs on longevity using the same method as inverse-variance meta-analysis two-sample Mendelian randomisation (MR) using summary statistics[168], which gives equivalent results. Prior to doing this, all effects observed in the external longevity studies were converted to hazard ratios using the *APOE* variant effect size as an empirical conversion factor, to allow the longevity studies to be meta-analysed despite their different study designs (and to be adjusted for sample overlap; see Methods).

Although the focus is on collective replication, our method has the advantage of transparency at an individual variant level, which is of particular importance for researchers seeking to follow-up individual loci. Remarkably, all lead lifespan variants show directional consistency with the independent longevity sample, and 4 SNPs or close proxies ($r^2 > 0.8$) reach nominal replication ($P < 0.05$, one-sided test) ([Figure 4—source data 1](#)). Of these, SNPs near *ABO*, *ZC3HC1*, and *IGF2R* are replicated for the first time, and thus appear to affect overall survival and survival to extreme age. The overall ratio of replication effect sizes to discovery effect sizes – excluding *APOE* – is 0.42 (95% CI 0.23–0.61; $P = 1.35 \times 10^{-5}$). The fact this ratio is significantly greater than zero indicates most lifespan SNPs are indeed longevity SNPs. However, the fact most SNPs have a ratio smaller than one indicates they may affect early mortality more than survival to extreme age, relative to *APOE* (which itself has a greater effect on late-life mortality than early mortality; [Figure 4](#)).

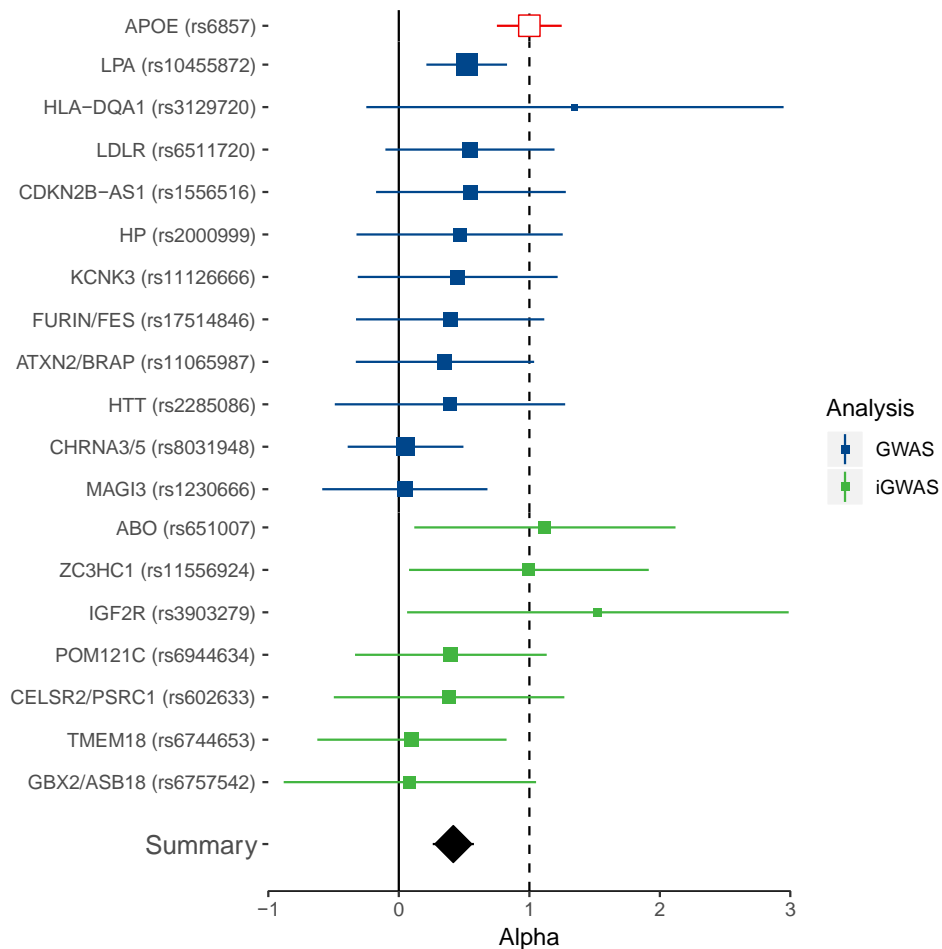


Figure 4: Collective replication of individual lifespan SNPs using GWAMAs for extreme long-livedness shows directional consistency in all cases. Forest plot of effect size ratios between genome-wide significant lifespan variants from our study and external longevity studies^{11,18}, having converted longevity effect sizes to our scale using APOE as benchmark (see Methods and [Figure 4—source data 1](#)). Alpha – ratio of replication to discovery effect sizes on the common scale and 95% CI (reflecting uncertainty in the numerator and denominator; P values are for one-sided test). A true (rather than estimated) ratio of 1 indicates the relationship between SNP effect on lifetime hazard and extreme longevity is the same as that of APOE, while a ratio of zero suggests no effect on longevity. A true ratio between 0 and 1 suggests a stronger effect on lifetime hazard than longevity relative to APOE. SNPs overlapping both 0 and 1 are individually underpowered. The inverse variance meta-analysis of alpha over all SNPs, excluding *APOE*, is 0.42 (95% 0.23 to 0.61; $P = 1.35 \times 10^{-5}$) for H_0 alpha = 0.

Sex- and age-specific effects

We stratified our UK Biobank sample (for which we had individual level data) by sex and age bands to identify age- and sex-specific effects for survival SNPs discovered and/or replicated in this study. Although power was limited, as we sought contrasts in small effect sizes, we find 5 SNPs with differential effects on lifespan when stratified (FDR 5% across the 24 variants considered).

The effect of the *APOE* variant increases with age: the $\epsilon 4$ log hazard ratio on individuals older than 70 years is around 3 times greater than those between ages 40–70. In contrast, the effect of lead variants near *CHRNA3/5*, *CDKN2B-AS1*, and *ABO* tends to decline after age 60, at least when expressed as hazard ratios ([Figure 5A](#)).

Independent of age, lead variants near *APOE* and *PSORS1C3* also show an effect around 3.6% greater in women (95% CI 1.3%–5.9%; 1.9%–5.6%, respectively), compared to men ([Figure 5B](#)). Notably, the SNP near *ZW10*, which was identified by Pilling *et al.*[159] in fathers, and which replicated in LifeGen fathers, may affect men and women equally (95% CI years gained per effect allele, men 0.17–0.42, women 0.04–0.31), as measured in our meta-analysis of UK Biobank and LifeGen.

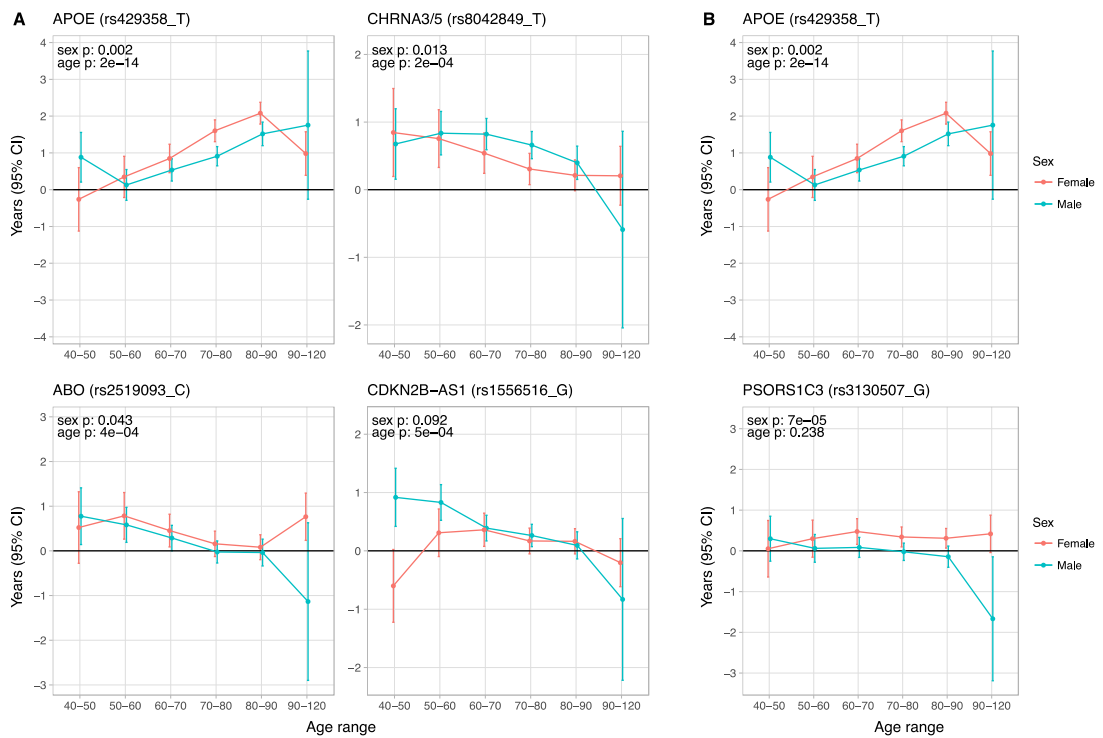


Figure 5: Age and sex specific effects on parent survival for 5 variants showing 5% FDR age- or sex-specificity of effect size from 23 lifespan-increasing variants. A) Variants showing age-specific effects; B) Variants showing sex-specific effects. Panel titles show the gene, cluster of genes, or cytogenetic band in close proximity to the lead lifespan variant, with this variant and lifespan-increasing allele in parentheses. Beta – $\log_e(\text{protection ratio})$ for one copy of effect allele in self in the age band (i.e. 2x observed due to 50% kinship). Note the varying scale of y-axis across panels. Age range: the range of ages over which beta was estimated. Sex p – nominal P value for association of effect size with sex. Age p – nominal P value for association of effect size with age.

Causal genes and methylation sites

We used SMR-HEIDI to look for causal effects of gene expression or changes in methylation on lifespan within the 24 loci discovered or replicated in our study. Using blood eQTL summary statistics from two studies[169,170], we suggest causal roles for expression of *PSRC1*, *SESN1*, *SH2B3*, *PSMA4*, *FURIN*, *FES*, and *KANK2* at 5% FDR ([Supplementary file 1](#)). GTEx tissue-wide expression data suggests further roles for 16 genes across 24 tissues, especially *FES* (9 tissues), *PMS2P3* (6 tissues) and *PSORS1C1* (4 tissues). Methylation data reveals roles for 44 CpG sites near 9 loci, especially near the *PSORS1C3* locus (21 sites), *APOE* locus (9 sites), and *HLA-DQA1* locus (4 sites) ([Supplementary file 2](#)).

We next used SOJO to perform conditional analysis on the same loci to find additional independent variants associated with lifespan. We find substantial allelic heterogeneity in several association intervals and identify an additional 335 variants, which increase out-of-sample explained variance from 0.095% to 0.169% (78% increase). *CELSR2/PSRC1*, *KCNK3*, *HLA-DQA1*, *LPA*, *ZW10*, *FURIN/FES*, and *APOE* are amongst the most heterogeneous loci with at least 25 variants per locus showing independent effects ([Supplementary file 3](#)).

Disease and lifespan

We next sought to understand the link between our lifespan variants and disease. We looked up known associations with our top hits and proxies ($r^2 > 0.6$) in the GWAS catalog[171] and PhenoScanner[172], excluding loci identified in iGWAS as these used disease associations to build the effect priors. We also excluded trait associations discovered solely in UK Biobank, as the overlap with our sample could result in spurious association due to correlations between morbidity and mortality. Under these restrictions, we find alleles which increase lifespan associate with a reduction in cardiometabolic, autoimmune, smoking-related, and neuropsychiatric disease and their disease risk factors ([Table 1](#); [Table 1—source data 3](#)). None of the loci show any association with cancer other than lung cancer. We then looked up associations of the 81 iGWAS SNPs (1% FDR) with the risk factor GWAMAs used to inform the prior. While associations are *a priori* limited to the risk factors included in the iGWAS, the pattern of association is still of interest. We find loci show strong clustering in either blood lipids or CVD, show moderate clustering of metabolic and neurological traits, and show weak but highly pleiotropic clustering amongst most of the remaining traits (see [Figure 3—figure supplement 1](#) for clustering of genome-wide significant SNPs).

In order to study the relative contribution of diseases to lifespan, we approached the question from the other end and looked up known associations for disease categories (CVD, type 2 diabetes, neurological disease, smoking-related traits, and cancers) in large numbers (>20 associations in each category) from the GWAS catalog[171] and used our GWAS to see if the disease loci associate with lifespan. Our measure was lifespan variance explained (LVE, years²) by the locus, which balances effect size against frequency, and is proportional to selection response and the GWAS test statistic and thus monotonic for risk of false positive lifespan associations. Taking each independent disease variant, we ordered them by LVE, excluding any secondary disease where the locus was pleiotropic.

The Alzheimer's disease locus *APOE* shows the largest LVE (0.23 years²), consistent with its most frequent discovery as a lifespan SNP in

GWAS[90,159,161,173]. Of the 20 largest LVE SNPs, 12 and 4 associate with CVD and smoking/lung cancer, respectively, while only 2 associate with other cancers (near *ZW10* and *NRG1*; neither in the top 15 LVE SNPs). Cumulatively, the top 20/45 LVE SNPs explain 0.33/0.43 years² through CVD, 0.13/0.15 years² through smoking and lung cancer, and 0.03/0.11 years² through other cancers (Figure 6).

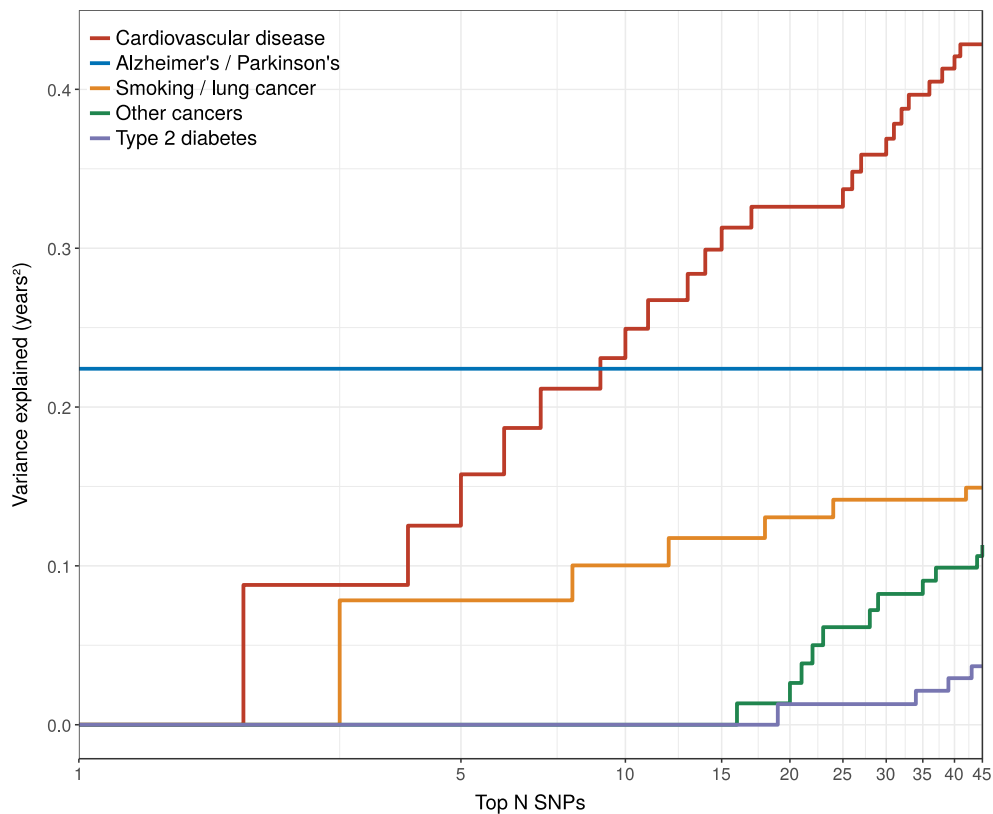


Figure 6: Disease loci explaining the most lifespan variance are protective for neurological disease, cardiovascular disease, and lung cancer. SNPs reported as genome-wide significant for disease in European population studies, ordered by their lifespan variance explained (LVE), show the cumulative effect of disease SNPs on variation in lifespan. An FDR cut-off of 1.55% is applied simultaneously across all diseases, allowing for 1 false positive association with lifespan among the 45 independent loci. Note the log scale on the X axis. Cardiovascular disease – SNPs associated with cardiovascular disease or myocardial infarction. Alzheimer's / Parkinson's – SNPs associated with Alzheimer's disease or Parkinson's disease. Smoking / lung cancer – SNPs associated with smoking behaviour, chronic obstructive pulmonary disease and lung adenocarcinomas. Other cancers – SNPs associated with cancers other than lung cancer (see Figure 7—source data 1 for a full list). Type 2 diabetes – SNPs associated with type 2 diabetes.

Strikingly, two of the three largest LVE loci for non-lung cancers (at or near *ATXN2/BRAP* and *CDKN2B-AS1*), show **increased** cancer protection associating with **decreased** lifespan (due to antagonistic pleiotropy with CVD), while the third (at or near *MAGI3*) also shows evidence of pleiotropy, having an association with CVD three times as strong as breast cancer, and in the same direction. In addition, 6 out of the 11 remaining cancer-protective loci which increase lifespan and pass FDR (near *ZW10*, *NRG1*, *C6orf106*, *HNF1A*, *C20orf187*, and *ABO*) also show significant associations with CVD but could not be tested for pleiotropy as we did not have data on the relative strength of association of every type of cancer against CVD, and thus (conservatively from the point of view of our conclusion) remain counted as cancer SNPs ([Figure 7](#); [Figure 7—source data 1](#)). Visual inspection also reveals an interesting pattern in the SNPs that did not pass FDR correction for affecting lifespan: cardio-protective variants associate almost exclusively with increased lifespan, while cancer-protective variants appear to associate with lifespan in either direction (grey dots often appear below the x-axis for other cancers).

Together, the disease loci included in our study with significant effects on lifespan explain 0.95 years², or less than 1% of the phenotypic variance of lifespan of European parents in UK Biobank (123 years²), and around 5% of the heritability.

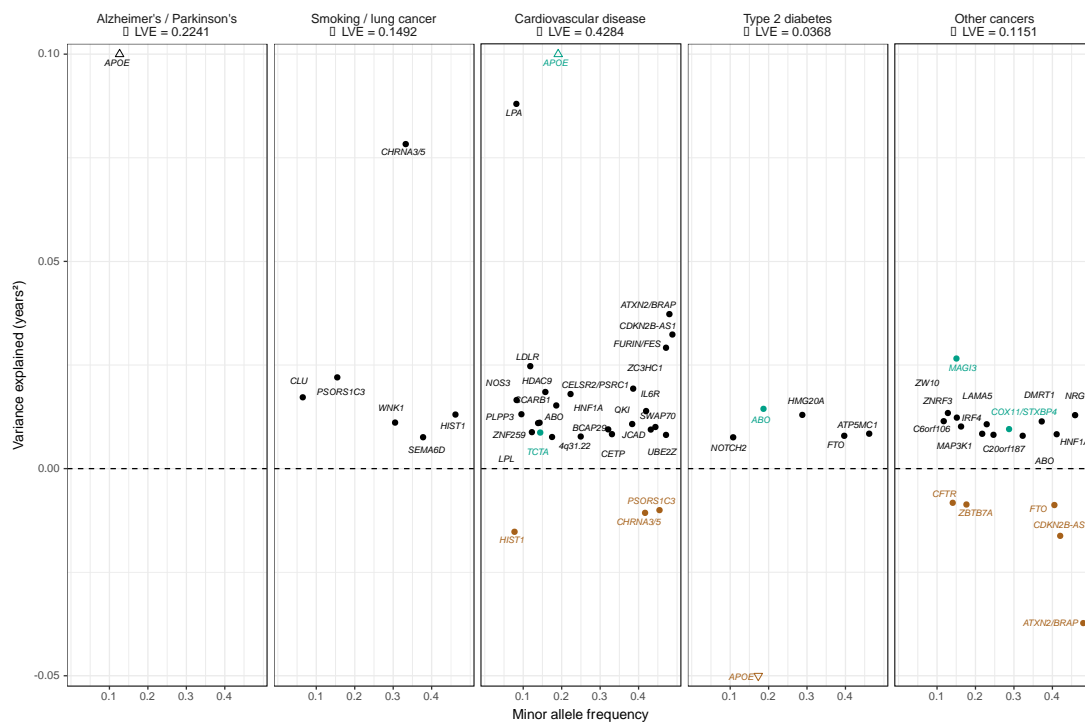


Figure 1: Lifespan variance explained by individual genome-wide significant disease SNPs within disease categories. Genome-wide significant disease SNPs from the GWAS catalog are plotted against the amount of lifespan variance explained (LVE), with disease-protective alleles signed positively when increasing lifespan and signed negatively when decreasing lifespan. SNPs with limited evidence of an effect on lifespan are greyed out: an FDR cut-off of 1.55% is applied simultaneously across all diseases, allowing for 1 false positive among all significant SNPs. Secondary pleiotropic SNPs (i.e. those associating more strongly with another one of the diseases, as assessed by PheWAS in UK Biobank) are coloured to indicate the main effect on increased lifespan seems to arise elsewhere. Of these, turquoise SNPs show one or more alternative disease associations in the same direction and at least twice as strong (double Z statistic – see Detailed Methods) as the principal disease, while brown SNPs show one or more significant associations with alternative disease in the opposite direction that explains the negative association of the disease-protective SNP with lifespan. The variance explained by all SNPs in black is summed (\sum LVE) by disease. Annotated are the gene, cluster of genes, or cytogenetic band near the lead SNPs. The Y axis has been capped to aid legibility of SNPs with smaller LVE: SNPs near APOE pass this cap and are represented by triangles. See [Figure 7—source data 1](#) for the full list of disease SNP associations.

Cell type and pathway enrichment

We used stratified LD-score regression to assess whether cell type-specific regions of the genome are enriched for lifespan variants. As this method derives its power from SNP heritability, we limited the analysis to genomically British individuals in UK Biobank, which showed the lowest heterogeneity and the highest SNP heritability. At an FDR < 5%, we find enrichment in SNP heritability in five categories: two histone and two chromatin marks linked to male and female foetal brain cells, and one histone mark linked to the dorsolateral prefrontal cortex (DLPC) of the brain. Despite testing other cell types, such as heart, liver, and immune cells, no other categories are statistically significant after multiple testing correction ([Supplementary file 4](#)).

We also determined which biological pathways could explain the associations between our genetic variants and lifespan using three different methods, VEGAS, PASCAL, and DEPICT. VEGAS highlights 33 gene sets at an FDR < 5%, but neither PASCAL nor DEPICT (with SNP thresholds at $P < 5 \times 10^{-8}$ and $P < 1 \times 10^{-5}$) identify any gene sets passing multiple testing correction. The 33 gene sets highlighted by VEGAS are principally for blood lipid metabolism (21), with the majority involving lipoproteins (14) or homeostasis (4). Other noteworthy gene sets are neurological structure and function (5) and vesicle-mediated transport (3). Enrichment was also found for organic hydroxy compound transport, macromolecular complex remodelling, signalling events mediated by stem cell factor receptor (c-kit), and regulation of amyloid precursor protein catabolism ([Supplementary file 5](#)).

Finally, we performed an analysis to assess whether genes that have been shown to change their expression with age[174] are likely to have a causal effect on lifespan itself. Starting with a set of independent SNPs affecting gene expression (eQTLs), we created categories based on whether gene expression was age-dependent and whether the SNP was associated with lifespan in our study (at varying levels of significance). We find eQTLs associated with lifespan are 1.69 to

3.39 times more likely to have age-dependent gene expression, depending on the P value threshold used to define the set of lifespan SNPs ([Supplementary file 6](#)).

Out-of-sample lifespan PRS associations.

We calculated polygenic risk scores (PRS) for lifespan for two subsamples of UK Biobank (Scottish individuals and a random selection of English/Welsh individuals), and one sample from the Estonian Biobank. The PRS were based on (recalculated) lifespan GWAS summary statistics that excluded these samples to ensure independence between training and testing datasets.

When including all independent markers, we find an increase of one standard deviation in PRS increases lifespan by 0.8 to 1.1 years, after doubling observed parent effect sizes to compensate for the imputation of their genotypes (see [Table 3—source data 1](#) for a comparison of performance of different PRS thresholds). Correspondingly – again after doubling for parental imputation – we find a difference in median survival for the top and bottom deciles of PRS of 5.6/5.6 years for Scottish fathers/mothers, 6.4/4.8 for English & Welsh fathers/mothers and 3.0/2.8 for Estonian fathers/mothers. In the Estonian Biobank, where data is available for a wider range of subject ages (i.e. beyond median survival age) we find a contrast of 3.5/2.7 years in survival for male/female subjects, across the PRS tenth to first deciles ([Table 3](#); [Figure 8](#)).

Sample Descriptives				Effect of polygenic score				Contrast age at death	
Population	Kin	N	Deaths	Beta	SE	Years	P	Men	Women
Scotland	Parents	46,936	33,196	0.107	0.011	1.07	4.2E-22	5.6	5.6
Scotland	Subjects	24,059	941	0.085	0.033	0.85	1.0E-02	-	-
E&W	Parents	58,070	39,347	0.133	0.010	1.33	7.3E-39	6.4	4.8
E&W	Subjects	29,815	760	0.098	0.037	0.98	7.1E-03	-	-
Estonia	Parents	61,728	29,660	0.099	0.012	0.99	2.5E-17	3.0	2.8
Estonia	Subjects	24,800	2,894	0.087	0.019	0.87	2.6E-06	3.5	2.7

Per standard deviation Top vs. bottom 10%

Table 3: Polygenic scores for lifespan associate with out-of-sample parent and subject lifespans. A polygenic risk score (PRS) was made for each subject using GWAS results that did not include the subject sets under consideration. Subject or parent survival information (age entry, age exit, age of death, if applicable) was used to test the association between polygenic risk score and survival as (a) a continuous score and (b) by dichotomising the top and bottom decile scores. Population – Population sample of test dataset, where E&W is England and Wales; Kin – Individuals tested for association with polygenic score; N – Number of lives used for analysis; Deaths – Number of deaths; Beta – Effect size per PRS standard deviation, in $\log_e(\text{protection ratio})$, doubled in parents to reflect the expected effect in cohort subjects. SE – Standard error, doubled in parents to reflect the expected error in cohort subjects; Years – Estimated years of life gained per PRS standard deviation; P – P value of two-sided test of association; Contrast age at death – difference between the median lifespan of individuals in the top and bottom deciles of the score in year of life (observed parent contrast is again doubled to account for imputation of their genotypes).

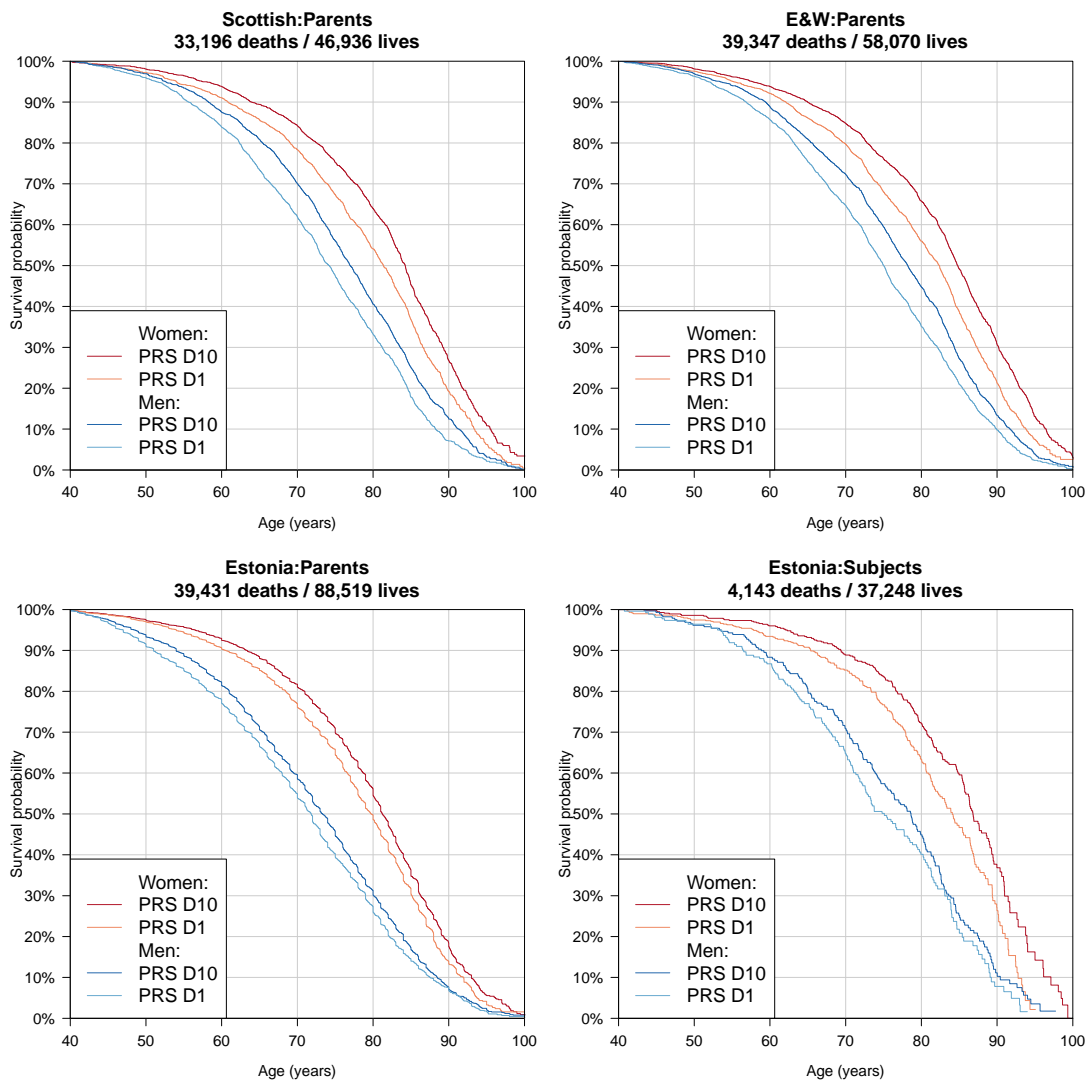


Figure 8: Survival curves for highest and lowest deciles of lifespan polygenic risk score. A polygenic risk score was made for each subject using GWAS results that did not include the subject sets under consideration. Subject or parent survival information (age entry, age exit, age of death (if applicable) was used to create Kaplan-Meier curves for the top and bottom deciles of score. In this figure (only) no adjustment has been made for the dilution of observed effects due to parent imputation from cohort subjects. Effect sizes in parent, if parent genotypes had been used, are expected to be twice that shown. E&W – England and Wales; PRS – polygenic risk score.

Finally, as we did for individual variants, we looked at the age- and sex-specific nature of the PRS on parental lifespan and then tested for associations with (self-reported) age-related diseases in subjects and their kin. We find a high PRS has a larger protective effect on lifespan for mothers than fathers in UK Biobank subsamples ($P = 0.0071$), and has a larger protective effect on lifespan in younger age bands ($P = 0.0001$) (Figure 9), although in both cases, it should be borne in mind that women and younger people have a lower baseline hazard, so a greater improvement in hazard ratio does not necessarily mean a larger absolute protection.

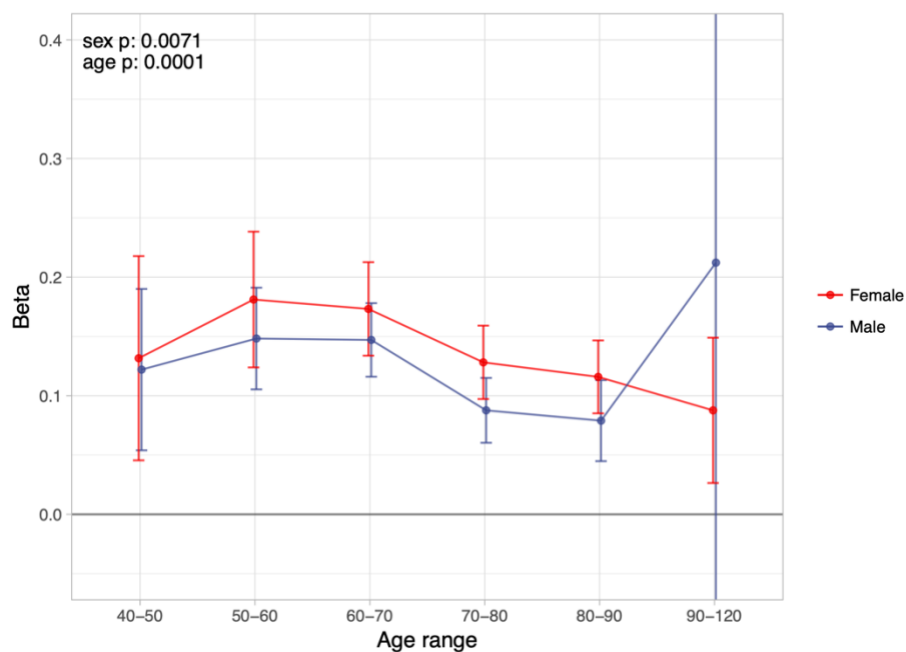


Figure 9: Sex and age specific effects of polygenic survival score (PRS) on parental lifespan in UK Biobank. The effect of out-of-sample PRS on parental lifespan stratified by sex and age was estimated for Scottish and English/Welsh subsamples individually (see Figure 9—figure supplement 1) and subsequently meta-analysed. The estimate for the PRS on father lifespan in the highest age range has very wide confidence intervals (CI) due to the limited number of fathers surviving past 90 years of age. The beta 95% CI for this estimate is -0.15 to 0.57 . Beta $- \log_e(\text{protection ratio})$ for 1 standard deviation of PRS for increased lifespan in self in the age band (i.e. 2 x observed due to 50% kinship), bounds shown are 95% CI; Age range – the range of ages over which beta was estimated; sex p – P value for association of effect size with sex; age p – P value for association of effect size with age.

We find that overall, higher PRS scores (i.e. genetically longer life) are associated with less heart disease, diabetes, hypertension, respiratory disease and lung cancer, but increased prevalence of Alzheimer’s disease, Parkinson’s disease, prostate cancer and breast cancer, the last three primarily in parents. We find no association between the score and prevalence of cancer in subjects ([Figure 10](#)).

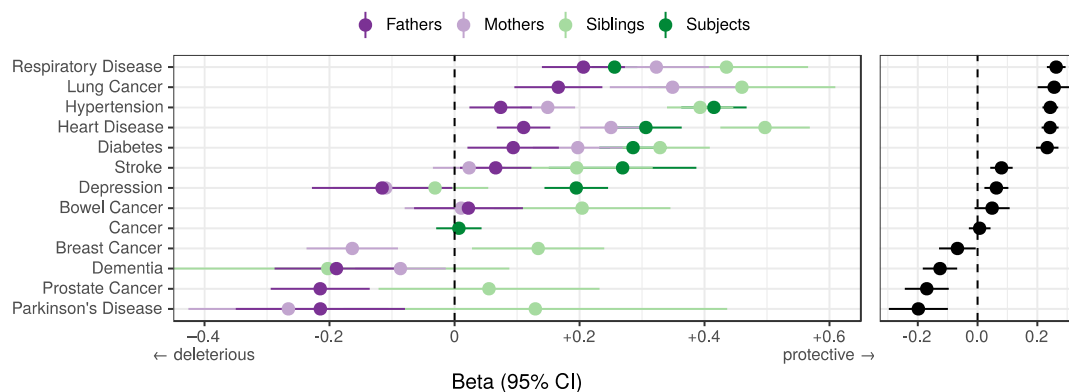


Figure 10: Associations between polygenic lifespan score and diseases of UK Biobank subjects and their kin. Logistic regression was performed on standardised polygenic survival score (all variants) and 21 disease traits reported by 24,059 Scottish and 29,815 English/Welsh out-of-sample individuals about themselves and their kin. For grouping of UK Biobank disease codes, see [Figure 10—source data 1](#). Displayed here are inverse-variance meta-analysed estimates of the diseases for which multiple sources of data were available (i.e. parents and/or siblings; see [Figure 10—figure supplement 1](#) for all associations). “Cancer” is only in subjects, whilst the specific subtypes are analysed for kin. The left panel shows disease estimates for each kin separately; the right panel shows the combined estimate, with standard errors adjusted for correlation between family members. Diseases have been ordered by magnitude of effect size (combined estimate). Beta – log odds reduction ratio of disease per standard deviation of polygenic survival score, where a negative beta indicates a deleterious effect of score on disease prevalence (lifetime so far), and positive beta indicates a protective effect on disease. Effect sizes for first degree relative have been doubled. Cancer – Binary cancer phenotype (any cancer, yes / no).

Discussion

Applying the kin-cohort method in a GWAS and mortality risk factor iGWAS across UK Biobank and the LifeGen meta-analysis, we identified 11 novel genome-wide significant associations with lifespan and replicated 6 previously discovered loci. We also replicated long-standing longevity SNPs near *APOE*, *FOXO3*, and 5q33.3/*EBF1* – albeit with smaller effect sizes in the latter two cases – but found evidence of no association (at effect sizes originally published) with lifespan for more recently published longevity SNPs near *IL6*, *ANKRD20A9P*, *USP42*, and *TMTC2*. Conversely, all individual variants identified in our analyses showed directionally consistent effects in a meta-analysis of two European-ancestry studies of extreme longevity, and a test of association of a polygenic risk score of the variants was highly significant in the longevity dataset ($P < 1.5 \times 10^{-5}$).

Our findings validate the results of a previous Bayesian analysis performed on a subset ($N = 116,279$) of the present study's discovery sample[164], which highlighted two loci which are now genome-wide significant in conventional GWAS in the present study's larger sample. iGWAS thus appears to be an effective method able to identify lifespan-associated variants in smaller samples than standard GWAS, albeit relying on known biology.

With the curious exception of a locus near *HTT* (the Huntington's disease gene), all lead SNPs are known to associate with autoimmune, cardiometabolic, neuropsychiatric, or smoking-related disease, and it is plausible these are the major pathways through which the variants affect lifespan. Whole-genome polygenic risk scores showed similar associations with disease, excluding late-onset disorders such as Alzheimer's and Parkinson's, where polygenic risk scores for extended lifespan increased risk (of survival to age at onset) of the disease.

Genetic variants affecting lifespan were enriched for pathways involving the transport, homeostasis and metabolism of lipoprotein particles, validating previous reports[164]. We also identified new pathways including vesicle transport, metabolism of acylglycerol and sterols, and synaptic and dendritic

function. We discovered genomic regions with epigenetic marks determining cell differentiation into foetal brain and DLPC cells were enriched for genetic variants affecting lifespan. Finally, we showed that we can use our GWAS results to construct a polygenic risk score, which makes 3 to 5 year distinctions in life expectancy at birth between individuals from the score's top and bottom deciles.

Despite studying over 1 million lives, our standard GWAS only identified 12 variants influencing lifespan at genome-wide significance. This contrasts with height (another highly polygenic trait) where a study of around 250,000 individuals by Wood *et al.*[175] found 423 loci. This difference can partly be explained by the much lower heritability of lifespan (0.12)[40] (cf. 0.8 for height[175]), consistent with evolution having a stronger influence on the total heritability of traits more closely related to fitness and limiting effect sizes. In addition, the use of indirect genotypes (the kin-cohort method) reduces the effective sample size to 1/4 for the parent-offspring design.

When considering these limitations, we calculate our study was equal in power to a height study of only around 23,224 individuals, were lifespan to have a similar genetic architecture to height (see Methods). Under this assumption, we would require a sample size of around 10 million parents (or equivalently 445,000 nonagenarian cases, with even more controls) to detect a similar number of loci as Wood *et al.* At the same time, our inability to replicate several previous borderline significant longevity and lifespan findings suggests research into survival in general requires substantial increases in power to robustly identify loci.

Meta-analysis of mothers and fathers, permitting common or sex-specific effect sizes, of course, doubled effective sample size, with slight attenuation to reflect the observed correlation (~10%) between father and mother traits (consistent with previous studies[40]). This correlation indicates the presence of assortative mating on traits which correlate with lifespan (as lifespan itself is of course not observed until later), or post-pairing environmental convergence. We note that in

principle, assortative mating could lead to allelic correlations at causal loci for the contributing traits, causing departures from Hardy-Weinberg equilibrium, and increasing the genotypic variance and thus power to detect association. However, in practice, at least for lifespan, the effects are too small for the effect to be material.

The association of lifespan variants with well-known, life-shortening diseases (cardiovascular, autoimmune, smoking-related diseases and lung cancer)[176] is not surprising, but the paucity of associations with other forms of cancer – without pleiotropic effects on CVD – is. This paucity suggests cancer deaths may often be due to (perhaps many) rarer variants or environmental exposures, although effect sizes might simply be slightly below our cut-off threshold to detect. Disappointingly, the variants and pathways we identified do not appear to underpin a generalised form of ageing independent of disease.

Our finding that lifespan genetics are enriched for lipid metabolism genes is in line with expectations, given lipid metabolites – especially cholesterol metabolites – have well-established effects on atherosclerosis, type-2-diabetes, Alzheimer's disease, osteoporosis, and age-related cancers[177]. Pilling *et al.*[159] implicated nicotinic acetylcholine receptor pathways in human lifespan, which we detected at nominal significance ($P = 2 \times 10^{-4}$) but not at 5% FDR correction ($q = 0.0556$). Instead we highlighted more general synapse and dendrite pathways and identified foetal brain and DLPC cells as important in ageing. The DLPC is involved in smoking addiction[178], dietary self-control[179], and is susceptible to neurodegeneration[180], which could explain why genetic variation for lifespan is specifically enriched in these cells, mediated through smoking-related, cardiometabolic, and neuropsychiatric disease.

Much work has been done implicating *FOXO3* as an ageing gene in model organisms[22,181], however we found the association in humans at that locus may be driven by expression of *SESN1* (admittedly a finding restricted to peripheral blood tissue). *SESN1* is a gene connected to the *FOXO3* promoter via

chromatin interactions and is involved in the response to reactive oxygen species and mTORC1 inhibition[182]. While fine-mapping studies have specifically found genetic variation within the locus causes differential expression of *FOXO3* itself[183,184], this does not rule out co-expression of *SESN1* effects. More powered tissue-specific expression data and experimental work on *SESN1* vs *FOXO3A* could elucidate. In the meantime the model organism results seem to leave the preponderance of evidence for *FOX3A*.

Our results suggest disease-associated lifespan variants reduce the chances of extreme long-livedness, but remain agnostic as to the more interesting two-part question: are there longevity variants that have little effect on lifespan in the normal range[160], and if so, do they control underlying ageing processes? We note, the genetic overlap between lifespan and extreme long-livedness is high (0.73), but not complete[164]. Regardless of this, only a small part of the heritability of both lifespan and longevity has thus far been explained by GWAS. It thus remains plausible that an enlarged long-livedness or lifespan study will find variants controlling the rate of ageing and associated pathways. Curiously, we find little evidence of SNPs of large deleterious effect on lifespan acting with antagonistic pleiotropy on other fitness and developmental component traits, despite long-standing theoretical suggestions to the contrary[185]. However, we did not examine mortality before the age of 40, or mortality of individuals without offspring (by definition as we were examining parental lifespans), which may exhibit this phenomenon. For the time being, our findings that the improved polygenic risk score for lifespan was associated with an increased prevalence of Alzheimer's disease, Parkinson's disease, and prostate and breast cancer, means we appear to be predominantly measuring a propensity for longer life through avoidance of early disease-induced mortality, rather than healthy ageing or fertility costs.

Whilst it has previously been shown that transcriptomic age calculated based on age-related genes is meaningful in the sense that its deviation from the chronological age is associated with biological features linked to ageing[174], the

role of these genes in ageing was unclear. A gene might change expression with age because (i) it is a biological clock (higher expression tracking biological ageing, but not influencing ageing or disease); (ii) it is a response to the consequences of ageing (e.g. a protective response to CVD); (iii) it is an indicator of selection bias: if low expression is life-shortening, older people with low expression tend to be eliminated from the study, hence the average expression level of older age groups is higher. However, our results now show that the differential expression of many of the age-related genes discovered by Peters *et al.*[174] are not only biomarkers of ageing, but are also enriched for direct effects on lifespan.

There is increasing interest in polygenic risk scores, and their potential clinical utility for some diseases appears to be similar to some Mendelian mutations (albeit such monogenic tests are usually only applied in the context of family history)[105]. At first sight, the magnitude of the distinctions in our genetic lifespan score (5 years of life between top and bottom deciles, for both the parent and subject generations) are quite small compared with variability in individual lifespans. However, these distinctions are potentially material at a group level, for example, actuarially. The implied distinction in price (14%; Methods) is greater than some recently reported annuity profit margins (8.9%)[186]. In our view, the legal and ethical frameworks (at least in the UK[187]) are presently underdeveloped for genome-wide scores, whether for disease or lifespan and this needs to be urgently addressed. At the same time, although material in isolation, our lifespan associations may only have practical utility in many applications if they provide additional information than that provided by conventional clinical risk measures (e.g. the Framingham score[188]). Such an assessment has been beyond the scope of this work, in part as such risk measures are not readily available for the parents (rather than subjects) studied.

One limitation of our study was the power reduction caused by the exclusion of relatives in our study, rather than linear mixed modelling (LMM) with a term for kinship as measured by the genomic relationship matrix (GRM) [159,189]. However, as the correct adjustment is not derivable under the kin-cohort method, we felt this was the best approach. To see that the normal adjustment is not correct, consider two siblings. The phenotypes under study are of course identical (as the parents are the same), but the expected correlation under the mixed model would only be 50% of the heritability. Simply excluding siblings, however, is not sufficient. For example, consider two offspring subjects who are first cousins descended from two full brothers. The GRM entry in this situation is 12.5% whilst the appropriate relatedness factor for the father trait is 50% and the mother 0%. Exclusion of relatives thus appears the most straightforward solution, although if a pedigree were available, not just a GRM, accurate LMM might have been feasible.

The analysis of parent lifespans has enabled us to probe mortality for a generation whose lives are mostly complete and attain increased power in a survival GWAS. However, changes in the environment (and thus the relative importance of each genetic susceptibility, for example following the smoking ban) inevitably mean we have less certainty about associations with prospective lifespans for the present generation of middle-aged people, or a different population (with perhaps different relative importance of disease or even overall heritability of lifespan). The 21% reduction in the effect size of the association between our PRS for the UK offspring generation supports this idea, although the estimated contrast in hazard ratios across the deciles was not reduced, which may be a statistical artefact or due to the different periods of life probed. The lower explanatory power of the PRS in Estonia may reflect the differing alleles and LD patterns between the UK training data and the Estonian test data, but also the different environments, in particular the sources of mortality in that country in the Soviet, and early post-Soviet era.

In conclusion, recent genomic susceptibility to death in the normal age range seems rooted in modern diseases: Alzheimer's, CVD and lung cancer; in turn arising from our modern – long-lived, obesogenic and tobacco-laden – environment, however the keys to the distinct traits of ageing and extreme longevity remain elusive. At the same time, genomic information alone can now make material distinctions at a group level in variations in expected length of life, although the limited individual accuracy of these distinctions is far from reaching genetic determinism of that most (self-) interesting of traits – your lifespan.

Methods

GWAS

For genetically British ancestry (as identified by UK Biobank using genomic PCA) and each self-reported European ethnicity in UK Biobank (including self-declared British but not genetically British ancestry), independent association analyses were performed between unrelated subjects' genotypes (MAF > 0.005; HRC imputed SNPs only; ~9 million markers) and parent survival using age and alive/dead status in residualised Cox models, as described in Joshi *et al.*[70]. To account for parental genotype imputation, effect sizes were doubled, yielding log hazard ratios for the allele in carriers themselves. These values were negated to obtain a measure of log protection ratio, where higher values indicate longer life. While methods exist to account for related individuals using linear mixed models, such as BOLT-LMM[189] , these are not accurate when trying to account for relatedness between parents (See Detailed Methods).

Mother and father survival information was combined in two separate ways, essentially assuming the effects were the same in men and women, or allowing for sex-specific effect sizes (SSE), with appropriate allowance for the covariance amongst the traits. For the first analysis we summed parental survival residuals; for the second analysis we used MANOVA, implemented in MultiABEL[190].

For LifeGen, where individual-level data was not available, parent survival summary statistics were combined using conventional fixed-effects meta-analysis, adjusted to account for the correlation between survival traits (estimated from summary-level data). For SSE, the same procedure was followed as for the UK Biobank samples, with correlation between traits again estimated from summary-level data. The GWAS statistics showed acceptable inflation, as measured by their LD-score regression intercept (<1.06, [Table 1—source data 2](#)).

Candidate SNP replication

Effect sizes from longevity studies were converted to our scale using an empirical conversion factor, based on the observed relationships between longevity and hazard ratio at the most significant variant at or near *APOE*, observed in the candidate SNPs study and our data[70]. These studies were then meta-analysed using inverse variance weighting and standard errors were inflated to account for sample overlap (see Detailed Methods)

Estimates reported in Pilling *et al.*[159] were based on rank-normalized Martingale residuals, unadjusted for the proportion dead, which – for individual parents – could be converted to our scale by multiplying by \sqrt{c}/c , where c is the proportion dead in the original study (see Detailed Methods for derivation). Combined parent estimates were converted using the same method as the one used for longevity studies.

The deletion reported by Ben-Avraham *et al.*[167] is perfectly tagged by a SNP that we used to assess replication. Assuming a recessive effect and parental imputation, we derived the expected additive effect to be $\hat{\beta}_C = \hat{\beta}_{CC} \frac{q^2}{q^2 + 2pq}$, where $\hat{\beta}_C$ is the effect we expect to observe under our additive model, $\hat{\beta}_{CC}$ is the homozygous effect reported in the original study, q is the C allele frequency, and p is $1 - q$. (see Detailed Methods for derivation)

iGWAS

58 GWAS on mortality risk factors were used to create Bayesian priors for the SNP effects observed in the CES study, as described in McDaid *et al.*[164]. Mendelian randomisation was used to estimate causal effects of independent risk factors on lifespan, and these estimates were combined with the risk factor GWAS to calculate priors for each SNP. Priors were multiplied with observed Z statistics and used to generate Bayes factors. Observed Z statistics were then permuted,

leading to 7.2 billion null Bayes factors (using the same priors), which were used to assess significance.

Sex and age stratified analysis

Cox survival models, adjusting for the same covariates as the standard GWAS, were used to test SNP dosage against survival of UK Biobank genomically British fathers and mothers, separately. The analysis was split into age bands, where any parent who died at an age younger than the age band was excluded and any parent who died beyond the age band was treated as alive. Using the R package “metafor”, moderator effects of sex and age on hazard ratio could be estimated while taking into account the estimate uncertainty (see Detailed Methods for formula).

Causal genes and methylation sites

SMR-HEIDI[191] tests were performed on CES statistics to implicate causal genes and methylation sites. Summary-level data from two studies on gene expression in blood[169,170] and data on gene expression in 48 tissues from the GTEx consortium[192] were tested to find causal links between gene expression and lifespan. Similarly, data from a genome-wide methylation study[193] was used to find causal links between CpG sites and lifespan. All results from the SMR test passing a 5% FDR threshold where the HEIDI test $P > 0.05$ were reported.

Conditional analysis

SOJO[194] was used to fine-map the genetic signals in 1 Mb regions around lead SNPs reaching genome-wide significance and candidate SNPs reaching nominal significance in our study. The analysis was based on CES statistics from UK Biobank genomically British individuals, using the LifeGen meta-analysis results to optimise the LASSO regression tuning parameters. For each parameter, a polygenic score was built and the proportion of predictable variance from the regional polygenic score in the validation sample was calculated.

Disease association analysis

The GWAS catalog[171] and PhenoScanner[172] were checked for known genome-wide significant associations with our GWAS hits and proxies ($r^2 > 0.6$) in European samples. Associations discovered in UK Biobank by Neale *et al.*[195] were omitted from the PhenoScanner database as the findings have not been replicated, and the large sample overlap with our own study could result in false positive associations, due to phenotypic correlations between morbidity and mortality. Triallelic SNPs and associations without effect sizes were excluded before near-identical traits were grouped together, discarding all but the strongest association and keeping the shortest trait name. For example, “Lung cancer”, “Familial lung cancer”, and “Small cell lung cancer” were grouped and renamed to “Lung cancer”. The remaining associations were classified into disease categories based on keywords and subsequent manual curation.

Lifespan variance explained by disease SNPs

The GWAS catalog[171] was checked for disease associations discovered in European ancestry studies, which were grouped into broad disease categories based on keywords and manual curation (see [Figure 7—source data 1](#) and Detailed Methods). Associations were pruned by distance (500kb) and LD ($r^2 < 0.1$), keeping the SNP most strongly associated with lifespan in the CES GWAS. Where possible this SNP was tested against diseases in UK Biobank subjects and their family to test for pleiotropy (see Detailed Methods). Significance of associations with lifespan was determined by setting an FDR threshold that allowed for 1 false positive among all independent SNPs tested ($q \leq 0.022$). Lifespan variance explained (LVE) was calculated as $2pqa^2$, where p and q are the frequencies of the effect and reference alleles in our lifespan GWAS, and a is the SNP effect size in years of life[196].

Cell type enrichment

Stratified LD-score regression[197] was used to test for cell type-specific enrichment in lifespan heritability. As the power of this method depends on SNP heritability, standard LD-score regression[198] was first used to check which of

our samples (UK Biobank, LifeGen, or the combined cohort) had the highest SNP heritability. Lifespan summary statistics from UK Biobank genomically British individuals were then analysed using the procedure described in Finucane *et al.*[197], and P values were adjusted for multiple testing using the Benjamin-Hochberg procedure.

Pathway enrichment

VEGAS2 v2.01.17[199] was used to calculate gene scores using SNPs genotyped in UK Biobank, based on summary statistics from the full CES cohort and the default software settings. VEGAS2Pathway was then used to check for pathway enrichment using those gene scores and the default list of gene sets[200].

DEPICT[201] was also used to map genes to lifespan loci and check for pathway enrichment in the combined cohort CES GWAS. Default analysis was run for regions with genome-wide significant ($P < 5e-8$) variants in the first analysis, and genome-wide suggestive ($P < 1e-5$) variants in the second analysis, excluding the MHC in both cases.

PASCAL[202] was used with the same summary statistics and gene sets as DEPICT, except the gene probabilities within the sets were dichotomized ($Z > 3$) as described in Marouli *et al.*[203]. For each software independently, pathway enrichment was adjusted for multiple testing using the Benjamin-Hochberg procedure.

Age-related eQTL enrichment

Combined cohort CES lifespan statistics were matched to eQTLs associated with the expression of at least one gene ($P < 10^{-3}$) in a dataset from the eQTLGen Consortium (31,684 individuals)[204]. Data on age-related expression[174] allowed eQTLs to be divided into 4 categories based on association with age and/or lifespan. Fisher's exact test was used check if age-related eQTLs were enriched for associations with lifespan.

Polygenic score analysis

Polygenic risk scores (PRS) for lifespan were calculated for two subsamples of UK Biobank (24,059 Scottish individuals and a random 29,815 English/Welsh individuals), and 36,499 individuals from the Estonian Biobank, using combined cohort CES lifespan summary statistics that excluded these samples. PRSice 2.0.14.beta[109] was used to construct the scores from genotyped SNPs in UK Biobank and imputed data from the Estonian Biobank, pruned by LD ($r^2 = 0.1$) and distance (250kb). Polygenic scores were Z standardised.

Cox proportional hazard models were used to fit parental survival against polygenic score, adjusted for subject sex; assessment centre; genotyping batch and array; and 10 principal components. Parental hazard ratios were converted into subject years of life as described in the GWAS method section.

Logistic regression models were used to fit polygenic score against the same self-reported UK Biobank disease categories used for individual SNPs. Effect estimates of first-degree relatives were doubled to account for imputation of genotypes and then meta-analysed using inverse variance weighting, adjusting for trait correlations between family members.

Postulation of equivalent sample size in height GWAS

The use of parent imputation, low trait heritability, and incomplete proportion dead all reduce the power to detect effect sizes. The equivalent sample size in a hypothetical, completely heritable trait with otherwise identical genetic architecture would be the original sample size, diluted (i.e. multiplied) by the heritability (0.122)[40], the r^2 of offspring genotype on parent genotype (0.250) and the proportion dead (0.602). This gives an equivalent sample size of 18,579 from the 1,012,240 parent lifespans. We then calculated sample size for height that would have the same properties, accounting for the heritability of height (0.8)[175]: 23,224 (i.e. $18,579/0.8$). We next calculated the P value that would have been reported by Wood *et al's* 250,000 sample size height GWAMA[175] for a SNP that was just significant in a hypothetical 23,224 sample height GWAMA: P

$< 1.8 \times 10^{-72}$. Six distinct loci passed this significance threshold in Wood *et al*'s results.

With 17,893 nonagenarians, Deelen et al[161] attained a P value of 2.33×10^{-26} at rs4420638. With 1.012m parents we attained a P value of 1.75×10^{-64} . Other things being equal a nonagenarian sample size of 44,500 thus appears to be equally powered to one million parents.

Data availability

The results that support our findings, in particular, the GWAS summary statistics for >1 million parental lifespans in this study are at <http://dx.doi.org/10.7488/ds/2463>. Gene expression data is being made available by the eQTLGen Consortium[204].

URLs

MultiABEL: <https://github.com/xiashen/MultiABEL/>

LDSC: <https://github.com/bulik/ldsc>

SMR/HEIDI: <https://cnsgenomics.com/software/smr/>

SOJO: <https://github.com/zhenin/sojo/>

DEPICT: <https://www.broadinstitute.org/mpg/depict/>

PASCAL: <https://www2.unil.ch/cbg/index.php?title=Pascal>

GTEEx: <https://gtexportal.org/home/datasets>

Acknowledgments

We thank the UK Biobank Resource, approved under application 8304; we acknowledge funding from the UK Medical Research Council Human Genetics Unit, Wellcome Trust PhD Training Fellowship for Clinicians - the Edinburgh Clinical Academic Track (ECAT) programme (204979/Z/16/Z), the Medical Research Council Doctoral Training Programme in Precision Medicine (MR/N013166/1) and the AXA research fund. We thank Tom Haller of the University of Tartu, for tailoring RegScan so we could use it with compressed

files[205]. We would also like to thank the researchers, funders and participants of the LifeGen consortium[70].

Contributions

PRJHT, NM, KL, KF, ZN, XF, AB, DC performed analyses.

TE, eQTLGen contributed data

XS, TE, KF, ZK, PKJ designed the experiments

PRJHT, NM, KL, KF, AB, XS, TE, ZK, JFW, PKJ wrote the manuscript

3.3 Conclusion

Genetic variants are involved in determining lifespan: this study highlighted 12 regions in the genome which strongly influence survival and confirmed an additional five candidate regions from previous studies. The concordance of effects of these regions between multiple studies, including lifespan-related traits such as longevity and cardiovascular disease, suggests the findings are robust. In contrast to our own discoveries, eight candidate loci highlighted in previous studies failed to replicate despite adequate power to detect an effect, while another 12 loci remain uncertain due to a lack of statistical power or data. These findings suggest there is a need for even larger studies on lifespan to detect effects with confidence. The majority of lifespan loci we have robust evidence for affect males and females equally, with regions near *APOE* and *PSORS1C3* being the only exceptions. Since we did not have data on sex chromosomes, we cannot exclude the possibility that there are variants with strong sex-specific effects which affect the discrepancy in male and female lifespan observed in Chapter 2.

Although each individual locus may only change an individual's lifespan by a couple of months, the polygenic survival scores created from SNPs across the genome can make meaningful distinctions—up to 5 years—between individuals in the top and bottom survival score deciles. The slight attenuation of the association of this score with survival in UK Biobank subjects and in Estonian subjects and their parents indicates there may be some population and generation-specific risk factors in UK Biobank parents which reduce predictive power in these other cohorts. Supporting this, we found higher scores were associated with lower levels of morbidity in subjects and parents, but the reductions in the odds of CVD and diabetes were greater for subjects. However, higher scores were also associated with increased odds of neurodegenerative disease and breast/prostate cancer suggesting the score is predicting early mortality rather than delayed ageing. While a study of the association of polygenic

survival score with the incidence of disease could show whether higher polygenic scores delay disease onset, UK Biobank individuals did not report the date of incidence of parental diseases, and very few individuals themselves have experienced late onset disease such as Alzheimer's disease and prostate cancer, which precluded such an analysis. For now, the main use of these polygenic survival scores will be stratifying individuals based on their risk of death within actuarial and healthcare settings. Within the former, they can be used to inform annuity pricing, while in the latter, they can be used to inform decisions to offset risk of early death in patients.

We also found common genetic variation influencing cancer susceptibility (unrelated to smoking) does not appear to have large effects on lifespan, contrasting with cancer mortality as shown in Chapter 2. In fact, we observed cancer-protective SNPs often shorten lifespan due to antagonistic pleiotropic effects on CVD. Whether these findings reflect true antagonistic effects or are artefacts of tissue-specific roles for these variants is unclear. Even if the antagonistic effects are real, this study is unable to determine whether they represent a biological trade-off, or are the result of the removal of one cause of death making the next cause of death more likely. Regardless, these contrasting effects complicate the search for therapeutic targets that could delay ageing and age-related disease, as lifespan genes with antagonistic or tissue-specific effects are poor drug targets due to their higher likelihood of side-effects, where knockdown of a gene or inactivation its gene product could decrease susceptibility to one disease but increase susceptibility to another. The next Chapter addresses this problem by searching for variants and genes with concordant effects on disease susceptibility and lifespan to identify better targets.

Lastly, this study linked the expression of 23 genes and the methylation of 44 CpG sites to parental survival and found enrichment of genetic signals in gene sets for lipid processing and synaptic signalling. When examining these results in the context of the strong links between lifespan variants and CVD and smoking-related disease, it is likely these genes and pathways capture the biological

processes underlying early mortality risk factors, such as obesity and smoking, rather than putative ageing processes, such as cellular maintenance and stress resistance. In conclusion, this study has shown survival can be predicted to some extent from DNA alone, but it appears the strongest genetic determinants of lifespan modify susceptibility to risk factors for early mortality, rather than the ageing process itself. Variants affecting the rate of ageing may require larger and more diverse samples to be detected and may require more health-related measures to be differentiated from mortality risk factors. In the next Chapter, I extend the current study with additional ageing-related GWAS to further explore this complexity.

Chapter 4: Genome-wide multivariate association of healthspan, lifespan, and longevity

4.1 Introduction

4.1.1 Context

In the previous Chapter, we performed a genome-wide association of parental lifespan and highlighted smoking-related disease and CVD pathways as important in determining lifespan, aligning closely with diseases highlighted in Chapter 2. However, there was a notable lack of cancer variants (other than lung cancer) influencing survival, raising the possibility that the genetic signals we identified were related to modern exposures influencing lifespan (e.g. smoking and obesity) rather than intrinsic ageing processes. These considerations do not necessarily invalidate the parental lifespan study: the polygenic survival scores demonstrate lifespan variants remain relevant in determining survival for individuals alive today. However, the biological pathways they implicate may not all be related to ageing but could rather reflect early, cause-specific mortality, and as such be better addressed by lifestyle changes and preventative healthcare.

On the other hand, knowledge of ageing pathways underlying multiple, if not all, age-related diseases would allow for interventions to be designed that could delay the burden of chronic disease and compress morbidity towards the end of life. The aim of the study in this Chapter was therefore to separate extrinsic determinants of lifespan (such as lifestyle and healthcare) from innate determinants of lifespan (such as molecular repair and homeostatic pathways). One way to achieve this is to study large populations with different behavioural and environmental exposures, and focus on genetic variants which determine health and lifespan regardless of external factors.

Fortunately, sharing of results of large genetic studies is becoming more common[71], allowing researchers to build on existing data. In addition to our own study of parental lifespan in UK Biobank and LifeGen, two other large studies

were recently performed on ageing-related phenotypes: a study on the healthspan of UK Biobank individuals[206], and a case-control study on the oldest old in Europe and the US[72]. Together, these studies span multiple generations and populations, examining different phenotypes using different methodologies. However, as I will show, despite the heterogeneity in cohort and study designs, there is a significant overlap in the genetics of these traits, wherein we can find the common mechanisms determining both age-related disease and lifespan.

4.1.2 Contributions

Peter Joshi conceived the idea of meta-analysing parental lifespan and exceptional longevity in a multivariate framework by using MultiABEL software developed by Xia Shen, which was also used in the parental lifespan analysis. Xia confirmed the software could combine GWAS of case-control and quantitative traits but was otherwise not involved in this study.

I built on the original idea, expanding the original multivariate analysis to include healthspan, and designing the downstream analyses. Peter Joshi and Joris Deelen provided feedback on the study design. Joris suggested expanding the gene prioritisation analysis to include tissue-specific GTEx data. All the data used in the study—summary statistics for healthspan, parental lifespan, longevity, and disease traits; eQTL data; and gene sets—were publicly available.

I retrieved the necessary data, performed all the analyses, compiled the results, and drew the figures. Joris wrote the first draft of the introduction and wrote the description of the longevity study. I wrote all other sections of the manuscript. All co-authors provided feedback on the draft manuscript.

4.2 Manuscript submitted to journal

What follows is a manuscript submitted to the journal *Nature Communications* on 6 January 2020, which is currently under review but has not yet been published. A copy of the manuscript as submitted to the journal is included below, with permission from the co-authors. Supplementary Figures can be found in the Appendix. Supplementary Tables are available as Excel Documents on request.

Multivariate genomic scan of human ageing traits reveals novel loci and identifies haem metabolism as a human ageing pathway

Paul R.H.J. Timmers¹, James F. Wilson^{1,2}, Peter K. Joshi¹, Joris Deelen^{3,4}

1) Centre for Population Health Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

2) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom

3) Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

4) Max Planck Institute for Biology of Ageing, Cologne, Germany

Abstract

Ageing phenotypes, such as years lived in good health (healthspan), total years lived (lifespan), and survival until an exceptional old age (longevity), are of interest to us all but require exceptionally large sample sizes to study genetically. By combining existing genome-wide association summary statistics for healthspan, lifespan, and longevity in a multivariate framework, we increased statistical power and showed that the traits share more than 50% of their underlying genetics. We identified 10 genomic loci which influence all three phenotypes, of which five (near *FOXO3*, *SLC44A7*, *LINC01259*, *ZW10*, and *FGD6*) are reported for the first time at genome-wide significance. The majority of these loci are associated with cardiovascular disease and several show signs of

antagonistic pleiotropy. Using gene expression data, we implicated the expression of 59 genes and found this set of genes to be enriched for ageing pathways previously highlighted in model organisms, such as the response to DNA damage, apoptosis, and homeostasis. Finally, we identify a new pathway worthy of further study: haem metabolism.

Introduction

Human ageing is characterised by a progressive decline in the ability to maintain homeostasis, leading to the onset of age-related diseases and eventually death. However, there is much variation between individuals, with some experiencing chronic disease early on and dying before age 60, while others are able to reach an exceptional old age, often free of disease until the last few years of life [207]. A long and healthy life is determined by many different factors, including lifestyle, environment, genetics, and pure chance. Recent estimates suggest the genetic component of both human lifespan (i.e. the number of years lived) and healthspan (the number of years lived in good health free of morbidities) is only around 10% [41,206], which makes genetic studies of these traits challenging, as noise tends to obscure effects unless sample sizes are large.

However, with sufficiently large samples, genome-wide association studies (GWAS) of lifespan traits have the potential to identify genes and pathways involved in the human ageing process. GWAS have attempted to identify loci and pathways related to healthspan [206,208], (parental) lifespan [70,88,209] and survival to exceptional old age (often called longevity) [72,166], with some overlap between findings. Multivariate analyses of correlated traits offers the prospect of increased power [210], especially where samples do not overlap, and offers the prospect of identifying variants influencing a common underlying ageing process.

Here, we assess the degree of genetic overlap between published GWAS of three different kinds of ageing phenotypes—healthspan, parental lifespan, and longevity (defined as survival to an age above the 90th percentile)—and perform

a multivariate meta-analysis to identify genetic variants related to healthy ageing. We subsequently characterise the sex- and age-specific effects of loci which affect all three lifespan traits and look up reported associations with age-related phenotypes and diseases. Finally, we link the observed signal in these loci to the expression of specific genes, including some that are currently studied in model organisms, and identify new pathways involved in healthy ageing.

Methods

We downloaded three publicly available sets of summary statistics on healthspan (Zenin et al. 2019; <http://doi.org/10.5281/zenodo.1302861>), parental lifespan (Timmers et al. 2019; <http://dx.doi.org/10.7488/ds/2463>), and longevity (Deelen et al. 2019; <https://www.longevitygenomics.org/downloads>), whose derivation is briefly described here.

The Healthspan GWAS consists of 300,477 unrelated, British-ancestry individuals from UK Biobank. The statistics were calculated by fitting Cox-Gompertz survival models with events defined as the first incidence of one of seven diseases (any cancer, diabetes, myocardial infarction, stroke, chronic obstructive pulmonary disease, dementia, and congestive heart failure) or death itself. Martingale residuals from this model were then regressed against HRC-imputed dosages. Of the 84,949 individuals who had experienced an event (and thus had complete healthspans), 51.3% experienced a cancer event, 18.0% a diagnosis of diabetes and 17.1% a myocardial event. Less than 5% of the individuals experienced their first event due to one of the remaining diseases. See Zenin et al. [206] for details. After removing single nucleotide polymorphisms (SNPs) with duplicate rsIDs (N = 19,386) summary statistics were available for 5,429,268 common (MAF \geq 0.05) and 5,860,562 rare (MAF < 0.05) SNPs.

The Parental Lifespan GWAS consists of unrelated, European-ancestry individuals reporting a total of 512,047 mother and 500,193 father lifespans, of which 60% were complete. The statistics for each participating cohort were calculated by fitting Cox survival models to father and mother survival separately, adjusted for

subject sex, at least 10 principal components, and study-specific covariates such as genotyping batch and array. Martingale residuals of the survival models were regressed against subject dosages (HRC-imputed). Father and mother results were combined into two separate ways: father and mother residuals from UK Biobank were combined before regression, while father and mother summary statistics from other cohorts were meta-analysed, adjusting for the phenotypic correlation between parents. See Timmers et al. [88] for details. Summary statistics were available for 5,526,246 common ($MAF \geq 0.05$) and 3,559,402 rare ($MAF < 0.05$) SNPs.

The Longevity GWAS consist of unrelated, European-ancestry individuals who lived to an age above the 90th survival percentile ($N_{cases} = 11,262$) or whose age at the last follow-up visit (or age at death) was at or before the 60th percentile age ($N_{controls} = 25,483$). The statistics for each of the participating cohorts were calculated using logistic regression and 1000G Phase 1 version 3-imputed dosages, adjusted for clinical site, known family relationships, and/or the first four principal components (if applicable) and subsequently combined using a fixed-effect meta-analysis. See Deelen et al. [72] for details. After removing SNPs with duplicate IDs ($N = 17,152$), summary statistics were available for 6,657,238 common ($MAF \geq 0.05$) and 2,181,962 rare ($MAF < 0.05$) SNPs.

We carried out a series of new, age-stratified GWAS using a sample of 325,614 unrelated, British-ancestry individuals from UK Biobank (as determined by genomic PCA and 3rd degree kinship or closer) [87], in order to calculate age band-specific effects of SNPs on lifespan. These individuals answered questions regarding their family history via touchscreen questionnaire, including their adoption status and parental age or age at death if deceased. Quality control was performed as in Timmers et al. [88], starting with 409,692 British-ancestry individuals and excluding subjects who were adopted, had two parents who died before age 40, or who did not provide information on parental age ($N = 12,406$; 3.0%). Additionally, we excluded individuals who had withdrawn their consent to participate as of 16 October 2018 and all but one of each related set of individuals

($N = 71,672$; 17.5%). Related individuals were excluded as mixed modelling is not well understood in the context of the kin-cohort method [88]. The remaining 325,614 individuals reported 312,088 and 322,672 father and mother lifespans, respectively, of which 67.7% were complete. Parent lifespans were then split into three age bands, 40–60, 60–80, and 80–120, excluding parents who died before the start of the age band and treating any parent who survived at least until the end of the age band as alive (i.e. right-censored). Cox proportional hazard models were fitted separately to each father and mother age band—six combinations in total—adjusted for subject sex, genotyping batch and array, and the first 40 genetic principal components.

$$h(x) = h_0(x)e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}$$

Where $h(x)$ is the hazard of the parent at age x , h_0 the baseline hazard, and $\beta_{1,2,\dots,n}$ the effect sizes (natural log of the hazard ratio) associated with the covariates $X_{1,2,\dots,n}$. Martingale residuals of these models were taken [69], divided by the proportion dead to scale effects to hazard ratios and doubled to account for parental genotype imputation [70], and then regressed against subject allelic dosage in an additive model using RegScan [211]. Individual parental lifespan statistics were combined using inverse-variance meta-analysis, inflating standard errors by $\sqrt{1 + r_p}$ to take into account the correlation between the parental phenotypes (r_p).

LD-score regression [198] was used to calculate genetic correlations between ageing trait GWAS, age-stratified parental lifespan (described above) and 27 European-ancestry GWAS of developmental, behavioural, and disease traits ([Table S1](#)). In line with recommendations [197], imperfectly imputed ($INFO < 0.9$) and low frequency ($MAF < 0.05$) SNPs, as well as those located in the Major Histocompatibility Complex, were discarded before merging the summary statistics with a HapMap3 reference panel to reduce statistical noise. An average of 1,086,952 SNPs (range 866,405–1,181,238) were used to calculate genetic

correlations per set of summary statistics, based on LD-score regression weights derived from European individuals.

Healthspan, parental lifespan, and longevity summary statistics were meta-analysed using MANOVA, while accounting for correlations between studies due to (limited) sample overlap and correlation amongst the traits, as implemented in MultiABEL v1.1-6 [210]. Correlations were calculated from summary statistics by taking the correlation in effect estimates from independent SNPs between studies (60,338 default SNPs provided by MultiABEL and shared between studies). These correlation estimates ranged from 0.013 between healthspan and longevity to 0.094 between healthspan and parental lifespan, reflecting a small degree of sample overlap and/or phenotypic correlation. Summary association statistics were calculated for the 7,320,282 SNPs shared between studies, of which 5,278,109 were common ($MAF \geq 0.05$) and 2,042,173 were rare ($MAF < 0.05$). These statistics represent the significance of each SNP affecting one or more of the traits, giving a P value against the null hypothesis that effect sizes are zero in all studies. The method does not provide a combined effect size.

Loci were defined as 500 kb regions flanking the lead genome-wide significant SNP in linkage equilibrium ($r^2 < 0.1$) with other lead SNPs. LD-score regression was used to assess inflation of the GWAS statistics, using 1,138,687 SNPs from the MANOVA and LD weights from European samples from the 1000 Genomes project. Loci with lead SNPs showing a nominally significant effect ($P < 0.05$) in all three datasets were considered more likely to capture intrinsic ageing pathways. We refer to them as “loci of interest” throughout this study.

Lead SNPs of loci of interest were looked up in individual father and mother survival statistics from Timmers et al. [88]. Differences in the parental effect sizes were tested using $(\beta_1 - \beta_2) / \sqrt{(\sigma_1^2 + \sigma_2^2)}$ which follows a Z distribution, assuming effects are independent.

Age-specific survival statistics were retrieved for the same loci from our age-stratified parental lifespan GWAS in UK Biobank. In order to standardise effects

for each locus, we expressed the age-specific effect as a fold change from the unstratified effect in UK Biobank, inflating standard errors using the Taylor series expansion to account for the uncertainty in the denominator:

$$\alpha = \frac{\beta_{band}}{\beta_{all}} - 1$$

$$SE_{\alpha} = \sqrt{\frac{SE^2_{band}}{\beta^2_{all}} + \frac{\beta^2_{band} SE^2_{all}}{(SE^2_{all})^2}}$$

Where α is the fold change in effect, β_{band} is the effect estimate of the age-specific band, β_{all} is the unstratified effect estimate, and SE is the standard error of the respective effects.

This provided a relative change in effect size per parental age band. We then calculated the median survival from Kaplan-Meier survival curves of each age band, allowing us to place the effects on a years-of-life scale. For each locus individually, effect sizes of age bands were regressed against median survival of the age band, inversely weighted by the variance of the effect estimate. Coefficients of the loci underpowered to detect a trend individually ($P > 0.05$) were meta-analysed, again weighted by the inverse of their variance, to provide a collective estimate. A sensitivity analysis examining the collective trend estimate using all loci of interest (instead of only underpowered loci) was performed using the “meta” R package and found substantial heterogeneity ($I^2 > 89\%$) driven by *APOE*, which represented almost 70% of the regression weights.

Lead SNPs from the multivariate GWAS and close proxies ($r^2_{EUR} > 0.6$) were looked up in the GWAS catalog (Buniello et al. 2019; 14 October 2019) and PhenoScanner [212]. All genome-wide associations were included except triallelic SNPs, associations without effect sizes, and associations with healthspan, lifespan, longevity, or medications. Similar traits were then grouped together using approximate string matching—verified manually—keeping only the strongest association and the shortest trait name. For example, “Body mass

index”, “Body mass index in smokers”, and “Body mass index in females greater than 50 years of age” were grouped and renamed to “Body mass index”. Associations were then categorised into seven disease phenotypes based on keywords and manual curation: Cardiovascular, Metabolic, Neuropsychiatric, Immune-related, Smoking-related, Cancer, and Age-related. Cardiovascular phenotypes included lipid levels, vascular traits, and diseases concerning the heart; Metabolic phenotypes included body (fat) mass and glycaemic traits; Neuropsychiatric phenotypes included neurodegenerative diseases and disorders of brain signalling such as restless leg syndrome and epilepsy; Immune-related phenotypes included measures of immune cells, and inflammatory and autoimmune disorders; Smoking-related phenotypes included smoking and lung function-related traits; Cancer included all neoplasms and carcinomas; Age-related phenotypes included traits typically associated with advancing age, such as age at menopause, male pattern baldness, age-related macular degeneration, hearing loss, and frailty. See [Table S2](#) for a list of all phenotypes within each category.

For each locus of interest, gene expression was tested for colocalisation with SNP effects within 500 kb of the lead SNP using SMR-HEIDI [191,213]. The gene expression studies included Westra (cis-eQTL), CAGE (cis-eQTL), Vosa (cis- and trans-eQTL), and GTEx v7 [169,192,204,214], the latter with eQTL $P < 10^{-5}$ only. Estimates of SNP effects are needed for SMR but are not directly provided by the multivariate analysis. Instead, we derived Z scores from multivariate P values and signed these based on the sign of the sum of underlying healthspan, parental lifespan, and longevity Z scores. The HEIDI statistic is dependent on the heterogeneity between effect estimates. We therefore recalculated standard errors and effect sizes based on allele frequency and sample size, using formula 6 from Zhu et al. [191]. For sample size, we used the sum of individual studies’ effective samples ($N = 709,709$) and performed a sensitivity test using the sum of all samples (regardless of their contribution to study power; $N = 1,349,432$). Differences in P_{HEIDI} between analyses were <0.0006 , i.e. had no practical effect on results. A Benjamini-Hochberg multiple testing correction was applied

separately to each eQTL dataset to account for the number of genes tested. Determining an optimal threshold for heterogeneity pruning is less straightforward: Wu et al. [213] consider 5% to be too conservative, especially when using summary-level data and SNPs with different sample sizes, and set a 1% threshold to correct for three colocalisation tests. We apply the same threshold, which may still be conservative in our study as we test many (albeit partially overlapping) tissues and we expect additional heterogeneity due to inferred Z scores (see Discussion).

Genes colocalising with loci of interest in cis or trans at $FDR < 5\%$ were tested for enrichment in 50 GO hallmark and 7350 biological process gene sets from the Molecular Signatures Database [215], using a procedure analogous to Gene2Func in FUMA [216]. First, we translated all unique gene symbols from the eQTL datasets to Entrez IDs ($N = 24,670$), and subsetted hallmark and GO biological process gene sets to only include genes for which eQTL were available. We then used a hypergeometric test to assess whether our genes were overrepresented in each pathway compared to all genes with eQTL. A minimum of three genes had to be present in a gene set for it to be tested for enrichment. Seven hallmark gene sets and 383 biological process gene sets met this requirement. Bonferroni correction was applied to account for multiple testing, separately for hallmark and biological process sets. Gene sets with $P_{\text{bonferroni}} < 5\%$ are reported.

Results

Genetic correlations between survival traits

We explored three public, European-ancestry GWAS of overlapping ageing traits: healthspan (N = 300,477 individuals, 28.3% no longer healthy), parental lifespan (N = 1,012,240 parents, 60% deceased), and longevity (N_{cases} = 11,262; N_{controls} = 25,483). The traits show substantial genetic correlations ($P < 5 \times 10^{-8}$) despite differences in age demographic, trait definition, and study design. Parental lifespan correlates strongly with both healthspan ($r_g = 0.70$; SE = 0.04) and longevity ($r_g = 0.81$; SE = 0.08), while healthspan and longevity show a weaker correlation with each other ($r_g = 0.51$; SE = 0.09) (Figure 1a). We performed an age-stratified GWAS of parental lifespan in UK Biobank to assess whether the genetic correlations between the traits are age-dependent, but our results showed no clear trend in the correlations between healthspan/longevity and age-stratified lifespan bands (Figure 1b).

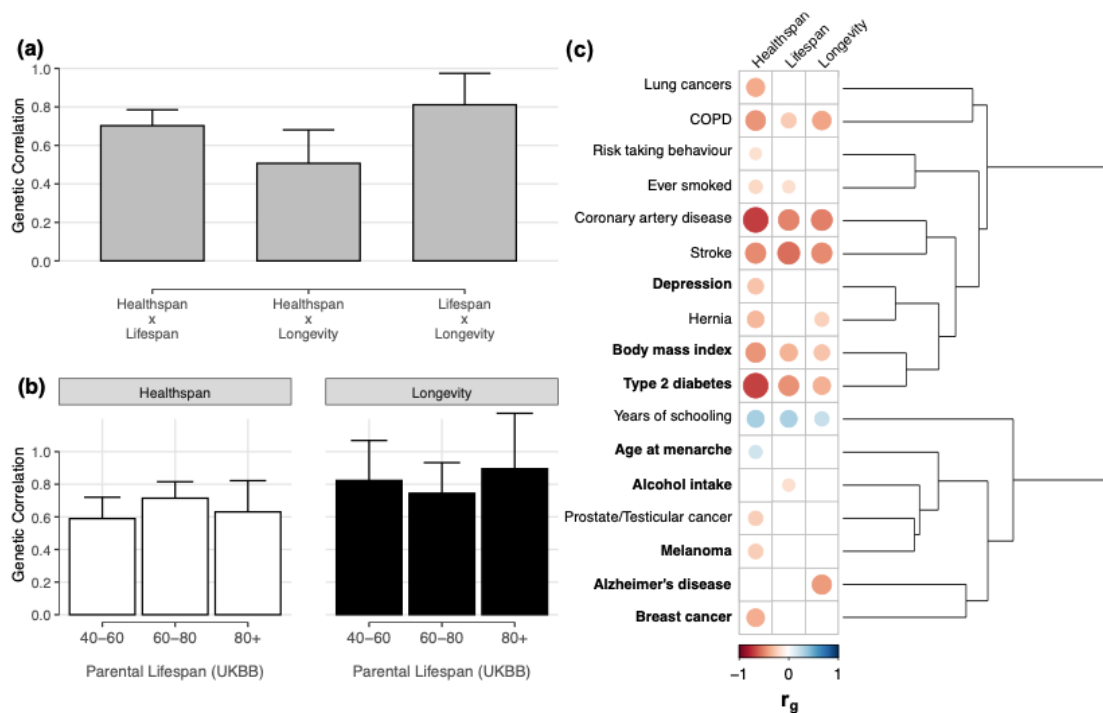


Figure 1: Healthspan, lifespan, and longevity are highly genetically correlated. a) Pairwise correlation between human ageing studies. b) Genetic correlations of age-stratified parental lifespan against healthspan and longevity. c) Genetic correlations of survival traits with traits related to development, behaviour, and disease. In bold are traits with heterogeneous correlations ($P_{het} < 0.05$). Displayed here are 17 traits which have at least one significant ($FDR < 5\%$) genetic

correlation with healthspan, lifespan, or longevity, out of the 27 traits tested. See [Table S3](#) for a full list of correlations. Blank squares represent correlations which did not pass multiple testing correction. Error bars represent 95% confidence intervals. COPD—Chronic Obstructive Pulmonary Disease.

We next tested whether differences in survival trait genetics could be explained by differences in genetic correlations with 27 other traits. We find all three survival traits show similar correlations ($P < 0.05/81$; $P_{\text{het}} > 0.05$) with coronary artery disease (range healthspan $r_g = -0.69$; SE = 0.07 to lifespan $r_g = -0.49$; SE = 0.10), stroke (range lifespan $r_g = -0.56$; SE = 0.11 to healthspan $r_g = -0.47$; SE = 0.06), chronic obstructive pulmonary disease (range healthspan $r_g = -0.45$; SE = 0.04 to lifespan $r_g = -0.26$; SE = 0.07), and years of schooling (range longevity $r_g = 0.24$; SE = 0.04 to healthspan $r_g = 0.34$; SE = 0.03). However, we also find evidence for differences in correlations across the traits ($P_{\text{het}} < 0.05$): healthspan correlated more strongly with metabolic traits (such as type 2 diabetes) than the other studies, and showed negative genetic correlations with depression and cancers, especially melanoma ($r_g = -0.25$; SE = 0.05), which were not observed in the other datasets. Conversely, parental lifespan correlated uniquely with alcohol intake ($r_g = -0.18$; SE = 0.06) and longevity showed a unique correlation with Alzheimer's disease ($r_g = -0.43$; SE = 0.11). ([Figure 1c](#); [Table S3](#)).

Genome-wide multivariate meta-analysis

Given the correlations amongst the traits, a combined MANOVA offered the prospect of increased power. We therefore performed a meta-analysis of GWAS of healthspan, parental lifespan, and longevity, which identified 24 loci at genome-wide significance ($P < 5 \times 10^{-8}$) ([Figure 2](#); [Table S4](#)). The combined statistics had an LD-score regression intercept of 1.064 (SE 0.009), suggesting limited inflation due to population stratification or relatedness. The *APOE* locus contained the most significant multivariate SNP ($P < 1 \times 10^{-126}$), associated with an average increase in lifespan of 12.7 months per allele (95% CI 11.4–14.0) and an increased odds ratio of reaching longevity of 1.66 (1.56–1.77). However, noting that <2% of the healthspan study sample experienced Alzheimer's disease, the

same allele was associated with an average healthspan increase of only around 50 days (2–98).

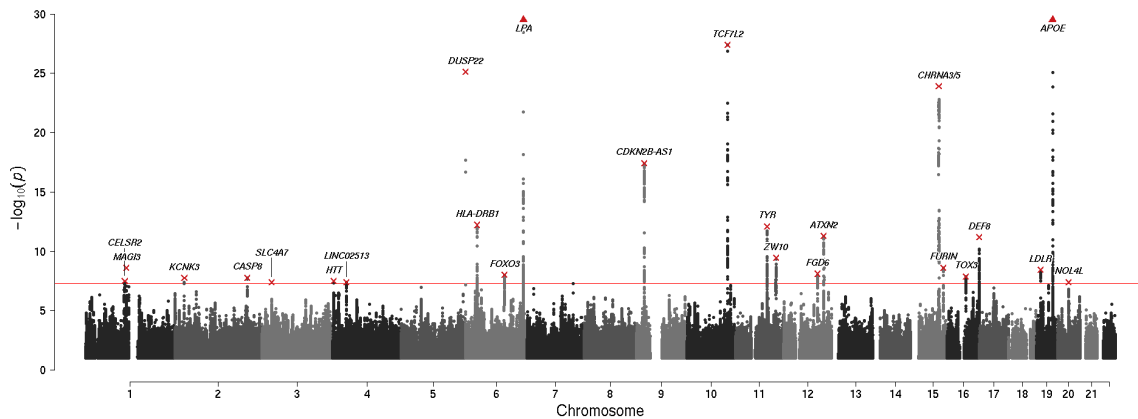


Figure 2: Twenty-four multivariate loci identified at genome-wide significance. Manhattan plot showing the strength of association $-\log_{10}(P \text{ value})$ on the y-axis against the chromosomal position of SNPs on the x-axis, where the null hypothesis is no association with healthspan, parental lifespan, and longevity. The red line represents the genome-wide significance threshold (5×10^{-8}). Annotated are the nearest gene(s) to the lead SNP (in red) of each locus. The y-axis has been capped at 5×10^{-30} to aid legibility; SNPs passing this cap are represented as triangles: *LPA* $P = 3.8 \times 10^{-30}$, *APOE* $P = 9.6 \times 10^{-127}$.

Twenty-one of the 24 multivariate GWAS loci reaching genome-wide significance had directionally consistent effects in the three studied datasets and 18 were nominally significant ($P < 0.05$) in two or more datasets ([Figure S1](#)). A look-up of the lead SNPs and close proxies in the GWAS catalog and PhenoScanner showed that healthspan-specific loci (i.e. $P < 0.05$ only in the healthspan dataset) were mostly associated with skin cancers and metabolic traits, while lifespan-specific loci were associated with smoking and risk taking ([Table S5](#)). Associations with these phenotypes suggests these variants influence (behaviours leading to) environmental exposures and thus likely capture extrinsic ageing processes. As we were primarily interested in genetic variation influencing the intrinsic ageing process, we focused the remainder of this study on genetic variants reaching nominal significance in all three datasets, which are less likely to be associated to study- or population-specific exposures.

Ten loci reached nominal significance ($P < 0.05$) in all ageing studies (Table 1). Five of these are of particular interest as they contain no genome-wide significant SNPs in any individual dataset. The lead multivariate SNP of these loci include rs2643826 (nearest gene *SLC4A7*), rs17499404 (*LINC01259*), rs1159806 (*FOXO3*), rs61905747 (*ZW10*), and rs12830425 (*FGD6*) (Figure S2-S6). The lead SNP near *FOXO3* is in moderate linkage disequilibrium (LD) ($r^2 > 0.4$) with rs2802292, a well-known candidate SNP from longevity studies [217].

Nearest Gene	rsID	Chr	Position	A1	Freq1	Healthspan			Lifespan			Longevity			MANOVA P
						Beta1	SE	P	Beta1	SE	P	Beta1	SE	P	
Novel															
<i>SLC4A7</i>	rs2643826	3	27562988	C	0.55	0.0210	4.9E-03	1.7E-05	0.0166	3.9E-03	2.2E-05	0.0451	0.0204	2.7E-02	3.95E-08
<i>LINC02513</i>	rs17499404	4	38385479	A	0.54	0.0165	4.9E-03	7.3E-04	0.0122	0.0039	1.6E-03	0.0837	1.9E-02	1.3E-05	3.94E-08
<i>FOXO3</i>	rs1159806	6	109006838	T	0.35	0.0144	0.0052	5.5E-03	0.0151	4.1E-03	2.2E-04	0.0953	2.0E-02	2.6E-06	9.83E-09
<i>ZW10</i>	rs61905747	11	113639842	A	0.82	0.0294	6.2E-03	2.0E-06	0.0237	4.9E-03	1.7E-06	0.0661	0.0259	1.1E-02	3.59E-10
<i>FGD6</i>	rs12830425	12	95580818	G	0.07	0.0436	9.3E-03	2.8E-06	0.0318	7.4E-03	1.8E-05	0.0774	0.0360	3.1E-02	7.85E-09
Known															
<i>LPA</i>	rs10455872	6	161010118	A	0.93	0.0574	9.0E-03	1.4E-10	0.0764	7.4E-03	8.5E-25	0.1236	0.0454	6.5E-03	3.80E-30
<i>CDKN2B-AS1</i>	rs7859727	9	22102165	C	0.51	0.0307	4.9E-03	2.6E-10	0.0250	3.9E-03	9.8E-11	0.0660	1.9E-02	5.8E-04	3.65E-18
<i>TOX3</i>	rs4783780	16	52571436	A	0.53	0.0233	4.9E-03	1.7E-06	0.0139	3.9E-03	3.0E-04	0.0524	0.0191	5.9E-03	1.33E-08
<i>LDLR</i>	rs6511720	19	11202306	T	0.12	0.0153	0.0075	4.0E-02	0.0339	6.0E-03	1.8E-08	0.0930	0.0301	2.0E-03	3.71E-09
<i>APOE</i>	rs429358	19	45411941	T	0.85	0.0137	0.0067	4.1E-02	0.1056	5.5E-03	3.1E-83	0.5098	3.2E-02	1.3E-56	9.61E-127

Table 1: Ten loci act across all three ageing traits, reaching nominal significance in each dataset.

Nearest gene—Gene closest to the index SNP; rsID—The SNP with the lowest P value in the multivariate analysis. Chr—Chromosome; Position—Base-pair position on chromosome (GRCh37); A1—the effect allele, increasing healthspan, lifespan, and odds to become long-lived; Freq1—Frequency of the A1 allele. Beta1—Effect size of the A1 allele, for healthspan and lifespan this is the negative log of the hazard ratio, for longevity this is the log odds of reaching an exceptional old age (90th percentile). SE—Standard error of the effect estimate. P—P value of the trait association. MANOVA P—P value against the null hypothesis of association with neither healthspan, lifespan, nor longevity. Novel loci contain SNPs that are not reported at genome-wide significance in any individual dataset. Known loci contain one or more genome-wide significant SNPs within 500 kb of the lead SNP in one of the individual datasets (Table S6).

Links with sex, age and age-related disease

We next tested whether loci of interest displayed varying effects on lifespan by sex, using sex-specific parental GWAS summary statistics from Timmers et al. [88]. We find evidence of sexual dimorphism for the ApoE $\epsilon 4$ allele ($\beta_{\text{fathers}} = 0.08$, $\beta_{\text{mothers}} = 0.13$, $P_{\text{diff}} < 1.5 \times 10^{-6}$) and evidence of no sexual dimorphism for lead SNPs near *LINC01259*, *SLC4A7*, *LPA*, *TOX3*, and *FOXO3* ($< 20\%$ difference or $P_{\text{diff}} > 0.50$). For the remaining loci near *CDKN2B-AS1*, *ZW10*, *FGD6* and *LDLR*, effect

size point estimates may differ by more than 20%, but we would need a larger sample size to be able to detect this difference with confidence ([Figure S7](#)).

Looking up the same SNPs in our age-stratified parental lifespan GWAS, we find that all loci, except *APOE* and *SLC4A7*, show a downward trend in effect size with parental age. This trend is highly significant for the *APOE* locus ($P = 8 \times 10^{-4}$), with the effect of the $\epsilon 4$ allele increasing by 32% (25%–39%) for every 10-year increase in parental survival. Conversely, the lead SNPs near *FOXO3* and *CDKN2B-AS1* show a nominally significant ($P < 0.05$) decrease in effect with age: for every 10-year increase in parental survival, SNP effects shrink by 38% (17%–60%) and 50% (19%–81%), respectively. While we are underpowered to confirm the trends for the remaining loci, we find that, collectively, the average effect of the protective alleles of these seven loci decreases by 14% (1%–27%; $P < 0.05$) for every 10-year increase in parental survival. ([Figure S8](#)).

We also found loci of interest had previously been associated at a genome-wide significant level with several age-related diseases and/or phenotypes. The life-extending allele of the majority of loci is associated with a reduction in cardiovascular disease phenotypes, including SNPs near the newly discovered ageing loci *SLC4A7*, *FGD6*, and *LINC01259*. Interestingly, protective variants near *FOXO3* are associated with a reduction in metabolic syndrome but also a reduction in cognitive ability. Life-extending SNPs near *APOE*, *FOXO3* and *FGD6* are all associated with increased measures of macular degeneration ([Figure S9](#); [Table S5](#)).

Ageing genes and pathways

Assessing the loci of interest for colocalisation with gene expression signals (eQTL), we find strong evidence ($FDR_{SMR} < 5\%$; $P_{HEIDI} > 1\%$; see Methods) of cis-acting eQTL colocalisation for eight out of 10 loci. In total, we highlight 28 unique genes acting across 32 tissues, especially whole blood (12 genes) and the tibial nerve (7 genes) ([Table S7](#)). In blood, higher expression levels of *BCL3* and *CKM* (near *APOE*); *CTC-510F12.2*, *ILF3*, *KANK2* and *PDE4A* (near *LDLR*); *USP28* and

ANKK1 (near *ZW10*); and *CDKN2B* are linked to an increase in multivariate lifespan traits, while the opposite is true for *EXOC3L2* (near *APOE*), *TTC12* (near *ZW10*), and *FOXO3*. For the multivariate signal near *SLC4A7* we find colocalisation of *NEK10* (liver); for the signal near *LPA* we find *SLC22A1/A3* (multiple tissues) and *MAP3K4* (pituitary); and for the signal near *FGD6* we find *FGD6* itself (adipose/arterial). Including trans-acting eQTL from blood while keeping the same thresholds for colocalisation, we additionally discover higher expression levels of *FOXO3B* colocalises with the life-extending signal near *FOXO3*. When we include genes which could not be tested for heterogeneity ($N_{eQTL} < 3$), we identify one additional cis-acting and 49 additional trans-acting genes (of which 10 colocalise with the signal near *LINC02513*) ([Table 2](#); [Table S7](#)).

Locus	Chr	Position	Cis-Genes	Trans-Genes
<i>SLC4A7</i>	3	27562988	<i>NEK10</i> -	
<i>LINC02513</i>	4	38385479		<i>EDAR</i> +, <i>MAL</i> +, <i>NOSIP</i> +, <i>CCR7</i> +, <i>ABLIM1</i> +, <i>KRT72</i> +, <i>FHIT</i> +, <i>MMP28</i> +, <i>EPHX2</i> +, <i>LEF1</i> +
<i>FOXO3</i>	6	109006838	<i>LINC00222</i> -, <i>FOXO3</i> -	<i>FOXO3B</i> +, <i>MEGF6</i> +, <i>CALCOCO1</i> +, <i>CYBRD1</i> +, <i>IGF1R</i> +, <i>PHF21A</i> +, <i>NDRG1</i> +, <i>KIAA1324</i> -, <i>FCHO2</i> +, <i>CNNM3</i> +
<i>LPA</i>	6	161010118	<i>SLC22A1</i> +, <i>SLC22A3</i> -, <i>AL591069.1</i> -, <i>MAP3K4</i> -	
<i>CDKN2B-AS1</i>	9	22102165	<i>CDKN2B</i> +	
<i>ZW10</i>	11	113639842	<i>USP28</i> +, <i>ANKK1</i> +, <i>TTC12</i> -, <i>RP11-159N11.4</i> -, <i>ANKK1</i> -	
<i>FGD6</i>	12	95580818	<i>RP11-256L6.3</i> +, <i>FGD6</i> -	
<i>LDLR</i>	19	11202306	<i>CTC-510F12.2</i> +, <i>KANK2</i> +, <i>SPC24</i> +, <i>SLC44A2</i> +, <i>ILF3</i> +, <i>ILF3-AS1</i> -, <i>DOCK6</i> -, <i>SMARCA4</i> -, <i>PDE4A</i> +	<i>AHSP</i> -, <i>SELENBP1</i> -, <i>EPB42</i> -, <i>SLC4A1</i> -, <i>HBD</i> -, <i>CA1</i> -, <i>FAM46C</i> -, <i>BLVRB</i> -, <i>TMOD1</i> -, <i>GYPB</i> -, <i>UBE2O</i> -, <i>BPGM</i> -, <i>TRIM58</i> -, <i>SNCA</i> -, <i>IFIT1B</i> -, <i>FECH</i> -, <i>GMPR</i> -, <i>EPB49</i> -, <i>RBM38</i> -, <i>TNS1</i> -, <i>MICAL2</i> -, <i>DCAF12</i> -, <i>RAB31L1</i> -, <i>PDZK1IP1</i> -, <i>HBM</i> -, <i>BCL2L1</i> -, <i>PLEK2</i> -, <i>E2F2</i> -, <i>TGM2</i> -
<i>APOE</i>	19	45411941	<i>EXOC3L2</i> -, <i>AC006126.4</i> +, <i>CKM</i> +, <i>BCL3</i> +, <i>PVRL2</i> +	<i>LDLR</i> -

Table 2: eQTL for 78 genes colocalise with the GWAS signal at 9 out of 10 loci of interest. Genes which showed a significant effect (FDR < 5%) of gene expression on ageing traits are displayed here. Locus—Nearest gene to lead variant in the multivariate analysis. Chr—Chromosome. Position—Base-pair position of lead variant (GRCh37). Cis-Genes—Genes in physical proximity (<500 kb) to the lead variant of the locus which colocalise with the multivariate signal. Trans-Genes—Genes located more than 500 kb from the lead variant of the locus. Gene names are annotated with the direction of effect, where “+” and “-” indicate whether the life-extending association of the locus is linked with higher or lower gene expression, respectively.

Finally, testing this list of cis- and trans-acting genes for gene set enrichment in 50 hallmark and 7350 biological process pathways, we find significant enrichment ($P_{\text{adjusted}} < 0.05$) in seven hallmark gene sets and 32 biological processes. The hallmark gene sets with the strongest enrichment include haem metabolism, hypoxia, and early oestrogen response (Figure 3). Enriched biological pathways cluster into categories involving apoptotic signalling, chemical homeostasis, and development of erythrocytes and myeloid cells, among others (Figure S10; Table S8).

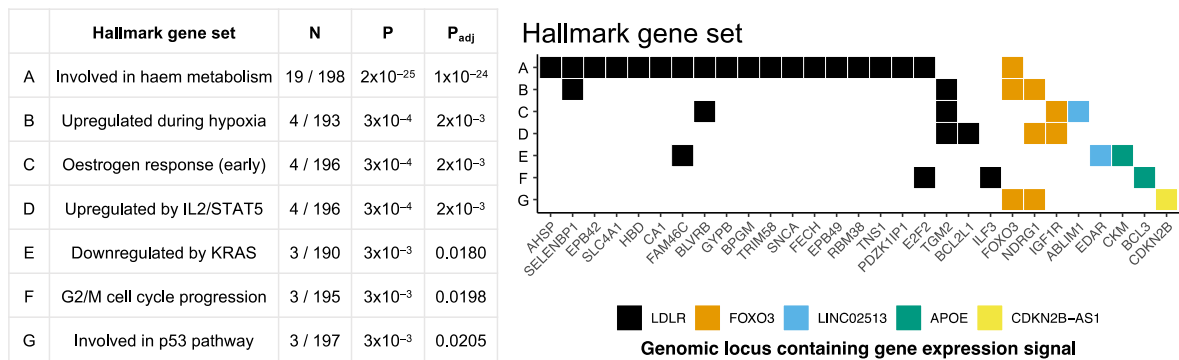


Figure 3: Seven hallmark gene pathways are enriched for ageing-related genes. N—number of genes of interest vs. total number of genes in the gene set for which eQTL are available. P—P value of the hypergeometric test for enrichment (against 24,670 background genes). P_{bonf} —Bonferroni-corrected P value for testing seven hallmark pathways (containing at least 3 genes). The figure shows individual genes on the x-axis and hallmark pathways on the y-axis, matching the order of the table. Squares represent the presence of a gene in the gene set.

Discussion

Genetic correlations between publicly available healthspan, parental lifespan, and longevity GWAS reveal these traits share 50% or more of their underlying genetics. Performing a multivariate meta-analysis on the GWAS summary statistics, we highlight 24 genomic regions influencing one or more of the traits. Ten regions are of particular interest as they associate with all three ageing traits and are as such likely candidates to capture intrinsic ageing processes, rather than extrinsic sources of ageing. Five of the loci of interest are not associated at a

genome-wide significant level in any individual dataset, including the region near *FOXO3* which has thus far only been highlighted in candidate gene association studies (reviewed in Sanese et al. 2019) and never at genome-wide significance. The effects of loci of interest on male and female lifespan are largely the same, although their effect on survival may be slightly stronger in middle age (40–60) compared to old age (80+). The ApoE ϵ 4 allele is exceptional in this regard as its effect is stronger in females and increases with age, likely due to its well-known link to Alzheimer’s disease [219]. Lastly, we find our loci of interest colocalise with the expression of 28 cis-genes and 50 trans-genes, which are enriched for seven hallmark gene sets (particularly haem metabolism) and 32 largely overlapping biological pathways (including apoptosis and homeostasis).

The antagonistic pleiotropy and hyperfunction theories of ageing predict the presence of genetic variants important for growth and development in early life with deleterious effects towards the end of the reproductive window [220,221]. While we are unable to directly capture the genetic effects on individuals before age 40 due to the study design of our datasets, we found the life-extending variant near *FOXO3* is associated with a delay in the age at menarche and a decrease in intracranial volume and cognitive abilities. Similarly, the lead life-extending SNP near *ZW10* shows an association with increased age at menarche, although not at a genome-wide significant level ($P = 6 \times 10^{-4}$) [222]. As such, it is clear that loci exhibiting antagonistic pleiotropy exist in humans. However, almost all loci of interest associate strongly with cardiovascular and blood cell phenotypes, without apparent antagonistic effects, in line with established knowledge that cardiovascular disease is a leading cause of mortality and morbidity worldwide [223].

The genes we identify are enriched for many pathways previously related to ageing in eukaryotic model organisms, including genomic stability, cellular senescence, and nutrient sensing [16]. For example, *FOXO3* and *IGF1R* are well-known players modulating survival in response to dietary restriction [224], but we also highlight many genes involved in the response to DNA damage and apoptosis, such as *CDKN2B*, *USP28*, *E2F2*, and *BCL3*. In addition to hallmarks

discovered in model organisms, we highlight the haem metabolism gene set as important in human ageing. This pathway includes genes involved in processing haem and differentiation of erythroblasts [215]. Although the enrichment is largely driven by genes linked to the *LDLR* locus, genes linked to other loci of interest (such as *FOXO3*, *CDKN2B*, *LINC02513*) are involved in similar biological pathways: myeloid differentiation, erythrocyte homeostasis, and chemical homeostasis.

Haem synthesis declines with age and its deficiency leads to iron accumulation, oxidative stress, and mitochondrial dysfunction [225]. In the brain, abnormal iron homeostasis is commonly seen in neurodegenerative diseases such as Alzheimer's, Parkinson's and multiple sclerosis [226]. Plasma ferritin concentration, a proxy for iron accumulation, has been associated with premature mortality in observational studies [227], and has been linked to liver disease, osteoarthritis, and systemic inflammation in Mendelian Randomisation studies [228,229].

A particular strength of this study is the ability to identify loci shared by multiple traits, without the need for additional sample collection. Comparing the strength of the multivariate association at our 10 loci of interest with the strength of association within each individual GWAS, we estimate the combined statistics are equivalent to a median sample size increase of 127% (95% CI 52%–728%; ~380,876 individuals) for the healthspan study, 76% (23%–146%; ~768,578 parents) for the parental lifespan study, and 415% (59%–620%; ~64,810 cases) for the longevity study. This gain in power is particularly important for the latter since the sample size of GWAS for longevity will likely not improve in the near future due to limited availability of data on long-lived people. Having demonstrated the advantages of jointly studying three ageing traits, we encourage future studies to incorporate additional large-scale age-related trait GWAS, such as a recent study on frailty in UK Biobank [230], to further improve power.

It is clear from the association of age-related diseases and the well-known ageing loci *APOE* and *FOXO3* that we are capturing the human ageing process to some extent; however, some judgment is involved in definitions. For one, there are currently no widely accepted standards for measuring healthspan [231]. Zenin et al. [206] define healthspan based on the incidence of the eight most common diseases increasing exponentially in incidence with age in their sample. As such, their trait is highly dependent on the characteristics of the UK Biobank cohort, who were aged 40–69 years when they were recruited in 2006–2010 and of which two-thirds have yet to experience an age-related disease. Therefore, loci with effects on diseases of middle age (cancer and heart disease) are likely overrepresented in our analysis. The lack of Alzheimer’s disease in the UK Biobank sample also explains the limited association of *APOE* in the healthspan GWAS, compared to the other ageing traits.

Multivariate analysis of traits does not provide a natural combined effect size or direction of effect. Colocalisation of eQTL with loci of interest requires effect directions to test for heterogeneity of instruments. As such, we used the direction of the sum of the Z scores of the underlying traits to assign a direction to Z scores derived from MANOVA P values. This works well for SNPs with concordant effects on ageing traits but is less accurate when SNPs have heterogeneous or antagonistic effects. For example, a SNP associated with an increase in healthspan and an equal decrease in lifespan—while likely rare—will have a large Z score in the MANOVA, but no clear direction of effect. This limitation will introduce some heterogeneity in the colocalisation analysis, and as a result inflate the HEIDI statistic. Furthermore, gene expression colocalisation is limited by the number of tissue eQTL (with some tissues being underpowered) and does not capture the effect of coding variation. There may be additional genes with highly tissue-specific effects or effects dependent on structure or splicing isoforms, which we are unable to detect.

The pathways we have highlighted are mostly biological processes for chemical and cellular homeostasis and are therefore likely to be generalisable across

populations; however, it is important to note that all GWAS summary statistics used in our study were derived from individuals from European ancestries and more follow-up work is necessary to validate our findings in individuals from other ethnic backgrounds. For example, certain population characteristics, such as levels of obesity and meat intake can affect the bioavailability of iron [232] and thus the relative importance of haem metabolism in ageing.

Importantly, the genes we have highlighted show natural variation in the human population and are therefore more likely candidates for therapeutic intervention. However, colocalisation of gene expression could be due to pleiotropy rather than causality, and there is a need to validate the effects of genetic variants in experimental models to confirm their role in disease aetiology. For example, we have found life-extending variants colocalise with decreased expression of *FOXO3* in blood, which aligns with previous work by Peters et al. [174] showing *FOXO3* expression increases with age, but experiments suggest the gene has many protective functions including detoxification of reactive oxygen species and DNA damage repair [218]. The observed inverse relationship between healthy life and *FOXO3* expression may reflect healthy individuals have less oxidative damage and require less *FOXO3* to mitigate this damage.

In conclusion, the challenge of studying ageing genetics in humans—low heritability and limited samples—can be overcome to some extent by combining large studies of closely related phenotypes that capture elements of ageing process. Focusing on the overlap between different populations and age-related traits has revealed that several ageing pathways discovered in model organisms also apply to humans, and has highlighted new pathways in humans which can now be further tested in model organisms. This study, and follow-up work on the genes we have highlighted, will eventually lead to new therapeutic targets that can reduce the burden of age-related diseases, extend the healthy years of life, and increase the chances of becoming long-lived without long periods of morbidity.

Acknowledgements

We would like to thank the authors of the many GWAS used in this work for making their summary statistics publicly available. We would also like to acknowledge funding from the Medical Research Council (PRHJT: MR/N013166/1, JFW: MC_UU_00007/10); the University of Edinburgh (PRHJT, PKJ); and the Alexander von Humboldt Foundation (JD).

Author Contributions

PRHJT: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing—Original draft preparation, Writing—Review & editing, Visualization. JFW: Supervision, Writing—Review & editing. PKJ: Conceptualization, Supervision, Project administration, Writing—Review & editing. JD: Conceptualization, Writing—Original draft preparation, Writing—Review & editing

4.3 Conclusion

Joint modelling of the genetics of healthspan, lifespan, and longevity has revealed the presence of genomic regions determining both the length and quality of life. We identified 24 loci affecting one or more ageing phenotypes, of which 10 had a significant effect on all three traits. Comparing the results with Chapter 3, we found 10 out of the 12 genomic regions highlighted for their role in determining parental lifespan also showed nominal evidence ($P < 0.05$) of a directionally concordant effect on healthspan or longevity. The exceptions were the regions near *CHRNA3/5* and *FURIN/FES*, which were associated with parental lifespan at genome-wide significance but showed evidence of no effect in the healthspan study and showed evidence of a directionally concordant albeit underpowered effect in the longevity study. As before, we tried looking for sex- and age-specific effects on parental lifespan, but did not find strong evidence for this beyond the *APOE* locus, which had a greater effect on older mothers. In the future, sex-specific data for healthspan and longevity may refine our estimate of the sex-specific effects we were underpowered to detect.

Loci significantly affecting one ageing trait but not the others appeared to associate with trait- and/or population-specific risk factors, such as obesity, skin cancer, smoking, and risk taking. Conversely, most of the loci significant for all three traits associated with CVD, once again highlighting the role of this disease in determining the end of (healthy) life. However, despite the overlap with Chapter 3 in CVD associations, the gene sets with evidence for enrichment in this study were quite different: the former mainly highlighted lipid-metabolism-related pathways, while this study highlighted pathways related to haem metabolism, apoptosis, and (intra-)cellular homeostasis. These findings are not necessarily contradictory, however, as lipid and iron levels are involved in CVD progression (which I discuss further in the next Chapter).

This study identified five genomic loci which had thus far not been discovered at genome-wide significance. While their role in health and survival is supported by associations with life-limiting diseases in independent studies, additional work is necessary to formally replicate our findings. The exception is the *FOXO3* locus, which has been highlighted as a candidate region numerous times but had never been discovered at genome-wide significance before. The statistical power of our multivariate analysis allowed us to perform a colocalisation analysis between our GWAS signal and gene expression signals in the region. This analysis confirmed roles for *FOXO3* and *IGF1R* expression, matching experimental work that initially highlighted the role of these genes in determining the lifespan of model organisms. Genes which associate with ageing traits in human studies and causally affect survival in model organisms have some of the most robust evidence for being biological determinants of human lifespan and represent the most promising targets for pharmaceutical intervention to date.

Finally, the discovery of haem metabolism as a putative pathway in human ageing highlights the importance of human-oriented genetic research on lifespan determinants. Whether hallmarks of ageing identified in model organisms apply equally to humans is hard to confirm without studies on humans, and there is no guarantee that interventions which extend model organism lifespan will have any effect on human survival. However, if biological pathways discovered in humans, such as haem metabolism, can be confirmed by follow-up work in model organisms, this would advance our understanding of which ageing processes are most relevant to our own lifespan, and how we can design interventions—be it dietary, lifestyle, medical, or otherwise—to affect these processes and reduce the burden of ageing and age-related disease.

Chapter 5: Discussion

I set out to investigate the determinants of human lifespan, both in terms of how trends in diseases have influenced population mortality rates, and in terms of how genetics can make certain individuals more susceptible to disease and death than others. In doing so, I hoped to deepen our understanding of the ageing process in humans and create new opportunities to address the growing burden of ageing and age-related disease in society.

5.1 Trends in disease and their effect on mortality

In order to understand determinants of population life expectancy, I first performed a survival analysis of almost the entire adult Scottish population in collaboration with National Health Service Scotland and University of Edinburgh researchers. We showed life expectancy has increased by 3.5 years for every decade of birth, matching a steady trend that started in the 1950s[233] and appears to be slowing down more recently[124]. Perhaps unsurprisingly, a large amount of the gains in life expectancy can be attributed to a reduction in the incidence of major diseases and improved recovery after diagnosis[91].

Cardiovascular disease (CVD) and cancer are the most common causes of death in high-income countries[234], and our study found the majority of improvements in mortality in Scotland could be attributed to improvements in diseases belonging to these categories, including ischaemic heart disease, respiratory organ cancers, and digestive organ cancers. For both CVD and cancers, we found a substantial decline in disease incidence (as measured by the incidence of hospital admission) whilst for cancer, we also observed substantial improvements in survival after disease hospitalisation. Across the 28 diseases we studied, we estimated around 60% of the total improvements in life expectancy

could be accounted for by improvements in survival after hospital admission, with the remainder attributable to falls in disease incidence.

In terms of disease incidence, the observed decrease in incidence of hospitalisation for CVD does not provide direct evidence of improved prevention or lifestyle intervention. However, it is likely the discovery and introduction of preventative CVD medication, such as antihypertensive drugs in 1967[235] and cholesterol-lowering medication in 1987[236] have decreased CVD incidence over time. Indeed, since the expiry of the simvastatin patent in the early 2000s, prescription of these preventative drugs continues to increase in Scotland[237], and average levels of total cholesterol and blood pressure continue to decline in the population[137]. Analogously, the incidence of cancers, especially lung cancer, could be falling in Scotland due to a decreasing prevalence of smokers and exposure to cigarettes (which is simultaneously linked to fewer CVD deaths[137,238]. This decline follows a 2005 ban on indoor smoking in public spaces[239], as well as a more recent strategic campaign aiming to shift attitudes towards smoking (with some success in younger individuals[238]), reduce exposure to tobacco advertising, and raise awareness of the harms of second-hand smoke[240].

Among improvements in disease survival, we saw particularly large advances in survival post-admission for breast and prostate cancers. Indeed, cancer treatments have undergone progressive improvements over time, from the rise of chemotherapy in the 1960s[241] to more recent use of combination therapies which include cancer-specific monoclonal antibodies[242]. At the same time, more timely diagnoses due to nationwide screening of breast[243] and bowel cancer[244], and earlier provision of end-of-life care[245] have prolonged cancer survival post-diagnosis. However, compared to other developed nations, such as Canada, Australia and Norway, UK cancer survival rates are still lagging behind[246].

The use of hospital admission records in our study means our analysis works best for diseases with a close link between hospital admission and disease incidence. Other studies have found recent increases in Alzheimer's disease and drug use disorders contributing to mortality in Scotland[120], diseases which are not accurately captured through hospital admissions and may be important determinants of lifespan for older and younger individuals, respectively. However, our study also identified worsening trends in infectious disease, independent from the population age, with influenza and pneumonia already accounting for the most disease-related deaths in Scotland, and should these trends continue, it is likely infectious disease will become an even larger contributor to mortality in the future. As infectious diseases are both more common and deadly in elderly individuals[247], it is likely this burden will become even more severe as population demographics shift to have more elderly individuals. As such, our study supports an increase in emphasis on infectious disease prevention and antibiotic resistance research in public health policy of high-income countries, a direction the UK recently started to take[248].

In all, our results are suggestive that the Scottish population has become more resilient over time, both with respect to the incidence of disease and to recovery after hospital admission. Whether improvements in lifestyle, environmental exposures, and preventative care have slowed down the rate of ageing and the associated incidence of age-related disease, or instead a plethora of changes each relating to a specific disease have collectively decreased the incidence of disease is hard for us to discern from the data. In all likelihood, the observed decline in disease incidence is a combination of both. Regardless, improvements in public health and disease survival have driven the remarkable improvements in life expectancy in early 21st century Scotland, but whether life expectancy will continue to rise in the future will depend on our ability to overcome slowdowns in the rate of improvement in CVD and cancers, and our ability to tackle the growing challenge of infectious disease.

5.2 The role of genetic factors on human lifespan

As recognised in the introduction, the heritability of lifespan is low but not immaterial. With exceptionally large samples, it is possible to detect genomic regions influencing survival. In an international collaboration with three other groups, I investigated the genetics of 1 million lifespans in the UK and Europe. We identified 12 genomic regions associated with lifespan in a univariate framework and 7 more when taking into account mortality risk factors. These loci influenced susceptibility to Alzheimer's disease, cardiovascular disease (CVD), and smoking-related disease. Comparing evidence between our study and previously published work, we raised the total number of lifespan loci with robust evidence for replication from four in 2017 (ref. [70]) to 14 at time of publishing[88]. Since then, another large GWAS of parental lifespan has been performed, this one using genotypes from the direct-to-consumer AncestryDNA cohort[249]. This study has provided additional evidence of replication for seven loci and identified another locus that could be confirmed by our own study. As a result, the growing body of literature regarding the genetic determinants of human survival now includes 22 genome-wide significant loci with robust evidence for replication ([Table 3](#)).

Locus	rsID	A1	Beta1	Phenotype	Disease	Discovery	Replication
<i>APOE</i>	rs429358	T	1.06	Longevity	Cardiometabolic Neuropsychiatric	Deelen et al. [250]	Joshi et al. [90]
<i>FOXO3</i>	rs3800231	A	0.17	Longevity	Cardiometabolic Neuropsychiatric	Flachsbar et al. [165]	Joshi et al. [70]
5q33.3/ <i>EBF1</i>	rs2149954	T	0.09	Longevity	Cardiometabolic	Deelen et al. [250]	Timmers et al. [88]
<i>LPA</i>	rs10455872	A	0.76	Lifespan	Cardiometabolic	Joshi et al. [70]	Joshi et al. [70]
<i>HLA-DQA1</i>	rs34967069	T	0.56	Lifespan	Immune-related	Joshi et al. [70]	Joshi et al. [70]
<i>CHRNA3/5</i>	rs8042849	T	0.44	Lifespan	Smoking-related	Joshi et al. [90]	Joshi et al. [90]
<i>LDLR</i>	rs142158911	A	0.36	Lifespan	Cardiometabolic	Timmers et al. [88]	Wright et al. [249]
<i>SH2B3/ATXN2</i>	rs11065979	C	0.28	Lifespan	Immune-related Cardiometabolic	Pilling et al. [209]	Timmers et al. [88]
<i>CDKN2B-AS1</i>	rs1556516	G	0.25	Lifespan	Cardiometabolic	Pilling et al. [209]	Timmers et al. [88]
<i>FURIN/FES</i>	rs6224	G	0.25	Lifespan	Cardiometabolic	Pilling et al. [209]	Timmers et al. [88]
<i>ABO</i>	rs651007	C	0.21	Lifespan	Cardiometabolic	Timmers et al. [88]	Timmers et al. [88]
<i>ZC3HC1</i>	rs11556924	T	0.20	Lifespan	Cardiometabolic	Timmers et al. [88]	Timmers et al. [88]
<i>MIR129-2</i>	rs4755202	A	0.13	Lifespan	Cardiometabolic	Wright et al. [249]	Timmers et al. [88]
13q21.31	rs61949650	C	0.53	Lifespan (Female)	Cancer	Pilling et al. [209]	Timmers et al. [88]
<i>PSORS1C3</i>	rs3130507	G	0.33	Lifespan (Female)	Immune-related	Pilling et al. [209]	Timmers et al. [88]
<i>IP6K1</i>	rs9872864	G	0.25	Lifespan (Female)	Cardiometabolic	Pilling et al. [209]	Wright et al. [249]
<i>SEMA6D</i>	rs4774495	G	0.31	Lifespan (Male)	Smoking-related	Pilling et al. [209]	Wright et al. [249]
<i>EPHX2</i>	rs7844965	G	0.30	Lifespan (Male)	Smoking-related	Pilling et al. [209]	Wright et al. [249]
<i>ZW10</i>	rs61905747	A	0.30	Lifespan (Male)	-	Pilling et al. [209]	Timmers et al. [88]
<i>CELSR2/PSRC1</i>	rs599839	G	0.29	Lifespan (Male)	Cardiometabolic	Pilling et al. [209]	Wright et al. [249]
<i>MICAB</i>	rs3131621	G	0.22	Lifespan (Male)	Immune-related	Pilling et al. [209]	Wright et al. [249]
<i>LPL</i>	rs15285	G	0.23	Lifespan (Male)	Cardiometabolic	Pilling et al. [209]	Wright et al. [249]

Table 3: Overview of genomic regions robustly associated with human survival. These regions have been associated with either longevity or (parental) lifespan at genome-wide

significance and have been replicated in at least one other study at nominal significance. Locus—Nearest gene or cytogenetic band; rsID—Lead variant in the region; A1—Effect allele, increasing survival; Beta1—Average years of life gained per effect allele (from Timmers et al. [88]); Phenotype—Survival phenotype for which the locus was originally discovered, where longevity refers to a case-control study and lifespan refers to a quantitative study of (parental) lifespan (phenotypes may differ between discovery and replication studies); Disease—Disease categories associated with the region; Discovery—Study which discovered the region at genome-wide significance; Replication—One of the studies which replicated the original association.

Empirical evidence from GWAS suggest the number of hits increases roughly proportionally to increases in the population sample size (i.e. double the sample, double the hits), after studies have reached a sample size able to detect genome-wide associations[251]. With the advent of sample sizes of hundreds of thousands to millions of individuals, it appears discovery of hits may even accelerate once sample sizes reach a critical threshold[252]. This accelerating discovery of hits appears to hold true for parental lifespan as well, where a sample of ~135,000 individuals (272,081 parents) detected two hits[90], a sample of ~300,000 individuals (606,059 parents) detected four hits[70], and a sample of ~500,000 individuals (1,012,240 parents) has now detected 12 hits[88]. Given the highly polygenic nature of lifespan, it is likely that hundreds of regions will eventually associate with the trait as sample sizes continue to increase and overcome statistical noise (as has been the case for height[253]). In fact, it is conceivable almost all genes will affect lifespan to some extent if disrupted[254]. However, similar to genetic studies of height, if hundreds of lifespan loci are discovered across the genome, these can still elucidate which pathways and cell types are central to the determining variation in the trait, and could improve the prediction accuracy of polygenic scores[255].

However, while GWAS have an outstanding track record of generating replicable findings[256], many findings from GWAS of lifespan-related traits have not stood the test of time, and as such any discoveries should be thoroughly evaluated before being accepted as real. For example, an early case-control GWAS of over

1,000 centenarians claimed to have found 70 hits[257], but was retracted after concerns over the genotyping array came to light[258] and it was discovered the hits were all false positives[259]. Similarly, GWAS of parental lifespan in LifeGen[70] has provided evidence against 8 candidate variants from previously published studies, and we drew another 8 candidate variants into question using the UK Biobank and LifeGen meta-analysis[88]. With the exception of *APOE*, common variants with large effects on lifespan or longevity have either been refuted[88], as was the case for the *d3-GHR* deletion predicted to increase male lifespan by 10 years[167], or were found to be replicable but at a much smaller effect sizes than originally estimated[70].

Despite these concerns, the variants discovered in our own lifespan GWAS can be considered reliable because they are supported by multiple lines of evidence. First, we observed directional concordance with results from a longevity GWAS performed in an independent population for each individual hit, and we were able to provide collective replication (even when excluding *APOE*) when considering the effect of all variants at once. Secondly, every single hit (except rs61348208 near *HTT*) had also been associated with one or more lifespan-limiting diseases in previous, unrelated studies. While the region near *HTT* did not show any association with disease under our strict inclusion criteria, SNPs within the region actually do associate strongly with Huntington's disease ($P = 4 \times 10^{-16}$) but were missed as no effect size was reported in the original disease study[260]. Lastly, the AncestryDNA parental lifespan study ($N \sim 482,000$ parents) did not directly compare their results with our study, but within the limited summary statistics they shared, we found independent evidence for known loci (*APOE*, *CHRNA3/5*, *LPA*, *SH2B3/ATXN2*, *CDKN2B-AST*) as well as the newly discovered *LDLR* locus[249].

At first sight it is disappointing to find each of these variants has only a small effect on survival, from a month to a year at most. However, this does not preclude GWAS of lifespan-related traits from being useful. For one, it appears magnitude of the effects of genetic variants is independent from their ability to provide

biological insights or predict viable drug targets[261]. So far, lifespan and longevity GWAS have consistently highlighted the role of lipid metabolism[88,164,209], suggesting genetic differences in the way individuals process lipids—whether from their diet or endogenous sources—could contribute to differences in survival. Well-established causal relationships between lipid metabolism and high-mortality age-related diseases such as Alzheimer’s[262] and CVD[263] provide further evidence for the link between lipids and longevity. It remains difficult to discern whether this association is mediated only through deadly disease, or if the lipid pathway also influences the rate of the ageing process itself as well as disease-specific mortality.

Another way lifespan variants can be useful (despite their small effects) is demonstrated by their ability to predict differences in survival between individuals when aggregated into polygenic survival scores. While the score derived from UK Biobank and LifeGen parents is the first of its kind, the score already shows a strong association with all-cause mortality in both the UK and Estonia, and it will undoubtedly be further refined by larger and more diverse samples in the future. Currently, only genetic tests for single, high-penetrance mutations are used to inform decisions in the clinic[264], but there is evidence that polygenic risk scores using common genetic variants may prove to be equally if not more informative for certain diseases[105]. Indeed, the NHS has expressed an interest in incorporating genomic information in its services to provide personalised care[265].

Genetic tests for progressive, incurable diseases are also currently performed, including a number of neurodegenerative diseases[266]. An accurate diagnosis explaining worrying symptoms can provide some relief[267], while also allowing individuals to take into account their condition while planning for the future. Similarly, a score for all-cause mortality could be useful in preparing individuals to mitigate their increased risk for multiple age-related diseases (in terms of lifestyle intervention or preventative medication), especially if interventions can be tailored. Genetic scores could also be used to make decisions regarding

pensions and life insurance. In an actuarial context, the possibility of adverse selection due to genetic risk is already recognised[268], and there are ethical considerations to be made whether polygenic survival scores should or should not be used to discriminate in such financial services[187]. However, a detailed review of the ethical and financial implications of genomic scores is beyond the scope of this work. From a clinical standpoint, I believe polygenic scores for specific diseases and cause-specific mortality will have more immediate benefits than a score for all-cause mortality. Nonetheless, future, well-powered polygenic survival scores may be used in clinical risk assessments, where they could identify resilient individuals, more likely to withstand adverse medical procedures such as surgery or chemotherapy.

5.3 Insights into the ageing process

I next analysed the genetics of healthspan (the number of years lived free from severe disease), lifespan (the total number of years lived), and longevity (being amongst the oldest 10%) with support from a researcher from the Max Planck Institute for Biology of Ageing. Together, we revealed these three ageing-related traits overlapped substantially in their genetics and we identified 24 genomic loci associating with one or more of the phenotypes. However, only 10 of these loci appeared to influence all three traits, and a number of loci associated exclusively with one trait but not the others. In theory, genes able to extend healthspan without affecting lifespan are of immense medical interest, as they could reveal biological pathways able to compress the period of morbidity, reducing the number of years lived in poor health. However, our observed trait-specific associations are likely the result of extrinsic risk factors unique to the populations under study, rather than molecular mechanisms compressing morbidity. For example, we found genomic loci associated with healthspan (but not lifespan or longevity) were strongly associated with tanning, sunburn, and skin cancers,

while genomic loci unique to lifespan were associated with smoking and risk taking[Chapter 4].

Ageing-related traits can be highly variable phenotypes, sensitive to the characteristics of the populations in which they are measured, as differences in common diseases and causes of death can influence which genetic variants have large effects. While the healthspan and lifespan GWAS we included in our study both relied on data from UK Biobank, the former studied the current generation[206] and the latter studied the parental generation[88]. Shifts in mortality and morbidity between generations could therefore have influenced the relative importance of genomic loci.

Two major changes between generations have been the fall in age-standardised incidence of lung cancer for UK men, which decreased 30% between 1980 and 2010, and the rise of age-standardised incidence of malignant melanoma, which tripled in the same period[269]. For women, lung cancer incidence remained comparatively low, but malignant melanoma incidence has also increased. Our own work on Scottish individuals supports an accelerating decrease in lung cancer incidence between decades of birth and suggests the contribution of lung cancer to mortality has fallen over time[91]. Therefore, it is highly plausible that smoking-related cancers and deaths played a much larger role in the parental generation and by extension the lifespan GWAS. In contrast, the contribution of skin cancer to death was less important in the parental generation, but its increasing incidence in the current generation has made it a common healthspan-ending event, and by extension an important factor in the healthspan GWAS. Smoking and sun exposure are population-specific, extrinsic risk factors. As such, genomic loci associating with these risk factors are less likely to provide insight into ageing mechanisms.

Analogously, genomic loci which affect multiple ageing traits at once, especially if these traits have been measured in multiple populations, are more likely to capture intrinsic features of the ageing process. In the context of our multivariate

analysis, we discovered that almost all genomic regions which were associated with an increase in healthspan, lifespan, and longevity also associated with decreased CVD or CVD risk factors, much like the loci discovered in previous studies on individual age-related traits[70,88,209]. However, unlike previous studies, the genes and pathways identified in this analysis aligned more closely with haem homeostasis and apoptosis rather than lipid metabolism. Indeed, not only do haem and iron play a role in vascular injury[270], chronic heart failure[271], and atherosclerosis[272], dysregulation of iron levels is also linked to other disease of old age, such as bacterial[273] and viral infections[274], cancer[275], type 2 diabetes[276], neurodegeneration[277], sarcopenia, chronic pain, and frailty[278].

Interestingly, two of the most promising longevity drug candidates, metformin and rapamycin, have each been implicated in iron metabolism. Metformin is an anti-diabetic drug currently being investigated for its effect on human lifespan in clinical trials[279] following promising work in model organisms showing both extended healthspan and lifespan, despite researchers not yet fully understanding its mechanism of action[115]. However, there is evidence which suggests metformin reduces serum iron levels[280] and influences intracellular iron levels by reducing the level of protein-bound iron, which can induce an iron deficiency-like stress response[281]. Similarly, rapamycin is an immunosuppressive drug shown to extend mouse lifespan[116] and improve human skin ageing at low doses[282], which has recently led to a clinical trial for lifespan in dogs[283]. Rapamycin use in humans has been linked to reduced serum iron levels[284,285] and has been shown to prevent iron accumulation in senescent cells[286].

Despite the multitude of studies suggesting a relationship between haem metabolism and the diseases of old age (and longevity drugs), some caution is needed in interpreting the link as causal. It is possible loss of iron homeostasis is a symptom of the progressive deterioration associated with ageing rather than a driving force. In the context of our multivariate study of ageing traits, observing

an enrichment in haem-related genes only confirms individuals carrying protective genetic variants have altered levels of haem-related gene expression, and does not reveal whether this expression causes their health- and/or life-extension. However, experimental perturbation of a number of genes highlighted in our study has already been shown to influence model organism lifespan, including decreased expression of the orthologues of *APOE*, *IGF1R*, and *MAP4K3* in mice[287–289], and the increased expression of the *FOXO3* orthologue in worms and flies[290]. Similar knockout or overexpression experiments using short-lived models may clarify the causal direction of the remaining genes. Genes which reliably extend model organism lifespan, rather than abrogate it, are of specific interest as these are the most likely to yield useful drug targets[19,291].

5.4 Sex and socioeconomic determinants of lifespan

While not the main factors investigated in this body of work, the role of sex and socioeconomic status on lifespan must be acknowledged. We found a difference of 3.5 years in mean life expectancy between Scottish men and women while accounting for socioeconomic differences[91]. This is very similar to the decade-on-decade improvements we observed in life expectancy and is comparable in magnitude to the difference in survival between top and bottom deciles of polygenic survival scores in Estonian individuals[88]. In other words, life expectancy for men in Scotland today is similar to that of women ten years ago, and women carrying mostly deleterious lifespan SNPs will have a similar life expectancy to men carrying mostly protective SNPs. Long-term records for the UK show the male-female mortality gap was smallest in the early 19th century (~2 years), increased sharply until it peaked around 1970 (~6 years), and has since started to diminish again[292]. The exact reasons for the existence and closure of the lifespan gap are not completely known, although they likely involve sex-specific behaviour, biology, and environmental exposures[293].

The prevalence of smoking in the UK has historically been much higher in men, bringing with it a larger burden of lung cancer mortality[294]. From around 1950 onwards, public health campaigns started to decrease the number of male smokers, although they continue to outnumber women[295]. Using Scottish hospital admissions, we observed both lung cancers and CVD resulted in greater burden of mortality for men than women, but we also found recent improvements in the incidence of these diseases were greater for men, which could explain the closing gap[91]. Similarly, Sundberg *et al.*[296] observed recent improvements in male lung cancer and CVD deaths in Swedish cause-of-death data, while female deaths from smoking and Alzheimer's disease continued to increase[296].

In terms of biology, differences in male and female lifespan could be partly driven by differences in sex chromosomes. There is evidence across species for a shorter lifespan in the heterogametic sex (e.g. XY in humans, ZW in birds), in line with the hypothesis that recessive deleterious mutations are more likely to be expressed in the heterogametic sex without a second copy of the same chromosome to mask the mutation[297]. Additionally, both X and Y chromosomes experience mosaic structural abnormalities over time, although the rate of mosaicism is higher for the Y chromosome and there is an association between the amount of Y mosaicism and early mortality[298–300]. In women, mosaic structural abnormalities occur more commonly in the inactivated X chromosome, likely masking some of the deleterious effects[299]. Thus far, studies of ageing-related traits have only examined autosomal chromosomes, where there is limited evidence for sex-specific effects of genetic variants affecting survival[Chapter 4]. Future genetic studies including sex chromosomes may be able to more accurately quantify the contribution of biological differences to the gender gap in lifespan.

Compared to the effect of sex and common genetic variants, socioeconomic status has a relatively large influence on lifespan. Scottish death records showed differences in median life expectancy of almost 10 years between the top and bottom decile of the Scottish Index of Multiple Deprivation (SIMD)[91], and the

Office for National Statistics found similar differences in England and Wales using a similar measure of deprivation[301]. While age-adjusted mortality in the most socially deprived individuals is higher, their relative mortality following hospital admission is similar to individuals from less deprived areas, suggesting differences in life expectancy in the UK is not due to differences in quality of treatment or end-of-life care[91]. Instead, our research shows almost all diseases with a high mortality burden (in terms of prevalence and deadliness) have a higher incidence in individuals from more deprived areas.

Deprivation is linked to lower levels of education[302], which itself results in lower awareness of mortality risk factors[303]. Indeed, rates of smoking and alcohol intake are higher and quality of sleep and physical activity are lower in more socioeconomically deprived individuals[304]. However, not only are these individuals more likely to have unhealthy lifestyles, the same lifestyle factors are associated with greater harm (in terms of CVD and mortality) in more deprived individuals[304]. This may be in part because long-term adversity, especially during childhood, can accelerate ageing and reduce resilience to harmful lifestyle factors in middle age[305]. With disparities in healthspan and lifespan between socioeconomic groups appearing to increase over time[301], improved health education and governmental support may be needed to reduce the burden of disease and death equally for everyone.

5.5 Future work

5.5.1 Lifespan or longevity?

Sample sizes will continue to increase for both epidemiological and genetic studies into human survival. Given multiple strategies have now been tried to identify ageing genes and pathways in genetic studies, we are able to assess their performance and make recommendations for future research.

The study of exceptional longevity has been the most widely used strategy to investigate the genetics of human ageing, with at least ten longevity GWAS being performed in the last decade[72,82–84,158,166,217,250,306,307]. However, these studies remain limited by their sample sizes. The largest study on longevity to date has collected an impressive 13,617 cases[72] after pooling together participants from 19 cohorts, some of which started their recruitment of elderly individuals almost 30 years ago[308,309]. However, the largest longevity sample is dwarfed by contemporary case-control samples for other diseases, many of which are arguably easier to study due to higher heritability and lower polygenicity. For example, a 2018 study into type 2 diabetes had over 62,000 cases and identified 143 independent loci[310]. In comparison, identification of new longevity loci has remained extremely challenging, with most GWAS unable to detect any loci at genome-wide significance beyond the well-known *APOE* locus. In addition, robust evidence of replication has thus far relied on parental lifespan GWAS in population cohorts, as most of the longevity sample is needed for the discovery phase[72,88].

Our own work suggests parental lifespan largely captures the same genetics as longevity, with genetic correlations between the traits exceeding 80% (95% CI 65%–97%)[Chapter 4]. However, unlike longevity cohorts, the study of parental lifespan has been possible in large population cohorts, which are easy to recruit across the world and are rapidly expanding[311]. Despite reductions in effective sample sizes due to parental genotype imputation and exclusion of related individuals, our analysis of 1 million parents was roughly equal in power to a longevity case-control study of 44,500 nonagenarians[88]. Future studies may take into account related individuals to improve precision of parental genotype imputation, as was recently suggested by Hwang *et al.*[312], which could mitigate some of the loss in effective sample size. Larger sample sizes and ease of recruitment indicate the study of the genetics of human ageing is most effectively done using the parental lifespan phenotype. Existing longevity data is useful for replication[88] and can increase power when combined with parental lifespan studies[Chapter 4], but without substantial increases in sample sizes, new

longevity studies are unlikely to yield significant advances in the discovery of common genetic variants determining lifespan. Similarly, whether these studies are able to detect rare genetic mutations with large effects on late-life survival (analogous to early-acting Mendelian disease mutations), may remain unknown for decades until enough individuals have outlived their birth cohorts to become centenarians.

However, the study of parental lifespan has a number of downsides as well, such as the lack of parental covariates and generational differences in disease and death. Our study of Scottish morbidities has shown large effects of socioeconomic status and smoking-related behaviour on mortality[91], which usually cannot be accounted for in parental lifespan studies due to lack of available data. As a result, there is a risk of finding genetic variants which affect lifespan through pathways unrelated to the ageing process. One example is the *CHRNA3/5* locus, which has one of the largest effects on mortality, but this is likely mediated through its effect on smoking behaviour and lung cancer[313]. Similarly, the effects of genetic variants in the parental generation may not translate perfectly to the current generation, as is demonstrated by the attenuated effect of polygenic survival scores—generated from parental survival statistics—on genotyped subjects themselves.

In the near future, these discrepancies may diminish when lifespan GWAS start to include deceased participants rather than parents from large genotyped cohorts, which have now reached sample sizes of thousands in some cohorts, including ~20,000 deaths in UK Biobank[99], ~30,000 deaths in Biobank Japan[314], and ~12,000 deaths in the Norwegian HUNT cohort[315]. These cohorts may also be able to quantify the bias introduced in the parental lifespan studies when unable to account for environmental mortality risk factors and allow gene-environment interactions for existing parental lifespan loci to be quantified. In this regard, future GWAS of subject lifespan, while significantly less powered than parental lifespan, will lead to more reliable and relevant biological findings and polygenic survival scores.

5.5.2 Trans-ethnic studies of survival

Research into the genetic determinants of human lifespan has largely been limited to samples of European ancestries, the exception being GWAS performed on (the same) Han Chinese centenarian sample[72,158,307], and one analysis of the parental lifespans of African ancestry individuals resident in the UK[70]. Due to the low heritability and high polygenicity of lifespan as a trait, very large sample sizes are required to detect the small effects of longevity variants, and non-European ancestry cohorts have thus far not reached sizes large enough to discover new genetic variants which could be replicated in other studies[70,88].

However, there are numerous advantages to the inclusion of individuals from multiple ancestries in GWAS, as long as analyses are stratified by ethnicity or can be adjusted for admixture to avoid false positives[316]. The most straightforward advantage is an increase in power to detect true associations with the trait of interest, as genetic effects across ethnicities tend to be directionally consistent, although estimates of their magnitude can vary[317]. This consideration can be extended to polygenic risk scores, which are most accurate in predicting traits in individuals from the same population they were generated from, but often have poor predictive power in other populations if they do not contain information from more diverse cohorts[318]. Furthermore, differences in ancestry result in differences in LD structure which allows genomic loci to be mapped more precisely: trans-ethnic analyses can narrow down the number of candidate causal variants in an LD block, yielding a clearer idea of the functional consequences of the causal variant(s)[319].

Despite these advantages, power to detect associations is not always increased when including more diverse cohorts. Differences in population risk factors for a disease, such as diet, lifestyle, and environmental exposures, or even differences in the prevalence and treatment of a disease can introduce heterogeneity in effect

size estimates for genetic variants[320]. For lifespan, heterogeneity in mortality risk factors between populations may actually be useful to differentiate genetic variants which associate with early death due to external, population-specific causes from variants associating with extended longevity due to slowdown of common ageing processes (if the two can be separated). However, differences in allele frequencies between populations—or at its extreme, monomorphism—can also result in *bona fide* genetic variants being missed in one population but not the other[321], although tools have been developed to take into account these forms of heterogeneity with a minimal loss of power, allowing trans-ethnic association analyses to be performed as long as sample sizes are large enough[322,323].

Indeed, the advent of population-specific genotyping arrays and multi-ancestry imputation reference panels such as TOPmed[324] has led to a rapid increase in the number and size of cohorts with individuals from non-European ancestries (for examples, see China Kadoorie Biobank[325], US Million Veterans Program[326], and Biobank Japan[327]. GWAS combining samples from multiple ancestries have already been performed on thousands of non-European individuals for traits like blood pressure[328], asthma[329] and haematological traits[62]. In the future, any cohort with information on mortality, be it individual or parental, could be included in work on lifespan to accelerate genetic discoveries, with trans-ethnic analysis being both more likely to fine-map causal variants and highlight pathways which affect ageing in all humans, not just Europeans.

5.5.3 Rare and recessive variants

Evolutionary theory predicts mutation-selection balance will eliminate genetic variants with large deleterious effects on survival, causing any such variants in the population to be rare and/or recessive[330]. Additionally, if the antagonistic pleiotropy theory of ageing holds, we expect variants with large positive effects

on survival (instead of negative) to be selected against as well, as they may reduce other elements of reproductive fitness[220]. Empirical evidence from a large study on human autozygosity supports a role for recessive variants in ageing, finding individuals with higher autozygosity (thus expressing more recessive variant effects) were more likely to display signs of ageing such as lower self-reported health, grip strength, and walking pace, and higher measures of facial ageing and hearing/vision loss[331]. Quantifying the contribution of recessive effects on lifespan, heritability studies suggest the additional variation in lifespan explained by dominance is smaller than the variation explained by additive genetic effects, but not insignificant ($\sim 4\%$)[40].

However, thus far genome-wide scans for lifespan variants have focused on identifying common, additive genetic effects[88]. Examining recessive variants requires larger sample sizes when the recessive allele is the minor allele in the population (as will likely be the case for most recessive lifespan variants), because of the low number of homozygous carriers. For a trait such as lifespan, the analysis is further complicated when regressing parental phenotypes against subject genotypes: observed effect sizes are halved in an additive model but are divided by twice the recessive allele frequency in a dominance model. To illustrate, detection of an additive genetic effect using parental phenotypes requires four times the subject sample, but detection of a recessive effect of a common variant with an allele frequency of 20% would require 100 times the sample. However, as discussed above, incorporating population allele frequencies and relatedness to improve parental genotype imputation would greatly reduce this burden[312], and the rapidly increasing number of cohorts with subject and parent mortality data may make a genome-wide analysis of recessive variants soon a possibility.

Rare variants may also harbour large effects on lifespan without suffering from the same parental imputation drawback as recessive variants. There is evidence for a number of polygenic traits that much of the heritability missed by common genetic variation may be hidden in rare variants of large effects[332]. Recently,

advances in next generation sequencing have facilitated the collection of high-quality exome and whole-genome data on large samples, soon including the whole of UK Biobank[333]. Rare variants pose additional problems as they require stringent quality control[334,335], may violate statistical approximations which rely on large samples to hold[336], and suffer from a larger multiple testing burden[337]. However, increasingly tools are being developed to address these statistical problems, such as aggregation tests which take into account multiple rare variants at once[338], and comparative and functional genomics which can focus analyses on deleterious variants[339]. It is clear whole-exome and -genome analyses are set to become widely used and more sophisticated over time, as has been the case for GWAS of common genetic variants[340]. With larger samples and better tools come opportunities to find the large effect genes which have been absent from human lifespan GWAS thus far, which will improve lifespan prediction and may provide additional insight into the ageing process.

5.5.4 Biomarkers of lifespan

One avenue of research which has become possible due to the advancements in understanding the underlying determinants of human lifespan is the identification and modification of ageing biomarkers. Multiple studies have looked for predictors of biological age, from blood-based markers[341–343] to epigenetic clocks[344,345], to indicators of frailty[346], which show higher rates of mortality and disease in individuals with higher predicted biological ages. While the prediction of disease and death using relatively inexpensive assays or measurements has obvious clinical benefits (see PRS discussion above for genetic markers), it also has the additional advantage of quantifying improvements from ageing interventions. For example, a recent study was able to measure—albeit in a small and uncontrolled, all-male sample—the effect of metformin, growth hormone, and dehydroepiandrosterone treatment on human ageing without having to wait for individuals to die or to rely on proxy measures such as blood

pressure or BMI[347]. However, current research on ageing biomarkers provides limited information on the causality of these markers. In order to identify causal targets, future work could apply a Mendelian randomisation (MR) framework to the latest lifespan GWAS results.

MR uses a genetic instrumental variable to establish a causal pathway between an exposure, such as LDL cholesterol, and an outcome, such as death. Its ability to establish causality relies on several crucial assumptions: 1) the instrumental variable must be robustly associated with the exposure, 2) the instrumental variable must affect the outcome only through the exposure (i.e. no horizontal pleiotropy), and 3) the instrumental variable must not associate with any confounding variables. A 2017 study used MR to investigate the causal effect of diseases and risk factors on mortality in LifeGen[70]. Unsurprisingly, it showed diseases such as Alzheimer's disease, CVD, type 2 diabetes, and lung/breast cancer shortened lifespan, but also provided causal estimates for clinical measures, including deleterious effects of LDL cholesterol, total cholesterol, triglycerides, fasting insulin, apolipoprotein B, and blood pressure on lifespan; and a protective effect of C-reactive protein and HDL cholesterol[70]. Our own risk factor-informed iGWAS of parental lifespan of the LifeGen and UK Biobank meta-analysis also used MR to estimate causal effects of mortality risk factors on lifespan and found these were highly concordant with the LifeGen results (although this is expected to some extent as the samples overlap)[88].

Genetic variants make for great instrumental variables when the exposure is a simple biomarker such as gene expression, as there is usually a single pathway from the variant to the exposure (e.g. alternative allele → transcription factor binding affinity → gene expression) and from there, the outcome. However, when there are many more biological steps between the genetic variant and the exposure (e.g. alternative allele → ... → type 2 diabetes), the likelihood increases that the variant affects the outcome through pathways other than the exposure, thus violating the second assumption. For example, both Joshi *et al.*[70] and our own study[88] observe a positive causal effect between education and lifespan,

but it is impossible to exclude the possibility that the instrumental variables affect lifespan through other pathways than education. Nonetheless, increasingly MR methods are being developed to mitigate[348] or even exploit[349] violations of these assumptions. Also, the success of MR in predicting which lipid metabolites would be effective targets in CVD drug trials raises the possibility the method will be successful for lifespan biomarkers as well[350].

Now, with the advent of large studies mapping the genetics of hundreds of biological markers, from transcriptomics[204] to proteomics[351] to metabolomics[352], it has even become possible to perform hypothesis-free MR analyses. Whereas previously only the suspected effects of a handful of exposures on an outcome were tested, large biomarker GWAS now allow for genome-wide MR studies of biomarkers on an outcome. Combined with highly powered genetics of the outcome, such as our parental lifespan GWAS of over 1 million individuals or our ageing GWAS meta-analysis, the search for targetable biomarkers is set to begin in earnest. For one, an MR analysis of haem-related phenotypes could test to what extent iron levels determine lifespan, while a multi-omics MR study of lifespan could identify new biological targets against which to develop therapeutics which delay the onset of age-related disease and decrease the burden of ageing.

5.7 Conclusion

It is clear human lifespan has been increasing in developed countries, bringing with it longer lives but also longer periods of age-related disease and disability. We discovered the steady increases in life expectancy—3.5 years per decade in Scotland—can be attributed mainly to reductions in mortality associated with cancers and CVD. Related studies showed these reductions are concurrent with a rise in Alzheimer's disease in the community. We predicted from our hospital admission data that, should current trends continue, infectious disease will also

become a major public health challenge in the near future, especially as these diseases disproportionately affect the elderly and individuals from more socioeconomically deprived backgrounds. Together, our findings suggest a need for healthcare policy to focus on preventing infectious diseases in those vulnerable groups while accelerating care in other high-mortality diseases, like Alzheimer's disease, CVD, and cancers, in order to have the greatest impact on future lifespan improvements.

Looking at differences in lifespan between people rather than over time, we found the largest differences in survival can be explained by an individual's sex (+3.5 years for females) and socioeconomic deprivation (-1 year per decile), largely reflecting lifestyle differences which influence disease incidence, such as smoking-related behaviour. Common genetic variation can also explain differences in survival, with individuals scoring in the top decile of polygenic score for survival living three to five years longer than those scoring in the bottom decile. These extra years of life are likely gained due to a reduction in disease, as high-scoring individuals have a lower incidence of Alzheimer's disease, CVD, and smoking-related cancer, although their likelihood of other cancers appears unaffected. The study of genetic variation using parental lifespans has been much more informative than the study of exceptionally old individuals due to small sample sizes in the latter. As population cohorts expand both in size and diversity, polygenic survival scores are set to become increasingly predictive—especially when incorporating rare variants—and may be used in the future to inform both clinical and financial decisions.

Examining the genetics of lifespan also has the potential to reveal new insights into the biology of ageing and identify therapeutics that could mitigate the increasing disease risk associated with advancing age. We and others have highlighted lipid metabolism as a key pathway in determining lifespan, but our comprehensive analysis comparing the genetics of different ageing-related traits additionally revealed the importance of haem metabolism in ageing. The implication of haem and iron homeostasis in myriad age-related diseases and

their link to the most promising longevity therapeutics to date provides encouraging evidence that haem metabolism is a pathway that can be targeted to extend healthy life. Additionally, information about the genetics of ageing-related traits can be used in the future to identify biomarkers with causal effects on survival, both yielding new targets for intervention as well as excellent markers to track the biological age and mortality risk of individuals over time. These markers are set to transform clinical trials on ageing interventions, where trials that would have taken a lifetime to complete in the past could be performed in months to a few years by monitoring changes in causal mortality markers.

Across the decades, human lifespan has progressed through multiple stages—from rapid growth due to vaccination and antibiotics, to steady growth due to CVD and cancer treatments, and now slowing growth due to the rise of Alzheimer’s disease and infections. I predict the next stage is on the horizon. Health inequalities are worsening, and our medical system is struggling under an ageing population. However, my work has shown that our understanding of the determinants of lifespan has advanced to a point where we have the tools and data necessary to direct healthcare policy, and where we can start to develop the pharmaceutical interventions needed to dramatically reduce the burden of ageing and improve the lives of all.

Bibliography

- 1 Cutler D, Deaton A, Lleras-Muney A. The Determinants of Mortality. *J Econ Perspect* 2006;**20**:97–120. doi:10.1257/jep.20.3.97
- 2 United Nations Department of Economic and Social Affairs Population Division. World Population Prospects 2019: Volume II: Demographic Profiles. 2019.
- 3 UK Office for National Statistics. Living longer and old-age dependency – what does the future hold? 2019.
- 4 Stafford M, Steventon A, Thorlby R, *et al*. Briefing: Understanding the health care needs of people with multiple health conditions. 2018;:1–26.
- 5 Partridge L, Deelen J, Slagboom PE. Facing up to the global challenges of ageing. *Nature* 2018;**561**:45–56. doi:10.1038/s41586-018-0457-8
- 6 UK Office for National Statistics. Living longer: how our population is changing and why it matters. 2018.
- 7 UK Office for National Statistics. Health state life expectancies, UK: 2014 to 2016. 2017.
- 8 Guzman-Castillo M, Ahmadi-Abhari S, Bandosz P, *et al*. Forecasted trends in disability and life expectancy in England and Wales up to 2025: a modelling study. *Lancet Public Heal* 2017;**2**:e307–13. doi:10.1016/S2468-2667(17)30091-9
- 9 Thorlby R, Gardner T, Turton C. NHS performance and waiting times Priorities for the next government. 2019.
- 10 Holman HR. Chronic disease and the healthcare crisis. *Chronic Illn* 2005;**1**:265–74. doi:10.1177/17423953050010040601
- 11 Kennedy BK, Berger SL, Brunet A, *et al*. Geroscience: Linking Aging to Chronic Disease. *Cell* 2014;**159**:709–13. doi:10.1016/j.CELL.2014.10.039
- 12 Gorre C. Global Alliance for Chronic Diseases. *Impact* 2017;**2017**:4–5. doi:10.21820/23987073.2017.4.4
- 13 UK Office for National Statistics. Period and cohort life expectancy explained: December 2019. 2019.
- 14 National Records of Scotland. Vital Events Reference Tables 2013. 2019.
- 15 UK Office for National Statistics. Vital statistics in the UK: births, deaths and marriages. 2019.
- 16 López-Otín C, Blasco MA, Partridge L, *et al*. The hallmarks of aging. *Cell* 2013;**153**:1194–217. doi:10.1016/j.cell.2013.05.039
- 17 Gompertz B. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philos Trans R Soc London* 1825;**115**:513–83.
- 18 Crumer AM. *Comparison Between Weibull and Cox Proportional Hazards Models*. 2008.
- 19 Klass MR. A method for the isolation of longevity mutants in the nematode *Caenorhabditis elegans* and initial results. *Mech Ageing Dev* 1983;**22**:279–86. doi:10.1016/0047-6374(83)90082-9
- 20 Friedman DB, Johnson TE. A mutation in the age-1 gene in *Caenorhabditis elegans* lengthens life and reduces hermaphrodite fertility. *Genetics* 1988;**118**:75–86. doi:10.1016/S0960-9822(00)00522-4

- 21 Morris JZ, Tissenbaum HA, Ruvkun G. A phosphatidylinositol-3-OH kinase family member regulating longevity and diapause in *Caenorhabditis elegans*. *Nature* 1996;**382**:536–9. doi:10.1038/382536a0
- 22 Kenyon C, Chang J, Gensch E, *et al.* A *C. elegans* mutant that lives twice as long as wild type. *Nature* 1993;**366**:461. doi:10.1038/366461a0
- 23 Kimura KD, Tissenbaum HA, Liu Y, *et al.* Daf-2, an insulin receptor-like gene that regulates longevity and diapause in *Caenorhabditis elegans*. *Science (80-)* 1997;**277**:942–6. doi:10.1126/science.277.5328.942
- 24 Ogg S, Paradis S, Gottlieb S, *et al.* The fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in *C. elegans*. *Nature* 1997;**389**:994–9. doi:10.1038/40194
- 25 Powers RW, Kaeberlein M, Caldwell SD, *et al.* Extension of chronological life span in yeast by decreased TOR pathway signaling. *Genes Dev* 2006;**20**:174–84. doi:10.1101/gad.1381406
- 26 Rogina B, Reenan RA, Nilsen SP, *et al.* Extended life-span conferred by cotransporter gene mutations in *Drosophila*. *Science* 2000;**290**:2137–40. doi:10.1126/science.290.5499.2137
- 27 Flurkey K, Papaconstantinou J, Miller RA, *et al.* Lifespan extension and delayed immune and collagen aging in mutant mice with defects in growth hormone production. *Proc Natl Acad Sci U S A* 2001;**98**:6736–41. doi:10.1073/pnas.111158898
- 28 Mayer PJ. Inheritance of longevity evinces no secular trend among members of six New England families born 1650-1874. *Am J Hum Biol* 1991;**3**:49–58. doi:10.1002/ajhb.1310030109
- 29 Sebastiani P, Sun FX, Andersen SL, *et al.* Families Enriched for Exceptional Longevity also have Increased Health-Span: Findings from the Long Life Family Study. *Front Public Heal* 2013;**1**:38. doi:10.3389/fpubh.2013.00038
- 30 Verweij KJH, Mosing MA, Zietsch BP, *et al.* Estimating heritability from twin studies. *Methods Mol Biol* 2012;**850**:151–70. doi:10.1007/978-1-61779-555-8_9
- 31 Herskind AM, McGue M, Holm N V, *et al.* The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900. *Hum Genet* 1996;**97**:319–23.
- 32 Skytthe A, Pedersen NL, Kaprio J, *et al.* Longevity Studies in GenomeUtwinn. *Twin Res* 2003;**6**:448–54. doi:10.1375/136905203770326457
- 33 Beckwith J, Morris CA. Twin studies of political behavior: Untenable assumptions? *Perspect. Polit.* 2008;**6**:785–91. doi:10.1017/S1537592708081917
- 34 Richardson K, Norgate S. The equal environments assumption of classical twin studies may not hold. *Br J Educ Psychol* 2005;**75**:339–50. doi:10.1348/000709904X24690
- 35 Gunderson EP, Tsai A-L, Selby J V, *et al.* Twins of Mistaken Zygosity (TOMZ): Evidence for Genetic Contributions to Dietary Patterns and Physiologic Traits. *Twin Res Hum Genet* 2006;**9**:540–9. doi:10.1375/twin.9.4.540
- 36 Mitchell BD, Hsueh WC, King TM, *et al.* Heritability of life span in the Old Order Amish. *Am J Med Genet* 2001;**102**:346–52.
- 37 Gögele M, Pattaro C, Fuchsberger C, *et al.* Fertility pattern and family

- structure in three Alpine settlements in South Tyrol (Italy): marriage cohorts from 1750 to 1949. *J Biosoc Sci* 2009;**41**:697–701. doi:10.1017/s0021932009003423
- 38 Gavrilova NS, Gavrilov LA, Evdokushkina GN, *et al.* Evolution, mutations, and human longevity: European royal and noble families. *Hum Biol* 1998;**70**:799–804.
- 39 Joshi PK. *Exploring the inheritance of complex traits in humans*. 2014.
- 40 Kaplanis J, Gordon A, Shor T, *et al.* Quantitative analysis of population-scale family trees with millions of relatives. *Science (80-)* 2018;**360**:171–5. doi:10.1126/science.aam9309
- 41 Ruby JG, Wright KM, Rand KA, *et al.* Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating. *Genetics* 2018;**210**:1109–24. doi:10.1534/genetics.118.301613
- 42 Gögele M, Pattaro C, Fuchsberger C, *et al.* Heritability Analysis of Life Span in a Semi-isolated Population Followed Across Four Centuries Reveals the Presence of Pleiotropy Between Life Span and Reproduction. *Journals Gerontol Ser A* 2011;**66A**:26–37. doi:10.1093/gerona/glq163
- 43 Greenwood J, Guner N, Kocharkov G, *et al.* Marry your like: Assortative mating and income inequality. *Am Econ Rev* 2014;**104**:348–53. doi:10.1257/aer.104.5.348
- 44 Rogot E, Sorlie PD, Johnson NJ. Life expectancy by employment status, income, and education in the National Longitudinal Mortality Study. *Public Health Rep* 1992;**107**:457–61.
- 45 Yengo L, Robinson MR, Keller MC, *et al.* Imprint of assortative mating on the human genome. *Nat. Hum. Behav.* 2018;**2**:948–54. doi:10.1038/s41562-018-0476-3
- 46 Segalowitz SJ. Why twin studies really don't tell us much about human heritability. *Behav. Brain Sci.* 1999;**22**:904–5. doi:10.1017/S0140525X99442207
- 47 Rose SPR. Commentary: Heritability estimates - Long past their sell-by date. *Int J Epidemiol* 2006;**35**:525–7. doi:10.1093/ije/dyl064
- 48 Gamma A, Liebrez M. Rethinking heritability. *F1000Research* 2019;**8**:1705. doi:10.12688/f1000research.20641.1
- 49 Burt CH, Simons RL. Heritability studies in the postgenomic era: The fatal flaw is conceptual. *Criminology* 2015;**53**:103–12. doi:10.1111/1745-9125.12060
- 50 Pulst SM. Genetic linkage analysis. *Arch Neurol* 1999;**56**:667–72. doi:10.1001/archneur.56.6.667
- 51 Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science (80-)*. 2008;**322**:881–8. doi:10.1126/science.1156409
- 52 Puca AA, Daly MJ, Brewster SJ, *et al.* A genome-wide scan for linkage to human exceptional longevity identifies a locus on chromosome 4. *Proc Natl Acad Sci U S A* 2001;**98**:10505–8. doi:10.1073/pnas.181337598
- 53 Reed T, Dick DM, Uniacke SK, *et al.* Genome-Wide Scan for a Healthy Aging Phenotype Provides Support for a Locus Near D4S1564 Promoting Healthy Aging. *Journals Gerontol Ser A Biol Sci Med Sci* 2004;**59**:B227–32. doi:10.1093/gerona/59.3.b227

- 54 Sherry ST. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;**29**:308–11. doi:10.1093/nar/29.1.308
- 55 Wang DG, Fan JB, Siao CJ, *et al.* Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. *Science (80-)* 1998;**280**:1077–82. doi:10.1126/science.280.5366.1077
- 56 Lander ES, Linton LM, Birren B, *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921. doi:10.1038/35057062
- 57 Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* 2005;**76**:449–62. doi:10.1086/428594
- 58 LaFramboise T. Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Res.* 2009;**37**:4181–93. doi:10.1093/nar/gkp552
- 59 Thomas SC, Hill WG. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* 2000;**155**:1961–72.
- 60 McCarthy S, Das S, Kretzschmar W, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;**48**:1279–83. doi:10.1038/ng.3643
- 61 Clarke L, Zheng-Bradley X, Smith R, *et al.* The 1000 Genomes Pproject: Data management and community access. *Nat. Methods.* 2012;**9**:1–4. doi:10.1038/nmeth.1974
- 62 Kowalski MH, Qian H, Hou Z, *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet* 2019;**15**:e1008500. doi:10.1371/journal.pgen.1008500
- 63 Aulchenko YS, Struchalin M V, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 2010;**11**:134. doi:10.1186/1471-2105-11-134
- 64 Loh PR, Kichaev G, Gazal S, *et al.* Mixed-model association for biobank-scale datasets. *Nat. Genet.* 2018;**50**:906–8. doi:10.1038/s41588-018-0144-6
- 65 Yang J, Lee SH, Goddard ME, *et al.* GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82. doi:10.1016/j.ajhg.2010.11.011
- 66 Svischcheva GR, Axenovich TI, Belonogova NM, *et al.* Rapid variance components-based method for whole-genome association analysis. *Nat Genet* 2012;**44**:1166–70. doi:10.1038/ng.2410
- 67 Aulchenko YS, De Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 2007;**177**:577–85. doi:10.1534/genetics.107.075614
- 68 Amin N, van Duijn CM, Aulchenko YS. A genomic background based method for association analysis in related individuals. *PLoS One* 2007;**2**:e1274. doi:10.1371/journal.pone.0001274
- 69 Therneau TM, Grambsch PM, Fleming TR. Martingale-Based residuals for Survival Models. *Biometrika* 1990;**77**:147–60. doi:10.1093/biomet/77.1.147

- 70 Joshi PK, Pirastu N, Kentistou KA, *et al.* Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nat Commun* 2017;**8**:910. doi:10.1038/s41467-017-00934-5
- 71 Buniello A, MacArthur JAL, Cerezo M, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;**47**:D1005–12. doi:10.1093/nar/gky1120
- 72 Deelen J, Evans DS, Arking DE, *et al.* A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat Commun* 2019;**10**:3669. doi:10.1038/s41467-019-11558-2
- 73 Perls TT, Wilmoth J, Levenson R, *et al.* Life-long sustained mortality advantage of siblings of centenarians. *Proc Natl Acad Sci U S A* 2002;**99**:8442–7. doi:10.1073/pnas.122587599
- 74 Newman AB, Glynn NW, Taylor CA, *et al.* Health and function of participants in the Long Life family study: A comparison with other cohorts. *Aging (Albany NY)* 2011;**3**:63–76. doi:10.18632/aging.100242
- 75 Westendorp RGJ, Van Heemst D, Rozing MP, *et al.* Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden longevity study. *J Am Geriatr Soc* 2009;**57**:1634–7. doi:10.1111/j.1532-5415.2009.02381.x
- 76 Hjelmberg JBV, Iachine I, Skytthe A, *et al.* Genetic influence on human lifespan and longevity. *Hum Genet* 2006;**119**:312–21. doi:10.1007/s00439-006-0144-y
- 77 He Y, Zhao Y, Yao Y, *et al.* Cohort profile: The China Hainan centenarian Cohort Study (CHCCS). *Int J Epidemiol* 2018;**47**:694-695H. doi:10.1093/ije/dyy017
- 78 Chatters R, Newbould L, Sprange K, *et al.* Recruitment of older adults to three preventative lifestyle improvement studies. *Trials* 2018;**19**:121. doi:10.1186/s13063-018-2482-1
- 79 Newman SJ. Supercentenarians and the oldest-old are concentrated into regions with no birth certificates and short lifespans. *bioRxiv* 2019;:704080. doi:10.1101/704080
- 80 Young R. Age 115 or more in the United States: Fact or fiction? Springer, Berlin, Heidelberg 2010. 247–84. doi:10.1007/978-3-642-11520-2_15
- 81 Sebastiani P, Bae H, Gurinovich A, *et al.* Limitations and risks of meta-analyses of longevity studies. *Mech Ageing Dev* 2017;**165**:139–46. doi:10.1016/j.mad.2017.01.008
- 82 Newman AB, Walter S, Lunetta KL, *et al.* A meta-analysis of four genome-wide association studies of survival to age 90 years or older: the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. *Journals Gerontol Ser A Biol Sci Med Sci* 2010;**65**:478–87. doi:10.1093/gerona/gkq028
- 83 Deelen J, Beekman M, Uh H-WW, *et al.* Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging Cell* 2011;**10**:686–98. doi:10.1111/j.1474-9726.2011.00705.x
- 84 Nebel A, Kleindorp R, Caliebe A, *et al.* A genome-wide association study

- confirms APOE as the major gene influencing survival in long-lived individuals. *Mech Ageing Dev* 2011;**132**:324–30. doi:10.1016/j.mad.2011.06.008
- 85 Lettre G, Jackson AU, Gieger C, *et al.* Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* 2008;**40**:584–91. doi:10.1038/ng.125
- 86 Leitsalu L, Haller T, Esko T, *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* 2015;**44**:1137–47. doi:10.1093/ije/dyt268
- 87 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9. doi:10.1038/s41586-018-0579-z
- 88 Timmers PRHJ, Mounier N, Lall K, *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances [dataset]. *Elife* 2019;**8**. doi:https://doi.org/10.7488/ds/2463
- 89 Pilling LC, Atkins JL, Bowman K, *et al.* Human longevity is influenced by many genetic variants: evidence from 75,000 UK Biobank participants. *Ageing (Albany NY)* 2016;**8**:547–60. doi:10.18632/aging.100930
- 90 Joshi PK, Fischer K, Schraut KE, *et al.* Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. *Nat Commun* 2016;**7**:11174. doi:10.1038/ncomms11174
- 91 Timmers PR, Kerssens JJ, Minton JW, *et al.* Trends in disease incidence and survival and their effect on mortality in Scotland: nationwide cohort study of linked hospital admission and death records 2001–2016. *BMJ Open* Published Online First: 2020. doi:http://dx.doi.org/10.1136/bmjopen-2019-034299
- 92 Kosova G, Abney M, Ober C. Heritability of reproductive fitness traits in a human population. *Proc Natl Acad Sci U S A* 2010;**107**:1772–8. doi:10.1073/pnas.0906196106
- 93 Aschard H, Vilhjálmsón BJ, Joshi AD, *et al.* Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet* 2015;**96**:329–39. doi:10.1016/j.ajhg.2014.12.021
- 94 Munafò MR, Tilling K, Taylor AE, *et al.* Collider scope: When selection bias can substantially influence observed associations. *Int J Epidemiol* 2018;**47**:226–35. doi:10.1093/ije/dyx206
- 95 ISD Scotland. National Data Catalogue. Natl. Datasets. 2016;:1.https://www.ndc.scot.nhs.uk/National-Datasets/ (accessed 14 Mar 2020).
- 96 National Records of Scotland. Mid-Year Population Estimates Scotland, Mid-2018. 2019.
- 97 UK Biobank. Genotyping and Quality Control of UK Biobank, a Large-Scale, Extensively Phenotyped Prospective Resource: Information for Researchers (Interim Data Release, 2015). *UK Biobank* 2015;:1–27.
- 98 Huang J, Howie B, McCarthy S, *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* 2015;**6**:8111. doi:10.1038/ncomms9111

- 99 UK Biobank. Data Showcase. Data Showc. User Guid. 2019.<http://biobank.ctsu.ox.ac.uk/crystal/index.cgi> (accessed 14 Mar 2020).
- 100 Mitt M, Kals M, Pärn K, *et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet* 2017;**25**:869–76. doi:10.1038/ejhg.2017.51
- 101 Reisberg S, Krebs K, Lepamets M, *et al.* Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genet Med* 2019;**21**:1345–54. doi:10.1038/s41436-018-0337-5
- 102 Levy HL, Albers S. Genetic Screening of Newborns. *Annu Rev Genomics Hum Genet* 2000;**1**:139–77. doi:10.1146/annurev.genom.1.1.139
- 103 Narod SA. BRCA mutations in the management of breast cancer: The state of the art. *Nat. Rev. Clin. Oncol.* 2010;**7**:702–7. doi:10.1038/nrclinonc.2010.166
- 104 Nance MA. Laboratory guidelines for huntington disease genetic testing. *Am J Hum Genet* 1998;**62**:1243–7. doi:10.1086/301846
- 105 Khera A V, Chaffin M, Aragam KG, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;**50**:1219. doi:doi:10.1038/s41588-018-0183-z
- 106 Cecile A, Janssens JW, Joyner MJ. Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better? *Clin Chem* 2019;**65**:609–11. doi:10.1373/clinchem.2018.296103
- 107 Vilhjálmsón BJ, Yang J, Finucane HK, *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* 2015;**97**:576–92. doi:10.1016/j.ajhg.2015.09.001
- 108 De La Vega FM, Bustamante CD. Polygenic risk scores: A biased prediction? *Genome Med* 2018;**10**:100. doi:10.1186/s13073-018-0610-x
- 109 Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics* 2015;**31**:1466–8. doi:10.1093/bioinformatics/btu848
- 110 Natarajan P, Young R, Stitzel NO, *et al.* Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* 2017;**135**:2091–101. doi:10.1161/CIRCULATIONAHA.116.024436
- 111 Mavaddat N, Pharoah PDP, Michailidou K, *et al.* Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst* 2015;**107**. doi:10.1093/jnci/djv036
- 112 Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 2004;**3**:711–5. doi:10.1038/nrd1470
- 113 Nelson MR, Tipney H, Painter JL, *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* 2015;**47**:856–60. doi:10.1038/ng.3314
- 114 Cao C, Moulton J. GWAS and drug targets. *BMC Genomics* 2014;**15**:S5. doi:10.1186/1471-2164-15-S4-S5
- 115 Martin-Montalvo A, Mercken EM, Mitchell SJ, *et al.* Metformin improves

- healthspan and lifespan in mice. *Nat Commun* 2013;**4**:2192. doi:10.1038/ncomms3192
- 116 Harrison DE, Strong R, Sharp ZD, *et al.* Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* 2009;**460**:392–5. doi:10.1038/nature08221
- 117 Mitchell SJ, Bernier M, Aon MA, *et al.* Nicotinamide Improves Aspects of Healthspan, but Not Lifespan, in Mice. *Cell Metab* 2018;**27**:667-676.e4. doi:10.1016/j.cmet.2018.02.001
- 118 Niccoli T, Partridge L. Ageing as a risk factor for disease. *Curr. Biol.* 2012;**22**:R741–52. doi:10.1016/j.cub.2012.07.024
- 119 Public Health England. A review of recent trends in mortality in England About Public Health England. 2018.
- 120 Mesalles-Naranjo O, Grant I, Wyper GMA, *et al.* Trends and inequalities in the burden of mortality in Scotland 2000–2015. *PLoS One* 2018;**13**:1–17. doi:10.1371/journal.pone.0196906
- 121 Sanders S, UK Office for National Statistics. National life tables, UK: 2015 to 2017. *Off Natl Stat Stat Bull* 2018;;1–11.
- 122 Lafortune, Gaetan; Balestat G. OECD Health Working Papers No.26. *Trends Sev Disabil Among Elder People* 2007;;81. doi:10.1787/217072070078 OECD
- 123 Hiam L, Dorling D, Harrison D, *et al.* What caused the spike in mortality in England and Wales in January 2015? *J R Soc Med* 2017;**110**:131–7. doi:10.1177/0141076817693600
- 124 UK Office for National Statistics, Office for National Statistics. Changing trends in mortality in England and Wales: 1990 to 2017 (Experimental Statistics). 2018.
- 125 Dyer O. US life expectancy falls for third year in a row. *BMJ* 2018;**363**:k5118. doi:10.1136/bmj.k5118
- 126 Palin J. Mortality improvements in decline. *Actuar* 2017.
- 127 Fenton L, Minton J, Ramsay J, *et al.* Recent adverse mortality trends in Scotland: comparison with other high-income countries. *bioRxiv* 2019;;542449. doi:10.1101/542449
- 128 Fenton L, Wyper G, McCartney G, *et al.* Socioeconomic inequality in recent adverse mortality trends in Scotland. *bioRxiv* 2019;;542472. doi:10.1101/542472
- 129 Nie X, Peng X. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *Chinese J Evidence-Based Med* 2017;**17**:475–87. doi:10.7507/1672-2531.201702009
- 130 Fleming M, Kirby B, Penny KI. Record linkage in Scotland and its applications to health research. *J Clin Nurs* 2012;**21**:2711–21. doi:10.1111/j.1365-2702.2011.04021.x
- 131 The Scottish Government. SIMD16 Technical Notes. 2016.
- 132 World Health Organisation. International Statistical Classification of Diseases and Related Health Problems Tenth Revision Volume 2. *World Heal Organ* Published Online First: 2004. doi:https://apps.who.int/iris/handle/10665/42980
- 133 Quaresma M, Coleman MP, Rachet B. 40-year trends in an index of survival

- for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: A population-based study. *Lancet* 2015;**385**:1206–18. doi:10.1016/S0140-6736(14)61396-9
- 134 Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B Stat Methodol* 1972;**34**:187.
- 135 Dubois A, Padovanib A, Scheltensc P, *et al*. Timely diagnosis for alzheimer’s disease: A literature review on benefits and challenges. *J Alzheimer’s Dis* 2015;**49**:617–31. doi:10.3233/JAD-150692
- 136 Smith Sehdev AE, Hutchins GM. Problems with proper completion and accuracy of the COD statement. *Arch Intern Med* 2001;**161**:277–84. doi:doi:10.1001/archinte.161.2.277
- 137 Hotchkiss JW, Davies CA, Dundas R, *et al*. Explaining trends in Scottish coronary heart disease mortality between 2000 and 2010 using IMPACTSEC model: Retrospective analysis using routine data. *BMJ* 2014;**348**:1–15. doi:10.1136/bmj.g1088
- 138 Torjesen I. Inequalities in life expectancy are widening, data confirm. *Bmj* 2018;**1017**:k1017. doi:10.1136/bmj.k1017
- 139 Bennett JE, Pearson-Stuttard J, Kontis V, *et al*. Contributions of diseases and injuries to widening life expectancy inequalities in England from 2001 to 2016: a population-based analysis of vital registration data. *Lancet Public Heal* 2018;**3**:e586–97. doi:10.1016/S2468-2667(18)30214-7
- 140 NHS Scotland. *Better Cancer Care: An Action Plan*. Edinburgh: 2008.
- 141 Department of Health. *Improving Outcomes: A Strategy for Cancer*. 2011;:101.
- 142 Harbeck N, Gnant M. Breast cancer. In: *Lancet (London, England)*. Elsevier 2017. 1134–50. doi:10.1016/S0140-6736(16)31891-8
- 143 O’Neill J. The Review on Antimicrobial Resistance. 2016. doi:10.1016/j.jpha.2015.11.005
- 144 Whyte B. Scottish mortality in a European Context 1950 – 2000. An analysis of comparative mortality trends. 2006.
- 145 Viña J, Borrás C, Miquel J. Theories of ageing. *IUBMB Life* 2007;**59**:249–54. doi:10.1080/15216540601178067
- 146 Williams GC. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution (N Y)* 1957;**11**:398. doi:10.2307/2406060
- 147 Medawar PB, Medawar PB. An Unsolved Problem of Biology. *Uniqueness Individ* 2019;:44–70. doi:10.4324/9780429299759-3
- 148 Blagosklonny M V. Aging and immortality: Quasi-programmed senescence and its pharmacologic inhibition. *Cell Cycle*. 2006;**5**:2087–102. doi:10.4161/cc.5.18.3288
- 149 Gems D, De La Guardia Y. Alternative perspectives on aging in caenorhabditis elegans: Reactive oxygen species or hyperfunction? *Antioxidants Redox Signal*. 2013;**19**:321–9. doi:10.1089/ars.2012.4840
- 150 de la Guardia Y, Gilliat AF, Hellberg J, *et al*. Run-on of germline apoptosis promotes gonad senescence in C. elegans. *Oncotarget* 2016;**7**:39082–96. doi:10.18632/oncotarget.9681
- 151 van den Berg N, Beekman M, Smith KR, *et al*. Historical demography and longevity genetics: Back to the future. *Ageing Res Rev* 2017;**38**:28–39.

- doi:<https://doi.org/10.1016/j.arr.2017.06.005>
- 152 Ljungquist B, Berg S, Lanke J, *et al.* The effect of genetic factors for longevity: a comparison of identical and fraternal twins in the Swedish Twin Registry. *J Gerontol A Biol Sci Med Sci* 1998;**53**:M441-6.
- 153 McGue M, Vaupel JW, Holm N, *et al.* Longevity is moderately heritable in a sample of Danish twins born 1870-1880. *J Gerontol* 1993;**48**:B237-44.
- 154 Young AI, Frigge ML, Gudbjartsson DF, *et al.* Relatedness disequilibrium regression estimates heritability without environmental bias. *Nat Genet* 2018;**50**:1304–10. doi:10.1038/s41588-018-0178-9
- 155 Sebastiani P, Solovieff N, DeWan AT, *et al.* Genetic Signatures of Exceptional Longevity in Humans. *PLoS One* 2012;**7**. doi:10.1371/journal.pone.0029848
- 156 Beekman M, Blanche H, Perola M, *et al.* Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study. *Aging Cell* 2013;**12**:184–93. doi:10.1111/accel.12039
- 157 Broer L, Buchman AS, Deelen J, *et al.* GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. *Journals Gerontol - Ser A Biol Sci Med Sci* 2015;**70**:110–8. doi:10.1093/gerona/glu166
- 158 Zeng Y, Nie C, Min J, *et al.* Novel loci and pathways significantly associated with longevity. *Sci Rep* 2016;**6**:21243. doi:10.1038/srep21243
- 159 Pilling LC, Kuo CL, Sicinski K, *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging (Albany NY)* 2017;**9**:2504–20. doi:10.18632/aging.101334
- 160 Sebastiani P, Nussbaum L, Andersen SL, *et al.* Increasing Sibling Relative Risk of Survival to Older and Older Ages and the Importance of Precise Definitions of ‘Aging,’ ‘Life Span,’ and ‘Longevity’. *J Gerontol A Biol Sci Med Sci* 2016;**71**:340–6. doi:10.1093/gerona/glv020
- 161 Deelen J, Beekman M, Uh H-W, *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum Mol Genet* 2014;**23**:4420–32. doi:10.1093/hmg/ddu139
- 162 Wacholder S, Hartge P, Struwing JP, *et al.* The kin-cohort study for estimating penetrance. *Am J Epidemiol* 1998;**148**:623–30.
- 163 Bycroft C, Freeman C, Petkova D, *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* Published Online First: 2017. doi:10.1101/166298
- 164 McDaid AF, Joshi PK, Porcu E, *et al.* Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nat Commun* 2017;**8**:15842. doi:10.1038/ncomms15842
- 165 Flachsbarth F, Caliebe A, Kleindorff R, *et al.* Association of FOXO3A variation with human longevity confirmed in German centenarians. *Proc Natl Acad Sci* 2009;**106**:2700–5. doi:10.1073/pnas.0809594106
- 166 Sebastiani P, Gurinovich A, Bae H, *et al.* Four Genome-Wide Association Studies Identify New Extreme Longevity Variants. *Journals Gerontol Ser A* 2017;**17**:6. doi:10.1093/gerona/glx027
- 167 Ben-Avraham D, Govindaraju DR, Budagov T, *et al.* The GH receptor exon 3 deletion is a marker of male-specific exceptional longevity associated with

- increased GH sensitivity and taller stature. *Sci Adv* 2017;**3**:e1602025. doi:10.1126/sciadv.1602025
- 168 Hemani G, Zheng J, Elsworth B, *et al.* The MR-base platform supports systematic causal inference across the human phenome. *Elife* 2018;**7**. doi:10.7554/eLife.34408
- 169 Westra H-J, Peters MJ, Esko T, *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* 2013;**45**:1238–43. doi:10.1038/ng.2756
- 170 Lloyd-Jones LR, Holloway A, McRae A, *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am J Hum Genet* 2017;**100**:371. doi:10.1016/j.ajhg.2017.01.026
- 171 MacArthur J, Bowler E, Cerezo M, *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* Published Online First: 2016. doi:10.1093/nar/gkw1133
- 172 Staley JR, Blackshaw J, Kamat MA, *et al.* PhenoScanner: A database of human genotype-phenotype associations. *Bioinformatics* 2016;**32**:3207–9. doi:10.1093/bioinformatics/btw373
- 173 Deelen J, Beekman M, Capri M, *et al.* Identifying the genomic determinants of aging and longevity in human population studies: progress and challenges. *Bioessays* 2013;**35**:386–96. doi:10.1002/bies.201200148
- 174 Peters MJ, Joehanes R, Pilling LC, *et al.* The transcriptional landscape of age in human peripheral blood. *Nat Commun* 2015;**6**:8570. doi:doi:10.1038/ncomms9570
- 175 Wood AR, Esko T, Yang J, *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014;**46**:1173–86. doi:10.1038/ng.3097
- 176 Mathers C, Stevens GA, Mahanani WR, *et al.* Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva: 2018.
- 177 Zarrouk A, Vejux A, Mackrill J, *et al.* Involvement of oxysterols in age-related diseases and ageing processes. *Ageing Res Rev* 2014;**18**:148–62. doi:10.1016/j.arr.2014.09.006
- 178 Hayashi T, Ko JH, Strafella AP, *et al.* Dorsolateral prefrontal and orbitofrontal cortex interactions during self-control of cigarette craving. *Proc Natl Acad Sci USA* 2013;**110**:4422–7. doi:10.1073/pnas.1212185110
- 179 Lowe CJ, Hall PA, Staines WR. The effects of continuous theta burst stimulation to the left dorsolateral prefrontal cortex on executive function, food cravings, and snack food consumption. *Psychosom Med* 2014;**76**:503–11. doi:10.1097/psy.0000000000000090
- 180 Morrison JH, Baxter MG. The ageing cortical synapse: hallmarks and implications for cognitive decline. *Nat Rev Neurosci* 2012;**13**:240–50. doi:10.1038/nrn3200
- 181 Hwangbo DS, Gersham B, Tu M-P, *et al.* Drosophila dFOXO controls lifespan and regulates insulin signalling in brain and fat body. *Nature* 2004;**429**:562. doi:doi:10.1038/nature02549
- 182 Donlon TA, Morris BJ, Chen R, *et al.* FOXO3 longevity interactome on chromosome 6. In: *Ageing Cell*. Department of ResearchGenetics

- LaboratoryHonolulu Heart Program/Honolulu-Asia Aging Study (HAAS)Kuakini Medical CenterHonoluluHawaiiJohn A. Burns School of MedicineUniversity of Hawaii ManoaHonoluluHawaii: 2017. 1016–25. doi:10.1111/accel.12625
- 183 Flachsbart F, Dose J, Gentschew L, *et al.* Identification and characterization of two functional variants in the human longevity gene FOXO3. *Nat Commun* 2017;**8**:2063. doi:doi:10.1038/s41467-017-02183-y
- 184 Grossi V, Medical Genetics D of BS, Human Oncology (DIMO) University of Bari Aldo Moro B 70124 I, *et al.* The longevity SNP rs2802292 uncovered: HSF1 activates stress-dependent expression of FOXO3 through an intronic enhancer. *Nucleic Acids Res* 2018;**46**:5587–600. doi:10.1093/nar/gky331
- 185 Williams GC. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution* (N. Y). 1957;**11**:398–411. doi:10.2307/2406060
- 186 Legal & General Group PLC. Interim Management Report. 2017.
- 187 HM Government. Concordat and Moratorium on Genetics and Insurance. UK Government and ABI 2014.
- 188 D'Agostino R. B. S, Vasan RS, Pencina MJ, *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;**117**:743–53. doi:10.1161/circulationaha.107.699579
- 189 Loh PR, Tucker G, Bulik-Sullivan BK, *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015;**47**:284–90. doi:10.1038/ng.3190
- 190 Shen X, Wang X, Ning Z, *et al.* Simple multi-trait analysis identifies novel loci associated with growth and obesity measures. *bioRxiv* Published Online First: 2015. doi:10.1101/022269
- 191 Zhu Z, Zhang F, Hu H, *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 2016;**48**:481–7. doi:10.1038/ng.3538
- 192 GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* 2017;**550**:204. doi:doi:10.1038/nature24277
- 193 McRae A, Marioni RE, Shah S, *et al.* Identification of 55,000 Replicated DNA Methylation QTL. Published Online First: 2017. doi:10.1101/166710
- 194 Ning Z, Lee Y, Joshi PK, *et al.* A Selection Operator for Summary Association Statistics Reveals Allelic Heterogeneity of Complex Traits. *Am J Hum Genet* 2017;**101**:903–12. doi:10.1016/j.ajhg.2017.09.027
- 195 Churchhouse C, Neale B. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank. 2017.<http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank> (accessed 2 Nov 2018).
- 196 Falconer DS, Mackay TFC, Frankham R. Introduction to quantitative genetics (4th edn). *Trends Genet* 1996;**12**:280.
- 197 Finucane HK, Bulik-Sullivan B, Gusev A, *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015;**47**:1228–35. doi:10.1038/ng.3404
- 198 Bulik-Sullivan BK, Loh P-R, Finucane HK, *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association

- studies. *Nat Genet* 2015;**47**:291. doi:doi:10.1038/ng.3211
- 199 Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res Hum Genet* 2015;**18**:86–91. doi:10.1017/thg.2014.79
- 200 Mishra A, MacGregor S. A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility. *Twin Res Hum Genet* 2017;**20**:1–9. doi:10.1017/thg.2016.100
- 201 Pers TH, Karjalainen JM, Chan Y, *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* 2015;**6**:5890. doi:10.1038/ncomms6890
- 202 Lamparter D, Marbach D, Rueedi R, *et al.* Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol* 2016;**12**:e1004714. doi:10.1371/journal.pcbi.1004714
- 203 Marouli E, Graff M, Medina-Gomez C, *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* 2017;**542**:186. doi:doi:10.1038/nature21039
- 204 Võsa U, Claringbould A, Westra H-J, *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 2018;;447367. doi:10.1101/447367
- 205 Haller T, Kals M, Esko T, *et al.* RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Br Bioinform* 2015;**16**:39–44. doi:10.1093/bib/bbt066
- 206 Zenin A, Tsepilov Y, Sharapov S, *et al.* Identification of 12 genetic loci associated with human healthspan. *Commun Biol* 2019;**2**:41. doi:10.1038/s42003-019-0290-0
- 207 Sebastiani P, Perls TT. The genetics of extreme longevity: Lessons from the new england centenarian study. *Front Genet* 2012;**3**:277. doi:10.3389/fgene.2012.00277
- 208 Walter S, Atzmon G, Demerath EW, *et al.* A genome-wide association study of aging. *Neurobiol Aging* 2011;**32**:2109.e15-2109.e28. doi:10.1016/j.neurobiolaging.2011.05.026
- 209 Pilling LC, Kuo C-LL, Sicinski K, *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging (Albany NY)* 2017;**9**:2504–20. doi:10.18632/aging.101334
- 210 Shen X, Klarić L, Sharapov S, *et al.* Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation. *Nat Commun* 2017;**8**:447. doi:10.1038/s41467-017-00453-3
- 211 Haller T, Kals M, Esko T, *et al.* RegScan: A GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Brief Bioinform* 2013;**16**:39–44. doi:10.1093/bib/bbt066
- 212 Kamat MA, Blackshaw JA, Young R, *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* 2019;**35**:4851–3. doi:10.1093/bioinformatics/btz469
- 213 Wu Y, Zeng J, Zhang F, *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* 2018;**9**:918. doi:10.1038/s41467-018-03371-0
- 214 Lloyd-Jones LR, Holloway A, McRae A, *et al.* The Genetic Architecture of

- Gene Expression in Peripheral Blood. *Am J Hum Genet* 2017;**100**:371. doi:10.1016/j.ajhg.2017.01.026
- 215 Liberzon A, Birger C, Thorvaldsdóttir H, *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* 2015;**1**:417–25. doi:10.1016/j.cels.2015.12.004
- 216 Watanabe K, Taskesen E, van Bochoven A, *et al.* Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017;**8**:1826. doi:10.1038/s41467-017-01261-5
- 217 Broer L, Buchman AS, Deelen J, *et al.* GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. *Journals Gerontol - Ser A Biol Sci Med Sci* 2015;**70**:110–8. doi:10.1093/gerona/glu166
- 218 Sanese P, Forte G, Disciglio V, *et al.* FOXO3 on the Road to Longevity: Lessons From SNPs and Chromatin Hubs. *Comput. Struct. Biotechnol. J.* 2019;**17**:737–45. doi:10.1016/j.csbj.2019.06.011
- 219 Strittmatter WJ, Roses AD. Apolipoprotein E and Alzheimer disease. *Proc Natl. Acad. Sci. U. S. A.* 1995;**92**:4725–7. doi:10.1073/pnas.92.11.4725
- 220 Williams GC. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution (N Y)* 1957;**11**:398–411. doi:10.2307/2406060
- 221 Blagosklonny M V. Answering the ultimate question ‘What is the proximal cause of aging?’ *Aging (Albany. NY).* 2012;**4**:861–77. doi:10.18632/aging.100525
- 222 Abbott L, Bryant S, Churchhouse C, *et al.* UK Biobank — Neale lab. 2018.<http://www.nealelab.is/uk-biobank/> (accessed 5 Dec 2019).
- 223 Institute for Health Metrics and Evaluation. Findings from the Global Burden of Disease Study 2017. Seattle, WA: 2018.
- 224 Kenyon CJ. The genetics of ageing. *Nature* 2010;**464**:504–12. doi:10.1038/nature08980
- 225 Atamna H, Killilea DW, Killilea AN, *et al.* Heme deficiency may be a factor in the mitochondrial and neuronal decay of aging. *Proc Natl Acad Sci U S A* 2002;**99**:14807–12. doi:10.1073/pnas.192585799
- 226 Ward RJ, Zucca FA, Duyn JH, *et al.* The role of iron in brain ageing and neurodegenerative disorders. *Lancet Neurol.* 2014;**13**:1045–60. doi:10.1016/S1474-4422(14)70117-6
- 227 Ellervik C, Marott JL, Tybjærg-Hansen A, *et al.* Total and cause-specific mortality by moderately and markedly increased ferritin concentrations: General population study and metaanalysis. *Clin Chem* 2014;**60**:1419–28. doi:10.1373/clinchem.2014.229013
- 228 Moen IW, Bergholdt HKM, Mandrup-Poulsen T, *et al.* Increased plasma ferritin concentration and low-grade inflammation—a mendelian randomization study. *Clin Chem* 2018;**64**:374–85. doi:10.1373/clinchem.2017.276055
- 229 Pilling LC, Tamosauskaite J, Jones G, *et al.* Common conditions associated with hereditary haemochromatosis genetic variants: Cohort study in UK Biobank. *BMJ* 2019;**364**:k5222. doi:10.1136/bmj.k5222
- 230 Atkins JL, Jylhava J, Pedersen N, *et al.* A Genome-Wide Association Study of the Frailty Index Highlights Synaptic Pathways in Aging. *medRxiv* 2019;:19007559. doi:10.1101/19007559

- 231 Kaeberlein M. How healthy is the healthspan concept? *GeroScience*. 2018;**40**:361–4. doi:10.1007/s11357-018-0036-9
- 232 Hurrell R, Egli I. Iron bioavailability and dietary reference values. *Am. J. Clin. Nutr.* 2010;**91**:1461S-1467S. doi:10.3945/ajcn.2010.28674F
- 233 National Records of Scotland. Expectation of life, by sex and selected age, Scotland, 1861 to 2011. 2012.
- 234 WHO. Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization, 2018. 2018.
- 235 Saklayen MG, Deshpande N V. Timeline of History of Hypertension Treatment. *Front. Cardiovasc. Med.* 2016;**3**:3. doi:10.3389/fcvm.2016.00003
- 236 Hajar R. Statins: Past and Present. *Hear Views* 2011;**12**:121. doi:10.4103/1995-705x.95070
- 237 Information Services Division Scotland. Scottish Heart Disease Statistics Year ending 31 March 2017. 2018. doi:10.1038/nature12509
- 238 Scottish Government. Tobacco use among adolescents in Scotland: profile and trends. 2016.
- 239 Scottish Parliament. *Smoking, Health and Social Care (Scotland) Act 2005*. Stationery Office 2005.
- 240 Reid G, Rennick L, Laird Y, *et al.* Review of ‘Creating a tobacco-free generation: A Tobacco Control Strategy for Scotland’. 2017.
- 241 Chabner BA, Roberts TG. Chemotherapy and the war on cancer. *Nat. Rev. Cancer.* 2005;**5**:65–72. doi:10.1038/nrc1529
- 242 Modjtahedi H, Ali S, Essapen S. Therapeutic application of monoclonal antibodies in cancer: Advances and challenges. *Br. Med. Bull.* 2012;**104**:41–59. doi:10.1093/bmb/lds032
- 243 Stockton D, Davies T, Day N, *et al.* Retrospective study of reasons for improved survival in patients with breast cancer in East Anglia: Earlier diagnosis or better treatment? *BMJ* 1997;**314**:472. doi:10.1136/bmj.314.7079.472
- 244 von Wagner C, Baio G, Raine R, *et al.* Inequalities in participation in an organized national colorectal cancer screening programme: Results from the first 2.6 million invitations in England. *Int J Epidemiol* 2011;**40**:712–8. doi:10.1093/ije/dyr008
- 245 Temel JS, Greer JA, Muzikansky A, *et al.* Early palliative care for patients with metastatic non-small-cell lung cancer. *N Engl J Med* 2010;**363**:733–42. doi:10.1056/NEJMoa1000678
- 246 Foot C, Harrison T. How to improve cancer survival Explaining England’s relatively poor rates. *Kings Fund* 2011;:1–32.
- 247 Yoshikawa TT. Epidemiology and Unique Aspects of Aging and Infectious Diseases. *Clin Infect Dis* 2000;**30**:931–3. doi:10.1086/313792
- 248 HM Government. Contained and controlled: the UK’s 20-year vision for antimicrobial resistance. 2019.
- 249 Wright KM, Rand KA, Kermany A, *et al.* A prospective analysis of genetic variants associated with human lifespan. *G3 Genes, Genomes, Genet* 2019;**9**:2863–78. doi:10.1534/g3.119.400448

- 250 Deelen J, Beekman M, Uh HWH-W, *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum Mol Genet* 2014;**23**:4420–32. doi:10.1093/hmg/ddu139
- 251 Visscher PM, Brown MA, McCarthy MI, *et al.* Five years of GWAS discovery. *Am. J. Hum. Genet.* 2012;**90**:7–24. doi:10.1016/j.ajhg.2011.11.029
- 252 Tam V, Patel N, Turcotte M, *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019;**20**:467–84. doi:10.1038/s41576-019-0127-1
- 253 Yengo L, Sidorenko J, Kemper KE, *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. *Hum Mol Genet* 2018;**27**:3641–9. doi:10.1093/hmg/ddy271
- 254 Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell.* 2017;**169**:1177–86. doi:10.1016/j.cell.2017.05.038
- 255 Guo MH, Hirschhorn JN, Dauber A. Insights and Implications of Genome-Wide Association Studies of Height. *J. Clin. Endocrinol. Metab.* 2018;**103**:3155–68. doi:10.1210/jc.2018-01126
- 256 Marigorta UM, Rodríguez JA, Gibson G, *et al.* Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet.* 2018;**34**:504–17. doi:10.1016/j.tig.2018.03.005
- 257 Sebastiani P, Solovieff N, Puca A, *et al.* Genetic Signatures of Exceptional Longevity in Humans. *Science (80-)* 2010;**210**. doi:10.1126/science.1190532
- 258 Alberts B. Editorial expression of concern. *Science (80-)*. 2010;**330**:912. doi:10.1126/science.330.6006.912-b
- 259 Sebastiani P, Solovieff N, Puca A, *et al.* Retraction. *Science (80-)*. 2011;**333**:404. doi:10.1126/science.333.6041.404-a
- 260 Lee JM, Gillis T, Mysore JS, *et al.* Common SNP-based haplotype analysis of the 4p16.3 Huntington disease gene region. *Am J Hum Genet* 2012;**90**:434–44. doi:10.1016/j.ajhg.2012.01.005
- 261 Hirschhorn JN. Genomewide association studies - Illuminating biologic pathways. *N Engl J Med* 2009;**360**:1699–701. doi:10.1056/NEJMp0808934
- 262 Di Paolo G, Kim TW. Linking lipids to Alzheimer's disease: Cholesterol and beyond. *Nat. Rev. Neurosci.* 2011;**12**:284–96. doi:10.1038/nrn3012
- 263 Linton MF, Yancey PG, Davies SS, *et al.* *The Role of Lipids and Lipoproteins in Atherosclerosis.* 2000.
- 264 National Health Service. NHS Commissioning: Medical Genetics. 2019. <https://www.england.nhs.uk/commissioning/spec-services/npc-crg/group-e/e01/> (accessed 15 Mar 2020).
- 265 NHS England. 100,000 Genomes Project: Paving the way to Personalised Medicine. *NHS Engl* 2016;:1–17.
- 266 Martin JB. Molecular basis of the neurodegenerative disorders. *N. Engl. J. Med.* 1999;**340**:1970–80. doi:10.1056/NEJM199906243402507
- 267 Craufurd D, MacLeod R, Frontali M, *et al.* Diagnostic genetic testing for huntington's disease. *Pract Neurol* 2015;**15**:80–4.

- doi:10.1136/practneurol-2013-000790
- 268 Howard RCWB. Genetic Testing Model for CI: If Underwriters of Individual Critical Illness Insurance Had No Access to Known Results of Genetic Tests. 2016.
- 269 Smittenaar CR, Petersen KA, Stewart K, *et al.* Cancer incidence and mortality projections in the UK until 2035. *Br J Cancer* 2016;**115**:1147–55. doi:10.1038/bjc.2016.304
- 270 Woollard KJ, Sturgeon S, Chin-Dusting JPF, *et al.* Erythrocyte hemolysis and hemoglobin oxidation promote ferric chloride-induced vascular injury. *J Biol Chem* 2009;**284**:13110–8. doi:10.1074/jbc.M809095200
- 271 Okonko DO, Mandal AKJ, Missouriis CG, *et al.* Disordered iron homeostasis in chronic heart failure: Prevalence, predictors, and relation to anemia, exercise capacity, and survival. *J Am Coll Cardiol* 2011;**58**:1241–51. doi:10.1016/j.jacc.2011.04.040
- 272 Morita T. Heme oxygenase and atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* 2005;**25**:1786–95. doi:10.1161/01.ATV.0000178169.95781.49
- 273 Cassat JE, Skaar EP. Iron in infection and immunity. *Cell Host Microbe*. 2013;**13**:509–19. doi:10.1016/j.chom.2013.04.010
- 274 Drakesmith H, Prentice A. Viral infection and iron metabolism. *Nat. Rev. Microbiol.* 2008;**6**:541–52. doi:10.1038/nrmicro1930
- 275 Wang Y, Yu L, Ding J, *et al.* Iron metabolism in cancer. *Int. J. Mol. Sci.* 2019;**20**. doi:10.3390/ijms20010095
- 276 Simcox JA, McClain DA. Iron and diabetes risk. *Cell Metab.* 2013;**17**:329–41. doi:10.1016/j.cmet.2013.02.007
- 277 Li K, Reichmann H. Role of iron in neurodegenerative diseases. *J. Neural Transm.* 2016;**123**:389–99. doi:10.1007/s00702-016-1508-7
- 278 Tamosauskaite J, Atkins JL, Pilling LC, *et al.* Hereditary hemochromatosis associations with frailty, sarcopenia and chronic pain: Evidence from 200,975 older UK biobank participants. *Journals Gerontol - Ser A Biol Sci Med Sci* 2019;**74**:337–42. doi:10.1093/gerona/gly270
- 279 Barzilai N, Crandall JP, Kritchevsky SB, *et al.* Metformin as a Tool to Target Aging. *Cell Metab.* 2016;**23**:1060–5. doi:10.1016/j.cmet.2016.05.011
- 280 Luque-Ramírez M, Álvarez-Blasco F, Botella-Carretero JI, *et al.* Increased body iron stores of obese women with polycystic ovary syndrome are a consequence of insulin resistance and hyperinsulinism and are not a result of reduced menstrual losses. *Diabetes Care* 2007;**30**:2309–13. doi:10.2337/dc07-0642
- 281 Stynen B, Abd-Rabbo D, Kowarzyk J, *et al.* Changes of Cell Biochemical States Are Revealed in Protein Homomeric Complex Dynamics. *Cell* 2018;**175**:1418-1429.e9. doi:10.1016/j.cell.2018.09.050
- 282 Chung CL, Lawrence I, Hoffman M, *et al.* Topical rapamycin reduces markers of senescence and aging in human skin: an exploratory, prospective, randomized trial. *GeroScience* 2019;**41**:861–9. doi:10.1007/s11357-019-00113-y
- 283 Urfer SR, Kaeberlein TL, Mailheau S, *et al.* A randomized controlled trial to establish effects of short-term rapamycin treatment in 24 middle-aged companion dogs. *GeroScience* 2017;**39**:117–27. doi:10.1007/s11357-017-

- 9972-z
- 284 Sofroniadou S, Kassimatis T, Goldsmith D. Anaemia, microcytosis and sirolimus-is iron the missing link? *Nephrol Dial Transplant* 2010;**25**:1667–75. doi:10.1093/ndt/gfp674
- 285 Maiorano A, Stallone G, Schena A, *et al.* Sirolimus interferes with iron homeostasis in renal transplant recipients. *Transplantation* 2006;**82**:908–12. doi:10.1097/01.tp.0000235545.49391.1b
- 286 Masaldan S, Clatworthy SAS, Gamell C, *et al.* Iron accumulation in senescent cells is coupled with impaired ferritinophagy and inhibition of ferroptosis. *Redox Biol* 2018;**14**:100–15. doi:10.1016/j.redox.2017.08.015
- 287 Holzenberger M, Dupont J, Ducos B, *et al.* IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature* 2003;**421**:182–7. doi:10.1038/nature01298
- 288 Chuang HC, Tan TH. MAP4K3/GLK in autoimmune disease, cancer and aging. *J. Biomed. Sci.* 2019;**26**:82. doi:10.1186/s12929-019-0570-5
- 289 Moghadasian MH, McManus BM, Nguyen LB, *et al.* Pathophysiology of apolipoprotein E deficiency in mice: relevance to apo E-related disorders in humans. *FASEB J* 2001;**15**:2623–30. doi:10.1096/fj.01-0463com
- 290 Hwangbo DS, Gersham B, Tu M-PP, *et al.* Drosophila dFOXO controls lifespan and regulates insulin signalling in brain and fat body. *Nature* 2004;**429**:562. doi:doi:10.1038/nature02549
- 291 Johnson TE, Lithgow GJ, Murakami S. Interventions that increase the response to stress offer the potential for effective life prolongation and increased health. *Journals Gerontol - Ser A Biol Sci Med Sci* 1996;**51**:B392–5. doi:10.1093/gerona/51A.6.B392
- 292 University of California Berkeley (USA), Max Planck Institute for Demographic Research (Germany). Human Mortality Database. <http://www.mortality.org/> (accessed 15 Mar 2020).
- 293 Rogers RG, Everett BG, Onge JMS, *et al.* Social, behavioral, and biological factors, and sex differences in mortality. *Demography* 2010;**47**:555–78. doi:10.1353/dem.0.0119
- 294 Peto R, Darby S, Deo H, *et al.* Smoking, smoking cessation, and lung cancer in the UK since 1950: Combination of national statistics with two case-control studies. *Br Med J* 2000;**321**:323–9. doi:10.1136/bmj.321.7257.323
- 295 Office for National Statistics. Adult smoking habits in the UK: 2017. *Stat Bull* 2018;:1–14.
- 296 Sundberg L, Agahi N, Fritzell J, *et al.* Why is the gender gap in life expectancy decreasing? The impact of age- and cause-specific mortality in Sweden 1997–2014. *Int J Public Health* 2018;**63**:673–81. doi:10.1007/s00038-018-1097-3
- 297 Xirocostas ZA, Everingham SE, Moles AT. The sex with the reduced sex chromosome dies earlier: a comparison across the tree of life. *Biol Lett* 2020;**16**:20190867. doi:10.1098/rsbl.2019.0867
- 298 Terao C, Momozawa Y, Ishigaki K, *et al.* GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nat Commun* 2019;**10**:4719. doi:10.1038/s41467-019-12705-5
- 299 MacHiela MJ, Zhou W, Karlins E, *et al.* Female chromosome X mosaicism is

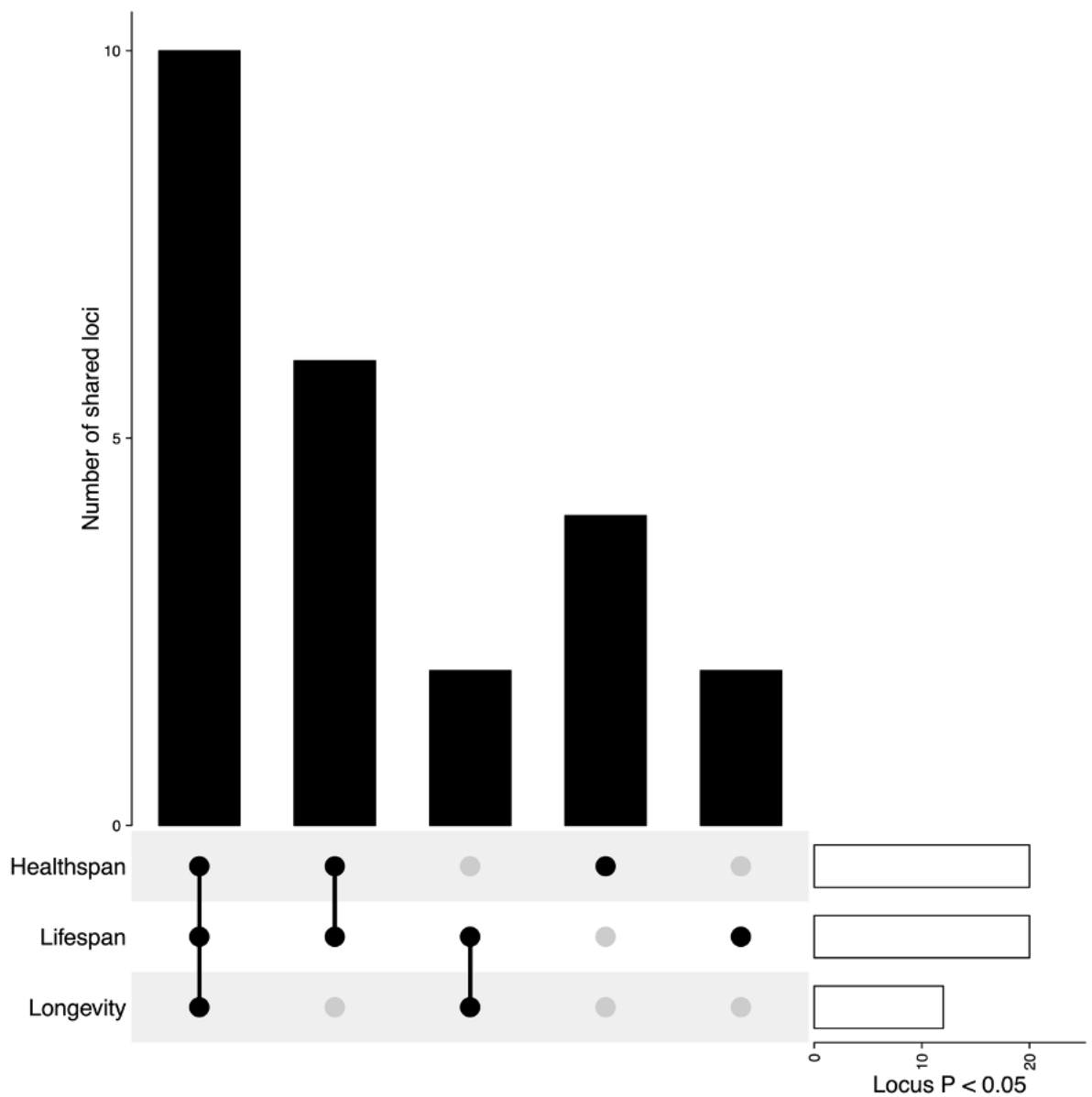
- age-related and preferentially affects the inactivated X chromosome. *Nat Commun* 2016;**7**:11843. doi:10.1038/ncomms11843
- 300 Forsberg LA, Rasi C, Malmqvist N, *et al.* Mosaic loss of chromosome y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat Genet* 2014;**46**:624–8. doi:10.1038/ng.2966
- 301 Office for National Statistics. Health state life expectancies by national deprivation deciles, England and Wales - Office for National Statistics. 2018.
- 302 OECD. *Education at a Glance 2016*. 2016. doi:https://doi.org/https://doi.org/10.1787/eag-2016-en
- 303 Sanderson SC, Waller J, Jarvis MJ, *et al.* Awareness of lifestyle risk factors for cancer and heart disease among adults in the UK. *Patient Educ Couns* 2009;**74**:221–7. doi:10.1016/j.pec.2008.08.003
- 304 Foster HME, Celis-Morales CA, Nicholl BI, *et al.* The effect of socioeconomic deprivation on the association between an extended measurement of unhealthy lifestyle factors and health outcomes: a prospective analysis of the UK Biobank cohort. *Lancet Public Heal* 2018;**3**:e576–85. doi:10.1016/S2468-2667(18)30200-7
- 305 Pool U. Socioeconomic inequalities in lifestyle-related health outcomes. *Lancet Public Heal*. 2019;**4**:e85. doi:10.1016/S2468-2667(19)30003-9
- 306 Flachsbart F, Ellinghaus D, Gentschew L, *et al.* Immunochip analysis identifies association of the RAD50/IL13 region with human longevity. *Aging Cell* 2016;**15**:585–8. doi:10.1111/acel.12471
- 307 Zeng Y, Nie C, Min J, *et al.* Sex Differences in Genetic Associations With Longevity. *JAMA Netw open* 2018;**1**:e181670. doi:10.1001/jamanetworkopen.2018.1670
- 308 Hoogendijk EO, Deeg DJH, Poppelaars J, *et al.* The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings. *Eur J Epidemiol* 2016;**31**:927–45. doi:10.1007/s10654-016-0192-0
- 309 Schächter F, Faure-Delanef L, Guénot F, *et al.* Genetic associations with human longevity at the APOE and ACE loci. *Nat Genet* 1994;**6**:29–32. doi:10.1038/ng0194-29
- 310 Xue A, Wu Y, Zhu Z, *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun* 2018;**9**:2941. doi:10.1038/s41467-018-04951-w
- 311 National Academies of Sciences Engineering and Medicine and others. Large Genetic Cohort Studies: A Background. In: *Applying an Implementation Science Approach to Genomic Medicine: Workshop Summary*. 2016.
- 312 Hwang L-D, Tubbs JD, Luong J, *et al.* Estimating indirect parental genetic effects on offspring phenotypes using virtual parental genotypes derived from sibling and half sibling pairs. *bioRxiv* 2020;:2020.02.21.959114. doi:10.1101/2020.02.21.959114
- 313 Ware JJ, van den Bree M, Munafò MR. From men to mice: CHRNA5/CHRNA3, smoking behavior and disease. *Nicotine Tob. Res.* 2012;**14**:1291–9. doi:10.1093/ntr/nts106
- 314 Sakaue S, Kanai M, Karjalainen J, *et al.* Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex

- traits with human lifespan. *bioRxiv* 2019;:856351. doi:10.1101/856351
- 315 Krokstad S, Ding D, Grunseit AC, *et al.* Multiple lifestyle behaviours and mortality, findings from a large population-based Norwegian cohort study - The HUNT Study. *BMC Public Health* 2017;**17**:58. doi:10.1186/s12889-016-3993-x
- 316 Hellwege JN, Keaton JM, Giri A, *et al.* Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet* 2017;**95**:1.22.1-1.22.23. doi:10.1002/cphg.48
- 317 Carlson CS, Matisse TC, North KE, *et al.* Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLoS Biol* 2013;**11**:e1001661. doi:10.1371/journal.pbio.1001661
- 318 Martin AR, Kanai M, Kamatani Y, *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;**51**:584–91. doi:10.1038/s41588-019-0379-x
- 319 Zaitlen N, Paşaniuc B, Gur T, *et al.* Leveraging Genetic Variability across Populations for the Identification of Causal Variants. *Am J Hum Genet* 2010;**86**:23–33. doi:10.1016/j.ajhg.2009.11.016
- 320 Wang X, Chua HX, Chen P, *et al.* Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Hum Mol Genet* 2013;**22**:2303–11. doi:10.1093/hmg/ddt064
- 321 Zanetti D, Weale ME. Transethnic differences in GWAS signals: A simulation study. *Ann Hum Genet* 2018;**82**:280–6. doi:10.1111/ahg.12251
- 322 Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* 2011;**88**:586–98. doi:10.1016/j.ajhg.2011.04.014
- 323 Morris AP. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* 2011;**35**:809–22. doi:10.1002/gepi.20630
- 324 Taliun D, Harris DN, Kessler MD, *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* 2019;:563866. doi:10.1101/563866
- 325 Chen Z, Chen J, Collins R, *et al.* China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* 2011;**40**:1652–66. doi:10.1093/ije/dyr120
- 326 Gaziano JM, Concato J, Brophy M, *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* 2016;**70**:214–23. doi:10.1016/j.jclinepi.2015.09.016
- 327 Nagai A, Hirata M, Kamatani Y, *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* 2017;**27**:S2–8. doi:10.1016/j.je.2016.12.005
- 328 Giri A, Hellwege JN, Keaton JM, *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet* 2019;**51**:51–62. doi:10.1038/s41588-018-0303-9
- 329 Demenais F, Margaritte-Jeannin P, Barnes KC, *et al.* Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet* 2018;**50**:42–50. doi:10.1038/s41588-017-0014-7

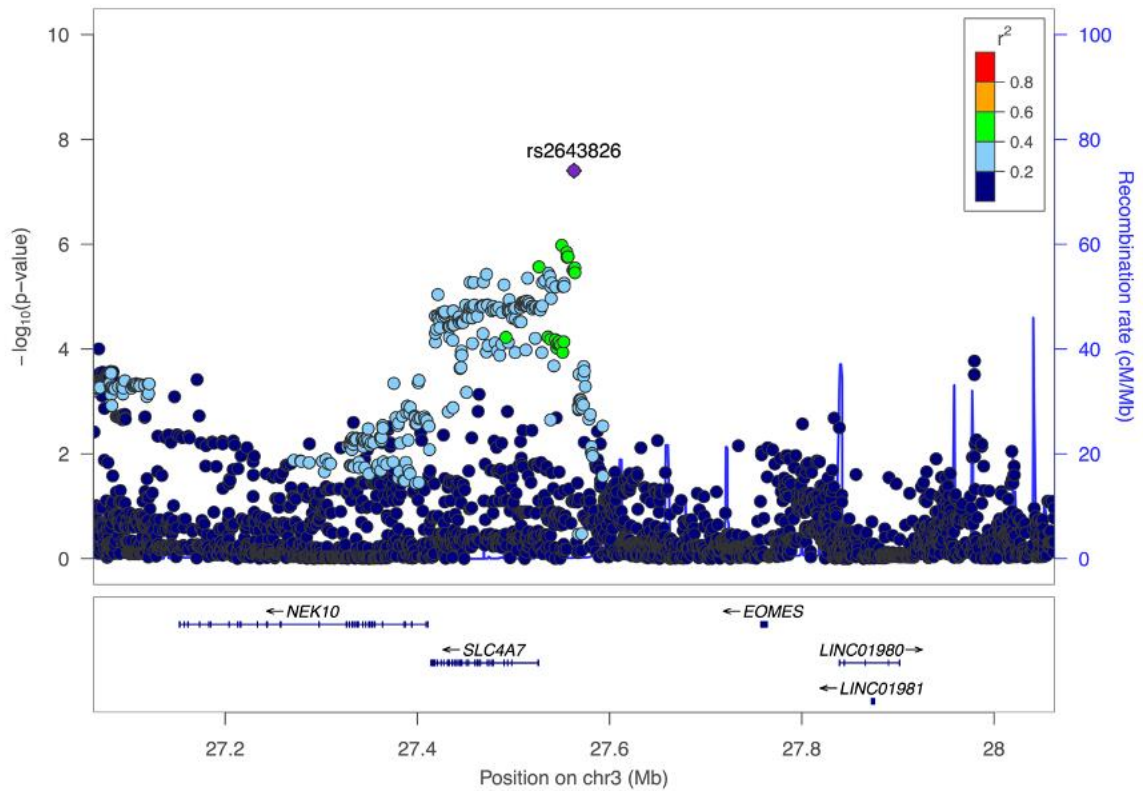
- 330 Charlesworth D, Willis JH. The genetics of inbreeding depression. *Nat. Rev. Genet.* 2009;**10**:783–96. doi:10.1038/nrg2664
- 331 Clark DW, Okada Y, Moore KHS, *et al.* Associations of autozygosity with a broad range of human phenotypes. *Nat Commun* 2019;**10**:1–17. doi:10.1038/s41467-019-12283-6
- 332 Wainschtein P, Jain DP, Yengo L, *et al.* Recovery of trait heritability from whole genome sequence data. *bioRxiv* 2019;:588020. doi:10.1101/588020
- 333 Hout CV Van, Tachmazidou I, Backman JD, *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv* 2019;:572347. doi:10.1101/572347
- 334 Carson AR, Smith EN, Matsui H, *et al.* Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Nephrol* 2014;**15**:125. doi:10.1186/1471-2105-15-125
- 335 Adelson RP, Renton AE, Li W, *et al.* Empirical design of a variant quality control pipeline for whole genome sequencing data using replicate discordance. *Sci Rep* 2019;**9**:16156. doi:10.1038/s41598-019-52614-7
- 336 Ma C, Blackwell T, Boehnke M, *et al.* Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet Epidemiol* 2013;**37**:539–50. doi:10.1002/gepi.21742
- 337 Pulit SL, de With SAJ, de Bakker PIW. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. *Genet Epidemiol* 2017;**41**:145–51. doi:10.1002/gepi.22032
- 338 Lee S, Abecasis GR, Boehnke M, *et al.* Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* 2014;**95**:5–23. doi:10.1016/j.ajhg.2014.06.009
- 339 Cooper GM, Shendure J. Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* 2011;**12**:628–40. doi:10.1038/nrg3046
- 340 Visscher PM, Wray NR, Zhang Q, *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 2017;**101**:5–22. doi:10.1016/j.ajhg.2017.06.005
- 341 Hertel J, Friedrich N, Wittfeld K, *et al.* Measuring Biological Age via Metabonomics: The Metabolic Age Score. *J Proteome Res* 2016;**15**:400–10. doi:10.1021/acs.jproteome.5b00561
- 342 Belsky DW, Caspi A, Houts R, *et al.* Quantification of biological aging in young adults. *Proc Natl Acad Sci U S A* 2015;**112**:E4104–10. doi:10.1073/pnas.1506264112
- 343 Krištić J, Vučković F, Menni C, *et al.* Glycans are a novel biomarker of chronological and biological ages. *Journals Gerontol - Ser A Biol Sci Med Sci* 2014;**69**:779–89. doi:10.1093/gerona/glt190
- 344 Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013;**14**:R115. doi:10.1186/gb-2013-14-10-r115
- 345 Hannum G, Guinney J, Zhao L, *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol Cell* 2013;**49**:359–67. doi:10.1016/j.molcel.2012.10.016
- 346 Kim S, Myers L, Wyckoff J, *et al.* The frailty index outperforms DNA

- methylation age and its derivatives as an indicator of biological age. *GeroScience* 2017;**39**:83–92. doi:10.1007/s11357-017-9960-3
- 347 Fahy GM, Brooke RT, Watson JP, *et al.* Reversal of epigenetic aging and immunosenescent trends in humans. *Aging Cell* 2019;**18**. doi:10.1111/accel.13028
- 348 Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* 2018;**27**:R195–208. doi:10.1093/hmg/ddy163
- 349 Cho Y, Haycock PC, Sanderson E, *et al.* Exploiting horizontal pleiotropy to search for causal pathways within a Mendelian randomization framework. *Nat Commun* 2020;**11**:1010. doi:10.1038/s41467-020-14452-4
- 350 Mokry LE, Ahmad O, Forgetta V, *et al.* Mendelian randomisation applied to drug development in cardiovascular disease: A review. *J Med Genet* 2015;**52**:71–9. doi:10.1136/jmedgenet-2014-102438
- 351 Sun BB, Maranville JC, Peters JE, *et al.* Genomic atlas of the human plasma proteome. *Nature* 2018;**558**:73–9. doi:10.1038/s41586-018-0175-2
- 352 Gallois A, Mefford J, Ko A, *et al.* A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat Commun* 2019;**10**:4788. doi:10.1038/s41467-019-12703-7

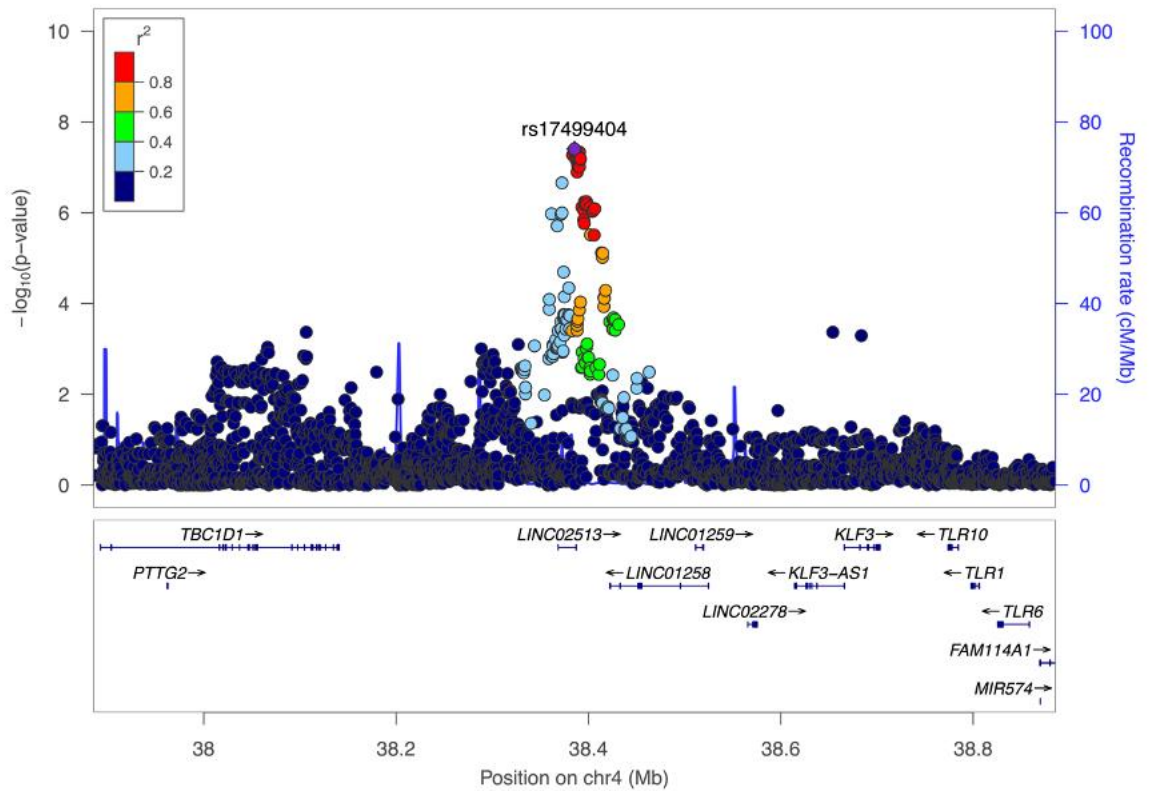
Appendix



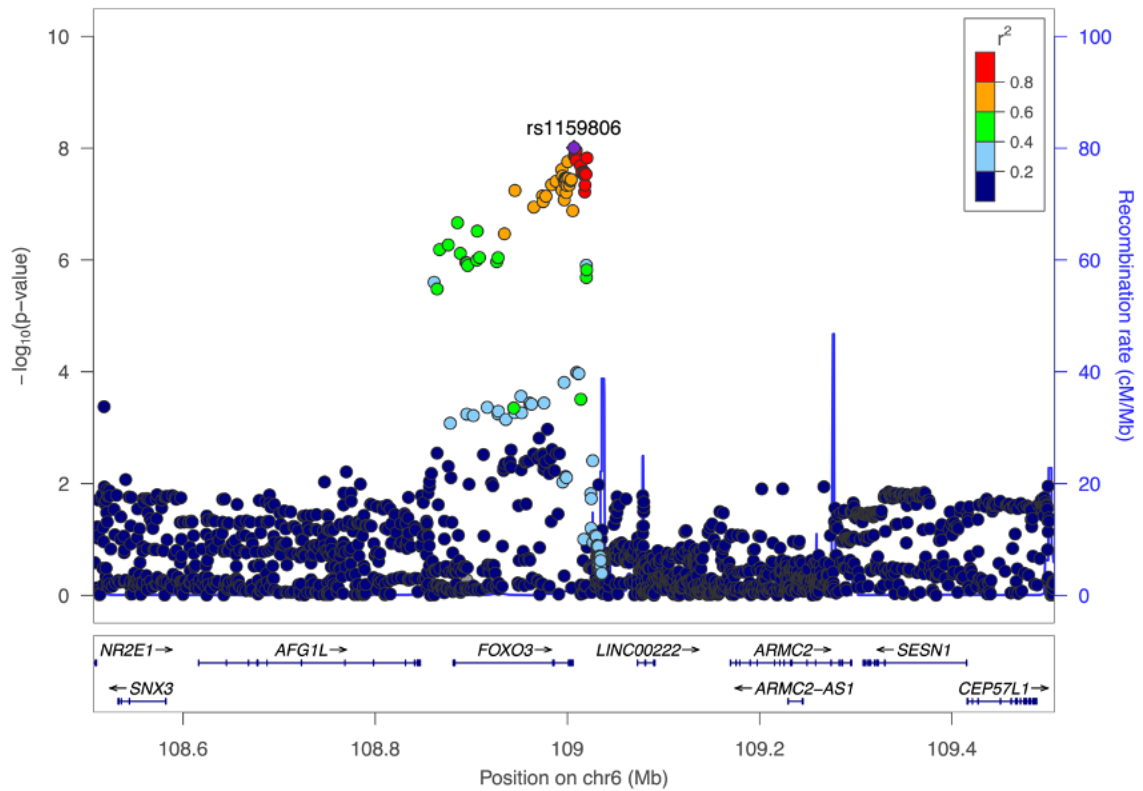
Supplementary Figure 1: Association of lead multivariate SNPs with ageing traits. Bars on the right represent the number of lead SNPs reaching nominal significance ($P < 0.05$) in each individual dataset, while bars on top represent the number of SNPs reaching nominal significance across datasets.



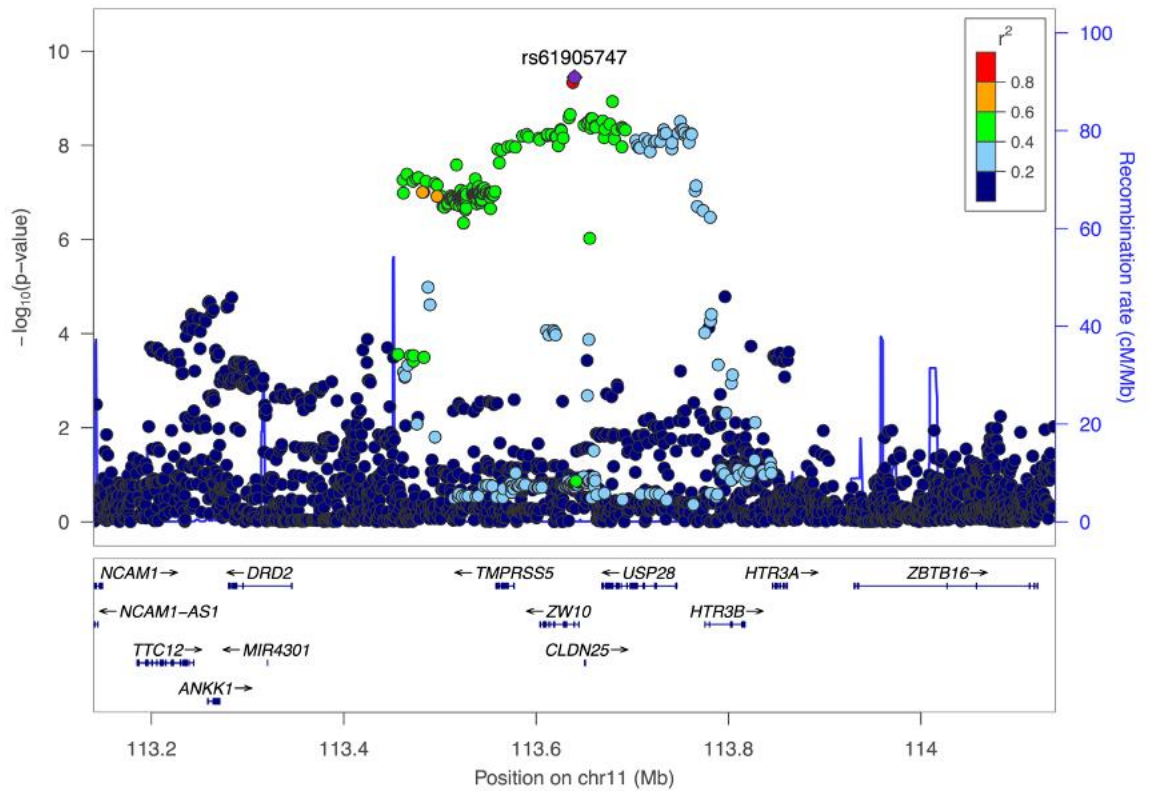
Supplementary Figure 2: LocusZoom plots of rs2643826 near *SLC4A7*. The x-axis shows the base-pair position of SNPs on chromosome 3 (GRCh37). The y-axis shows the strength of the MANOVA association. SNPs are coloured by their degree of linkage disequilibrium with the lead SNP, based on 1000 Genomes European ancestry individuals.



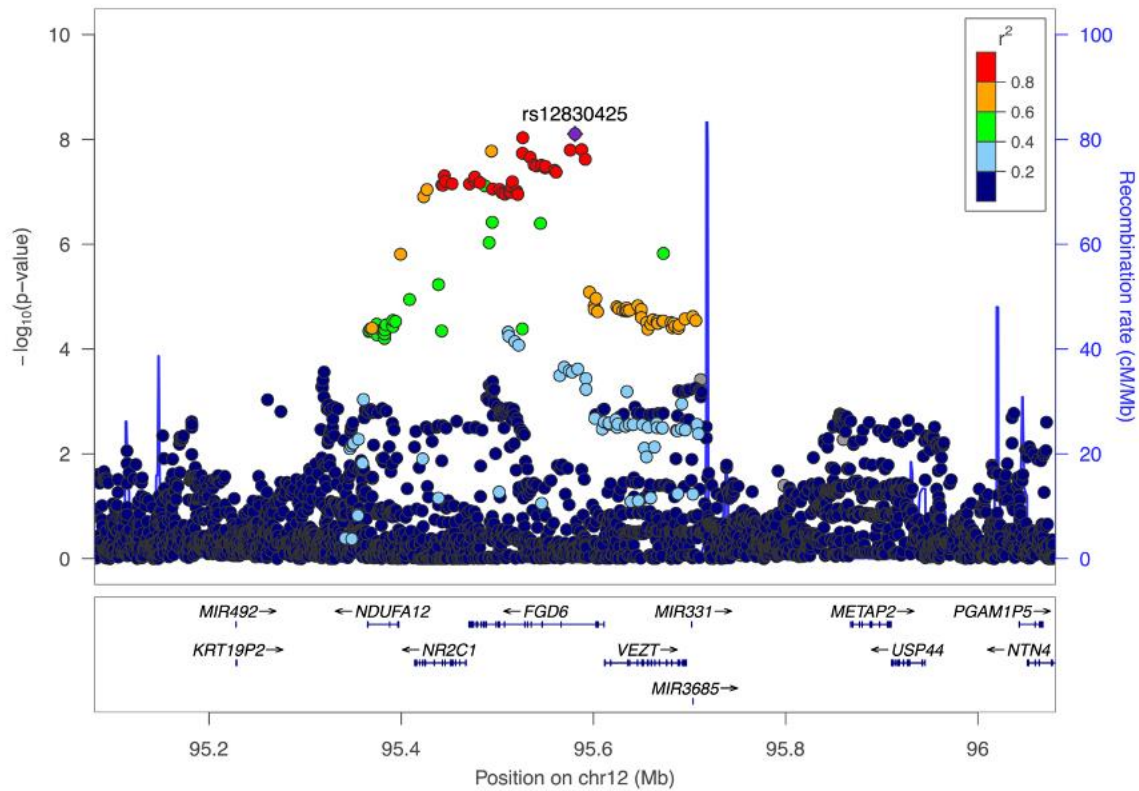
Supplementary Figure 3: LocusZoom plots of rs17499404 near *LINC02513*. The x-axis shows the base-pair position of SNPs on chromosome 4 (GRCh37). The y-axis shows the strength of the MANOVA association. SNPs are coloured by their degree of linkage disequilibrium with the lead SNP, based on 1000 Genomes European ancestry individuals.



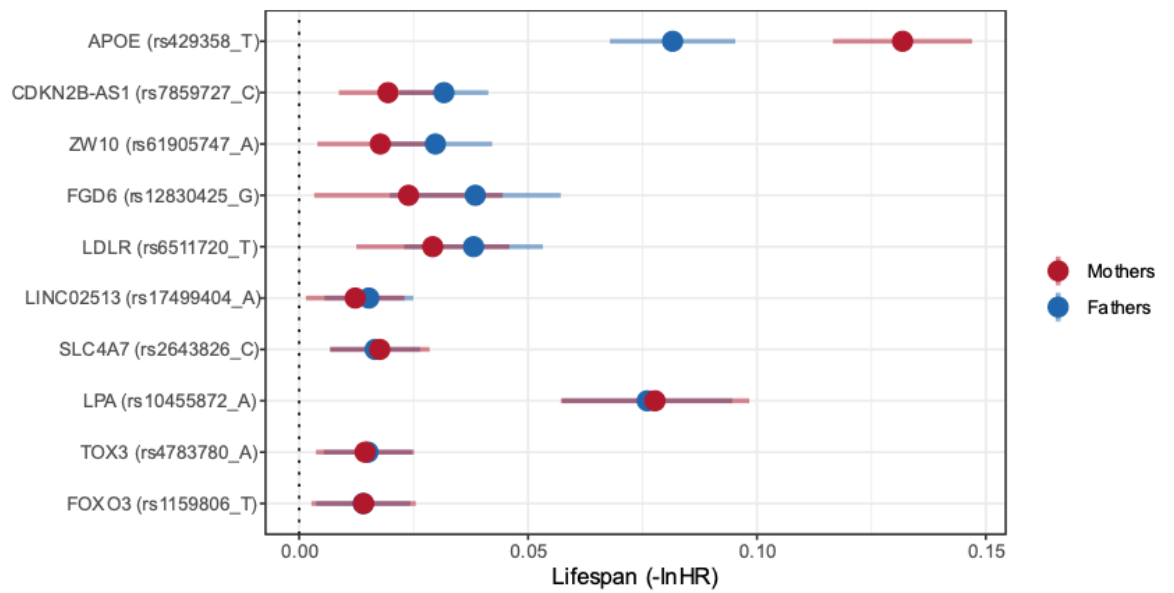
Supplementary Figure 4: LocusZoom plots of rs1159806 near *FOXO3*. The x-axis shows the base-pair position of SNPs on chromosome 6 (GRCh37). The y-axis shows the strength of the MANOVA association. SNPs are coloured by their degree of linkage disequilibrium with the lead SNP, based on 1000 Genomes European ancestry individuals.



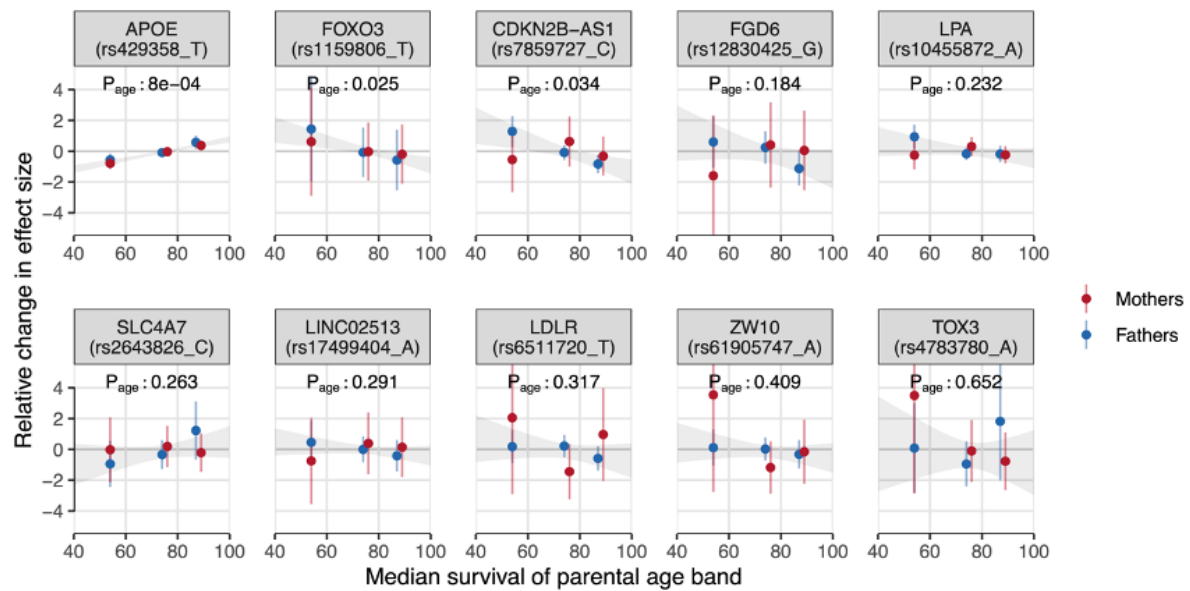
Supplementary Figure 5: LocusZoom plots of rs61905747 near *ZW10*. The x-axis shows the base-pair position of SNPs on chromosome 11 (GRCh37). The y-axis shows the strength of the MANOVA association. SNPs are coloured by their degree of linkage disequilibrium with the lead SNP, based on 1000 Genomes European ancestry individuals.



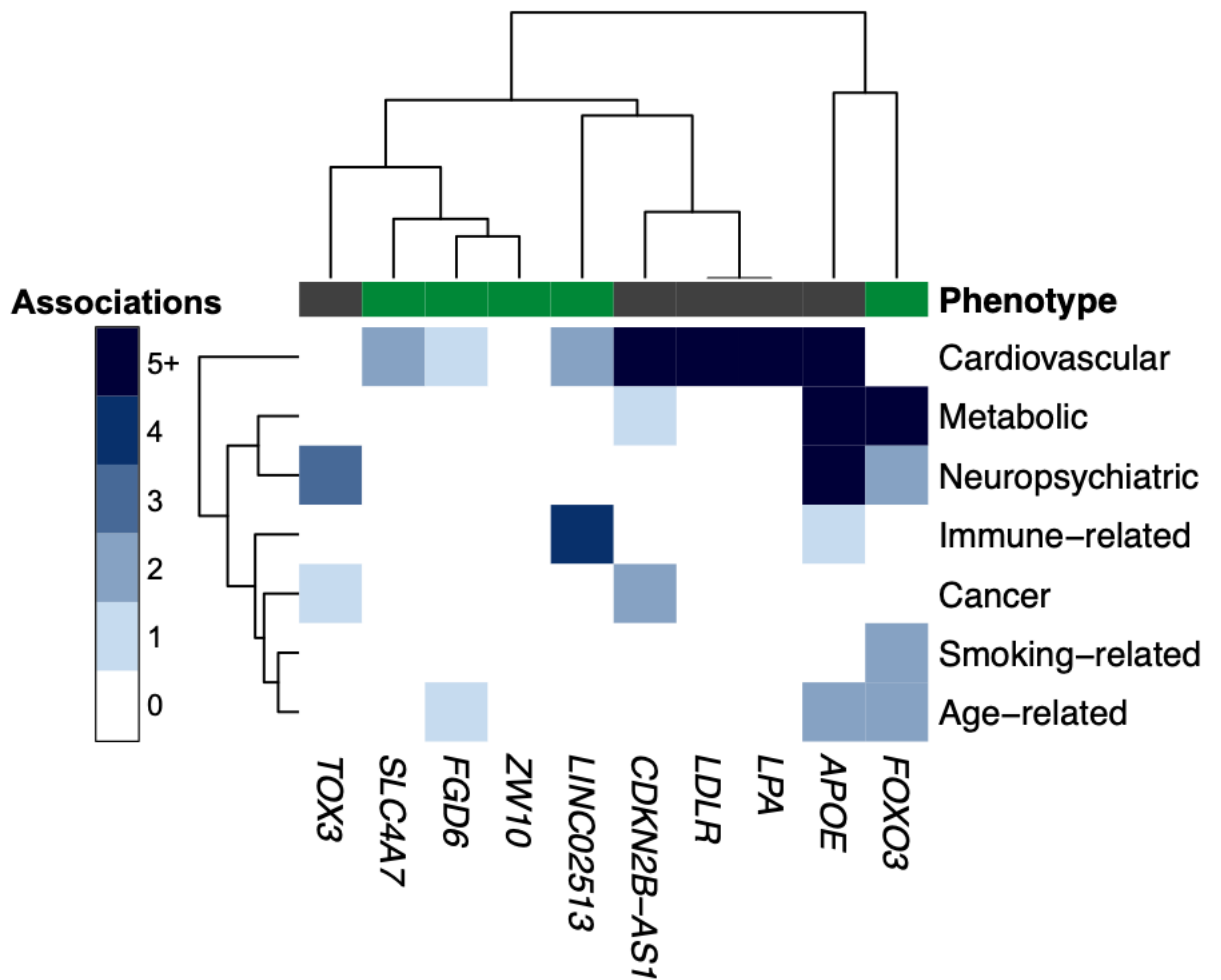
Supplementary Figure 6: LocusZoom plots of rs12830425 near *FGD6*. The x-axis shows the base-pair position of SNPs on chromosome 12 (GRCh37). The y-axis shows the strength of the MANOVA association. SNPs are coloured by their degree of linkage disequilibrium with the lead SNP, based on 1000 Genomes European ancestry individuals.



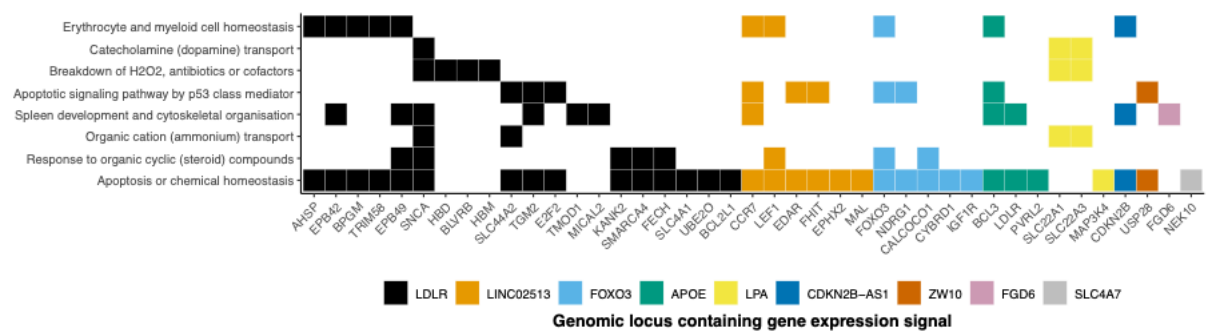
Supplementary Figure 7: Apart from *APOE*, loci of interest show limited evidence of sex specificity. Mother and father lifespan effect estimates from Timmers et al. [88] for each locus are shown in red and blue, respectively. Annotated are the nearest gene, index SNP and lifespan-increasing allele. Lines represent 95% confidence intervals.



Supplementary Figure 8: Apart from *APOE*, most loci of interest tend to decrease in effect with age. Each panel contains the age-stratified estimates on parental survival the lead SNP in each locus of interest, annotated with the nearest gene, lead SNP ID and its protective allele. The y-axis shows the effect sizes of age-stratified analyses (parents survival 40-60, 60-80, and 80+) relative to the unstratified analysis, for mothers and fathers separately. Lines represent 95% confidence intervals. The x-axis shows the median survival of each parental age band, calculated from Kaplan-Meier curves. Grey ribbons indicate the regression of father and mother estimates against median survival, weighted by the inverse variance of each estimate. The annotated P_{age} value is the significance of the coefficient of median survival in this regression.



Supplementary Figure 9: Loci of interest have previously been linked to cardiovascular traits. This heatmap shows the number of genome-wide significant associations reported in the GWAS catalog and PhenoScanner for lead SNPs and close proxies ($r_{2EUR} > 0.6$) of each locus of interest. Loci annotated with green bars are novel; those annotated with grey bars are known.



Supplementary Figure 10: Genes of interest are enriched ($P_{\text{bonferroni}} < 0.05$) for biological processes related to apoptosis and chemical homeostasis. Genes colocalizing with loci of interest in cis or trans are listed on the x-axis; GO biological processes gene sets from the Molecular Signatures Database, grouped into 8 broad categories using k-means clustering, are listed on the y-axis. See [Supplementary Table 8](#) for the full list of 32 biological process pathways with $P < 0.05/383$, where 383 is the number of gene sets passing the inclusion criteria. Squares represent the presence of a gene within one or more gene sets contained in the broad category. Squares are coloured based on their colocalisation with loci of interest.